Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Calibrated automatic relevance determination: finite-sample false discovery rate control for variable selection in Gaussian process regression

Jia Cai

jcai8@gmu.edu

George Mason University

Research Article

Keywords: Multiple testing, e-values, Conformal p-values, Exchangeability, Knockoffs, Kernel methods

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11319814/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Calibrated automatic relevance determination: finite-sample false discovery rate control for variable selection in Gaussian process regression

Jia Cai[1*]

[1*]Department of Statistics, George Mason University, Fairfax, Virginia, USA.

Corresponding author(s). E-mail(s): jcai8@gmu.edu;

Abstract

Automatic relevance determination (ARD) endows Gaussian process (GP) regression with a per-input relevance score—the reciprocal lengthscale—widely used to rank and select explanatory variables. In practice a variable is declared relevant by thresholding these scores with an informal rule, often calibrated against "inert" reference variables known to be inactive; such rules carry no guarantee on the proportion of false selections. We formalize the inert-reference device into a procedure with provable finite-sample false discovery rate (FDR) control. The central observation is an exchangeability argument: under a symmetric (independent-column) design, the relevance scores of genuinely null inputs are exchangeable with the scores of synthetic reference variables, yielding valid conformal-style $\boldsymbol{p}$-values and e-values from a single GP fit. Feeding these e-values to the e-Benjamini–Hochberg procedure controls FDR under arbitrary dependence; the Benjamini–Hochberg procedure applies when references are scored by independent per-input fits. A salient feature is robustness to non-identifiability: because the guarantee rests on column symmetry of the relevance map rather than consistent estimation of the lengthscales, it sidesteps their well-known weak identifiability, which affects power but not validity. The procedure adds a calibration layer to one GP fit; we establish its exchangeability lemma, FDR theorems, a derandomized variant and selection consistency, and analyse its computational cost, scalability and implementation. Simulations confirm finite-sample control where the uncorrected reference rule fails, and a model-X knockoff variant handles correlated designs. An application to National Health and Nutrition Examination Survey data, with planted nulls and a semi-synthetic response, illustrates error control and power.

Keywords: Multiple testing, e-values, Conformal p-values, Exchangeability, Knockoffs, Kernel methods

MSC Classification: 62J15 , 62G08 , 62F15 , 60G15

1 Introduction

Gaussian process (GP) regression provides a flexible nonparametric prior over functions with closed-form posteriors and calibrated uncertainty (Rasmussen and Williams, 2006; Williams and Rasmussen, 1996). When the covariance kernel is equipped with a separate lengthscale ℓ_d for each input coordinate—the automatic relevance determination (ARD) parameterization (Neal, 1996; MacKay, 1996)—the fitted lengthscales summarize how strongly the response varies along each input. A short lengthscale signals a rapidly varying, and therefore relevant, coordinate; a long lengthscale signals an inert one. The reciprocal lengthscale $r_d = 1/\hat{\ell}_d$ is accordingly used throughout the applied literature as a *relevance score* for ranking inputs and for selecting an active subset (Rasmussen and Williams, 2006; Piironen and Vehtari, 2017).

The difficulty is calibration. To turn the scores $\{r_d\}_{d=1}^{D}$ into a selected set $\hat{S} \subseteq \{1, \ldots, D\}$ one must choose a threshold, and the scores live on an uninterpretable scale that depends on the kernel variance, the noise level, and the (often poorly identified) lengthscale geometry. A popular device, introduced by Linkletter et al (2006) for computer experiments, augments the design with *inert reference variables*—synthetic inputs known to be inactive—and declares a real input relevant only if its score exceeds what the inert references achieve by chance. This is an appealing idea: the references supply an empirical null distribution for the relevance score. But as originally formulated the device is informal. It compares each input to the references one at a time, with no correction for the multiplicity of inputs, and it offers no control over the expected fraction of selected inputs that are in fact null. As the number of candidate inputs grows, the probability that some null input exceeds the reference floor by chance accumulates, and the realized error rate drifts upward (we document this in Section 6).

We supply the missing calibration. The device of comparing real inputs to added null variables is not unique to Gaussian processes: it underlies the Boruta selector (Kursa and Rudnicki, 2010), which confirms a feature when its importance exceeds the maximum among permuted "shadow" features, and the pseudovariable approach to the false selection rate (Wu et al, 2007; Boos et al, 2009), which calibrates forward selection against added phony predictors and connects to an adaptive Benjamini–Hochberg rule. What has been missing for the GP relevance map specifically is an exchangeability analysis that turns this heuristic into a procedure with *finite-sample false discovery rate* (FDR) control (Benjamini and Hochberg, 1995). The false discovery rate is the expected proportion of false selections,

$$\mathrm{FDR} \;=\; \mathbb{E}\left[\frac{|\hat{S} \cap S_0^c|}{|\hat{S}| \vee 1}\right], \tag{1}$$

where S_0 is the set of truly active inputs and S_0^c its complement. Controlling (1) at a user-specified level q guarantees that, on average, no more than a fraction q of the reported inputs are spurious—exactly the assurance an applied analyst needs when handing a variable list to a domain expert.

1.1 Contributions

Our approach rests on a single structural observation and develops its consequences.

1. **Exchangeability of relevance scores (Section 4, Lemma 4.2).** Under a symmetric design in which the non-signal columns are exchangeable given the response and the active columns—which holds, for example, when the inputs are drawn with independent coordinates, as in space-filling computer experiments—the relevance scores of the genuinely null inputs are exchangeable with the scores of synthetic reference (dummy) variables. The argument requires only that the relevance map be equivariant under relabeling of input columns, a property the ARD-GP fit satisfies.
2. **Valid p-values and e-values from one fit (Propositions 4.3 and the construction of Section 3).** Exchangeability turns the rank of a real input's score among the dummy scores into a conformal-style p-value that is super-uniform under the null, and into an e-value via a calibrator. All scores come from a single augmented GP fit.
3. **FDR control (Theorems 4.4–4.6).** Passing the e-values to the e-Benjamini–Hochberg (e-BH) procedure (Wang and Ramdas, 2022) controls FDR at level q under *arbitrary* dependence among the scores; passing the p-values to Benjamini–Hochberg (BH) controls FDR under independent reference sets; and averaging e-values over independent reference draws yields a derandomized procedure that retains control while reducing the randomness of the selected set (Ren and Barber, 2024).
4. **Robustness to non-identifiability (Section 4.3).** The ARD lengthscales are only weakly identified: in the infill regime only a microergodic combination of variance and lengthscale is consistently estimable (Zhang, 2004). Our guarantee is immune to this. It uses the relevance score solely as an exchangeable label, never as an estimate of anything, so non-identifiability degrades power but not validity. This converts a much-discussed obstruction into a non-issue *for validity*—though not for power, which the weak signal in the lengthscales still limits.
5. **Honest treatment of the design assumption (Sections 6–8).** The clean guarantee requires exchangeable non-signal columns, which independent dummies satisfy under independent designs but not under arbitrary correlation. We report a setting in which the joint GP fit empirically tolerates correlation, and a complementary adversarial setting (near-duplicate nulls) in which independent dummies fail and a model-X knockoff construction (Candès et al, 2018) restores control.
6. *Computation and reproducibility.* The calibration adds only $O(D + K)$ work to a single GP fit, so CARD inherits the $O(IN^3)$ cost of the host model. We give its complexity and measured running times, the column-symmetry condition under which sparse approximations preserve validity, and a reproducible implementation (Online Resource 1).

1.2 Related work

Relevance and selection in GP regression. Beyond raw lengthscale thresholding, GP variable selection has been pursued through spike-and-slab and shrinkage priors on the lengthscales (Savitsky et al, 2011; Dance and Paige, 2022), sparse axis-aligned subspace priors for Bayesian optimization (Eriksson and Jankowiak, 2021), projection-predictive and sensitivity-based scores (Paananen et al, 2019)—recently extended to discrete-response families (Weber et al, 2025)—and robust priors tailored to emulation (Gu, 2019). In a Bayesian formulation, posterior variable-selection consistency has been established for rescaled GP priors with stochastic inclusion under fixed-domain, random-design asymptotics when the truth has finite smoothness (Jiang and Tokdar, 2021). The difficulty of optimizing GP lengthscales in high dimensions has itself motivated alternatives: Appriou et al (2024) build high-dimensional kriging surrogates by combining sub-models with randomly drawn lengthscales, sidestepping likelihood-based hyperparameter optimization, which echoes our observation that the relevance map need not be a consistent estimate to be useful. These methods target posterior inclusion probabilities or point selections; none of them provides finite-sample frequentist FDR control, and all leave the relevance threshold to be chosen by the analyst. Our contribution is orthogonal and complementary: it is a calibration layer that converts *any* column-symmetric relevance score—reciprocal lengthscale, posterior sensitivity, or otherwise—into an FDR-controlled set.

Reference, shadow, and pseudo variables. The idea of calibrating selection against added inactive variables has several independent lineages. In computer experiments, Linkletter et al (2006) add synthetic inert inputs and compare posterior lengthscale distributions. In machine learning, the Boruta algorithm (Kursa and Rudnicki, 2010) augments the design with permuted "shadow" copies of every feature and confirms a feature when its importance exceeds the maximum shadow importance, iterating and judging the resulting hit counts by a binomial test. In statistics, the false-selection-rate methods of Wu et al (2007) and Boos et al (2009) add phony predictors to calibrate forward selection, the latter making explicit the link to an adaptive Benjamini–Hochberg rule. Our e-value (7) is exactly the Boruta "exceed the maximum shadow" event, and our derandomized average is its hit-counting analogue; what we add for this family is a finite-sample FDR guarantee, in place of a binomial heuristic or an uncorrected per-input comparison, together with an exchangeability proof tailored to the jointly fitted GP relevance map.

Multiple testing, conformal selection, e-values, and knockoffs. Our machinery is the modern apparatus of FDR control. The BH procedure controls FDR under independence and positive dependence (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001); e-values and the e-BH procedure control FDR under arbitrary dependence (Vovk and Wang, 2021; Wang and Ramdas, 2022; Grünwald et al, 2024; Ramdas et al, 2023). Model-based procedures instead exploit a posited dependence structure—for instance hidden semi-Markov models for large-scale dependent testing (Wang and Wang, 2024)—whereas the e-value route we take is agnostic to the dependence entirely. The e-BH procedure has since been shown to be uniformly improvable for free, via the closed e-BH procedure built on a closure principle for FDR control (Xu et al, 2025),

and the choice of p-to-e calibrator has been studied for knockoff-assisted selection (Biswas and Ramdas, 2026); we connect to both in Section 3.4. The rank p-value we use, its discreteness, the indicator p-to-e calibrator, and the routing of the resulting dependent p-values through e-BH are the standard apparatus of *conformal selection* (Bates et al, 2023; Jin and Candès, 2023; Bashari et al, 2023); our p-values coincide with one-sided conformal p-values against a shared calibration pool (Vovk et al, 2005; Lei et al, 2018). Knockoffs (Barber and Candès, 2015; Candès et al, 2018) construct negative-control copies of each feature and have been derandomized through e-values (Ren and Barber, 2024); related model-free constructions include Gaussian mirrors (Xing et al, 2023) and data-splitting FDR control (Dai et al, 2023a). Knockoff methodology continues to be actively developed for complex designs: latent Gaussian copula constructions sample valid knockoffs for skewed, mixed-type covariates (Vásquez et al, 2024), model-X knockoffs have been carried to the accelerated-failure-time model for survival data (He et al, 2024), and—closest to our derandomized variant—a truncated D-vine copula knockoff has itself been *derandomized through e-values* to control FDR (Vásquez et al, 2025). Closest in spirit to a finite-sample nonparametric guarantee is the kernel-knockoffs selector of Dai et al (2023b), which *does* control FDR for any sample size in nonparametric additive models via random features and subsampling; we do not claim primacy for finite-sample nonparametric FDR. Our procedure differs in that it operates directly on the ARD-GP relevance map, requires no feature randomization, and produces shared references that feed BH/e-BH rather than the pairwise knockoff filter. The contribution is thus a synthesis—conformal-selection calibration applied to GP relevance—together with the GP-specific exchangeability lemma and the non-identifiability analysis below, rather than a new multiple-testing engine.

The remainder of the paper is organized as follows. Section 2 fixes notation for ARD-GP regression and recalls the FDR apparatus. Section 3 presents the calibrated procedure, which we call CARD (calibrated ARD). Section 4 develops the exchangeability lemma and the FDR theorems and treats non-identifiability. Section 5 analyses time complexity, the column-symmetry condition under which scalable approximations preserve validity, and reproducibility. Sections 6 and 7 report simulations and a real-data illustration. Section 8 collects concluding remarks, limitations, and extensions. Proofs are gathered in the Appendix.

2 Background and problem formulation

2.1 Gaussian process regression with ARD

We observe data $\mathcal{D}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with inputs $\boldsymbol{x}_i \in \mathbb{R}^D$ and scalar responses $y_i \in \mathbb{R}$. Collect the inputs in the design matrix $X \in \mathbb{R}^{N\times D}$ with columns $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_D$ (so $\boldsymbol{X}_d \in \mathbb{R}^N$ is the vector of the dth coordinate across observations) and the responses in $\boldsymbol{y} \in \mathbb{R}^N$. We posit

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \qquad f \sim \mathcal{GP}(0, k_\theta), \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_n^2), \tag{2}$$

with the squared-exponential ARD kernel

$$k_\theta(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\tfrac{1}{2} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\ell_d^2} \right), \qquad \theta = \left(\sigma_f^2, \ell_1, \ldots, \ell_D, \sigma_n^2\right). \tag{3}$$

The hyperparameters are fit by maximizing the log marginal likelihood

$$\log p(\boldsymbol{y} \mid X, \theta) = -\tfrac{1}{2} \boldsymbol{y}^\top (K_\theta + \sigma_n^2 I)^{-1} \boldsymbol{y} - \tfrac{1}{2} \operatorname{logdet}(K_\theta + \sigma_n^2 I) - \tfrac{N}{2} \log 2\pi, \tag{4}$$

where $(K_\theta)_{ij} = k_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Write $\hat{\theta} = \arg\max_\theta \log p(\boldsymbol{y} \mid X, \theta)$ for the maximizer returned by the optimizer and $\hat{\ell}_d$ for its lengthscale components. The *relevance score* of input d is

$$r_d = 1/\hat{\ell}_d. \tag{5}$$

Coordinates with large r_d vary quickly and are deemed relevant. Any strictly monotone transform of $1/\hat{\ell}_d$ (for instance $1/\hat{\ell}_d^2$) leaves the analysis below unchanged, because the procedure depends on the scores only through their ranks. We work with a generic relevance map T in Section 4, of which (5) is the running example; sensitivity-based scores (Paananen et al, 2019) are equally admissible.

Maximization of (4) is non-convex, and a symmetric (isotropic) initialization can be trapped at a saddle in which all lengthscales coincide. As is standard in ARD, we initialize the lengthscales at a common large value, so that all coordinates begin nearly inactive and the optimizer "switches on" only those that improve the fit; this initialization is identical across coordinates and hence preserves the column symmetry required below.

2.2 The selection problem and the false discovery rate

Let f_0 denote the regression function generating the data. Input d is *null* if f_0 does not depend on its dth coordinate, i.e. $f_0(\boldsymbol{x})$ is constant in x_d; let S_0^c be the set of null inputs and $S_0 = \{1, \ldots, D\} \setminus S_0^c$ the active set, with $m_0 = |S_0^c|$. A selection rule returns a set $\hat{S}$, and we measure its quality by the false discovery proportion $\mathrm{FDP} = |\hat{S} \cap S_0^c| / (|\hat{S}| \vee 1)$ and its expectation, the FDR (1). Our goal is a rule with $\mathrm{FDR} \le q$ for a prescribed $q \in (0, 1)$, while selecting as many active inputs as possible (power, the expected fraction of S_0 recovered).

2.3 *p*-values, e-values, and step-up procedures

We recall the two engines we use. Given p-values $p_1, \ldots, p_D$ with p_d super-uniform under its null ($\mathbb{P}(p_d \le t) \le t$ for null d), the BH procedure at level q orders $p_{(1)} \le \cdots \le p_{(D)}$, finds $k^\star = \max\{k : p_{(k)} \le qk/D\}$, and rejects the hypotheses with the $k^\star$ smallest p-values; it controls FDR at $q\, m_0/D$ under independence or positive regression dependence (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001).

An e-value is a nonnegative statistic e_d with $\mathbb{E}[e_d] \le 1$ under the null (Vovk and Wang, 2021; Ramdas et al, 2023). The e-BH procedure at level q orders $e_{(1)} \ge \cdots \ge e_{(D)}$, finds $k^\star = \max\{k : e_{(k)} \ge D/(kq)\}$, and rejects the corresponding $k^\star$ hypotheses.

Its defining property is robustness: e-BH controls FDR at $q\,m_0/D \le q$ under *arbitrary* dependence among the e-values (Wang and Ramdas, 2022). A super-uniform p-value yields a valid e-value through the calibrator $e = \mathbf{1}\{p \le \alpha_0\}/\alpha_0$ for any fixed $\alpha_0 \in (0,1)$, since $\mathbb{E}[\mathbf{1}\{p \le \alpha_0\}/\alpha_0] = \mathbb{P}(p \le \alpha_0)/\alpha_0 \le 1$. Averages of valid e-values are valid e-values, which is the basis of derandomization (Ren and Barber, 2024).

3 The calibrated ARD procedure

We now describe the procedure, which we abbreviate CARD. The idea is to manufacture an empirical null distribution for the relevance score by augmenting the design with synthetic reference variables, and then to convert each real input's score into a p-value (and an e-value) by its rank among the references.

3.1 Reference augmentation

Fix a number K of reference variables. Draw a reference (dummy) design $Z \in \mathbb{R}^{N\times K}$ with columns $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K$ and form the augmented design $\widetilde{X} = [\,X \mid Z\,] \in \mathbb{R}^{N\times(D+K)}$. Fit the ARD-GP (2)–(4) to $(\widetilde{X}, \boldsymbol{y})$ and read off the relevance scores of all $D+K$ columns,

$$r_1, \dots, r_D \quad \text{(real inputs)}, \qquad r_1^z, \dots, r_K^z \quad \text{(references)}.$$

Because all scores come from one joint fit, the references compete with the real inputs on equal footing and absorb the same kernel-variance and noise nuisances. How the references are drawn governs the validity argument (Section 4): for independent designs it suffices to draw each $\boldsymbol{Z}_k$ from the common coordinate law (e.g. independent draws matching the marginal of an input column); for correlated designs one draws structured references, as discussed in Section 8.

3.2 Rank p-values and e-values

For each real input d, compare its score to the reference pool and form the conformal-style rank p-value

$$p_d \;=\; \frac{1 + \#\{\,k \in \{1,\dots,K\} : r_k^z \ge r_d\,\}}{K+1}. \tag{6}$$

A relevant input, whose score sits above the reference cloud, receives a small p_d; a null input, whose score is statistically indistinguishable from the references, receives a p_d that is approximately uniform. The attainable resolution is $1/(K+1)$: the smallest possible p-value is achieved when r_d exceeds every reference.

To obtain e-values we apply the calibrator $e_d = \mathbf{1}\{p_d \le \alpha_0\}/\alpha_0$. The canonical, tuning-free choice takes $\alpha_0 = 1/(K+1)$, which flags an input precisely when it dominates the entire reference pool:

$$e_d \;=\; (K+1)\,\mathbf{1}\Big\{ r_d > \max_k r_k^z \Big\}. \tag{7}$$

This is a valid e-value because, under the null exchangeability established below, the probability that r_d exceeds all K references equals $1/(K+1)$, so $\mathbb{E}[e_d] = (K+1)\cdot\frac{1}{K+1} =$

1. Larger K raises both the magnitude $(K+1)$ of the flagged e-value and the stringency of the flag; we discuss the trade-off in Section 8.

3.3 Selection

The scores, p-values and e-values feed standard step-up rules, giving three variants:

- CARD-**e**: apply e-BH at level q to $\{e_d\}_{d=1}^D$ from (7). Valid under arbitrary dependence (Theorem 4.4).
- CARD-**BH**: apply BH at level q to $\{p_d\}_{d=1}^D$ from (6). Provably valid when the references are scored by independent per-input fits (Theorem 4.5); for the single shared-reference fit used by default, control is supported empirically (Remarks 2 and 1) but is not guaranteed.
- CARD-**BY**: apply BH at the corrected level q/H_D, $H_D = \sum_{j=1}^D 1/j$, guaranteeing control under any dependence (Benjamini and Yekutieli, 2001), at the cost of conservativeness.

A derandomized variant, dCARD-**e**, averages the e-values over M reference draws before applying e-BH (Theorem 4.6). Algorithm 1 summarizes the default (CARD-e) procedure.

3.4 Choice of calibrator and recent e-BH refinements

The calibrator $e_d = \mathbf{1}\{p_d \le \alpha_0\}/\alpha_0$ is one of a family: any nonincreasing $f : [0,1] \to [0,\infty]$ with $\int_0^1 f \le 1$ is an admissible p-to-e calibrator (Vovk and Wang, 2021), including the bounded and continuous calibrators recently exploited for knockoff-assisted selection (Biswas and Ramdas, 2026). It is tempting to expect a continuous calibrator—which preserves the magnitude by which an active score exceeds the reference floor—to outperform the binary indicator. For the e-BH *step-up* procedure, the opposite holds, and for a structural reason. e-BH rejects the hypotheses with the largest e-values clearing the data-dependent bar $D/(kq)$; what it rewards is therefore the *concentration* of evidence at the top, not its spread. Among all calibrators with $\int_0^1 f \le 1$, the indicator at $\alpha_0 = 1/(K+1)$ is the one that places the maximal possible mass, $K+1$, on the smallest attainable p-value, and is in this sense the power-optimal calibrator for the top rank. A continuous calibrator such as $f_a(p) = (1-a)p^{-a}$ necessarily assigns the top input a smaller e-value, $(1-a)(K+1)^a < K+1$, which can fall below the e-BH bar and forfeit the rejection entirely. In a representative configuration ($D = 20$, $s = 8$, $K = 40$, SNR $= 8$, $q = 0.2$) the indicator calibrator attains power 0.49 while the continuous family f_a attains power 0 for every $a \in \{0.3, 0.5, 0.7, 0.9\}$: the spreading of evidence collapses the procedure. (The power gains reported for continuous calibrators in Biswas and Ramdas, 2026 arise in the *weighted*-BH framework, where the calibrated value modulates a p-value weight rather than serving as a standalone test statistic; that mechanism is distinct from the e-BH step-up used here.) We therefore retain the indicator calibrator throughout.

Proposition 3.1 (A continuous calibrator can forfeit every rejection)**.** *Draw K references so that the attainable rank p-values lie on the grid $\{1/(K+1), 2/(K+1), \ldots, 1\}$,*

and let $f : [0,1] \to [0,\infty]$ be any nonincreasing p-to-e calibrator. If its largest attainable value falls short of $1/q$, i.e. $f\big(1/(K+1)\big) < 1/q$, then e-BH at level q applied to $e_d = f(p_d)$ returns $\hat{S} = \varnothing$ for every *realization of the scores. Consequently the power family $f_a(p) = (1-a)p^{-a}$ forfeits all rejections whenever $(1-a)(K+1)^a < 1/q$, whereas the indicator calibrator attains the maximal value $f(1/(K+1)) = K+1$ and rejects every input exceeding all references as soon as their number is at least $D/((K+1)q)$.*

Proof. Since f is nonincreasing and every attainable $p_d \geq 1/(K+1)$, each e-value obeys $e_d = f(p_d) \leq f(1/(K+1))$. The e-BH step-up makes a nonempty rejection iff some rank k satisfies $e_{(k)} \geq D/(kq)$, equivalently $k\, e_{(k)} \geq D/q$. As $e_{(k)} \leq f(1/(K+1))$ and $k \leq D$, we have $k\, e_{(k)} \leq D\, f(1/(K+1)) < D/q$, so no such k exists and $\hat{S} = \varnothing$. Evaluating f at $1/(K+1)$ gives the two special cases; for the indicator, T inputs exceeding all references yield T e-values equal to $K+1$, and $k = T$ clears the bar precisely when $T \geq D/((K+1)q)$. □

This makes CARD-e the most conservative of our variants—a known characteristic of e-BH—and invites the closed e-BH procedure of Xu et al (2025) as a cost-free improvement. Built from the e-closure principle, closed e-BH retains finite-sample FDR control under arbitrary dependence and *uniformly dominates* e-BH: its rejection set always contains that of e-BH and is sometimes strictly larger (Xu et al, 2025). It applies verbatim to the valid e-values of Proposition 4.3, so substituting it for the final step of Algorithm 1 can only add discoveries while preserving the bound of Theorem 4.4. In the present setting, however, the gain is structurally capped, and it is worth seeing why. The indicator calibrator makes every e-value binary, $e_d \in \{0, K+1\}$; the closure improvement extracts extra rejections from the *fractional slack* between an e-value and the step-up bar, and binary e-values carry no intermediate slack—an input either attains the maximal value $K+1$ or contributes 0, and no intersection e-value built from a coordinate with individual value 0 can clear a positive threshold. Closed e-BH therefore coincides with e-BH on the maximal-value set and cannot promote a zero-valued input, so the two agree except in degenerate ties, consistent with the negligible empirical difference. We accordingly keep standard e-BH as the default—exact, transparent, and losing essentially nothing under the binary calibrator—while recommending closed e-BH as the drop-in replacement whenever a non-binary calibrator is adopted (where, by Proposition 3.1, care with the calibrator scale is already required). The higher-power operating point in our experiments is instead supplied by CARD-BH.

4 Theoretical guarantees

We work with a generic relevance map and isolate the symmetry it must possess.

Definition 4.1 (Column-equivariant relevance map)**.** *A relevance map T assigns to a design with p columns and a response $\boldsymbol{y}$ a vector of scores $T(\widetilde{X}, \boldsymbol{y}) \in \mathbb{R}^p$. It is* column-equivariant *if for every permutation π of $\{1, \ldots, p\}$ and every $(\widetilde{X}, \boldsymbol{y})$, $T(\widetilde{X}_\pi, \boldsymbol{y}) = \pi\big(T(\widetilde{X}, \boldsymbol{y})\big)$, where $\widetilde{X}_\pi$ permutes the columns of $\widetilde{X}$ by π.*

Equivariance says only that the scoring procedure does not depend on the order in which columns are presented: relabeling the inputs relabels the scores identically. The ARD marginal likelihood (4) is invariant to permutations of the input coordinates,

Algorithm 1 CARD-e: calibrated ARD selection with e-BH

Require: data $(X, \boldsymbol{y})$; reference count K; target FDR level q; (optional) draws M

for $m = 1, \dots, M$ **do**
  draw reference design $Z^{(m)} \in \mathbb{R}^{N \times K}$ from the column law
  fit ARD-GP to $[\, X \mid Z^{(m)} \,]$; obtain scores r_d and references $r_1^{z,(m)}, \dots, r_K^{z,(m)}$
  $e_d^{(m)} \leftarrow (K+1)\, \mathbf{1}\{ r_d > \max_k r_k^{z,(m)} \}$ for $d = 1, \dots, D$
end for
$\bar{e}_d \leftarrow \frac{1}{M} \sum_{m=1}^{M} e_d^{(m)}$ for $d = 1, \dots, D$ ▷ $M = 1$ recovers the single-draw rule
order $\bar{e}_{(1)} \geq \cdots \geq \bar{e}_{(D)}$; $k^\star \leftarrow \max\{\, k : \bar{e}_{(k)} \geq D/(kq) \,\}$ (set $k^\star = 0$ if none)
return $\hat{S} = \{\, d : \bar{e}_d \geq D/(k^\star q) \,\}$ if $k^\star \geq 1$, else $\hat{S} = \varnothing$ ▷ step-up rejection set; for single-draw binary e_d this equals the $k^\star$ largest

and with a coordinate-symmetric initialization and optimizer the resulting score map (5) is column-equivariant. We make this our standing requirement and state the design condition under which references are valid negative controls.

Two caveats on the equivariance hypothesis are in order, since the entire guarantee rests on it. First, equivariance is a property of the realized *fitting procedure*, not merely of the likelihood: it holds for a deterministic optimizer started from a coordinate-symmetric initialization with a symmetrically computed gradient (the configuration we use), in which case the optimizer's trajectory is permutation-equivariant in exact arithmetic; floating-point non-associativity perturbs it only at the level of numerical error. A user who instead perturbs coordinates *independently* across random restarts breaks the symmetry, and this invalidates the exchangeability of Lemma 4.2—hence the FDR guarantee itself, not only its power. Restarts that apply a common (scalar) shift to all lengthscales preserve equivariance and are safe. Second, a weaker distributional form of equivariance, $T(\widetilde{X}_\pi, \boldsymbol{y}) \overset{d}{=} \pi(T(\widetilde{X}, \boldsymbol{y}))$ jointly with the design symmetry, would suffice for the argument and would accommodate symmetric *randomized* optimizers; we state and use the deterministic form for concreteness.

Assumption (A) (Exchangeable non-signal columns)**.** *Conditional on* $\big(\boldsymbol{y},\, \{\boldsymbol{X}_d\}_{d \in S_0}\big)$*, the collection of non-signal columns*

$$\big\{ \boldsymbol{X}_d : d \in S_0^c \big\} \;\cup\; \big\{ \boldsymbol{Z}_k : k = 1, \dots, K \big\}$$

is exchangeable; that is, its joint law is invariant under permutations within the collection.

Assumption (A) is the crux, and it is worth understanding when it holds. It holds whenever the columns $\boldsymbol{X}_1, \dots, \boldsymbol{X}_D, \boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K$ are independent and identically distributed and the response depends on the design only through the active columns. Indeed, in that case $\boldsymbol{y}$ is a function of $\{\boldsymbol{X}_d\}_{d \in S_0}$ and exogenous noise, so conditioning on $(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d \in S_0})$ leaves the remaining columns—true nulls and references alike—as i.i.d. draws from the common column law, hence exchangeable. Independent-coordinate designs are the norm in computer experiments, where space-filling plans (Latin hypercube, factorial, Sobol') are constructed to have uncorrelated columns, and they are precisely the setting in which the inert-reference device was first proposed

(Linkletter et al, 2006). The assumption can fail when null inputs are correlated with active ones; we analyze that case empirically in Section 6 and theoretically in Section 8.

4.1 Exchangeability of relevance scores

Lemma 4.2 (Exchangeability of scores)**.** *Let T be column-equivariant and suppose Assumption (A) holds. Then, conditional on $\big(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}\big)$, the scores*

$$\big\{r_d : d \in S_0^c\big\} \;\cup\; \big\{r_k^z : k = 1, \dots, K\big\}$$

are exchangeable. In particular, for each null input $d \in S_0^c$, the score r_d is exchangeable with the reference scores $r_1^z, \dots, r_K^z$.

The proof, given in Appendix A, conditions additionally on the unordered multi-set of non-signal columns: Assumption (A) makes the assignment of labels to these columns uniformly random, and equivariance turns that uniform labeling into an exchangeable score vector. The lemma is a statement about distributional symmetry; it makes no reference to whether $\hat{\ell}_d$ estimates any population quantity. We return to this point in Section 4.3.

Proposition 4.3 (Valid p-values)**.** *Under the hypotheses of Lemma 4.2, the rank p-value (6) is super-uniform under the null, conditionally and marginally: for every null input $d \in S_0^c$ and every $t \in [0, 1]$,*

$$\mathbb{P}\big(p_d \le t \;\big|\; \boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}\big) \le t, \qquad \textit{hence} \qquad \mathbb{P}(p_d \le t) \le t.$$

Consequently the e-value (7) satisfies $\mathbb{E}[e_d \mid \boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}] \le 1$ and $\mathbb{E}[e_d] \le 1$ for null d.

4.2 False discovery rate control

The conditional super-uniformity of Proposition 4.3 is exactly what the FDR engines require. We state three guarantees of increasing robustness and decreasing assumptions.

Theorem 4.4 (e-BH control under arbitrary dependence)**.** *Suppose Assumption (A) holds and T is column-equivariant. Let $\{e_d\}_{d=1}^D$ be the e-values (7) from a single reference draw, and let $\hat{S}$ be the output of e-BH at level q. Then*

$$\mathrm{FDR} \;=\; \mathbb{E}\left[\frac{|\hat{S} \cap S_0^c|}{|\hat{S}| \vee 1}\right] \;\le\; q\,\frac{m_0}{D} \;\le\; q,$$

regardless of the dependence among the scores $\{r_d\} \cup \{r_k^z\}$ induced by the joint fit.

Theorem 4.4 is our headline guarantee. Because e-BH controls FDR for *any* collection of e-values that are marginally valid under their nulls (Wang and Ramdas, 2022), and Proposition 4.3 supplies exactly that marginal validity, the dependence among the jointly-fit scores—which is intricate and difficult to characterize—never needs to be analyzed. This is the practical payoff of routing the inert-reference device through e-values.

Theorem 4.5 (BH control under an independent-score scheme)**.** *Suppose the relevance scores are produced by a scheme under which, conditional on* $\big(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}\big)$*, the null rank p-values* $\{p_d\}_{d\in S_0^c}$ *are (i) each super-uniform and (ii) mutually independent. Then BH at level* q *applied to* $\{p_d\}_{d=1}^{D}$ *controls* FDR $\le q\, m_0/D$*.*

A concrete scheme satisfying (i)–(ii) under an independent-coordinate design is the following per-input *construction: for each input* d*, draw an independent reference block* $Z^{(d)}$ *and fit a column-equivariant relevance map separately to the reduced design* $\big[\,\{\boldsymbol{X}_j\}_{j\in S_0} \mid \boldsymbol{X}_d \mid Z^{(d)}\,\big]$*—the active columns, the single test column* d*, and its own references—reading off* p_d *as the rank of* r_d *among the reference scores. With* $\{(\boldsymbol{X}_d, Z^{(d)})\}_{d\in S_0^c}$ *mutually independent given* $\big(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}\big)$*, conditions (i)–(ii) hold.*

Remark 1 (The single-fit procedure of Algorithm 1 does not meet condition (ii))**.** *Theorem 4.5 does* not *cover the shared-reference, single-fit* CARD-*BH of Algorithm 1. There a single joint fit makes every* p_d *a function of the* entire *augmented design through the inversion of* K_θ*, so the null p-values share the random null columns and are dependent; conditioning on* $(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0})$ *does not fix the remaining null columns, and independent reference* draws *do not decouple scores that are coupled through the shared real design. Moreover super-uniformity (Proposition 4.3) requires the null test column to be random and exchangeable with the references, whereas conditional independence across tests requires the design held fixed; for a full-design joint fit these requirements are incompatible. Consequently the only rigorous guarantee for the single-fit procedure is the e-BH bound of Theorem 4.4; the FDR control of single-fit* CARD-*BH is an empirical finding (Section 6), consistent with the PRDS-like behaviour discussed in Remark 2 but not established as a theorem.*

Remark 2 (Shared references, positive dependence, and why e-values are the right tool)**.** *With a single shared reference pool the p-values (6) are dependent, since they share the references; they coincide with one-sided conformal p-values computed against a common calibration set. For conformal p-values whose scores are computed* independently *per test point against the shared pool, this dependence is positively regression dependent on the null set (PRDS) (Bates et al, 2023), the condition under which BH retains control (Benjamini and Yekutieli, 2001). The ARD scores, however, are produced by a* joint *fit: the relevance of one column depends on all others through the inversion of* K_θ*, and the resulting "explaining-away" between correlated columns can induce negative dependence that PRDS does not accommodate. We do not claim PRDS holds—nor, conversely, that it provably fails—for jointly fitted relevances; establishing either appears delicate, and our own experiments (Section 6) show* CARD-*BH controlling FDR empirically even under correlation, consistent with PRDS-like behavior in practice but short of a guarantee. This ambiguity is precisely what makes e-values the appropriate instrument here: the e-BH guarantee of Theorem 4.4 requires* only *marginal validity of each null e-value and is indifferent to the sign and structure of the dependence. We therefore designate* CARD-*e the primary, assumption-free procedure, present* CARD-*BH as a higher-power option whose shared-reference control we verify empirically, and retain* CARD-*BY as a dependence-agnostic but conservative p-value fallback.*

Theorem 4.6 (Derandomization)**.** *Draw* M *reference designs* $Z^{(1)}, \dots, Z^{(M)}$*, conditionally independent and identically distributed given the data* $(\boldsymbol{y}, X)$*, and let* $e_d^{(m)}$ *be*

the e-value (7) *from draw* m*. The average* $\bar{e}_d = M^{-1}\sum_{m=1}^{M} e_d^{(m)}$ *is a valid e-value for null* d *(i.e.* $\mathbb{E}[\bar{e}_d] \le 1$*), and e-BH applied to* $\{\bar{e}_d\}_{d=1}^{D}$ *at level* q *controls* FDR $\le q\, m_0/D \le q$*. Writing* $\mu_d(\boldsymbol{y}, X) = \mathbb{E}[e_d^{(1)} \mid \boldsymbol{y}, X]$ *and* $v_d(\boldsymbol{y}, X) = \mathrm{Var}(e_d^{(1)} \mid \boldsymbol{y}, X)$*, the marginal variance of the averaged e-value is*

$$\mathrm{Var}(\bar{e}_d) \;=\; \underbrace{\mathrm{Var}\big(\mu_d(\boldsymbol{y}, X)\big)}_{\textit{data-induced}} \;+\; \frac{1}{M}\;\underbrace{\mathbb{E}\big[v_d(\boldsymbol{y}, X)\big]}_{\textit{reference-induced}}\;, \tag{8}$$

so that $\mathrm{Var}(\bar{e}_d) \downarrow \mathrm{Var}(\mu_d(\boldsymbol{y}, X))$ *as* $M \to \infty$*.*

Equation (8) corrects a naive "$1/M$" reading: because $\boldsymbol{y}$ and the active columns are shared across all M fits, the draws $e_d^{(1)}, \ldots, e_d^{(M)}$ are positively dependent (only conditionally i.i.d. given the data), and averaging removes the reference-induced (conditional) variance but not the data-induced variance $\mathrm{Var}(\mu_d(\boldsymbol{y}, X))$, which is the irreducible floor. Derandomization thus stabilizes the selection with respect to the auxiliary randomness of the references—the source of run-to-run irreproducibility that motivates it—without claiming to drive the total variance to zero. A related use of e-values to derandomize a selection procedure—there for the knockoff filter rather than for relevance scores—appears in Vásquez et al (2025).

4.3 Robustness to non-identifiability

A recurring concern about lengthscale-based relevance is that the ARD lengthscales are only weakly identified. In the infill (fixed-domain) regime, the SE-kernel parameters σ_f^2 and ℓ_d are not separately consistently estimable; only a microergodic combination is (Zhang, 2004; Stein, 1999). One might fear that this undermines any procedure built on $\hat{\ell}_d$.

It does not undermine *ours*. The guarantees of Theorems 4.4–4.6 flow entirely from Lemma 4.2, whose hypotheses are (i) column-equivariance of the relevance map and (ii) exchangeability of the non-signal columns. Neither hypothesis mentions consistency, identifiability, or the existence of a population target for $\hat{\ell}_d$. The score r_d enters only as an exchangeable label: what matters is that the fitting procedure treats a genuinely null input and a reference symmetrically, which equivariance guarantees no matter how poorly $\hat{\ell}_d$ pins down a "true" lengthscale. Formally, even if $\hat{\ell}_d$ is an arbitrarily bad estimator of any putative $\ell_d^\star$, the *realized value* returned by the deterministic-given-data optimizer is exchangeable with the reference values, so p_d is super-uniform and FDR is controlled.

Non-identifiability instead manifests as a *power* phenomenon. If the relevance score separates active inputs from the null floor only weakly—because the identifiable signal in the lengthscales is faint—then active p-values are not small and few discoveries are made, but no false-discovery guarantee is violated. The same robustness covers two further ARD ambiguities. First, the score is defined up to a strictly monotone transform: $1/\hat{\ell}_d$ and $1/\hat{\ell}_d^2$ induce identical ranks and hence identical p-values (6), so the arbitrary choice of relevance scale is immaterial. Second, an overall confounding between σ_f^2 and the lengthscales rescales all coordinates together and cancels in the rank comparison. In short, the exchangeability route purchases validity with symmetry

rather than with estimation, which is why it is well suited to a model whose parameters are notoriously hard to estimate.

Proposition 4.7 (Selection consistency under separation)**.** *Consider* CARD-*e with reference count K and let $\eta = \max_k r_k^z$ be the reference maximum, with D, K, and the active set S_0 held fixed as $N \to \infty$. Suppose that (i) every active input satisfies $\mathbb{P}(r_d > \eta) \to 1$ as $N \to \infty$ (signal separation), and (ii) the sparsity obeys $|S_0| \geq D/((K+1)q)$. Then $\mathbb{P}(S_0 \subseteq \hat{S}) \to 1$ while* FDR $\leq q$*; the procedure is selection-consistent with asymptotic FDR control.*

Condition (i) is the natural requirement that the relevance score eventually places active inputs above the reference cloud—the regime in which ARD selection is meaningful at all—and condition (ii) states that the reference count is large enough, relative to sparsity, for the e-BH threshold to admit the active set. The proof (Appendix A) follows from the e-value mechanics: under (i), every active input attains $e_d = K + 1$ with probability tending to one, and under (ii) the e-BH cutoff $D/(kq)$ is cleared at $k = |S_0|$.

Remark 3 (On the separation condition)**.** *Condition (i) is a signal-strength ("beta-min") condition of the kind that is standard—and, in general, unavoidable—in the variable-selection-consistency literature, where consistency of the lasso, of knockoffs, and of GP selection alike is established only under an explicit lower bound on the strength of the active signals (Jiang and Tokdar, 2021; Candès et al, 2018). It is therefore a hypothesis, not a tautology, and it is deliberately stated at the level of the relevance* scores *rather than the lengthscales. A sufficient primitive condition that does* not *invoke lengthscale consistency is the following: if there exist constants $\eta^* < \rho^*$ such that the reference relevance distribution has support eventually below η^* (so $\max_{k \leq K} r_k^z \leq \eta^*$ with probability tending to one, K fixed) while each active relevance satisfies $\mathbb{P}(r_d \geq \rho^*) \to 1$, then (A1)-style domination gives (i) by the continuous mapping theorem. This requires only that active scores be* bounded away from *null scores in the limit—a far weaker demand than $\hat{\ell}_d \to \ell_d^*$, and one compatible with the non-identifiability of the lengthscales themselves.*

One could *instead try to derive (i) from estimator consistency: in a regime where the lengthscales* are *consistently estimable—for example under classical increasing-domain asymptotics for stationary covariance models, a setting distinct both from the sparse ARD selection problem and from the fixed-domain regime that motivates Section 4.3—active relevances would converge to positive constants while null and reference relevances vanish, yielding separation. (The Bayesian selection-consistency result of Jiang and Tokdar (2021), by contrast, holds in a fixed-domain random-design regime and is a statement about posterior inclusion, not about consistency of the marginal-likelihood lengthscale estimates; it does not supply this route.) We do not adopt the consistency-based justification as the primitive one, for two reasons. First, it is regime-specific: in the fixed-domain regime the lengthscales are* not *consistently estimable (Zhang, 2004), yet separation—and hence power—may still hold through the microergodic signal. Second, and more importantly, grounding power in lengthscale consistency would re-import exactly the identifiability assumptions that the validity*

guarantee was constructed to avoid. Keeping the two on separate footings is deliberate: validity *(Theorems 4.4–4.6) holds with no separation and no consistency, whereas* power *is governed by a transparent separation condition, however it is realized.*

5 Computational complexity and scalability

Because CARD wraps a calibration layer around a single ARD-GP fit rather than introducing a new estimator, its cost is governed by the host GP and is straightforward to characterize, and its validity interacts with scalable approximations and finite-precision arithmetic in ways a practitioner must respect.

5.1 Time complexity

Each augmented fit evaluates the marginal likelihood (4) and its gradient at every optimizer step. The cost is dominated by the Cholesky factorization of the $N \times N$ matrix $K_\theta + \sigma_n^2 I$, an $O(N^3)$ operation; given the factorization, the solve for $K_\theta^{-1}\boldsymbol{y}$ is $O(N^2)$ and each of the $D+K$ partial derivatives costs $O(N^2)$, for an $O((D+K)N^2)$ gradient on top of the factorization. Over I optimizer iterations a single fit is thus $O(I\,N^3)$, the cubic term dominating whenever $N \gtrsim D+K$. Single-draw CARD-e and CARD-BH inherit this $O(I\,N^3)$ cost; the M-draw derandomized variant is $O(M\,I\,N^3)$ with the M fits embarrassingly parallel; and the model-X variant fits a $2D$-column design at the same $O(I\,N^3)$ order, preceded by a one-time $O(D^3)$ covariance factorization and, for a general positive-semidefinite constraint, a semidefinite program for the knockoff vector s (avoided by the closed-form equicorrelated choice). The reference count enters only through the design width $D+K$ and the $O((D+K)N^2)$ gradient term, so the recommended $K = 2D$ leaves the $O(N^3)$ bottleneck untouched. The calibration layer itself—reference draws, the rank p-values (6), the indicator e-values (7), and the step-up—adds only $O(D+K)$ sorting and counting per draw and is negligible. Exact fitting is therefore confined to N in the low thousands, the standard exact-GP regime. Table 1 reports measured times: one fit takes seconds at $N = 500$ and under a minute at $N = 2{,}000$ with the matrix-product implementation, the growth from $N = 1{,}000$ to $N = 2{,}000$ (a factor of 6.5–7) matching the cubic term, and the model-X knockoff competitor costs less per fit only because its $2D$-column design is narrower than the $D+K$-column augmented design.

5.2 Scalable approximations and column symmetry

The exchangeability guarantee of Lemma 4.2 is a statement about the *symmetry* of the relevance map, not about exact matrix inversion, so it survives any approximation that treats the input coordinates symmetrically. Variational sparse GPs with inducing points (Titsias, 2009) and nearest-neighbor GPs (Datta et al, 2016) replace K_θ by a structured surrogate and lower the per-fit cost to $O(N M_{\text{ind}}^2)$ for M_{ind} inducing points or neighbors. The caveat is sharp and must be checked rather than assumed: *the surrogate must be invariant to relabeling of the input columns.* Inducing-point locations optimized jointly with the lengthscales, or neighbor sets fixed by an input-dependent ordering, can inject a coordinate asymmetry that breaks Definition 4.1 and with it

Table 1 Wall-clock seconds for one fit on a synthetic design with $D = 20$ inputs, $K = 40$ references and a nonlinear response, on one core of an Intel Xeon 2.10 GHz processor with a single BLAS thread. "Original" is the pairwise-difference implementation used in Section 6; "matrix product" the algebraically identical implementation used in Section 7; the model-X knockoff competitor constructs Gaussian knockoffs and fits the $2D = 40$-column design with the matrix-product implementation. Single-draw CARD-e and CARD-BH cost one fit; the M-draw derandomized variant costs M fits

N	CARD fit, original	CARD fit, matrix product	model-X knockoff
250	1.3	0.9	0.4
500	4.3	2.7	1.8
1,000	20.5	8.0	5.5
2,000	132.6	55.8	20.7

the FDR guarantee—not merely the power. Constructions that place and update the surrogate by a coordinate-symmetric rule (for instance shared inducing inputs initialized and optimized identically across coordinates, or a neighbor rule defined through a permutation-invariant distance) retain equivariance, and for these Lemma 4.2 and Theorems 4.4–4.6 hold verbatim, extending exact finite-sample control to $N = 10^5$ and beyond. A full empirical study of sparse CARD is left to future work; the point here is the design constraint, easy to satisfy but fatal to ignore.

5.3 Exact equivariance under finite-precision arithmetic

Validity requires the *realized* relevance map to be column-equivariant, and in finite precision this is an idealization worth stating plainly. With a coordinate-symmetric initialization and a symmetrically computed gradient—the configuration we use—the optimizer trajectory is permutation-equivariant in exact arithmetic. Floating-point non-associativity in the underlying BLAS/LAPACK routines can nonetheless make the factorization of a column-permuted design differ from the permutation of the original factorization at the level of machine epsilon, so two columns with identical conditional law can receive infinitesimally different lengthscale estimates. Three practices keep this immaterial. First, converge the optimizer to a strict gradient tolerance, so residual asymmetries fall far below the spacing of the rank statistics that enter (6)—only the *order* of the scores matters, and generic separations dwarf perturbations of size 10^{-10}. Second, if random restarts are used, apply each perturbation symmetrically across coordinates (a common scalar shift of the log-lengthscales), which preserves equivariance exactly; coordinate-wise independent restarts must be avoided, as they break validity rather than merely degrade power. Third, fix the linear-algebra backend and thread count so the factorization is deterministic across the permutations compared within a run. Under these practices the rank p-values are reproducible and the guarantees of Section 4 hold to numerical precision.

5.4 Implementation and reproducibility

The accompanying implementation fits the ARD-GP with a deterministic quasi-Newton optimizer (L-BFGS-B) from a coordinate-symmetric initialization on the

CPU, forming the marginal likelihood and gradient through a Cholesky factorization, and then applies the calibration layer described above. All randomness—reference draws, simulated designs and responses, and any restart perturbations—is governed by explicit seeds recorded with each experiment, and the number of BLAS threads is pinned so that the factorizations, and hence the rank statistics, are reproducible within a run. Each reported figure and table is generated by a single script from its recorded seed. We keep the inference path on the CPU by design: a GPU or automatic-differentiation backend would require the corresponding device-level determinism controls—deterministic kernels and a fixed floating-point reduction order—to preserve the exact column-equivariance on which validity rests, a guarantee that is materially harder to enforce amid the massively parallel, non-associative accumulation of GPU linear algebra. The code implementing CARD-e, CARD-BH and the derandomized dCARD-e, together with the scripts that regenerate every experiment and the data illustration and their recorded outputs, is provided as Online Resource 1 and will be released as a documented open-source package upon publication. An AI assistant (Claude, Anthropic) was used to draft and revise parts of the text, to write and run the analysis code for Section 7, and to check proofs and references; the author verified all content, code and results and takes full responsibility for the manuscript.

6 Simulation studies

We assess finite-sample behavior on synthetic data with known active sets. Inputs are drawn on $[0,1]^D$ with independent coordinates (Experiments 1–2) or with correlation (Experiments 3–4); the regression function is additive over the active inputs, $f_0(\boldsymbol{x}) = \sum_{j \in S_0} \sin(2\pi x_j)$, with the active set $S_0 = \{1, \ldots, s\}$, and Gaussian noise is added at a prescribed signal-to-noise ratio (SNR), defined as the ratio of the variance of f_0 to the noise variance. We standardize $\boldsymbol{y}$. The ARD-GP is fit by maximizing (4) with a coordinate-symmetric large-lengthscale initialization, using an analytic-gradient L-BFGS optimizer. Unless noted we use $K = 2D$ references, target level $q = 0.2$, and 40 replications, reporting empirical FDR and power with Monte Carlo standard errors shown as error bars. We compare the three CARD variants against three baselines computed from the same fits: the uncorrected reference rule, which we label RDVS (reference-distribution variable selection), declaring input d relevant when $p_d \leq 0.05$ with no multiplicity correction, in the spirit of Linkletter et al (2006); the ARD "elbow" rule, which thresholds at the largest gap in the sorted scores; and a model-X knockoff filter (Candès et al, 2018; Barber and Candès, 2015) using the ARD relevance as importance statistic. The shadow-feature selector of Kursa and Rudnicki (2010) confirms a feature when its relevance exceeds the maximum reference relevance and judges the resulting hit counts by a binomial test; this is the uncalibrated antecedent of CARD-e, which replaces the binomial rule by the e-BH step-up and thereby targets a prescribed FDR rather than a fixed per-feature error. Figure 1 illustrates the calibration on a single fit: the active inputs sit well above the reference cloud, the null inputs within it, and the reference maximum supplies the calibrated cutoff. Table 2 consolidates the headline numbers across representative cells; the figures that follow trace the full curves.

Table 2 Empirical FDR and power (target $q = 0.2$, 40 replications) at representative cells. Experiment 1 varies SNR at $D = 20$; Experiment 2 varies D at SNR $= 8$. The uncorrected RDVS rule breaches q as D grows (boldface), while all CARD variants remain controlled. Monte Carlo standard errors (40 replications) are at most 0.06 for FDR and 0.08 for power across these cells; per-cell error bars are shown in the corresponding figures.

	Experiment 1 ($D = 20$)				Experiment 2 (SNR $= 8$)			
	SNR $= 2$		SNR $= 8$		$D = 40$		$D = 60$	
Method	FDR	Power	FDR	Power	FDR	Power	FDR	Power
CARD-e	0.06	0.28	0.06	0.64	0.09	0.41	0.07	0.32
CARD-BH	0.11	0.49	0.15	0.91	0.16	0.63	0.13	0.57
CARD-BY	0.00	0.00	0.01	0.05	0.01	0.03	0.00	0.00
Model-X knockoff	0.14	0.42	0.16	0.74	0.12	0.45	0.16	0.60
RDVS (uncorrected)	0.10	0.52	0.08	0.75	**0.19**	0.81	**0.25**	0.91
ARD elbow	0.03	0.27	0.00	0.31	0.00	0.29	0.00	0.29

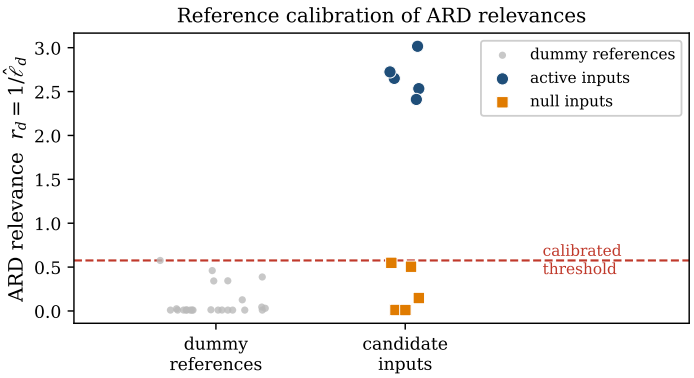


Fig. 1 Reference calibration on a single ARD-GP fit ($N = 200$, $D = 12$, $s = 5$, $K = 24$). Synthetic reference (dummy) scores form an empirical null cloud; active inputs (circles) exceed it, null inputs (squares) fall within it, and the reference maximum (dashed line) is the calibrated cutoff underlying the e-value (7)

6.1 Experiment 1: error control and power versus signal strength

We fix $D = 20$, $s = 8$, $N = 200$, $K = 40$ and vary SNR over $\{1, 2, 4, 8, 16\}$. Figure 2 reports the results. All three CARD variants control FDR at every SNR: CARD-e holds the realized FDR between 0.04 and 0.06, CARD-BH between 0.08 and 0.15, and CARD-BY near zero. Power rises with SNR, and a clear ordering emerges among the controlled procedures. CARD-BH is the most powerful, reaching power 0.91 at SNR 8 while holding FDR at 0.15; CARD-e is more conservative (power 0.64 at SNR 8, FDR 0.06), the price of the assumption-free e-BH guarantee; and CARD-BY is markedly conservative (power below 0.05), as expected from the logarithmic BY penalty at $D = 20$. The ARD elbow rule cannot adapt: its power saturates near 0.31 regardless of SNR because the largest-gap heuristic is insensitive to the target level. The knockoff filter controls FDR

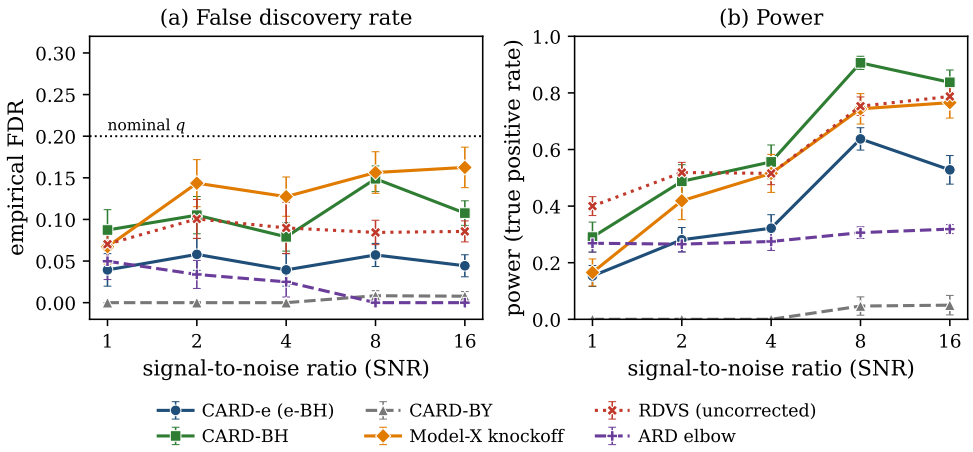


Fig. 2 Experiment 1 ($D = 20$, $s = 8$, $K = 40$, $q = 0.2$). (a) Empirical FDR and (b) power versus signal-to-noise ratio. All CARD variants control FDR (dotted line) at every SNR; CARD-BH attains the highest power among controlled procedures, while the uncorrected RDVS rule and the ARD elbow do not target the level

but is the least powerful at low SNR (power 0.17 at SNR 1), reflecting the well-known difficulty of the knockoff filter when the number of true signals is small relative to $1/q$.

6.2 Experiment 2: robustness to dimension and failure of the uncorrected rule

This experiment isolates the contribution of multiplicity correction. We fix $s = 8$, SNR $= 8$, $N = 220$ and grow the number of candidate inputs $D \in \{15, 20, 30, 40, 60\}$ with $K = 2D$, so that the number of null inputs increases while the signal is held fixed. Figure 3 shows the consequence. The uncorrected RDVS rule degrades exactly as multiple-testing theory predicts: its realized FDR climbs monotonically from 0.035 at $D = 15$ to 0.090, 0.128, 0.187, and 0.249 at $D = 60$, breaching the target $q = 0.2$ once the null count is large enough. The CARD procedures, by contrast, hold the line at every dimension: CARD-e between 0.035 and 0.090 and CARD-BH between 0.077 and 0.161, both always at or below q. Power declines gracefully with dimension for all methods—CARD-BH from 0.90 to 0.57, CARD-e from 0.69 to 0.32—as the fixed signal is diluted among more candidates and the GP fit in higher dimension becomes harder. The contrast in panel (a) is the empirical case for the paper: the informal reference device and the calibrated one have similar power, but only the calibrated one controls the error rate as the problem scales.

6.3 Experiments 3–4: correlated designs

Assumption (A) requires exchangeable non-signal columns, which independent references guarantee only for independent designs. We probe the consequences of correlation in two complementary regimes; Figure 4 summarizes both.

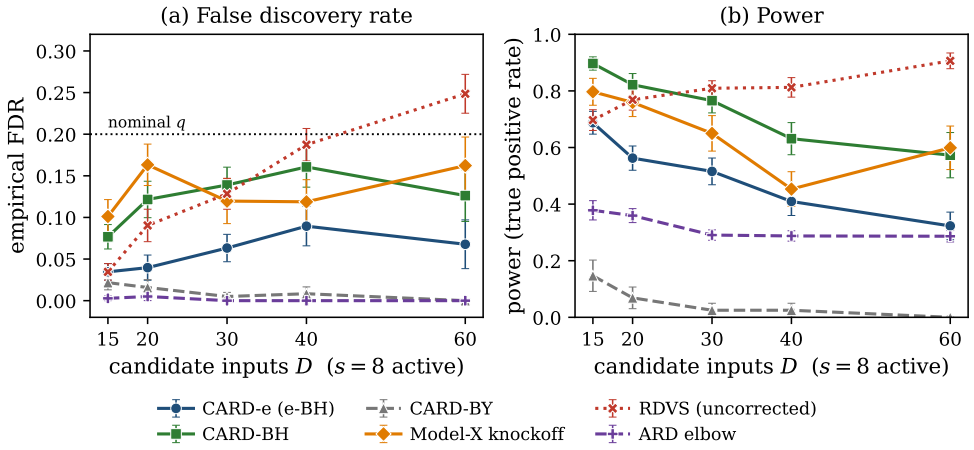


Fig. 3 Experiment 2 ($s = 8$, SNR $= 8$, $K = 2D$, $q = 0.2$). (a) The uncorrected RDVS rule breaches the target FDR as the number of null inputs grows, reaching 0.25 at $D = 60$, whereas CARD-e and CARD-BH remain controlled. (b) Power declines gracefully with dimension for all methods

Experiment 3 uses an autoregressive design, $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{jk} = \rho^{|j-k|}$ and a contiguous active block, and varies $\rho \in \{0, 0.3, 0.5, 0.7, 0.9\}$. Here CARD-e with independent references remains controlled across the whole range, with FDR between 0.027 and 0.055 (Figure 4a), even though correlation formally violates Assumption (A). The reason is instructive: in a *joint* GP fit, a null input that is correlated with an active one but conditionally uninformative given that active input is "explained away"—it receives low relevance because the active input already accounts for the response—so it does not exceed the reference cloud. Joint ARD thus provides empirical robustness to mere correlation that marginal screening would lack.

This robustness is not unconditional. Experiment 4 constructs an adversarial design in which $s = 8$ active inputs are accompanied by 8 near-duplicate proxies, $\boldsymbol{Z}_j = \sqrt{r}\,\boldsymbol{X}_j + \sqrt{1-r}\,\xi_j$ for active $\boldsymbol{X}_j$ and independent noise ξ_j, plus 8 independent nulls, and varies the proxy correlation r. A near-duplicate of an active input is conditionally null but carries almost the same predictive information as its active twin, so the joint fit can no longer cleanly explain it away, and the proxy's relevance can exceed that of independent references. Figure 4b shows CARD-e (with independent references) losing control as r grows: its FDR rises from 0.063 at $r = 0.5$ to 0.148 at $r = 0.8$ and 0.231 at $r = 0.95$, breaching $q = 0.2$ in the near-duplicate regime. The remedy is to abandon independent references for *structured* ones that reproduce the conditional law of the design—that is, model-X knockoffs (Candès et al, 2018). Replacing the independent references with Gaussian knockoffs constructed from the (here known) covariance restores control across the entire range, with FDR 0.197, 0.134 and 0.099 at $r = 0.5, 0.8, 0.95$, at the expected cost of reduced power (from 0.93 to 0.32 as the proxies make the active twins harder to distinguish). The honest reading is that independent-reference CARD is a cheap, single-fit procedure that is valid for

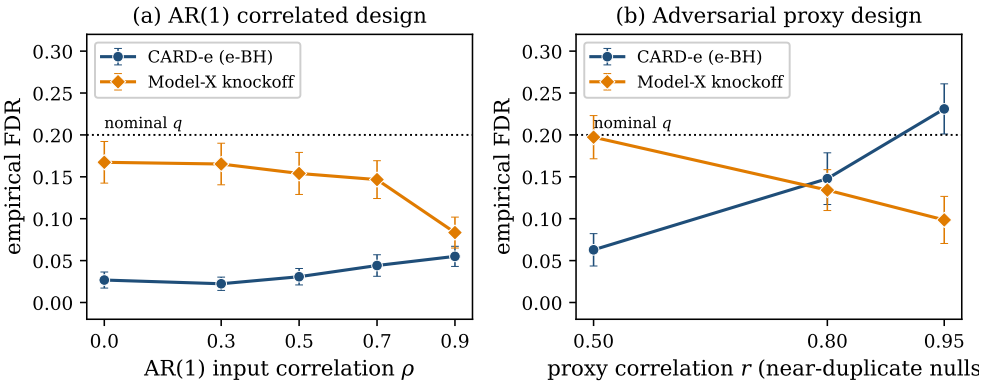


Fig. 4 Correlated designs ($D = 20$, $s = 8$, SNR $= 8$, $q = 0.2$). (a) Experiment 3: under AR(1) correlation, CARD-e with independent references stays controlled because joint ARD explains away correlated-but-null inputs. (b) Experiment 4: with near-duplicate proxy nulls, independent references fail as the proxy correlation r grows, while model-X knockoff references restore control

independent designs and empirically forgiving of moderate correlation, but that guaranteed control under strong correlation requires the heavier knockoff construction of Section 8.

7 Real data illustration

Data and design. We illustrate the procedure on the U.S. National Health and Nutrition Examination Survey (NHANES) 2017–March 2020 pre-pandemic files, released by the National Center for Health Statistics in 2021 and in the public domain (Centers for Disease Control and Prevention, National Center for Health Statistics, 2021; Stierman et al, 2021). The response is log glycohemoglobin (HbA1c, %), a marker of long-term glycaemic control, and the $D = 10$ candidate inputs are age, sex, body-mass index (BMI), mean systolic and diastolic blood pressure (SBP, DBP; the average of up to three oscillometric readings), and five measurements from the standard serum biochemistry profile: total cholesterol (TC), log triglycerides (TG), uric acid (urate), and the log liver enzymes alanine aminotransferase (ALT) and γ-glutamyl transferase (GGT). The design thus pairs demographics, adiposity, blood pressure and serum chemistry, as in the classical diabetes data of Efron et al (2004), but in a contemporary cohort. We retain adults aged 20 or older who were not pregnant, reported no insulin or oral glucose-lowering medication and have complete data, leaving 5,996 participants. Waist circumference is omitted because it is nearly collinear with BMI ($r = 0.91$); the largest remaining correlation is 0.64 (SBP with DBP). Because the aim is methodological, survey weights are not used, and no result below is a nationally representative estimate. In each of 100 repetitions we draw a subsample of $N = 1{,}000$ participants, standardize the inputs and the response, append $P = 10$ planted null variables drawn independently of the data—genuine nulls for which Assumption (A)

holds by construction—and add $K = 40$ references, giving 20 candidates. Every repetition uses a fresh subsample, fresh planted nulls and fresh references. The GP is fitted with the same objective, gradient, initialization and optimizer settings as in Section 6, using an algebraically identical but faster implementation that forms the kernel matrix and the gradient from matrix products rather than from the tensor of pairwise differences.

Real response. Figure 5a shows the mean ARD relevances. Age leads (mean relevance 0.79), followed by log GGT (0.56), urate (0.52) and BMI (0.40), while log ALT, DBP and TC are the least relevant inputs (0.18, 0.16 and 0.15). Sex has a high but erratic mean relevance (0.60, standard error 0.16), as the lengthscale of a binary input is weakly determined, and it exceeds every reference in only 19% of repetitions. The signal is weak overall. The fitted noise variance collapses towards zero in 41% of the fits, a symptom of the weak identifiability of the ARD hyperparameters discussed in Section 4.3, and the largest of the 40 reference relevances averages 0.66, above the mean relevance of every input except age. Because an input receives a nonzero e-value only by exceeding every reference, CARD-e is correspondingly cautious. At $q = 0.2$ it makes selections in 43 of the 100 repetitions, choosing log GGT, urate and age most often (in 31%, 27% and 24% of repetitions), as Figure 5b shows; at $q = 0.1$ it selects in 14 repetitions. These selections agree with the clinical literature: circulating GGT (Kunutsor et al, 2014) and serum uric acid (Kodama et al, 2009) are established prospective risk markers of type 2 diabetes in meta-analyses of cohort studies, and age is a dominant correlate of glycaemia in adults. Validity, by contrast, is directly visible in the planted nulls. By Lemma 4.2 and Proposition 4.3, each planted null should exceed all 40 references with probability $1/(K+1) = 0.024$; the observed rate is 0.022. CARD-e selects 0.17 planted nulls per repetition at $q = 0.2$, and the planted nulls' share of its selections, which is a lower bound on the realized false discovery proportion, averages 0.041. The uncorrected reference rule selects 0.43 planted nulls per repetition, and planted nulls make up 11% of its selections on average.

Semi-synthetic response with known truth. Because the real response has no known active set, we measure power by keeping the real inputs and replacing the response by $y = f + \varepsilon$, where, on the standardized inputs,

$$f \;=\; 0.8\,\mathrm{age} + \sin(1.5\,\mathrm{BMI}) + 0.6\,(\log \mathrm{TG})(\log \mathrm{GGT}) - 0.4\,(\log \mathrm{GGT})^2,$$

and the Gaussian noise ε is scaled so that $\mathrm{SNR} = \mathrm{Var}(f)/\sigma^2 \in \{1, 4\}$. The other six real inputs and the ten planted variables are nulls ($m_0 = 16$ of $D = 20$). Several real nulls are correlated with the signals—log ALT with log GGT ($r = 0.59$), SBP with age ($r = 0.44$)—so, as in Experiment 3, these nulls fall outside Assumption (A). Table 3 reports FDR and power over 100 repetitions per setting. At $q = 0.2$, CARD-e holds the FDR at 0.08–0.09, well below the target, with power 0.97 at SNR = 1 and 1.00 at SNR = 4. CARD-BH reaches power 0.98–1.00 at an FDR of 0.16, just under the target. The uncorrected rule reaches full power at an FDR of 0.14–0.15 in this low-dimensional design, but its false selections grow with the number of nulls (Experiment 2), and the ARD elbow rule avoids false selections only by missing about two thirds of the signals (power 0.34–0.35).

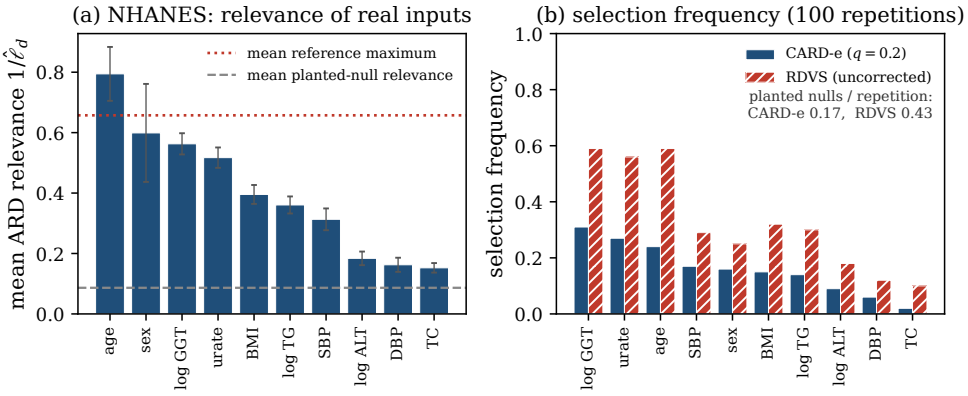


Fig. 5 NHANES 2017–March 2020, real response log HbA1c ($N = 1{,}000$ per repetition, ten real inputs, ten planted nulls, $K = 40$ references, 100 repetitions). (a) Mean ARD relevance of the real inputs ($\pm$ one standard error); the dotted line is the mean of the per-repetition maximum reference relevance, the level an input must exceed to earn a nonzero e-value, and the dashed line is the mean relevance of the planted nulls. (b) Selection frequencies at $q = 0.2$ for CARD-e and for the uncorrected reference rule; planted nulls are selected 0.17 and 0.43 times per repetition, respectively

Table 3 Semi-synthetic study on the NHANES inputs ($N = 1{,}000$, $K = 40$, 100 repetitions per setting). The response depends on age, BMI, log TG and log GGT; the other six real inputs and ten planted variables are nulls ($m_0 = 16$ of $D = 20$). Entries are empirical FDR and power, with Monte Carlo standard errors in parentheses; SNR $= \mathrm{Var}(f)/\sigma^2$. At $q = 0.1$ the two CARD variants coincide in every repetition.

		SNR = 1		SNR = 4	
Method	q	FDR	Power	FDR	Power
CARD-e	0.2	0.08 (0.01)	0.97 (0.01)	0.09 (0.01)	1.00 (0.00)
CARD-BH	0.2	0.16 (0.02)	0.98 (0.01)	0.16 (0.02)	1.00 (0.00)
CARD-e and CARD-BH	0.1	0.08 (0.01)	0.33 (0.05)	0.09 (0.01)	0.37 (0.05)
RDVS (uncorrected)	–	0.14 (0.02)	1.00 (0.00)	0.15 (0.02)	1.00 (0.00)
ARD elbow	–	0.00 (0.00)	0.34 (0.02)	0.00 (0.00)	0.35 (0.02)

At $q = 0.1$ both CARD variants lose most of their power (power 0.33–0.37), as the theory predicts. With $D = 20$ and $K = 40$ the e-BH bar requires at least $D/((K + 1)q) = 4.9$ rejections (Proposition 4.7), so the four signals can be rejected only when a fifth input also exceeds every reference. At SNR $= 4$ all four signals exceed every reference in 99 of the 100 repetitions, yet CARD-e rejects only in the 37 repetitions in which a fifth input does too. The two variants coincide at this level, because the smallest attainable rank p-value, $1/(K + 1)$, imposes the same bar on BH. When few signals are expected at a stringent level, a larger reference count lowers the bar.

Summary. On a contemporary cohort the calibration behaves as the theory predicts. Planted nulls exceed the references at the rate that exchangeability dictates,

false selections stay within budget both for independent planted nulls and for correlated real nulls, and when the signal is weak the procedure declines to select rather than guess.

8 Concluding remarks

We have shown that the long-standing inert-reference heuristic for ARD relevance, properly formalized through an exchangeability argument and routed through e-values, yields variable selection with finite-sample FDR control at essentially the cost of one extra GP fit. The guarantee is attractively modular: it attaches to any column-symmetric relevance score, it is agnostic to the dependence among scores when e-BH is used, and—because it relies on symmetry rather than estimation—it is immune to the weak identifiability that complicates inference about ARD lengthscales.

Choosing the references. Two design choices govern the procedure: the reference count K and, implicitly, the calibrator level α_0. The tuning-free e-value (7) fixes $\alpha_0 = 1/(K+1)$, so K alone controls resolution. Larger K makes the flagged e-value $(K+1)$ larger, which lowers the e-BH admission threshold and lets the procedure act on sparser signals, but it also makes "exceeding all references" more stringent. The rule $K = 2D$ used throughout balances these forces and renders the e-BH bar essentially dimension-free: with binary e-values the bar at $k = |S_0|$ is $D/(|S_0|q) \le K + 1 = 2D + 1$, which rearranges to the requirement $|S_0| \ge D/((K+1)q) \approx 1/(2q)$, independent of D. Proposition 4.7 formalizes this requirement $K \gtrsim D/(q\,|S_0|)$ for power. In practice we recommend K a small multiple of D and, when selection stability matters, the derandomized variant of Theorem 4.6, which averages over reference draws.

Correlated designs and a model-X construction. The principal limitation is Assumption (A). Independent references are exchangeable with null columns only under independent designs; Experiment 4 shows that near-duplicate nulls can defeat them. The remedy preserves the conditional law of the design by drawing *structured* references—model-X knockoffs (Candès et al, 2018). Concretely, for a centered design with column covariance Σ, a Gaussian knockoff matrix is

$$\tilde{X} = X\big(I - \Sigma^{-1}\mathrm{diag}(s)\big) + E, \qquad E \sim \mathcal{N}\big(0,\; 2\,\mathrm{diag}(s) - \mathrm{diag}(s)\,\Sigma^{-1}\mathrm{diag}(s)\big), \quad (9)$$

with E drawn independently of X and the vector $s \ge 0$ chosen so that the covariance in (9) is positive semidefinite (the equicorrelated choice $s_j = \min\{2\lambda_{\min}(\Sigma), 1\}$, or a semidefinite program). By construction $(X, \tilde{X})$ is invariant in distribution under swapping any null column $\boldsymbol{X}_d$ with its knockoff $\tilde{\boldsymbol{X}}_d$, conditional on the response (Candès et al, 2018). This swap-invariance is exactly the hypothesis of Lemma 4.2 applied to the pair $(\boldsymbol{X}_d, \tilde{\boldsymbol{X}}_d)$: the column-equivariant relevance map then makes $(r_d, \tilde{r}_d)$ exchangeable under the null, so the signed statistic $W_d = r_d - \tilde{r}_d$ is symmetric about zero and the knockoff filter—or its derandomized e-value form (Ren and Barber, 2024)—controls FDR. Experiment 4 confirms that (9) recovers control where independent references fail, at the model-X cost of a known or well-estimated input distribution and, empirically, reduced power. Computationally the knockoff route adds only a one-time $O(D^3)$

covariance factorization and, for a general constraint, a semidefinite program for s before any GP fit (Section 5.1); since these are dominated by the $O(I\,N^3)$ GP cost whenever $N \gtrsim D$, the choice is driven by design collinearity rather than runtime—independent references suffice unless near-duplicate columns render Assumption (A) untenable, the regime Experiment 4 isolates. Bridging the cheap independent-reference regime and the robust knockoff regime—for instance through partially structured references that match only low-order moments—is a natural direction.

Computational footprint. Section 5 treats time complexity, the column-symmetry condition under which sparse approximations preserve validity, and exact equivariance in finite precision; the practical message is that CARD inherits the host GP's $O(I\,N^3)$ cost and that any scalable surrogate must be coordinate-symmetric for the guarantee to survive.

Limitations. We collect the assumptions on which the guarantee rests. (i) *Validity is conditional on column-equivariance of the realized relevance map*, which is a property of the fitting procedure, not only of the likelihood; an asymmetric optimizer or coordinate-wise random restarts break it (Section 4). (ii) The clean guarantee requires *exchangeable non-signal columns* (Assumption (A)), which independent references satisfy under independent designs but not under strong correlation; near-duplicate nulls require the heavier model-X construction, at a known-distribution cost and reduced power. (iii) Only CARD-*e (e-BH) is rigorous under the default single shared-reference fit*; the higher-power CARD-BH is provably valid only under the per-input independent-score scheme of Theorem 4.5, and its single-fit control is empirical (Remark 1). (iv) The procedure controls *false discoveries, not power*: weak identifiability of the lengthscales and signal dilution in high dimension reduce power, and the separation/sparsity conditions of Proposition 4.7 are genuine hypotheses; with a weak signal the procedure may select nothing at all (Section 7). (v) Exact fitting inherits the $O(N^3)$ GP bottleneck; scalable surrogates retain validity only if they preserve coordinate symmetry. None of these threatens the finite-sample FDR guarantee within its stated scope, but each delimits it.

Non-stationary or additive kernels would let the relevance score localize relevance to regions of input space, with the exchangeability argument applied conditionally. Sensitivity-based relevance scores (Paananen et al, 2019) slot directly into the framework and may separate active from null inputs more sharply than raw lengthscales, improving power without affecting validity. Finally, because the construction is generic in the relevance map, the same calibration applies to other kernel methods that expose a per-feature importance, offering a route to error-controlled selection beyond Gaussian processes.

Acknowledgments. The author thanks colleagues for helpful discussions.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing interests. The author has no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate. This study is a secondary analysis of de-identified, publicly available NHANES data. The NHANES protocol was approved by the National Center for Health Statistics Research Ethics Review Board, and written informed consent was obtained from all survey participants; no additional ethics approval was required.

Data availability. The NHANES 2017–March 2020 pre-pandemic data analysed in Section 7 are in the public domain and freely available from the National Center for Health Statistics (Centers for Disease Control and Prevention, National Center for Health Statistics, 2021) (files P_DEMO, P_BMX, P_BPXO, P_BIOPRO, P_GHB and P_DIQ); their use is governed by the NCHS Data User Agreement. The accompanying code constructs the analysis cohort from these files and records their checksums. The simulated data sets analysed in Section 6 are generated by the accompanying code from the random seeds recorded therein and contain no personal data.

Code availability. Python code implementing the CARD procedure, the scripts used to produce all simulation results, figures and the NHANES illustration (including the script that builds the analysis cohort from the public CDC files), and their recorded outputs are provided as Online Resource 1 (ESM_1.zip).

Author contributions. J. Cai is the sole author and is responsible for all aspects of the work, including conceptualization, methodology, theory, implementation, experiments, and writing.

Appendix A Proofs

Proof of Lemma 4.2. Condition throughout on $G = \big(\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d\in S_0}\big)$, and let $\mathcal{W} = S_0^c \cup \{D+1, \dots, D+K\}$ index the $m_0 + K$ non-signal columns (true nulls together with references). Let $\mathfrak{S}_{\mathcal{W}}$ denote the subgroup of permutations of $\{1, \dots, D+K\}$ that fix every active index in S_0 and permute the indices in $\mathcal{W}$ among themselves; for $\pi \in \mathfrak{S}_{\mathcal{W}}$ write $\widetilde{X}_\pi$ for the design with columns reordered by π. Two facts combine.

(a) Distributional invariance. By Assumption (A), the family of non-signal columns is exchangeable given G, i.e. for every $\pi \in \mathfrak{S}_{\mathcal{W}}$ and every Borel set $B \subseteq \mathbb{R}^{N\times(D+K)}$,

$$\mathbb{P}\big(\widetilde{X} \in B \mid G\big) = \mathbb{P}\big(\widetilde{X}_\pi \in B \mid G\big),$$

since π leaves the active columns (which are G-measurable) untouched and merely permutes the exchangeable non-signal columns.

(b) Equivariance. Here $T(\,\cdot\,, \boldsymbol{y})$ is taken to be a fixed measurable selection from the set of maximizers of (4)—well defined because the deterministic, coordinate-symmetric optimizer of Section 4 returns a single root under a permutation-invariant tie-breaking rule—so that Definition 4.1 holds pointwise even where the non-convex likelihood admits several optima. By Definition 4.1, $T(\widetilde{X}_\pi, \boldsymbol{y}) = \pi\big(T(\widetilde{X}, \boldsymbol{y})\big)$ for every π. Since $\pi \in \mathfrak{S}_{\mathcal{W}}$ maps $\mathcal{W}$ bijectively onto itself, $\pi^{-1}(j) \in \mathcal{W}$ for every $j \in \mathcal{W}$; restricting the score vector to the coordinates $\mathcal{W}$, the sub-vector $T_{\mathcal{W}} := \big(T(\widetilde{X}, \boldsymbol{y})_j\big)_{j\in\mathcal{W}}$ therefore satisfies $T_{\mathcal{W}}(\widetilde{X}_\pi, \boldsymbol{y}) = \pi\restriction_{\mathcal{W}}\big(T_{\mathcal{W}}(\widetilde{X}, \boldsymbol{y})\big)$, where $\pi\restriction_{\mathcal{W}}$ is the induced permutation of $\mathcal{W}$.

Combining (a) and (b): for any bounded measurable h and any $\pi \in \mathfrak{S}_{\mathcal{W}}$, and using that $\boldsymbol{y}$ is G-measurable so that $h(T_{\mathcal{W}}(\cdot, \boldsymbol{y}))$ is a fixed measurable functional under the conditioning,

$$\mathbb{E}\big[h(T_{\mathcal{W}}(\widetilde{X}, \boldsymbol{y})) \mid G\big] \overset{(a)}{=} \mathbb{E}\big[h(T_{\mathcal{W}}(\widetilde{X}_\pi, \boldsymbol{y})) \mid G\big] \overset{(b)}{=} \mathbb{E}\big[h(\pi \restriction_{\mathcal{W}} T_{\mathcal{W}}(\widetilde{X}, \boldsymbol{y})) \mid G\big],$$

so the conditional law of $T_{\mathcal{W}}$ given G is invariant under every $\pi \restriction_{\mathcal{W}}$ as π ranges over $\mathfrak{S}_{\mathcal{W}}$, i.e. under the full symmetric group on $\mathcal{W}$. That is exactly exchangeability of $\{r_d : d \in S_0^c\} \cup \{r_k^z : k = 1, \dots, K\}$ conditional on G. The marginal statement for a single null d follows by taking h to depend only on the coordinate d and the reference coordinates. □

Proof of Proposition 4.3. Fix a null input $d \in S_0^c$ and condition on $G = (\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d \in S_0})$. By Lemma 4.2 the values $V_0 := r_d$ and $V_1, \dots, V_K := r_1^z, \dots, r_K^z$ are exchangeable. For each $i \in \{0, \dots, K\}$ define the upper rank $R_i = 1 + \sum_{j \neq i} \mathbf{1}\{V_j \geq V_i\}$, so that $(K+1)p_d = R_0$ by (6). We claim that for every realization of $(V_0, \dots, V_K)$ and every integer $c \in \{1, \dots, K+1\}$,

$$\#\{\, i : R_i \leq c \,\} \leq c. \tag{A1}$$

Indeed, $R_i \leq c$ means $\#\{j : V_j \geq V_i\} \leq c$ (counting i itself), i.e. at most c of the $K+1$ values are $\geq V_i$. Order the distinct values $u_1 > u_2 > \cdots$ with multiplicities $n_1, n_2, \dots$. A value equal to u_ℓ has $\#\{j : V_j \geq u_\ell\} = n_1 + \cdots + n_\ell$, so it satisfies $R_i \leq c$ only if $n_1 + \cdots + n_\ell \leq c$. Hence $\{i : R_i \leq c\}$ is a union of complete ties $u_1, \dots, u_L$ with $n_1 + \cdots + n_L \leq c$, and its cardinality is exactly $n_1 + \cdots + n_L \leq c$, proving (A1). This argument makes no genericity assumption and so covers the deterministic ties that can arise when the optimizer of (4) returns equal lengthscales for several coordinates.

By exchangeability, $\mathbb{P}(R_0 \leq c \mid G) = \mathbb{P}(R_i \leq c \mid G)$ for every i, whence

$$\mathbb{P}(R_0 \leq c \mid G) = \frac{1}{K+1} \sum_{i=0}^{K} \mathbb{P}(R_i \leq c \mid G) = \frac{1}{K+1}\, \mathbb{E}\big[\#\{i : R_i \leq c\} \mid G\big] \leq \frac{c}{K+1},$$

the last step by (A1). For $t \in [0, 1]$, taking $c = \lfloor (K+1)t \rfloor$ gives

$$\mathbb{P}(p_d \leq t \mid G) = \mathbb{P}\big(R_0 \leq \lfloor (K+1)t \rfloor \,\big|\, G\big) \leq \frac{\lfloor (K+1)t \rfloor}{K+1} \leq t,$$

and marginalizing over G yields $\mathbb{P}(p_d \leq t) \leq t$. For the e-value, with $\alpha_0 = 1/(K+1)$ and $c = 1$,

$$\mathbb{E}[e_d \mid G] = \frac{\mathbb{P}(R_0 = 1 \mid G)}{\alpha_0} = (K+1)\, \mathbb{P}(R_0 = 1 \mid G) \leq (K+1) \frac{1}{K+1} = 1,$$

and $\mathbb{E}[e_d] \leq 1$ follows. Note $\{R_0 = 1\} = \{r_d > \max_k r_k^z\}$ exactly: $R_0 = 1$ requires *no* reference to tie or exceed r_d, which is the strict-domination event in (7); ties at the maximum make $e_d = 0$, so the indicator is conservatively valid. □

Proof of Theorem 4.4. By Proposition 4.3, each null e-value satisfies $\mathbb{E}[e_d] \le 1$. The e-BH procedure controls the FDR at level $q\,m_0/D \le q$ for any set of e-values whose null members are marginally valid, with no restriction on their joint dependence (Wang and Ramdas, 2022, Theorem 2). Applying this with the null set S_0^c yields the bound. The dependence among $\{r_d\} \cup \{r_k^z\}$ induced by the joint fit is therefore immaterial. □

Proof of Theorem 4.5. Validity of the engine. For independent p-values whose null members are super-uniform and whose non-null members are arbitrary, BH at level q controls FDR $\le q\,m_0/D$ (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Conditioning on $G = (\boldsymbol{y}, \{\boldsymbol{X}_d\}_{d \in S_0})$ and taking expectations transfers the conditional bound to the marginal FDR. It therefore suffices to verify (i) and (ii) for the per-input scheme.

Super-uniformity (i). Fix a null input $d \in S_0^c$. In the fit for d, the non-signal columns are $\{d\} \cup \{\text{references } Z^{(d)}\}$ and the signal columns are $\{\boldsymbol{X}_j\}_{j \in S_0}$. Under an independent-coordinate design with references drawn from the common coordinate law, Assumption (A) holds for this reduced design—the single test column and its references are i.i.d. given G—and the relevance map is column-equivariant by hypothesis. Lemma 4.2 then makes r_d exchangeable with $\{r_k^{z,(d)}\}_{k=1}^K$ conditional on G, and Proposition 4.3 gives $\mathbb{P}(p_d \le t \mid G) \le t$.

Conditional independence (ii). By construction p_d is a measurable function of $(G, \boldsymbol{X}_d, Z^{(d)})$ alone, because no other null column enters the fit for d. Under the independent-coordinate design the null columns $\{\boldsymbol{X}_d\}_{d \in S_0^c}$ are independent of one another and of $\boldsymbol{y}$ given the active columns, and the reference blocks $\{Z^{(d)}\}$ are drawn independently; hence the pairs $\{(\boldsymbol{X}_d, Z^{(d)})\}_{d \in S_0^c}$ are mutually independent given G, and so are the null p-values $\{p_d\}_{d \in S_0^c}$. Conditions (i)–(ii) hold, completing the proof. □

Proof of Theorem 4.6. Each $e_d^{(m)}$ is a valid null e-value by Proposition 4.3, so by linearity $\mathbb{E}[\bar{e}_d] = M^{-1} \sum_m \mathbb{E}[e_d^{(m)}] \le 1$; the average is a valid e-value. Theorem 4.4's invocation of Wang and Ramdas (2022) requires only marginal validity, so e-BH on $\{\bar{e}_d\}$ controls FDR at $q\,m_0/D$. For the variance, condition on the data $(\boldsymbol{y}, X)$. Given $(\boldsymbol{y}, X)$ the augmented fit is a deterministic function of the reference design, so $e_d^{(1)}, \dots, e_d^{(M)}$ are i.i.d. functions of the conditionally i.i.d. draws $Z^{(1)}, \dots, Z^{(M)}$; hence $\mathbb{E}[\bar{e}_d \mid \boldsymbol{y}, X] = \mu_d(\boldsymbol{y}, X)$ and $\mathrm{Var}(\bar{e}_d \mid \boldsymbol{y}, X) = v_d(\boldsymbol{y}, X)/M$. The law of total variance gives

$$\mathrm{Var}(\bar{e}_d) = \mathbb{E}\big[\mathrm{Var}(\bar{e}_d \mid \boldsymbol{y}, X)\big] + \mathrm{Var}\big(\mathbb{E}[\bar{e}_d \mid \boldsymbol{y}, X]\big) = \frac{1}{M}\mathbb{E}[v_d(\boldsymbol{y}, X)] + \mathrm{Var}(\mu_d(\boldsymbol{y}, X)),$$

which is (8). As $M \to \infty$ the first term vanishes, leaving the data-induced floor $\mathrm{Var}(\mu_d(\boldsymbol{y}, X))$. □

Proof of Proposition 4.7. Under separation (i), $\mathbb{P}(r_d > \eta) \to 1$ for each active d, where $\eta = \max_{1 \le k \le K} r_k^z$ is the (random) reference maximum; because condition (i) is stated relative to η itself, no separate bound on the reference cloud is required—the comparison is between each active score and the entire reference set simultaneously. Because the active set S_0 is fixed and finite, a union bound gives $\mathbb{P}\big(\exists\, d \in S_0 : r_d \le \eta\big) \le$

$\sum_{d\in S_0}\mathbb{P}(r_d \le \eta) \to 0$, so the event $\mathcal{E}_N = \{\min_{d\in S_0} r_d > \eta\}$ has $\mathbb{P}(\mathcal{E}_N) \to 1$; on $\mathcal{E}_N$ *every* active input simultaneously attains the maximal e-value $e_d = K+1$ (it exceeds all references). On that event there are at least $|S_0|$ inputs with $e_d = K+1$. Reading e-BH through its step-up rejection set $\hat{S} = \{d : \bar{e}_d \ge D/(k^\star q)\}$ (Algorithm 1), the cutoff index satisfies $k^\star = \max\{k : e_{(k)} \ge D/(kq)\}$; at $k = |S_0|$ the condition $e_{(|S_0|)} = K+1 \ge D/(|S_0|q)$ holds by sparsity condition (ii), so $k^\star \ge |S_0|$ and the threshold $D/(k^\star q) \le D/(|S_0|q) \le K+1$. Every input with $e_d = K+1$ therefore clears the threshold and is selected, giving $\mathbb{P}(S_0 \subseteq \hat{S}) \to 1$. (For the single-draw binary e-values the step-up index lands exactly at the count of maximal e-values, so the "$k^\star$ largest" and threshold rules coincide; the threshold form is stated to cover the derandomized $\bar{e}_d$, which can tie at intermediate levels.) FDR control at q holds throughout by Theorem 4.4. □

References

Appriou T, Rullière D, Gaudrie D (2024) Combination of optimization-free kriging models for high-dimensional problems. Computational Statistics 39(6):3049–3071. https://doi.org/10.1007/s00180-023-01424-7

Barber RF, Candès EJ (2015) Controlling the false discovery rate via knockoffs. The Annals of Statistics 43(5):2055–2085

Bashari M, Epstein A, Romano Y, et al (2023) Derandomized novelty detection with FDR control via conformal e-values. Advances in Neural Information Processing Systems 36

Bates S, Candès E, Lei L, et al (2023) Testing for outliers with conformal p-values. The Annals of Statistics 51(1):149–178

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57(1):289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4):1165–1188

Biswas A, Ramdas A (2026) Improving the adjusted Benjamini–Hochberg method using e-values in knockoff-assisted variable selection. arXiv preprint arXiv:260211610

Boos DD, Stefanski LA, Wu Y (2009) Fast FSR variable selection with applications to clinical trials. Biometrics 65(3):692–700

Candès E, Fan Y, Janson L, et al (2018) Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B 80(3):551–577

Centers for Disease Control and Prevention, National Center for Health Statistics (2021) National health and nutrition examination survey data, 2017–march 2020 pre-pandemic. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD, URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020

Dai C, Lin B, Xing X, et al (2023a) False discovery rate control via data splitting. Journal of the American Statistical Association 118(544):2503–2520

Dai X, Lyu X, Li L (2023b) Kernel knockoffs selection for nonparametric additive models. Journal of the American Statistical Association 118(543):2158–2170

Dance H, Paige B (2022) Fast and scalable spike and slab variable selection in high-dimensional Gaussian processes. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS), pp 7976–8002

Datta A, Banerjee S, Finley AO, et al (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association 111(514):800–812

Efron B, Hastie T, Johnstone I, et al (2004) Least angle regression. The Annals of Statistics 32(2):407–499

Eriksson D, Jankowiak M (2021) High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In: Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI), pp 493–503

Grünwald P, de Heide R, Koolen WM (2024) Safe testing. Journal of the Royal Statistical Society: Series B 86(5):1091–1128

Gu M (2019) Jointly robust prior for Gaussian stochastic process in emulation, calibration and variable selection. Bayesian Analysis 14(3):857–885

He B, Xia D, Pan Y (2024) High dimensional controlled variable selection with model-X knockoffs in the AFT model. Computational Statistics 39(4):1993–2009. https://doi.org/10.1007/s00180-023-01426-5

Jiang S, Tokdar ST (2021) Variable selection consistency of Gaussian process regression. The Annals of Statistics 49(5):2491–2505

Jin Y, Candès EJ (2023) Selection by prediction with conformal p-values. Journal of Machine Learning Research 24(244):1–41

Kodama S, et al (2009) Association between serum uric acid and development of type 2 diabetes. Diabetes Care 32(9):1737–1742

Kunutsor SK, Abbasi A, Adler AI (2014) Gamma-glutamyl transferase and risk of type II diabetes: an updated systematic review and dose-response meta-analysis. Annals

of Epidemiology 24:809–816. https://doi.org/10.1016/j.annepidem.2014.09.001

Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. Journal of Statistical Software 36(11):1–13

Lei J, G'Sell M, Rinaldo A, et al (2018) Distribution-free predictive inference for regression. Journal of the American Statistical Association 113(523):1094–1111

Linkletter C, Bingham D, Hengartner N, et al (2006) Variable selection for Gaussian process models in computer experiments. Technometrics 48(4):478–490

MacKay DJC (1996) Bayesian methods for backpropagation networks. In: Domany E, van Hemmen JL, Schulten K (eds) Models of Neural Networks III. Springer, New York, p 211–254

Neal RM (1996) Bayesian Learning for Neural Networks, Lecture Notes in Statistics, vol 118. Springer, New York

Paananen T, Piironen J, Andersen MR, et al (2019) Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), pp 1743–1752

Piironen J, Vehtari A (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics 11(2):5018–5051

Ramdas A, Grünwald P, Vovk V, et al (2023) Game-theoretic statistics and safe anytime-valid inference. Statistical Science 38(4):576–601

Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA

Ren Z, Barber RF (2024) Derandomized knockoffs: Leveraging e-values for false discovery rate control. Journal of the Royal Statistical Society: Series B 86(1):122–154

Savitsky T, Vannucci M, Sha N (2011) Variable selection for nonparametric Gaussian process priors: Models and computational strategies. Statistical Science 26(1):130–149

Stein ML (1999) Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York

Stierman B, Afful J, Carroll MD, et al (2021) National health and nutrition examination survey 2017–march 2020 prepandemic data files—development of files and prevalence estimates for selected health outcomes. National Health Statistics Reports 158, National Center for Health Statistics, Hyattsville, MD, https://doi.org/10.15620/cdc:106273

Titsias MK (2009) Variational learning of inducing variables in sparse Gaussian processes. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 5:567–574

Vásquez AR, Márquez Urbina JU, González Farías G, et al (2024) Controlling the false discovery rate by a latent Gaussian copula knockoff procedure. Computational Statistics 39(3):1435–1458. https://doi.org/10.1007/s00180-023-01346-4

Vásquez AR, Márquez Urbina JU, González Farías G, et al (2025) Derandomized truncated D-vine copula knockoffs with e-values to control the false discovery rate. Computational Statistics 40(7):3843–3866. https://doi.org/10.1007/s00180-024-01587-x

Vovk V, Wang R (2021) E-values: Calibration, combination and applications. The Annals of Statistics 49(3):1736–1754

Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer, New York

Wang J, Wang P (2024) Large-scale dependent multiple testing via hidden semi-Markov models. Computational Statistics 39(2):1093–1126. https://doi.org/10.1007/s00180-023-01367-z

Wang R, Ramdas A (2022) False discovery rate control with e-values. Journal of the Royal Statistical Society: Series B 84(3):822–852

Weber F, Glass Ä, Vehtari A (2025) Projection predictive variable selection for discrete response families with finite support. Computational Statistics 40(1):701–721. https://doi.org/10.1007/s00180-024-01506-0

Williams CKI, Rasmussen CE (1996) Gaussian processes for regression. Advances in Neural Information Processing Systems 8:514–520

Wu Y, Boos DD, Stefanski LA (2007) Controlling variable selection by the addition of pseudovariables. Journal of the American Statistical Association 102(477):235–243

Xing X, Zhao Z, Liu JS (2023) Controlling false discovery rate using Gaussian mirrors. Journal of the American Statistical Association 118(541):222–241

Xu Z, Fischer L, Ramdas A (2025) Bringing closure to FDR control: Beating the e-Benjamini–Hochberg procedure. arXiv preprint arXiv:250411759

Zhang H (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association 99(465):250–261