Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

Model ranking and confidence shortlists under grouped regression noise

Submission ID	265459628
Article Type	Original Article
Keywords	model comparison, grouped data, ranking uncertainty, random effects, subset selection, simultaneous inference
Authors	Zafar Turakulov, Jaloliddin Eshbobaev, Mirjalol Yusupov, Olimjon Uralov, Takhir Kuchkarov, Elyor Samadov, Elbek Ortikov

For any queries please contact:

LSSP-peerreview@journals.tandf.co.uk

Note for Reviewers:

To submit your review please visit https://mc.manuscriptcentral.com/LSSP

Model ranking and confidence shortlists under grouped regression noise

Zafar Turakulov[1,*], Jaloliddin Eshbobaev[1], Mirjalol Yusupov[1], Olimjon Uralov[2], Takhir Kuchkarov[2], Elyor Samadov[2], Elbek Ortikov[3]

[1] Department of Automation and Digital Control, Tashkent Institute of Chemical Technology, Tashkent, Uzbekistan

[2] Faculty of Digital Economy and Information Technologies, Tashkent State University of Economics, Tashkent, Uzbekistan

[3] Department of Industrial Automation, Tashkent State Technical University, Tashkent, Uzbekistan

* Corresponding author: Zafar Turakulov, webdastur@gmail.com

Abstract

When test observations share group effects, such as a common batch, site or process run, a small test set can rank regression models unreliably. We study this ranking uncertainty conditional on the fitted predictors and the test design. Under squared loss, the paired loss difference is affine in the test noise, and its variance is a quadratic form in the prediction contrast. Under Gaussian noise this gives a closed-form reversal probability and expected selection regret. An alignment coefficient shows that, at fixed marginal noise variance, group effects can inflate or deflate this variance, depending on how prediction contrasts sum within groups. The same variance yields a Bonferroni shortlist that contains every conditional risk minimizer with probability at least $1 - \alpha$ when the noise covariance is known, without knowledge of the mean function. In a simulation with two response functions, five regression procedures, 30 independent realization blocks and 2000 noise draws for each of 1080 test contexts, empirical reversal frequencies agreed with the analytic values to within Monte Carlo error. At $\alpha = 0.05$ the shortlist contained the best fitted predictor in 99.98% of draws and held 1.81 models on average, so the guarantee is conservative. An additive spline ridge baseline had the lowest mean test error for both response functions.

Keywords: model comparison; grouped data; ranking uncertainty; random effects; subset selection; simultaneous inference; squared loss

1 Introduction

A held-out test set that played no part in model fitting can still separate two regression procedures poorly when it is small. The problem is worse when test observations share a process run, site or batch: their errors are then correlated, and the number of rows overstates the information available for the comparison. Two questions follow. How often does test noise reverse the ordering of two fitted predictors, and how much expected loss is incurred when it does?

Several sources of uncertainty enter a model comparison. Reusing evaluation data for tuning biases reported performance [4], leakage can create an apparent improvement that does not exist [14], and the overlap between cross-validation folds rules out treating fold errors as independent replicates [3, 17]. Bates et al. [2] show that cross-validation estimates the average error of a fitting procedure over training sets drawn from the same population, rather than the error of the particular model fitted to the data at hand. We separate these issues by design. Configurations are chosen on an independent pilot, the frozen procedures are refitted on new training groups, and the analysis conditions on the fitted predictors and the test design, so that the only remaining randomness is in the test outcomes. This target is narrower than a population ranking of learning algorithms, but it admits an exact finite-sample analysis.

For Peer Review

Inference on differences in predictive loss has a long history. Diebold and Mariano [5] test equal predictive accuracy from a series of loss differentials, Nadeau and Bengio [17] and Bengio and Grandvalet [3] show why resampling-based variance estimates for such differences are delicate, and Hothorn et al. [12] treat benchmark comparisons as designed experiments. When the data cannot single out one model, a set that contains the best with controlled probability can be reported instead. This idea goes back to subset selection [8] and multiple comparisons with the best [13], and appears in modern form as the model confidence set [9] and cross-validation with confidence [15]. Grouped data add a further complication, because rows from the same site or batch are not exchangeable; grouped cross-validation is recommended for such data [20].

Our contribution is narrower and specific to paired squared losses on independent test groups. We give the exact conditional variance of a paired loss difference as a quadratic form in the prediction contrast, an alignment coefficient that shows when shared group effects inflate or deflate this variance, closed-form reversal probabilities and expected regret under Gaussian noise, and a Bonferroni shortlist with a finite-sample retention guarantee under known covariance. The Gaussian and union-bound arguments are standard. What the analysis adds is an explicit account of how group structure acts on differences between fitted predictions, which is what determines the reliability of a ranking.

We examine these results in a simulation study in which the mean response and the noise covariance are known, organized by the aims, data-generating mechanisms, estimands, methods and performance measures (ADEMP) framework [16]. Two prespecified response functions are combined with independent or grouped covariates. Meta-simulation approaches that learn an approximate generator from limited data address a related question of method selection [1]; here the generators are fixed mathematical functions. The six inputs and their units are modeled on the variables of a fungal solid-state feed treatment, so that effect sizes and noise levels have an interpretable scale. The response functions are test surfaces and are not presented as models of that process.

The paper has three aims: to derive the conditional distribution of paired ranking errors under grouped test noise; to construct a model shortlist whose retention guarantee does not depend on the mean response; and to check both on a reproducible grid of grouped simulations. The simulations include simple polynomial and spline comparators, so that an advantage due to the shape of a particular response surface is not mistaken for general superiority of a method. Section 2 develops the theory, Section 3 describes the simulation design, Section 4 reports the results and Section 5 discusses limitations and extensions.

2 Conditional ranking uncertainty

Target and assumptions

Let C collect the training data, all tuning decisions, the fitted predictors, and the test covariates and group labels. Throughout this section C is fixed. The test groups are independent of the training groups. For n test observations, let the conditional mean vector be μ and the noise vector be u, with zero conditional mean and covariance Σ. In the random-intercept specialization, a group effect and an observation effect are independent, and both are independent of the covariates:

$$Y_i = \mu_i + u_i,\ \ u_i = \zeta_{g_i} + \varepsilon_i,\ \ \zeta_g \sim N\left(0,\sigma_\zeta^2\right),\ \ \varepsilon_i \sim N\left(0,\sigma_\varepsilon^2\right) \quad (1)$$

The index a refers to a fitted predictor, not to a learning algorithm. Let w_i be nonnegative weights, fixed in C and summing to one, and evaluate every predictor on the same outcomes. The observed squared loss and its conditional expectation are

$$L_a = \sum_i w_i\left(Y_i - f_a(x_i)\right)^2,\ \ R_a = E(L_a \mid C) \quad (2)$$

For Peer Review

All experiments use equal weights; the theory allows any weights fixed before the test outcomes are seen. Because every model is scored on the same rows, mean squared error and its square root give the same ordering, but the variance and regret formulas below refer to the difference in squared loss, not in root mean squared error.

Paired loss distribution

For a pair of predictors a and b, let a_i and b_i denote their errors relative to the conditional mean, c_i their prediction contrast and δ_{ab} the conditional expected loss difference, which is positive when b is better:

$$a_i = f_a(x_i) - \mu_i, \;\; b_i = f_b(x_i) - \mu_i, \;\; c_i = a_i - b_i \quad (3)$$

$$\Delta_{ab} = L_a - L_b, \;\; \delta_{ab} = R_a - R_b = \sum_i w_i({a_i}^2 - {b_i}^2), \;\; d_i = w_i c_i \quad (4)$$

Proposition 1. For any zero-mean test noise with finite covariance, the paired loss difference has the representation and moments below. If the noise vector is jointly Gaussian, Δ_{ab} is normal with this mean and variance.

$$\Delta_{ab} = \delta_{ab} - 2\sum_i d_i u_i, \;\; E(\Delta_{ab} \mid C) = \delta_{ab}, \;\; V_{ab} = Var(\Delta_{ab} \mid C) = 4 d^T \Sigma d \quad (5)$$

Proof. Expanding the two squared errors cancels the squared noise term in every row. The remaining random term is minus twice the weighted prediction contrast times the noise. Its mean is zero and its variance is the covariance quadratic form in (5). A linear combination of a Gaussian vector is Gaussian. No independence between the two models' losses is assumed.

For independent random-intercept test groups, the quadratic form becomes

$$V_{ab} = 4\left(\sigma_\varepsilon^2 \sum_i {w_i}^2 {c_i}^2 + \sigma_\zeta^2 \sum_g \left(\sum_{i \in g} w_i c_i\right)^2\right) \quad (6)$$

A shared group deviation therefore enters only through the sum of the prediction contrasts within that group, and contrasts of opposite sign cancel even though the outcomes themselves are positively correlated. The expected loss has the usual decomposition into noise variance and squared approximation error:

$$R_a = \tau^2 + \sum_i w_i {a_i}^2, \;\; \tau^2 = \sigma_\zeta^2 + \sigma_\varepsilon^2 \quad (7)$$

Here τ^2 is the marginal noise variance. Note that τ is not the expected value of a finite-sample root mean squared error, and a realized test score can fall below it.

Reversal probability and regret

Corollary 1. Under Gaussian test noise, if $\delta_{ab} \neq 0$ and $V_{ab} > 0$, the probability that the observed paired losses reverse the conditional expected ordering is q_{ab}, and the expected excess conditional loss from selecting the observed winner is r_{ab}:

$$q_{ab} = P(\Delta_{ab}\delta_{ab} < 0 \mid C) = \Phi\left(-\frac{|\delta_{ab}|}{\sqrt{V_{ab}}}\right), \;\; r_{ab} = |\delta_{ab}| q_{ab} \quad (8)$$

The first expression follows by standardizing the normal variable Δ_{ab}, and the second because a wrong choice costs exactly $|\delta_{ab}|$. A high reversal probability therefore matters little when the gap is small. We report regret divided by τ^2, which makes it unit-free. If $V_{ab} = 0$ and $\delta_{ab} \neq 0$, no reversal can occur; if $\delta_{ab} = 0$, the two predictors have equal conditional risk, a reversal is undefined and the regret is zero.

For Peer Review

Contrast alignment under fixed marginal variance

Corollary 2. For equal weights, a nonzero contrast vector and positive marginal noise variance, define the alignment coefficient κ and the group fraction ρ as follows. Holding τ^2 fixed, compare V with the variance under independent noise of that same marginal variance:

$$\kappa_{ab} = \frac{\sum_g (\sum_{i \in g} c_i)^2}{\sum_i c_i^2}, \quad \rho = \frac{\sigma_\zeta^2}{\tau^2}, \quad \frac{V_{ab}}{V_{ab,ind}} = 1 - \rho + \rho\kappa_{ab} \qquad (9)$$

If s_{max} is the largest test-group size, the Cauchy–Schwarz inequality gives $0 \le \kappa_{ab} \le s_{max}$, because each squared group sum is at most the group size times the group's sum of squared contrasts. For fixed predictions and design, $\kappa_{ab} < 1$ reduces the variance of the comparison relative to independent noise and $\kappa_{ab} > 1$ increases it. Equal contrasts in groups of common size s attain $\kappa_{ab} = s$, and contrasts that sum to zero within every group give $\kappa_{ab} = 0$. These statements concern the variance of a paired comparison; they say nothing about the accuracy of either predictor.

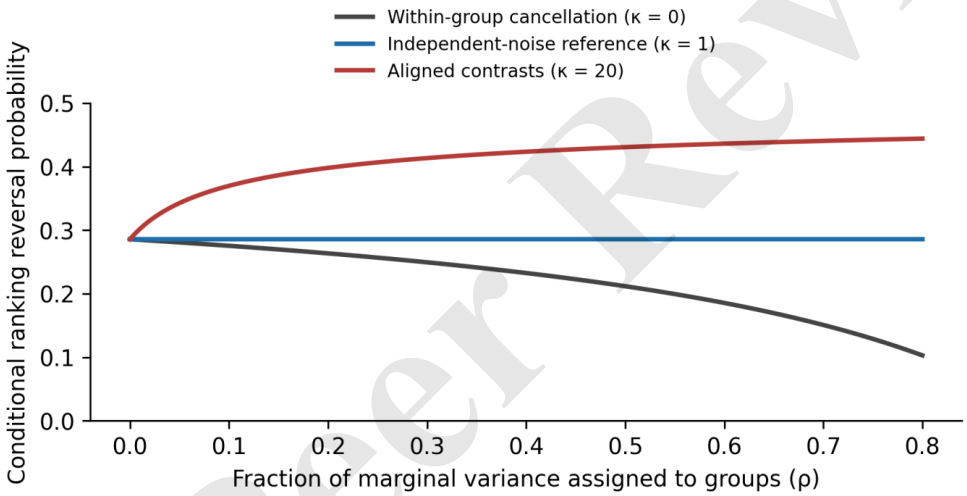


Figure 1. Exact conditional reversal probabilities under different contrast alignment. The illustrative design has 200 rows in ten groups of size 20, unit-magnitude prediction contrasts, an expected loss gap of 0.05 squared percentage points, and $\tau = 0.6247$ percentage points. The three curves hold the gap and contrast norm fixed while varying the fraction of noise variance assigned to groups. They represent separate contrast configurations, not refitted models.

Equation (8) also gives a precision requirement for design: a reversal probability no larger than $\beta < 1/2$ requires $|\delta_{ab}| \ge \Phi^{-1}(1-\beta)\sqrt{V_{ab}}$. The requirement involves the gap as well as κ_{ab}, so adding test groups need not reduce the reversal risk monotonically when it also changes the covariates and hence the gap.

Simultaneous confidence shortlists

The reversal probability depends on the unknown mean response through δ_{ab}. The variance V_{ab} does not: it depends only on the prediction contrasts, the weights and the noise covariance. When Σ is known and the noise is Gaussian, (5) therefore gives a confidence interval for each expected loss difference without knowledge of μ. For m fitted candidates, let K be the number of unordered pairs and α the familywise error level:

$$K = \frac{m(m-1)}{2}, \quad z = \Phi^{-1}\left(1 - \frac{\alpha}{2K}\right), \quad I_{ab} = \left[\Delta_{ab} - z\sqrt{V_{ab}}, \Delta_{ab} + z\sqrt{V_{ab}}\right] \qquad (10)$$

For Peer Review

For reversed pairs, $\Delta_{ba} = -\Delta_{ab}$ and $V_{ba} = V_{ab}$. Candidate a is excluded only if, for some b, the lower confidence limit for δ_{ab} is positive. The shortlist is

$$S = \{a: \Delta_{ab} - z\sqrt{V_{ab}} \leq 0 \text{ for every } b \neq a\} \quad (11)$$

Proposition 2. Under the conditions above, all conditional risk minimizers are retained in S with probability at least 1 − α. No independence assumption between pairwise differences is required.

$$P(\arg\min_a R_a \subseteq S \mid C) \geq 1 - \alpha \quad (12)$$

Proof. Each nondegenerate interval in (10) has marginal coverage 1 − α/K; a zero-variance pair has an exact degenerate interval. The union bound gives joint coverage of all K differences of at least 1 − α. On this event, the lower limit cannot be positive when the true difference is nonpositive. A minimizer has a nonpositive difference relative to every other candidate and therefore cannot be excluded. The rule is a conservative subset-selection procedure in the sense of Gupta [8], with the same aim as the model confidence set [9] and cross-validation with confidence [15]. Its guarantee is conditional on a finite test design and a known covariance; it is not a statement about the best learning algorithm in a population.

With five candidates and α = 0.05, K = 10 and z = 2.8070. An unresolved comparison leaves both candidates in the shortlist; it is not evidence that their risks are equal. The shortlist is never empty, because a model with the smallest observed loss has nonpositive observed differences against every other candidate. If Σ is estimated from the same test sample, the guarantee no longer holds exactly.

Remark 1. Proposition 2 controls all K two-sided intervals, which is more than retention of a minimizer requires. Conditional on C, a unique minimizer a* is a fixed index, and it is retained whenever $\Delta_{a*b} - z'\sqrt{V_{a*b}} \leq \delta_{a*b}$ for every $b \neq a*$. These are m − 1 one-sided events, so the union bound gives retention probability at least 1 − α with $z' = \Phi^{-1}\{1 - \alpha/(m-1)\}$, which equals 2.2414 for m = 5 and α = 0.05. Because the joint covariance of all paired differences is known (Section S5), z′ can be replaced by the largest over candidates a of the (1 − α) quantile of $\max_{b \neq a} \frac{\Delta ab - \delta ab}{\sqrt{Vab}}$, in the spirit of multiple comparisons with the best [13]. The results in Section 4 use the Bonferroni rule (11), so the reported coverage and retention are conservative relative to these alternatives.

Scope beyond Gaussian test noise

The representation and variance in (5) do not require Gaussian noise. Cantelli's one-sided inequality then bounds the reversal probability when $\delta_{ab} \neq 0$:

$$q_{ab} \leq \frac{V_{ab}}{V_{ab} + \delta_{ab}^2} \quad (13)$$

A proof is given in Section S5. The bound can be loose, and it does not extend the coverage guarantee of the shortlist to non-Gaussian noise. Conditioning on C also removes training variability: over new training sets and test designs, $\mathrm{Var}(\Delta_{ab}) = E(V_{ab}) + \mathrm{Var}(\delta_{ab})$, and the unconditional reversal probability is $E(q_{ab})$. The simulation therefore uses independent realization blocks in addition to repeated noise draws on fixed fitted models.

3 Simulation design and reproducibility

Response functions and predictor designs

The six predictors are a loading variable D (0.5–8% w/w), initial pH (4–7), temperature T (22–34 °C), treatment time t (24–240 h), moisture M (55–85%) and initial crude protein CP_0; the simulated target is final crude protein in percent dry matter. CP_0 equals 19.7 + 0.211Z, where Z is a standard normal coordinate winsorized at ±3. The two response means are $\mu_k(x) = CP_0 + F_k(x)$ for $k \in \{A, B\}$.

For Peer Review

The smooth function A combines Gaussian-shaped optima in temperature, pH and moisture, saturating effects of loading and time, a three-way interaction, and quadratic penalties for excess loading and excess time. The threshold function B combines trapezoidal responses in temperature, pH and moisture, piecewise-linear loading and time effects with ramps, plateaus and declines, and a temperature-by-pH interaction. All coefficients were fixed before the simulations reported here; their values, transformations and breakpoints are listed in Section S1. Responses are not clipped.

To vary the clustering of covariates while keeping their margins fixed, each latent coordinate is the sum of independent standard normal group and row terms with weights $\sqrt{\eta}$ and $\sqrt{(1-\eta)}$. Applying the standard normal distribution function and rescaling gives uniform margins for the first five predictors; the sixth coordinate gives winsorized initial protein. Coordinates are mutually independent. With $\eta = 0.5$, rows in the same group share a latent component, and $\eta = 0$ gives independent rows; η is a correlation on the latent normal scale, not after transformation. Training and test groups are generated independently, and group labels are not supplied to any fitted model.

The baseline standard deviations are $\sigma_\zeta = 0.2587$ and $\sigma_\varepsilon = 0.5686$ percentage points, so $\tau^2 = 0.3902$ and the baseline group fraction is $\rho = 0.1715$ (exact values in Section S1). The stress scenarios use $\rho = 0$, 0.1715 or 0.5 with τ^2 held fixed, and training and test data share the scenario's noise allocation. Because group labels are not used, each fitted predictor targets a new group with zero expected group deviation, not a realized group intercept.

Independent pilot and fitting procedures

Hyperparameters are chosen once on an independent pilot of 250 rows in ten groups ($\eta = 0$, baseline ρ). For each candidate configuration, five-fold group cross-validation is run for each response function on the same pilot design, and the configuration with the smallest mean of the ten validation mean squared errors is selected for each procedure. No main-study data enter this step. The selected configurations are then frozen for all scenarios and training sizes, so the study compares fixed fitting procedures conditional on the pilot; it does not evaluate procedures that are retuned for each dataset. For four of the five procedures at least one selected value lies at the edge of its pilot grid (Table S3); Section 5 discusses the consequences.

Linear and polynomial ridge provide simple baselines [11]. The polynomial contains all degree-two terms in standardized predictors. Spline ridge is an additive cubic B-spline expansion followed by standardization and ridge regression; it has the additive structure of a generalized additive model [10] but penalizes the coefficients instead of the roughness of the fitted functions. Gaussian process regression uses a rational-quadratic kernel [19]. Gradient boosting uses squared-error regression trees [7]. All preprocessing is fitted on training rows only, and the Gaussian process kernel is fixed after the pilot, with no later likelihood optimization.

Table 1. Fitting procedures and configurations selected on the independent pilot. Complete grids and software settings are in Section S2.

Procedure	Predictor representation	Frozen configuration
Linear ridge	Standardized original predictors	Ridge penalty 100
Polynomial ridge	Standardized predictors and 27 degree-two terms	Ridge penalty 1
Spline ridge	Additive cubic B-splines, then standardization	4 knots per predictor; ridge penalty 0.1
Gaussian process	Standardized predictors; rational-quadratic kernel	Length scale 4; kernel shape 1; white-noise level 0.05
Gradient boosting	Original predictors; squared-error trees	200 trees; learning rate 0.05; depth 3; minimum leaf size 10

For Peer Review

Stress grid and conditional noise replication

Each of 30 independent realization blocks contains both response functions, both η values and all three ρ values. Every training set has 200 rows in ten groups. The five fitted procedures are evaluated on three new 200-row test designs with five, ten or twenty groups of equal size, generated independently of the training data. This gives 360 training sets, 1080 fitted-model and test-design contexts and 10,800 pair contexts. Within a block, random-number streams are shared across response functions and noise profiles to allow matched comparisons, so contexts within a block are not independent.

For each test context, 2000 noise vectors are drawn from the specified covariance and added to the known mean, with predictions and covariates held fixed; all five models are scored on the same outcome vector in each draw. Empirical reversal frequencies are compared with (8), and the shortlist (10)–(12) is assessed by simultaneous coverage, retention of the oracle best predictor, shortlist size and the frequency of a single-model shortlist. The observed winner is the model with the smallest observed squared loss. The oracle is used only for evaluation; the shortlist itself uses only observed losses, predictions and the specified covariance.

Learning curves and uncertainty summaries

A separate seed family generates 30 independent learning-curve blocks for each response function, with $\eta = 0$ and baseline ρ. Training sizes are 50, 100, 250, 500 and 1000 rows, always in ten groups, and are nested within a block by taking the first rows of each group. All sizes share one independent 200-row test set in ten new groups, so comparisons between procedures and between sizes are paired within a block. Hyperparameters stay at their pilot values, while preprocessing and coefficients are refitted to each training set.

For the learning curves we report the mean and standard deviation of test root mean squared error (RMSE) over blocks, with Monte Carlo standard error (MCSE) equal to the standard deviation divided by $\sqrt{30}$ and approximate 95% intervals based on the t distribution with 29 degrees of freedom. Matched comparisons use the 30 within-block differences. These intervals describe the precision of simulation means. For the stress-grid summaries, quantities are first averaged within each block, and standard errors and intervals are computed from the 30 block means.

Within a context, an empirical reversal frequency has binomial standard error $\sqrt{[q(1 - q)/2000]} \leq 0.0112$. Calibration intervals use 5000 bootstrap resamples of whole realization blocks [6], so rows, pairs and noise draws are not treated as independent replicates. The seed map and all numerical outputs are included in the data archive.

Software and computational record

The computations use Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0 [18], Matplotlib 3.10.8 and threadpoolctl 3.6.0. The study performs 240 pilot fits and 3300 main fits. The data archive includes every generated training and test row, group identifier, mean response, saved outcome and fitted prediction; additional noise draws are recreated from their seeds. Three scripts reproduce the study, the tables, the shortlist audit and the figures, and a README documents the commands and column definitions.

4 Results

Analytic probability audit

Across the 10,800 pair contexts, the mean difference between empirical and analytic reversal probabilities is −0.000004 (95% interval from block means, −0.000098 to 0.000090). The root mean squared discrepancy is 0.00370 and the largest absolute discrepancy 0.03910. Many pairs have negligible reversal probability; among the 2366 pairs with analytic probability between 0.01 and 0.49, the mean difference is 0.000015 and

For Peer Review

the root mean squared discrepancy 0.00779. Both values are of the order expected from binomial error with 2000 draws per context. Because (8) is exact for the simulated Gaussian noise, this comparison verifies the implementation; it does not test the formula under other noise distributions.

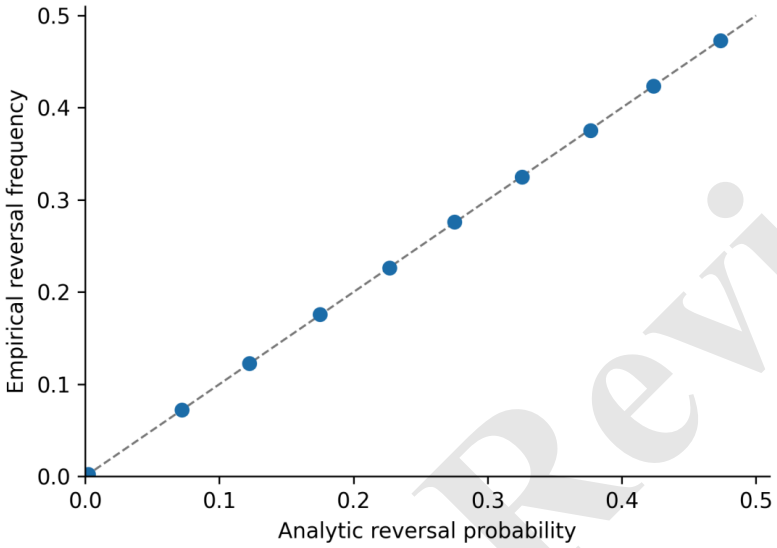


Figure 2. Analytic and empirical conditional reversal probabilities. Points average pair contexts in ten probability bins across the balanced stress grid. Vertical bars show 95% realization-block bootstrap intervals for empirical minus analytic probability, shifted onto the probability scale; they are small at the displayed resolution. The dashed line marks exact agreement. All 2000 noise draws for a context keep its fitted predictions and test design fixed.

Group alignment and ranking sensitivity

Observed alignment coefficients range from 0.039 to 31.69, within the group-size bound. With $\rho = 0.5$ and independent covariates, mean variance ratios are close to one, with some contrasts attenuated and others amplified. With grouped covariates the ratios are much larger (Table 2 and Figure 3): for five test groups they average 4.64, against 1.14 with independent covariates. When covariates are clustered, rows in the same group receive similar predictions and therefore similar contrasts, so the shared group error falls more heavily on the comparison.

For Peer Review

Table 2. Alignment and conditional variance at ρ = 0.5. Summaries average both response functions and all ten pairs within each realization block. Parentheses give 95% Monte Carlo intervals from 30 blocks. At ρ = 0.5 the variance ratio equals (1 + κ)/2, so its mean is determined by the mean κ.

Latent η	Test groups	Mean κ	Mean variance ratio (95% interval)	Contrasts with κ < 1 (%)
0	5	1.276	1.138 (1.057, 1.219)	42.0
0	10	1.182	1.091 (1.039, 1.142)	37.7
0	20	0.993	0.997 (0.965, 1.028)	55.0
0.5	5	8.287	4.643 (3.736, 5.551)	1.0
0.5	10	4.575	2.787 (2.552, 3.023)	1.2
0.5	20	2.664	1.832 (1.738, 1.926)	0.5

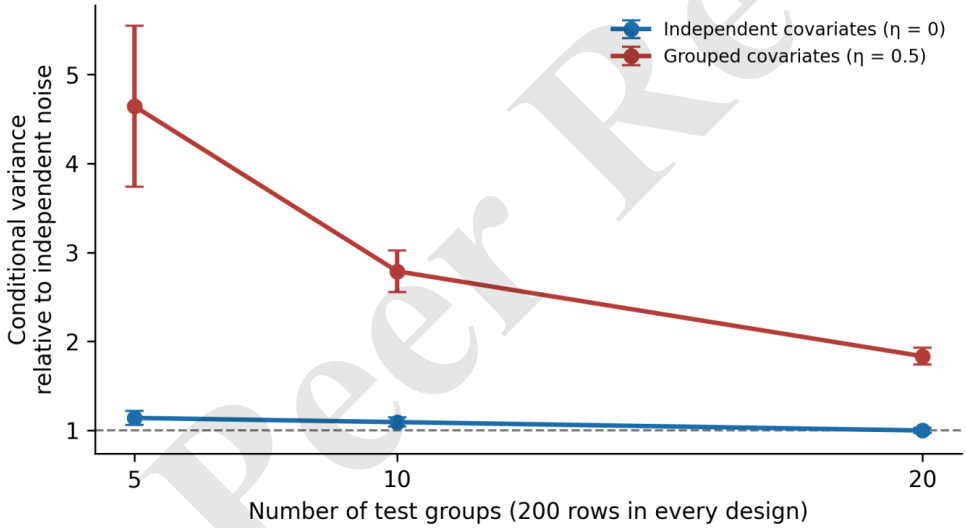


Figure 3. Conditional paired-loss variance relative to independent test noise with the same marginal variance, at ρ = 0.5. Every test design has 200 rows. Means first aggregate the two response functions and ten model pairs within a realization block; bars are 95% Monte Carlo intervals across 30 independent blocks. The horizontal line is a variance ratio of one.

In the grouped-covariate scenarios the mean variance ratio falls as the number of test groups increases, but the reversal probability also depends on the expected loss gap. For the Gaussian process and gradient boosting pair under η = ρ = 0.5, mean reversal probabilities range from 0.146 to 0.242 (Table 3) and do not decrease with the number of groups, because each group count uses a different test design and hence a different gap. The alignment identity holds for each design separately and implies no monotone pattern across designs.

For Peer Review

Table 3. Gaussian process versus gradient boosting under grouped covariates and noise ($\eta = \rho = 0.5$). Each row averages 30 fitted-model and test-design realizations. Intervals describe the mean conditional probability; normalized regret is $|\delta|q/\tau^2$.

Response	Test groups	Mean reversal probability (95% interval)	Mean normalized regret	Mean κ
A	5	0.227 (0.173, 0.281)	0.0321	6.058
A	10	0.185 (0.129, 0.241)	0.0216	3.283
A	20	0.242 (0.180, 0.304)	0.0168	2.028
B	5	0.146 (0.093, 0.199)	0.0220	4.758
B	10	0.169 (0.113, 0.226)	0.0207	3.382
B	20	0.220 (0.155, 0.284)	0.0162	1.911

Ranking uncertainty and loss consequence

Across the stress grid, 6.56% of pair contexts have a reversal probability of at least 0.25; their median normalized regret is 0.0146. Figure 4 shows why the probability alone is a poor guide to the consequences of selection. A nearly tied pair can switch order often at little cost, and the largest regrets occur when a sizeable gap coincides with a noisy comparison. The regret refers to a choice between two fitted candidates on a given test design and excludes the cost of tuning or data collection.

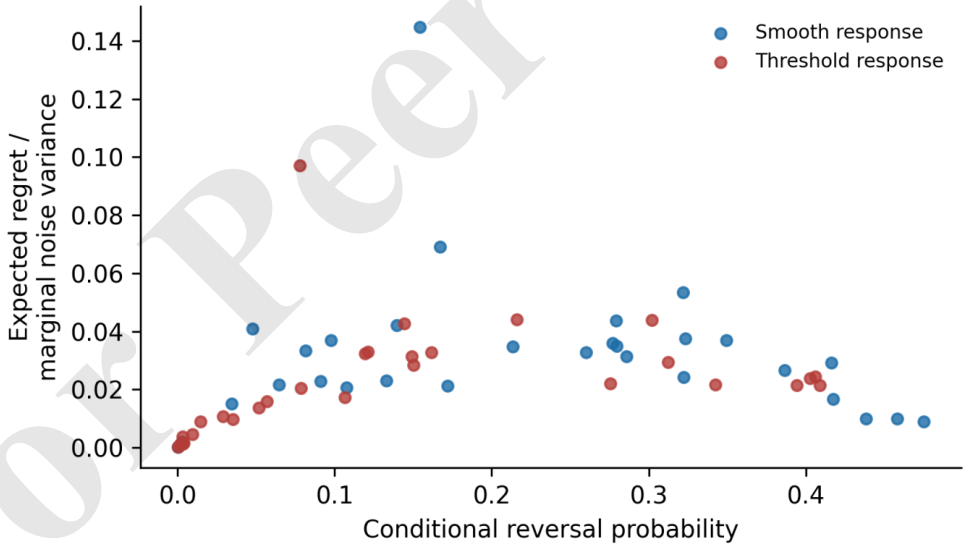


Figure 4. Reversal probability and its expected loss consequence for Gaussian process versus gradient boosting. Each point is one fitted-model and test-design realization with $\eta = \rho = 0.5$ and five test groups. The 30 smooth and 30 threshold cases show that a high probability of switching order can accompany a small conditional regret. Regret is normalized by τ^2. The mean response is known here for simulation auditing.

For Peer Review

Learning curves and matched comparisons

Spline ridge has the lowest mean test RMSE at every training size for both response functions (Figure 5 and Table 4). Its mean RMSE is 0.6741 for A and 0.6964 for B at 250 rows, and 0.6474 and 0.6681 at 1000 rows, against a noise level of $\tau = 0.6247$. Both response functions are dominated by low-order additive effects, which suits an additive spline basis. The result shows the value of strong simple baselines. It does not identify a generally best nonlinear method, particularly because the Gaussian process and boosting configurations were not retuned for each training size.

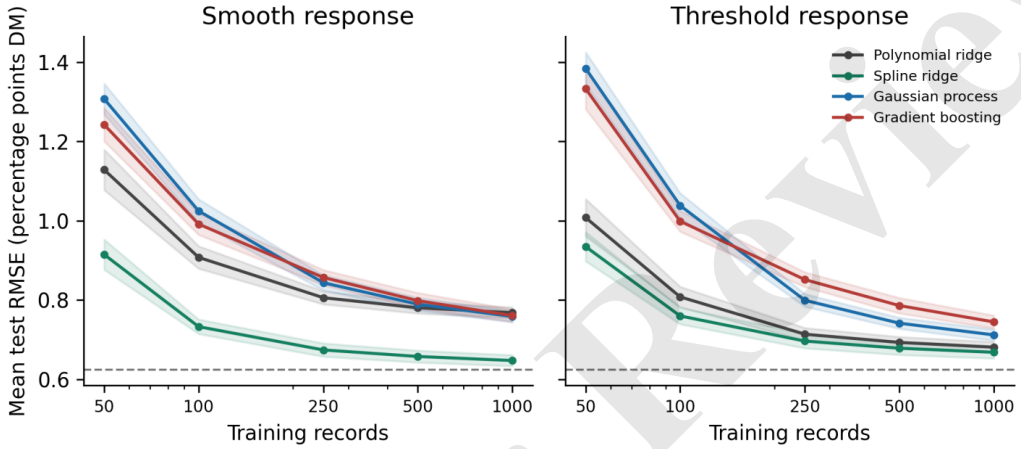


Figure 5. Learning curves for the four nonlinear procedures under $\eta = 0$ and baseline $\rho = 0.1715$. Curves show mean test RMSE across 30 independent blocks, with approximate 95% Monte Carlo intervals shaded. DM, dry matter. Training sizes are nested within a block, and each block's 200-row test set is shared across sizes. The dashed line marks the noise standard deviation τ, which is not a finite-sample lower bound. Linear ridge (Tables 4 and S6) is omitted to keep the vertical scale readable.

Table 4. Test RMSE in percentage points dry matter at two training sizes. Entries are the mean (standard deviation) across 30 independent learning-curve blocks. All sizes and Monte Carlo errors are in Table S6.

Response	Procedure	250 training rows	1000 training rows
A	Linear ridge	1.8112 (0.0788)	1.8047 (0.0721)
A	Polynomial ridge	0.8058 (0.0443)	0.7680 (0.0370)
A	Spline ridge	0.6741 (0.0448)	0.6474 (0.0371)
A	Gaussian process	0.8435 (0.0514)	0.7584 (0.0382)
A	Gradient boosting	0.8567 (0.0509)	0.7607 (0.0423)
B	Linear ridge	1.9593 (0.0828)	1.9508 (0.0763)
B	Polynomial ridge	0.7137 (0.0427)	0.6810 (0.0357)
B	Spline ridge	0.6964 (0.0481)	0.6681 (0.0417)
B	Gaussian process	0.7992 (0.0426)	0.7119 (0.0379)
B	Gradient boosting	0.8513 (0.0499)	0.7447 (0.0437)

At 250 rows the matched comparison of Gaussian process regression and gradient boosting depends on the response function (Table 5). The mean RMSE difference, boosting minus Gaussian process, is 0.0132 for A (95% interval −0.0041 to 0.0305) and 0.0521 for B (0.0371 to 0.0672). Both procedures are clearly worse than spline ridge, so a study that compared only these two would have missed the strongest method.

For Peer Review

Table 5. Matched mean RMSE differences at 250 training rows. Positive values favor the procedure after "minus". Approximate 95% Monte Carlo intervals use the 30 within-block differences. The selected comparisons are descriptive and are not a familywise hypothesis-testing analysis.

Response	Comparison	Mean difference	95% interval
A	Gradient boosting minus Gaussian process	0.0132	(−0.0041, 0.0305)
A	Gaussian process minus Spline ridge	0.1694	(0.1570, 0.1819)
A	Polynomial ridge minus Spline ridge	0.1317	(0.1212, 0.1422)
B	Gradient boosting minus Gaussian process	0.0521	(0.0371, 0.0672)
B	Gaussian process minus Spline ridge	0.1028	(0.0881, 0.1175)
B	Polynomial ridge minus Spline ridge	0.0173	(0.0071, 0.0275)

Confidence shortlist audit

With the specified Gaussian covariance and $\alpha = 0.05$, the simultaneous intervals cover all ten conditional pairwise gaps in 97.10% of draws over the stress grid. The oracle best predictor is retained in 99.977% of draws (95% interval 99.973–99.982%). The mean shortlist size is 1.81 (1.75–1.87), and 39.5% of shortlists contain a single model. Coverage and retention exceed the guaranteed 95% by a wide margin. This reflects both the conservatism of the Bonferroni rule (Remark 1) and the size of the gaps between the best predictor and the others in these scenarios.

Table 6. Confidence shortlist summaries by covariate and noise grouping. Each realization-block value first averages both response functions and all three test-group counts. Intervals use 30 independent blocks. Oracle retention is a simulation audit measure; constructing a shortlist does not require the oracle.

Latent η	Noise ρ	Oracle retained (%)	Mean shortlist size (95% interval)	Observed winner is oracle (%) (95% interval)
0	0.0000	99.979	1.603 (1.563, 1.644)	89.83 (87.67, 91.99)
0	0.1715	99.978	1.623 (1.581, 1.666)	89.13 (86.97, 91.29)
0	0.5000	99.978	1.685 (1.631, 1.739)	88.60 (86.17, 91.02)
0.5	0.0000	99.981	1.636 (1.559, 1.713)	90.44 (88.40, 92.48)
0.5	0.1715	99.981	1.945 (1.805, 2.084)	88.40 (86.23, 90.57)
0.5	0.5000	99.968	2.384 (2.216, 2.553)	84.88 (81.78, 87.99)

Choosing the observed lowest-loss model selects the oracle best in 88.55% of draws (95% interval 87.15–89.95%), with mean normalized regret 0.00770 (0.00578–0.00962). This is a different target from shortlist retention, since a shortlist can contain the best model together with inferior ones. Shortlists are largest in the profiles with grouped covariates and grouped noise, which are also the profiles in which the observed winner is least often the oracle best (Table 6; full results in Table S5).

5 Discussion

The intraclass correlation of the outcomes does not by itself determine how uncertain a paired model ranking is. At fixed marginal noise variance, the effect of grouping depends on the within-group sums of the prediction contrasts: alignment ($\kappa > 1$) inflates the variance of the comparison relative to independent noise, and cancellation ($\kappa < 1$) reduces it. For a given covariance and given fitted predictors this statement is exact.

For Peer Review

How κ is distributed over new datasets depends on the fitting procedures and the covariate design; in our simulations, clustered covariates produced strong alignment.

The reversal probability and the regret require the mean response, so in practice they are diagnostics for simulation studies with a known generator. The shortlist needs only the predictions and the test covariance, and gives a finite-sample retention statement for the conditional risks on the test design. Its main practical limitation is the assumption that Σ is known. When Σ must be estimated, two routes seem natural. Variance components could be estimated from an independent source, such as replicate runs or historical data from the same process, with the additional uncertainty propagated into the critical value. Alternatively, because Δ_{ab} is a sum of independent group-level contributions, a cluster-level variance estimate or a paired t interval over test groups would avoid modeling Σ altogether, at the price of conservativeness when group-level expected differences vary and of low precision when there are few groups. We have not evaluated either approach.

The shortlist applies standard simultaneous-comparison ideas [8, 13] to a specific setting and does not replace the model confidence set [9] or cross-validation with confidence [15], which address different targets. It also differs from tests of predictive accuracy under serial dependence [5]. The analysis is exact here because test groups are independent, the fitted candidates are fixed and the loss is quadratic, so the loss difference is affine in the noise. The quadratic form (5) covers any known covariance, including heteroscedastic or random-slope structures, but our simulations use random intercepts only. The one-sided and max-type critical values in Remark 1 would give smaller shortlists with the same guarantee and are a direct next step.

The learning-curve results depend on the response surfaces and on the tuning protocol and should be read with both in mind. Both surfaces are largely additive, and spline ridge performs well even on the threshold surface. The frozen pilot makes the comparison reproducible and rules out tuning on test data, but it does not show how each method would perform after a fresh search at each training size. The Gaussian process kernel was isotropic and was not re-optimized, and for four of the five procedures the selected values lie at the edge of the pilot grid (Table S3), so the results may understate what Gaussian process regression and boosting can achieve. Other response functions, interaction strengths, covariate distributions or tuning budgets could change the ordering.

Other limitations concern the scope of the evidence. Thirty blocks give moderate precision for small mean differences, and the 10,800 pair contexts come from only 30 independent blocks. All simulations use Gaussian noise with known scale and equal group sizes; for non-Gaussian noise we have only the moment bound (13), not the shortlist guarantee. Finally, the study contains no application to real data, so the practical size of alignment effects in applied grouped data remains to be established.

For benchmarking, the observed winner, the uncertainty of the ranking and the expected cost of a wrong choice answer different questions. Reporting all three, with independent replication and strong simple baselines, keeps a small difference in test error from being given more weight than the design supports. When the covariance can be justified, a shortlist is a direct way to report an unresolved comparison with a stated coverage guarantee.

6 Conclusions

For predictors fitted independently of the test groups, the difference in squared test loss between two predictors is affine in the test noise. This gives closed-form Gaussian reversal probabilities and expected selection regret, and an alignment identity that shows when group structure inflates or deflates the variance of a comparison. The same variance yields a shortlist that contains every conditional risk minimizer with at least the nominal probability when the Gaussian noise covariance is known. Simulations confirm these properties and show that a simple additive spline baseline can outperform more flexible methods on largely additive surfaces. Extending the guarantee to estimated covariances and to population-level rankings is the main open problem.

For Peer Review

Acknowledgments

None.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflicts of interest.

Ethics approval and consent

This simulation study involved no human participants or animals. Ethics approval and consent were not required.

Author contributions

Conceptualization: J.E. and Z.T. Methodology: J.E., Z.T. and M.Y. Software: J.E. and Z.T. Validation: J.E., T.K., O.U. and E.S. Formal analysis: J.E. and Z.T. Investigation: E.S., E.O., O.U., T.K., M.Y., Z.T. and J.E. Resources: E.S., E.O., T.K., O.U., M.Y. and Z.T. Data curation: J.E., M.Y., E.S. and Z.T. Writing-original draft: J.E. and Z.T. Writing-review and editing: J.E. and Z.T. Visualization: E.S. and M.Y. Supervision: J.E. and Z.T. Project administration: Z.T.

Use of generative artificial intelligence

OpenAI ChatGPT was used to revise the manuscript. Reported numerical results were produced by the Python scripts supplied in the data archive, and no language-model output was used as data. The authors take full responsibility for all arguments, code and results.

Data availability

All generated datasets, fitted predictions, numerical results, simulation and analysis code, and figure sources are provided in the supplementary file.

Supporting information

The Supporting Information contains the full generator formulas, predictor distributions, pilot grids, software and seed details, complete scenario and learning-curve summaries, additional proofs and the data dictionary. Vector and high-resolution raster versions of all figures are supplied with the data files.

References

[1] Alwash, M., Al Hajj, G. S., Grytten, I., and Sandve, G. K. (2025). Meta simulation approach for evaluating machine learning method selection in data limited settings. Scientific Reports, 15, 40766. https://doi.org/10.1038/s41598-025-24627-y

[2] Bates, S., Hastie, T., and Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? Journal of the American Statistical Association, 119, 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

[3] Bengio, Y., and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. Journal of Machine Learning Research, 5, 1089–1105.

For Peer Review

[4] Cawley, G. C., and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.

[5] Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13, 253–263. https://doi.org/10.1080/07350015.1995.10524599

[6] Efron, B., and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall, New York.

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 1189–1232. https://doi.org/10.1214/aos/1013203451

[8] Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. Technometrics, 7, 225–245. https://doi.org/10.1080/00401706.1965.10490251

[9] Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. Econometrica, 79, 453–497. https://doi.org/10.3982/ECTA5771

[10] Hastie, T. J., and Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall, London.

[11] Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634

[12] Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics, 14, 675–699. https://doi.org/10.1198/106186005X59630

[13] Hsu, J. C. (1996). Multiple Comparisons: Theory and Methods. Chapman & Hall, London.

[14] Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4, 100804. https://doi.org/10.1016/j.patter.2023.100804

[15] Lei, J. (2020). Cross-validation with confidence. Journal of the American Statistical Association, 115, 1978–1997. https://doi.org/10.1080/01621459.2019.1672556

[16] Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38, 2074–2102. https://doi.org/10.1002/sim.8086

[17] Nadeau, C., and Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52, 239–281. https://doi.org/10.1023/A:1024068626366

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[19] Rasmussen, C. E., and Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

[20] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40, 913–929. https://doi.org/10.1111/ecog.02881

For Peer Review

Supporting information for Model ranking and confidence shortlists under grouped regression noise

S1 Complete generator specification

Predictor margins and group structure

Table S1. Predictor margins. The first five variables are transformed from latent normal coordinates through the standard normal distribution function Φ. The sixth is a winsorized normal coordinate. Ranges are simulation choices.

Predictor	Symbol	Margin or transformation	Unit
Loading	D	Uniform on [0.5, 8]	% w/w
Initial pH	pH	Uniform on [4, 7]	pH units
Temperature	T	Uniform on [22, 34]	°C
Treatment time	t	Uniform on [24, 240]	h
Moisture	M	Uniform on [55, 85]	%
Initial crude protein	CP_0	19.7 + 0.211 clip(Z, −3, 3)	% dry matter

For each coordinate j, group g and row i, Z and E are independent standard normal draws, also independent across coordinates. The transformed latent coordinate is

$$Z^*_{ij} = \sqrt{\eta} Z_{g(i),j} + \sqrt{1-\eta} E_{ij}, \;\; U_{ij} = \Phi\left(Z^*_{ij}\right) \quad \text{(S1)}$$

The first five coordinates are their lower bound plus the range length times U. Initial protein uses the sixth Z*. Latent within-group correlation is η, equal to 0 or 0.5 in the stress study. Covariate coordinates are independent of all outcome noise. Groups are balanced. The design function always draws group and row normal arrays, including when one weight is zero, so the saved seed fully specifies the random-number sequence.

Smooth response function A

For both functions, the mean is initial protein plus a deterministic increment. In A, the main response components are

$$f_T = \exp\left(-\frac{1}{2}\left(\frac{T-28}{3}\right)^2\right), \;\; f_P = \exp\left(-\frac{1}{2}\left(\frac{pH-5.55}{0.85}\right)^2\right) \quad \text{(S2)}$$

$$f_M = \exp\left(-\frac{1}{2}\left(\frac{M-69.5}{8}\right)^2\right), \;\; s_D = 1-\exp\left(-\frac{D}{2.5}\right), \;\; s_t = 1-\exp\left(-\frac{t}{80}\right) \quad \text{(S3)}$$

$$P_D = \gamma_D\left(\max(D-6,0)\right)^2, \;\; P_t = \gamma_t\left(\frac{\max(t-180,0)}{60}\right)^2 \quad \text{(S4)}$$

$$F_A = \beta_0 + \beta_m(f_T + f_P + f_M) + \beta_I f_T f_P f_M + \beta_D s_D + \beta_t s_t - P_D - P_t \quad \text{(S5)}$$

Threshold response function B

Define a trapezoid by the following compact expression. The arguments satisfy a < b ≤ c < d; its value is zero outside [a,d], one on [b,c] and linear on each flank.

$$\mathrm{trap}(q;a,b,c,d) = \max\left(0,\min\left(1,\frac{q-a}{b-a},\frac{d-q}{d-c}\right)\right) \quad \text{(S6)}$$

$$g_T = \mathrm{trap}(T;22,26,30,34), \;\; g_P = \mathrm{trap}(pH;4,5,6,7) \quad \text{(S7)}$$

For Peer Review

$$g_M = \mathrm{trap}(M;55,65,74,85) \quad \text{(S8)}$$

$$g_D = \begin{cases} 0, & \text{if D} \le 1 \\ \frac{D-1}{3}, & \text{if } 1 < \text{D} < 4 \\ 1, & \text{if } 4 \le \text{D} \le 6 \\ \max(0,1-0.22(D-6)), & \text{if D} > 6 \end{cases} \quad \text{(S9)}$$

$$g_t = \begin{cases} 0, & \text{if t} < 24 \\ \frac{t-24}{96}, & \text{if } 24 \le \text{t} < 120 \\ 1, & \text{if } 120 \le \text{t} \le 180 \\ \max\left(0,1-0.55\frac{t-180}{60}\right), & \text{if t} > 180 \end{cases} \quad \text{(S10)}$$

$$F_B = \beta_0 + \beta_T g_T + \beta_P g_P + \beta_M g_M + \beta_D g_D + \beta_t g_t + \beta_{TP} g_T g_P \quad \text{(S11)}$$

The time response equals 0.45 at 240 h. The coefficients in Table S2 are fixed scenario constants and are reported to full stored precision so that the generator can be reproduced exactly.

Table S2. Complete coefficients for the two response functions. Parameters follow the order of the components in (S5) and (S11).

Function	Coefficient component	Value
A	Intercept	−1.4612859169
A	Sum of smooth main effects	3.1706098522
A	Triple interaction	0.4390182454
A	Loading saturation	1.6733970057
A	Time saturation	1.5394535529
A	Loading penalty	0.1648389508
A	Time penalty	2.2421740551
B	Intercept	−0.4893655287
B	Temperature	3.3406364735
B	pH	2.3672715118
B	Moisture	2.4994364831
B	Loading	1.0062364665
B	Time	0.9574278917
B	Temperature by pH interaction	0.6759080007

Upper-case temperature and lower-case time subscripts are distinct in equations (S5) and (S11). Table S2 adds descriptions where needed. The code is the executable specification; coefficients, component shapes and breakpoints are also recorded in generator_specification.json in the data archive.

Noise generator

$$Y_i = \mu_k(x_i) + \tau\sqrt{\rho}B_{g(i)} + \tau\sqrt{1-\rho}\varepsilon_i, \;\; B_g, \varepsilon_i \sim N(0,1) \quad \text{(S12)}$$

S2 Pilot selection and software settings

The pilot has 250 rows in ten groups with η = 0 and baseline ρ. Each configuration is evaluated using five GroupKFold splits for A and five for B, sharing the pilot predictor design and noise. Preprocessing is fitted within each training fold. The arithmetic mean of ten validation mean squared errors determines the selected configuration. Candidate order is deterministic, and the first minimum would resolve a tie. The pilot does not contribute to the reported main-study scores.

For Peer Review

Table S3. Pilot grids, selected configurations and mean validation squared errors. Hyperparameter selection is performed once, independently of the main study.

Procedure	Candidate grid	Selected values	Pilot MSE
Linear ridge	Penalty: 0.1, 1, 10, 100	Penalty 100	3.55223
Polynomial ridge	Penalty: 0.1, 1, 10, 100	Penalty 1	0.40601
Spline ridge	Knots: 4, 6; penalty: 0.1, 1, 10	4 knots; penalty 0.1	0.33828
Gaussian process	Length: 1, 2, 4; white noise: 0.05, 0.15	Length 4; white noise 0.05	0.51419
Gradient boosting	Depth: 2, 3; minimum leaf: 3, 10	Depth 3; minimum leaf 10	0.61805

Linear ridge uses a StandardScaler followed by Ridge. Polynomial ridge uses a StandardScaler, PolynomialFeatures with degree 2 and no bias column, then Ridge; the six inputs generate 27 terms. The intercept is fitted by Ridge and is not represented by a polynomial bias term. Ridge fits use the scikit-learn defaults except for the recorded penalty.

Spline ridge uses SplineTransformer with degree 3, four uniformly spaced knots in each training predictor range, include_bias=False and linear extrapolation. Its 30 output basis columns are standardized before Ridge. The penalty is a coefficient ridge penalty, not an integrated squared-derivative penalty. Knot locations and scaling are recomputed from training rows at each fit, while their number and degree are fixed.

The Gaussian process kernel is a unit ConstantKernel times an isotropic RationalQuadratic kernel with shape parameter 1 and length 4, plus WhiteKernel with noise level 0.05. All kernel bounds are fixed, optimizer=None, normalize_y=True and numerical jitter alpha=10^{-10}. Predictors are standardized within training data. White-noise level is expressed on the normalized-target scale; it is not the known physical noise variance τ^2. The returned predictive mean is used; Gaussian process predictive intervals are not part of this study.

GradientBoostingRegressor uses loss="squared_error", 200 estimators, learning rate 0.05, maximum depth 3, minimum leaf size 10 and random_state=42; other settings are the scikit-learn 1.8.0 defaults. It uses the original numeric predictors. All procedures receive the same six predictors and no group label.

Software versions: Python 3.12.14; NumPy 2.3.5; pandas 2.2.3; SciPy 1.17.0; scikit-learn 1.8.0; Matplotlib 3.10.8; threadpoolctl 3.6.0. BLAS and OpenMP threads are limited to one per worker, and the fitting study uses six process workers. Fit times recorded in the archive depend on the computing environment and are not analyzed.

Seed map

The random generator is NumPy default_rng. Pilot design and noise seeds are 730001 and 730002. For stress realization r = 1,…,30, define s = 100000 + 1000r. Training design and noise use s + 1 and s + 2. A test design with G groups uses s + 100 + G. Conditional test noise uses the separate seed 2000000 + 100000r + 1000e + 100h + G, with e = 0 or 5 for $\eta = 0$ or 0.5, and h = 0, 1 or 2 for $\rho = 0$, baseline or 0.5. These seeds are distinct across realization blocks and across profiles, with intentional sharing between A and B within a profile. The seed is saved in every pair record. Noise arrays are drawn as a 2000 by G group array followed by a 2000 by 200 observation array.

For learning realization r, define s = 900000 + 1000r. Maximum training design, training noise, test design and test noise use s + 1 through s + 4. Training sizes select the first n/10 rows of each of the ten groups from the 1000-row maximum design. Both functions share underlying designs and noise within a realization. This matching does not make them independent replicates. Bootstrap calibration intervals use seed 998877; the separate algebra check uses seed 4321.

For Peer Review

S3 Complete focal ranking scenarios

Table S4 reports the Gaussian process versus gradient boosting comparison for every profile. A profile averages 30 independent realization blocks. Mean q is the exact conditional reversal probability averaged over blocks; its MCSE and approximate t interval quantify the precision of this mean, not the binomial error of a single 2000-draw audit. Regret is $|\delta|q/\tau^2$. Profile labels list response function, latent covariate correlation η, noise allocation ρ and test-group count G. Every test set has 200 rows and every fitted training set has 200 rows in ten groups.

Table S4. Complete focal ranking summaries. Probability intervals are intersected with [0,1]. Full results for all ten pairs appear in conditional_ranking.csv.

Profile k / η / ρ / G	Mean q	MCSE	95% interval	Normalized regret	Mean κ
A / 0 / 0.0000 / 5	0.17105	0.02724	(0.11534, 0.22676)	0.01396	1.347
A / 0 / 0.0000 / 10	0.15178	0.02746	(0.09562, 0.20793)	0.01266	1.241
A / 0 / 0.0000 / 20	0.15648	0.02278	(0.10989, 0.20306)	0.01534	1.097
A / 0 / 0.1715 / 5	0.17940	0.02611	(0.12601, 0.23280)	0.01517	1.424
A / 0 / 0.1715 / 10	0.16545	0.02422	(0.11593, 0.21498)	0.01511	1.281
A / 0 / 0.1715 / 20	0.16406	0.02538	(0.11216, 0.21596)	0.01415	1.065
A / 0 / 0.5000 / 5	0.19956	0.02547	(0.14746, 0.25166)	0.01762	1.529
A / 0 / 0.5000 / 10	0.19036	0.02596	(0.13726, 0.24345)	0.01572	1.328
A / 0 / 0.5000 / 20	0.18938	0.02962	(0.12880, 0.24995)	0.01274	1.054
A / 0.5 / 0.0000 / 5	0.15397	0.02932	(0.09400, 0.21395)	0.01159	5.663
A / 0.5 / 0.0000 / 10	0.12528	0.02175	(0.08079, 0.16977)	0.01411	3.562
A / 0.5 / 0.0000 / 20	0.12149	0.02383	(0.07275, 0.17023)	0.01226	1.940
A / 0.5 / 0.1715 / 5	0.17381	0.02609	(0.12044, 0.22717)	0.01869	5.890
A / 0.5 / 0.1715 / 10	0.15763	0.02734	(0.10171, 0.21355)	0.01588	3.357
A / 0.5 / 0.1715 / 20	0.19120	0.02646	(0.13709, 0.24530)	0.01661	1.965
A / 0.5 / 0.5000 / 5	0.22662	0.02644	(0.17254, 0.28069)	0.03205	6.058
A / 0.5 / 0.5000 / 10	0.18534	0.02744	(0.12923, 0.24146)	0.02156	3.283
A / 0.5 / 0.5000 / 20	0.24185	0.03045	(0.17957, 0.30412)	0.01679	2.028
B / 0 / 0.0000 / 5	0.14763	0.02932	(0.08766, 0.20760)	0.01038	1.010
B / 0 / 0.0000 / 10	0.11444	0.02562	(0.06204, 0.16683)	0.01021	1.081
B / 0 / 0.0000 / 20	0.11002	0.02413	(0.06068, 0.15936)	0.01023	1.021
B / 0 / 0.1715 / 5	0.14366	0.02847	(0.08544, 0.20188)	0.01084	1.136
B / 0 / 0.1715 / 10	0.12935	0.02605	(0.07606, 0.18263)	0.01127	1.183
B / 0 / 0.1715 / 20	0.11200	0.02314	(0.06468, 0.15933)	0.01075	1.027
B / 0 / 0.5000 / 5	0.14006	0.02572	(0.08745, 0.19267)	0.01256	1.288
B / 0 / 0.5000 / 10	0.14975	0.02579	(0.09700, 0.20250)	0.01309	1.236
B / 0 / 0.5000 / 20	0.11771	0.02602	(0.06449, 0.17094)	0.01030	1.020
B / 0.5 / 0.0000 / 5	0.06488	0.01846	(0.02713, 0.10263)	0.00774	4.793
B / 0.5 / 0.0000 / 10	0.12770	0.02749	(0.07148, 0.18392)	0.01025	3.577
B / 0.5 / 0.0000 / 20	0.14134	0.02905	(0.08193, 0.20076)	0.01038	1.856
B / 0.5 / 0.1715 / 5	0.10645	0.02431	(0.05674, 0.15616)	0.01155	4.374
B / 0.5 / 0.1715 / 10	0.17792	0.03161	(0.11327, 0.24256)	0.01395	3.372
B / 0.5 / 0.1715 / 20	0.22085	0.03425	(0.15081, 0.29089)	0.01186	1.861
B / 0.5 / 0.5000 / 5	0.14607	0.02606	(0.09278, 0.19937)	0.02203	4.758
B / 0.5 / 0.5000 / 10	0.16948	0.02756	(0.11311, 0.22585)	0.02066	3.382
B / 0.5 / 0.5000 / 20	0.21958	0.03157	(0.15501, 0.28416)	0.01624	1.911

For Peer Review

S4 Complete confidence shortlist scenarios

Five fitted candidates give ten unordered comparisons. Using $\alpha = 0.05$, the normal critical value is 2.8070337683. An observed loss difference minus this critical value times its exact standard deviation must be positive before the first candidate is excluded. Table S5 summarizes the 2000-draw audit for each profile, averaged across 30 blocks. Simultaneous coverage checks all ten expected pair gaps; oracle retention checks the unique best fitted predictor. Retaining the best in a set and choosing it as a single winner are different events.

Table S5. Complete shortlist audit. Every row uses known Gaussian covariance. Profile notation matches Table S4. Percentages average conditional Monte Carlo frequencies; uncertainties for each column are provided in shortlist_summary.csv.

Profile k / η / ρ / G	Joint coverage (%)	Oracle retained (%)	Mean size	Singleton (%)	Observed winner is oracle (%)
A / 0 / 0.0000 / 5	96.91	100.000	1.134	87.90	99.97
A / 0 / 0.0000 / 10	96.95	100.000	1.167	85.56	99.97
A / 0 / 0.0000 / 20	96.86	100.000	1.109	90.11	100.00
A / 0 / 0.1715 / 5	97.01	100.000	1.165	85.52	99.98
A / 0 / 0.1715 / 10	97.12	100.000	1.192	83.31	99.92
A / 0 / 0.1715 / 20	96.93	100.000	1.127	88.59	99.99
A / 0 / 0.5000 / 5	97.04	100.000	1.263	77.83	99.92
A / 0 / 0.5000 / 10	97.02	100.000	1.254	78.47	99.80
A / 0 / 0.5000 / 20	97.07	100.000	1.168	85.18	99.98
A / 0.5 / 0.0000 / 5	97.05	99.990	1.241	82.72	98.05
A / 0.5 / 0.0000 / 10	97.05	100.000	1.226	81.15	99.87
A / 0.5 / 0.0000 / 20	96.89	99.998	1.309	78.21	98.80
A / 0.5 / 0.1715 / 5	97.29	100.000	1.681	53.97	98.39
A / 0.5 / 0.1715 / 10	96.95	99.998	1.611	57.48	96.13
A / 0.5 / 0.1715 / 20	97.05	100.000	1.450	68.47	98.74
A / 0.5 / 0.5000 / 5	97.50	99.998	2.366	21.66	93.48
A / 0.5 / 0.5000 / 10	97.34	99.987	2.170	31.70	94.89
A / 0.5 / 0.5000 / 20	97.09	99.998	1.710	51.80	97.63
B / 0 / 0.0000 / 5	97.08	99.963	2.093	6.13	80.64
B / 0 / 0.0000 / 10	97.06	99.952	2.063	6.39	78.63
B / 0 / 0.0000 / 20	96.90	99.960	2.053	6.98	79.78
B / 0 / 0.1715 / 5	97.04	99.947	2.113	6.32	77.06
B / 0 / 0.1715 / 10	97.17	99.960	2.082	5.67	78.12
B / 0 / 0.1715 / 20	97.07	99.960	2.061	6.92	79.70
B / 0 / 0.5000 / 5	97.12	99.942	2.179	4.81	76.81
B / 0 / 0.5000 / 10	97.08	99.975	2.154	4.74	78.49
B / 0 / 0.5000 / 20	96.97	99.952	2.089	6.58	76.59
B / 0.5 / 0.0000 / 5	97.06	99.963	2.124	12.28	80.94
B / 0.5 / 0.0000 / 10	97.09	99.963	1.945	14.05	81.72
B / 0.5 / 0.0000 / 20	97.00	99.970	1.972	16.41	83.24
B / 0.5 / 0.1715 / 5	97.42	99.955	2.519	5.22	78.69
B / 0.5 / 0.1715 / 10	97.21	99.963	2.247	8.13	77.66
B / 0.5 / 0.1715 / 20	97.22	99.967	2.161	10.33	80.80
B / 0.5 / 0.5000 / 5	97.60	99.922	2.982	1.86	69.95
B / 0.5 / 0.5000 / 10	97.30	99.945	2.690	3.68	75.69
B / 0.5 / 0.5000 / 20	97.19	99.955	2.388	5.27	77.66

Across the balanced grid, joint coverage is 0.971035, oracle retention 0.999773, mean shortlist size 1.812757 and singleton frequency 0.394826. These averages first aggregate contexts within each realization

For Peer Review

block. The lowest context-level retention is 0.9955 and the lowest context-level joint coverage 0.9580; both are 2000-draw estimates for single contexts and lie above the guaranteed 0.95.

S5 Additional checks and learning results

Moment bound without Gaussianity

Let $X = \Delta - \delta$ have mean zero and variance V. If $\delta > 0$, reversal requires $X < -\delta$. For any $h \geq 0$, on that event $(X - h)^2 \geq (\delta + h)^2$. Markov's inequality gives

$$P(X < -\delta) \leq \frac{V+h^2}{(\delta+h)^2} \quad \text{(S13)}$$

Taking $h = V/\delta$ reduces the right-hand side to $V/(V + \delta^2)$. For $\delta < 0$, apply the same argument to $-X$ and $|\delta|$. If V is zero the noise contrast vanishes almost surely, and a nonzero expected gap cannot reverse. This proves the bound using only finite second moments.

Unconditional variance and dependence across comparisons

Removing the conditioning on training data and test design gives $E(\Delta) = E(\delta)$ and $Var(\Delta) = E(V) + Var(\delta)$, by the laws of total expectation and variance. The probability of reversing the conditional expected ordering is E(q), provided oracle ties are dealt with separately. It is generally not equal to a normal tail computed from only the two unconditional moments, because the unconditional contrast is a mixture over changing gaps and variances.

For two ordered pairs (a,b) and (c,d), let their weighted prediction-contrast vectors be d_{ab} and d_{cd}. Their conditional covariance follows from the same linear representation:

$$\mathrm{Cov}(\Delta_{ab}, \Delta_{cd} \mid C) = 4 d_{ab}{}^{T} \Sigma d_{cd} \quad \text{(S14)}$$

This covariance can be nonzero even for pairs without a model in common, since all comparisons use the same test outcomes. The simultaneous shortlist proof uses the union bound and does not treat these pairwise contrasts as independent. In non-Gaussian settings a conservative alternative could be constructed from moment inequalities, but such a procedure is not evaluated here.

Numerical algebra checks

The supplied run_study.py includes a separate 100,000-draw variance check on 40 rows in four groups. Its theoretical variance is 0.057484608, compared with empirical variance 0.058062573; the relative discrepancy is 1.01%. The check also verifies $\kappa = 10$ for unit contrasts aligned within groups of size ten and $\kappa = 0$ for alternating contrasts summing to zero within every group. The shortlist audit separately verifies, for every pair context, that direct differences of squared losses match the linear contrast formula to numerical tolerance 10^{-9}.

These checks are designed to catch implementation errors in sign, normalization and group aggregation. Raw check values are in analytic_checks.json.

Complete learning curve summaries

Table S6 reports all five procedures at all five training sizes, with $\eta = 0$ and baseline noise allocation. Mean and SD are across 30 independent blocks. MCSE is $SD/\sqrt{30}$, and intervals use the Student t quantile with 29 degrees of freedom. Training sizes and test outcomes are paired within a block; comparisons across sizes are not independent observations. Units of RMSE and its uncertainty summaries are percentage points dry matter.

For Peer Review

Table S6. Complete learning-curve results. k identifies the response function and n is the number of training rows. Data and fitted predictions for every block and size are supplied.

k	n	Procedure	Mean RMSE	SD	MCSE	95% interval
A	50	Linear ridge	1.83989	0.07337	0.01339	(1.81250, 1.86729)
A	50	Polynomial ridge	1.12824	0.13634	0.02489	(1.07733, 1.17915)
A	50	Spline ridge	0.91454	0.10196	0.01862	(0.87647, 0.95261)
A	50	Gaussian process	1.30691	0.10404	0.01900	(1.26806, 1.34576)
A	50	Gradient boosting	1.24192	0.11216	0.02048	(1.20004, 1.28381)
A	100	Linear ridge	1.82578	0.08054	0.01470	(1.79571, 1.85585)
A	100	Polynomial ridge	0.90748	0.07573	0.01383	(0.87920, 0.93576)
A	100	Spline ridge	0.73275	0.04875	0.00890	(0.71455, 0.75095)
A	100	Gaussian process	1.02442	0.07958	0.01453	(0.99470, 1.05413)
A	100	Gradient boosting	0.99136	0.07219	0.01318	(0.96441, 1.01832)
A	250	Linear ridge	1.81123	0.07884	0.01440	(1.78179, 1.84067)
A	250	Polynomial ridge	0.80575	0.04426	0.00808	(0.78923, 0.82228)
A	250	Spline ridge	0.67407	0.04482	0.00818	(0.65733, 0.69080)
A	250	Gaussian process	0.84350	0.05137	0.00938	(0.82432, 0.86269)
A	250	Gradient boosting	0.85671	0.05089	0.00929	(0.83771, 0.87572)
A	500	Linear ridge	1.80647	0.07344	0.01341	(1.77905, 1.83389)
A	500	Polynomial ridge	0.78036	0.03773	0.00689	(0.76627, 0.79445)
A	500	Spline ridge	0.65759	0.04033	0.00736	(0.64253, 0.67265)
A	500	Gaussian process	0.78930	0.04299	0.00785	(0.77325, 0.80535)
A	500	Gradient boosting	0.79826	0.05287	0.00965	(0.77852, 0.81801)
A	1000	Linear ridge	1.80466	0.07210	0.01316	(1.77774, 1.83158)
A	1000	Polynomial ridge	0.76800	0.03701	0.00676	(0.75418, 0.78182)
A	1000	Spline ridge	0.64736	0.03715	0.00678	(0.63349, 0.66123)
A	1000	Gaussian process	0.75843	0.03818	0.00697	(0.74417, 0.77268)
A	1000	Gradient boosting	0.76071	0.04233	0.00773	(0.74491, 0.77652)
B	50	Linear ridge	1.99353	0.08354	0.01525	(1.96234, 2.02473)
B	50	Polynomial ridge	1.00761	0.12538	0.02289	(0.96079, 1.05442)
B	50	Spline ridge	0.93437	0.09797	0.01789	(0.89779, 0.97095)
B	50	Gaussian process	1.38356	0.11014	0.02011	(1.34243, 1.42468)
B	50	Gradient boosting	1.33255	0.13790	0.02518	(1.28105, 1.38404)
B	100	Linear ridge	1.97633	0.08346	0.01524	(1.94516, 2.00749)
B	100	Polynomial ridge	0.80781	0.06870	0.01254	(0.78216, 0.83347)
B	100	Spline ridge	0.75983	0.05570	0.01017	(0.73903, 0.78063)
B	100	Gaussian process	1.03807	0.08423	0.01538	(1.00662, 1.06952)
B	100	Gradient boosting	0.99861	0.06984	0.01275	(0.97254, 1.02469)
B	250	Linear ridge	1.95928	0.08277	0.01511	(1.92837, 1.99019)
B	250	Polynomial ridge	0.71369	0.04274	0.00780	(0.69773, 0.72965)
B	250	Spline ridge	0.69639	0.04807	0.00878	(0.67844, 0.71434)
B	250	Gaussian process	0.79917	0.04261	0.00778	(0.78326, 0.81508)
B	250	Gradient boosting	0.85132	0.04991	0.00911	(0.83268, 0.86996)
B	500	Linear ridge	1.95346	0.07849	0.01433	(1.92415, 1.98276)
B	500	Polynomial ridge	0.69301	0.03774	0.00689	(0.67892, 0.70711)
B	500	Spline ridge	0.67842	0.04480	0.00818	(0.66169, 0.69515)
B	500	Gaussian process	0.74153	0.03747	0.00684	(0.72754, 0.75552)
B	500	Gradient boosting	0.78594	0.05217	0.00952	(0.76646, 0.80542)
B	1000	Linear ridge	1.95079	0.07626	0.01392	(1.92231, 1.97926)
B	1000	Polynomial ridge	0.68103	0.03573	0.00652	(0.66769, 0.69438)
B	1000	Spline ridge	0.66809	0.04171	0.00762	(0.65251, 0.68366)
B	1000	Gaussian process	0.71193	0.03790	0.00692	(0.69778, 0.72608)
B	1000	Gradient boosting	0.74473	0.04372	0.00798	(0.72841, 0.76106)

For Peer Review

S6 Data dictionary and reproduction

Table S7. Main row-data fields. File names identify experiment, response function, covariate and noise profiles, realization, training size or test-group count.

Field	Meaning
loading, pH, temperature, time, moisture, initial_protein	The six numeric model inputs in the units of Table S1.
group	Zero-based group identifier, local to one training or test dataset. The same numeric label in another file does not imply the same realized group.
conditional_mean	The known response function evaluated at the row's inputs; present in main-study row files.
noise	The saved realized group plus observation noise.
response	Conditional mean plus the saved noise.
prediction_linear_ridge, prediction_polynomial_ridge, prediction_spline_ridge, prediction_gaussian_process, prediction_gradient_boosting	Predictions from the five procedures fitted on the associated training file; present only in test files.
target_A, target_B	Pilot response columns from the two response functions; their design and noise are shared.

For Peer Review