HILBERT SPLIT: GEOMETRY-AWARE DATA SPLITTING FOR GENERALIZATION ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Data splitting is a commonly used technique for assessing how well predictive models generalize. The challenge is to preserve the distribution of the data relevant to prediction while limiting train–test similarities that can make evaluation overly optimistic. We introduce Hilbert Split, which identifies response-relevant predictors through sensitivity analysis. The local regression fits underlying this analysis provide a geometry for separating observations. A Hilbert ordering of the selected predictors and response organizes the data into local strata, within which this geometry guides test-point selection. This construction preserves distributional coverage while reducing train-test similarity. For randomized within-stratum selection, we establish sharper concentration of summary averages than simple random splitting and Wasserstein control of distributional discrepancy, with rates governed by the selected dimension.

1 INTRODUCTION

Splitting data into training and testing sets is a fundamental procedure for validating models in machine learning, in both supervised and unsupervised settings. Despite its importance, practitioners often default to a simple random split because of its ease of implementation and computational efficiency. The resulting variation in the allocation can affect the reliability and reproducibility of empirical evaluations (Snee, 1977; Picard & Berk, 1990).

Several deterministic methods have been proposed, beginning with DUPLEX (Snee, 1977). More recent algorithms, including SPlit (Joseph & Vakayil, 2022) and Twinning (Vakayil & Joseph, 2022), seek statistical similarity between subsets through Energy distance. Although applications may use training, validation, and test sets, these can be constructed by successive binary partitions. We therefore consider the binary train–test problem. In this work, we pursue two complementary objectives.

1. **Deviation control under performance metrics.** Models are evaluated by appropriate performance metrics, such as squared error loss in regression. Large differences between subset averages can make model comparisons sensitive to the allocation. We seek to control these differences and retain the distribution relevant to prediction.
2. **Separation for generalization assessment.** Data leakage can arise when repeated or dependent observations carry shared information across the training–test boundary (Roberts et al., 2017; Kapoor & Narayanan, 2023). Test error can then reward a model for exploiting that overlap rather than predicting new conditions. Controlling training–test similarity helps the test set distinguish candidate models by their generalization performance.

Evaluating a model by a chosen metric can be viewed as estimating its expected loss on each subset. Controlling the deviation therefore leads to bounds on differences between these estimators. We use a Hilbert ordering (Gerber et al., 2019) to form disjoint local strata. Within these strata, randomized selection has tighter concentration than random splitting. This provides the probabilistic framework for the subsequent geometric allocation.

The gain from Hilbert stratification weakens as the ordering dimension increases. Irrelevant predictors therefore reduce this gain without adding response information. They also complicate separation: high-dimensional distances can lose contrast (Beyer et al., 1999; Aggarwal et al., 2001), and observations with similar predictive content can appear far apart. Hilbert Split addresses both issues through

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

local-linear fits (Anonymous Authors, 2026a). Their residual-risk contrasts identify response-relevant predictors for the joint predictor–response ordering. On the selected coordinates, fitted slopes define a local diagonal geometry that guides test-point selection within the resulting Hilbert strata.

Related work also addresses the effect of similarity on evaluation (Roberts et al., 2017; Rabinowicz & Rosset, 2022; Kapoor & Narayanan, 2023). MaxTwin combines representation and separation (Anonymous Authors, 2026b). DataSAIL controls similarity across folds subject to size and stratification constraints (Joeres et al., 2025; 2026). SAE targets a distribution of test-to-training similarity (Zhang et al., 2025), and SPECTRA examines evaluation across overlap levels (Ektefaie et al., 2024). Hilbert Split combines distributional control with separation in a response-relevant local geometry.

We begin in section 2 with the statistical properties of Hilbert stratification. The dependence of these properties on dimension motivates the variable selection and local distance in section 3. Section 4 brings these together in a sequential splitting procedure, which we evaluate in section 5 through distributional representation, model assessment, and generalization after model selection. The overall framework is illustrated in figure 1 below.

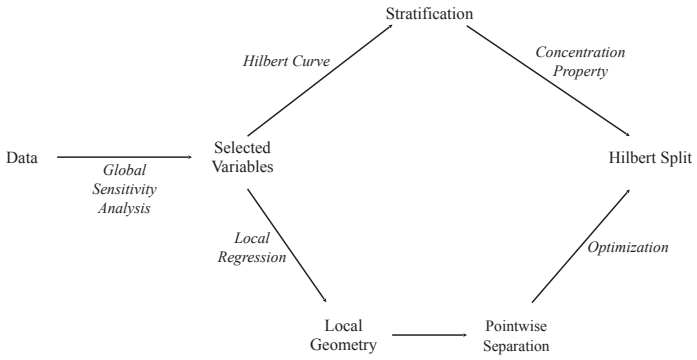

Figure 1: Construction of Hilbert Split. Selected variables define Hilbert strata and a local regression geometry for sequential test-point selection.

2 Statistical Properties of Hilbert Split

In machine learning, a performance metric is minimized on the training set and evaluated on the test set. We first study the difference between the corresponding subset averages. For theoretical convenience, let $D_1, \ldots, D_n \in [0,1]^d$ be fixed, let T and $R = T^c$ denote test and training, and put $|T| = m$ and $\rho = m/n \leq 1/2$. For a fixed summary ϕ, write $\bar{\phi}_S = |S|^{-1} \sum_{i \in S} \phi(D_i)$ and $\bar{\phi}_D = n^{-1} \sum_i \phi(D_i)$. The representation discrepancy is

$$\Delta_\phi(T, R) = |\bar{\phi}_T - \bar{\phi}_R|. \tag{1}$$

We compare random splitting with Hilbert stratification while holding ϕ fixed. Probabilities below are conditional on the sample and its representation.

Proposition 2.1 (Random splitting). *If T is uniform over the m-subsets of $\{1, \ldots, n\}$, then $\mathbb{E}\bar{\phi}_T = \mathbb{E}\bar{\phi}_R = \bar{\phi}_D$. With $\sigma_D^2 = n^{-1} \sum_i \{\phi(D_i) - \bar{\phi}_D\}^2$,*

$$\mathrm{Var}(\bar{\phi}_T - \bar{\phi}_R) = \frac{n^2 \sigma_D^2}{m(n-m)(n-1)}.$$

For every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$|\bar{\phi}_T - \bar{\phi}_R| \leq \frac{1}{\sqrt{(1-\rho)\delta}} \left\{ \frac{\sigma_D^2}{m} \frac{n}{n-1} \right\}^{1/2}. \tag{2}$$

2.1 HILBERT ORDERING AND DISJOINT STRATA

The Hilbert curve provides a locality-preserving scalar ordering of a multivariate sample (Hilbert, 1891; Moon et al., 2001). We use this order to construct disjoint strata. Let $H_d : [0,1] \to [0,1]^d$ be a Hilbert curve and h_d a one-to-one Borel-measurable right inverse (Gerber et al., 2019). The locality bound of He & Owen (2016) is

$$\|H_d(u) - H_d(v)\|_2 \le 2\sqrt{d+3}\,|u-v|^{1/d}. \tag{3}$$

Sort the observations by $h_d(D_i)$. With $s_b = \lfloor bn/m \rfloor$ and $s_0 = 0$, let C_b contain ranks $s_{b-1} + 1, \dots, s_b$ and $p_b = |C_b| \ge 2$. The randomized design independently draws $J_b \sim \mathrm{Unif}(C_b)$ and sets $T = \{J_1, \dots, J_m\}$. For unequal block sizes, use

$$\widehat{\phi}_T = \frac{1}{n}\sum_b p_b \phi(D_{J_b}), \qquad \widehat{\phi}_R = \frac{1}{n}\sum_b \frac{p_b}{p_b - 1} \sum_{i \in C_b \setminus \{J_b\}} \phi(D_i). \tag{4}$$

These equal the ordinary test and training averages when m divides n.

2.2 CONCENTRATION BOUNDS FOR HILBERT STRATIFICATION

Proposition 2.2 (Conditional unbiasedness)**.** *Independent uniform selection within strata gives* $\mathbb{E}\widehat{\phi}_T = \mathbb{E}\widehat{\phi}_R = \bar{\phi}_D$.

We adapt the Hilbert-ordered resampling arguments of Li et al. (2022) to complementary training and test sets. Within-stratum cancellation and Hilbert locality yield the following bounds.

Theorem 2.3 (Variance and concentration)**.** *Let* $d \ge 2$ *and let* $\phi : [0,1]^d \to [0,1]$ *be* L_ϕ*-Lipschitz. Then*

$$\mathrm{Var}(\widehat{\phi}_T),\ \mathrm{Var}(\widehat{\phi}_R) \le (1+2\rho)^2 (d+3) L_\phi^2 m^{-1-2/d}. \tag{5}$$

For every $t > 0$*,*

$$\mathbb{P}\{|\widehat{\phi}_T - \widehat{\phi}_R| \ge t\} \le 4\exp\left[-\frac{n^2 t^2/8}{(\rho^{-1}+2)^2(d+3)L_\phi^2 m^{1-2/d} + (nt/6)(\rho^{-1}+2)}\right]. \tag{6}$$

Corollary 2.4 (Hölder-continuous functions)**.** *Under the remaining assumptions of theorem 2.3, if* $|\phi(x) - \phi(z)| \le L_\phi \|x - z\|_2^\alpha$ *for* $\alpha \in (0,1]$*, then* (5) *holds with* $m^{-1-2\alpha/d}$*, and* (6) *holds with* $m^{1-2\alpha/d}$ *in its variance term.*

For Lipschitz functions, the deviation scale at fixed confidence improves from $m^{-1/2}$ under random splitting to $m^{-1/2-1/d}$ under Hilbert stratification. This gain decreases as the dimension increases. Proofs appear in appendix A, together with ordering-rate optimality (proposition A.1) and Wasserstein control (theorem A.2). We therefore identify response-relevant predictors before stratification, reducing the effective dimension.

3 RESPONSE-RELEVANT GEOMETRY

The Hilbert construction controls differences between subset averages, but its usefulness depends on the coordinates being ordered. Predictors unrelated to the response increase the dimension without adding useful information. They can also make two observations with similar predictive content appear far apart. We use local regression to overcome both parts of this problem: residual risks identify the coordinates to retain, and fitted slopes determine how distances should be measured within those coordinates.

Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, with $X_i \in \mathbb{R}^p$ and $\mathbb{E}(Y^2) < \infty$. Min–max scaling each coordinate of (X_i, Y_i) to $[0,1]$ gives the full joint representation D_i of section 2, with $d = p + 1$. Write $f_0(X) = \mathbb{E}(Y \mid X)$ for the regression mean.

3.1 IDENTIFYING THE RELEVANT COORDINATES

For regression, the coordinates used for splitting should retain the conditional mean. A set $A \subseteq \{1, \ldots, p\}$ is *mean-sufficient* if

$$\mathbb{E}(Y \mid X_A) = f_0(X) \qquad \text{almost surely.} \tag{7}$$

Such a set represents the regression mean using X_A alone. We seek a small set satisfying this criterion so that the Hilbert construction operates in a low-dimensional, response-relevant space.

We use the local-linear screening procedure developed for PRIME (Anonymous Authors, 2026a). It assesses whether adding or removing a predictor changes the estimated residual prediction risk. Denote the selected set by $\widehat{A}$ and its size by $a = |\widehat{A}| \geq 1$. The estimator and implementation settings can be found in appendix B. Other model-free selectors can also be used. Using the selected predictors and response reduces the ordering dimension from d to $a + 1$. Under the assumptions of theorem 2.3 in the respective representations, this reduces the randomized-stratification deviation scale from $m^{-1/2-1/d}$ to $m^{-1/2-1/(a+1)}$.

3.2 CONSTRUCTING THE LOCAL DISTANCE

The effect of a selected predictor can vary across the input space. We use local-linear regression to measure this variation. Write $\widetilde{X}_{\widehat{A}}$ for the selected predictors standardized by their sample means and standard deviations. Let $\widehat{g}(x)$ denote the local-linear slope vector fitted in these coordinates at x. The fitting details are given in appendix B. The contribution of coordinate j to a small displacement $z - x$ is $\widehat{g}_j(x)(z_j - x_j)$. We measure the size of a displacement by adding the squared coordinatewise contributions. Opposing signs then cannot cancel one another in the distance.

For a nonzero slope vector, define

$$D(x) = a\,\frac{\operatorname{diag}\{\widehat{g}_1(x)^2, \ldots, \widehat{g}_a(x)^2\}}{\|\widehat{g}(x)\|_2^2}, \qquad \Phi_x(z) = D(x)^{1/2}(z - x). \tag{8}$$

Thus a difference along a locally influential coordinate contributes more to $\|\Phi_x(z) - \Phi_x(w)\|_2$. Squared derivatives are used in sensitivity analysis (Sobol' & Kucherenko, 2009). Here the fitted slopes weight coordinate differences according to their local effects on the response. The trace normalization fixes the total weight at a and removes response units. Implementation conventions for zero slopes and one-dimensional representations are given in appendix B.2.

Residual prediction risks guide variable selection for Hilbert ordering. On the retained coordinates, the local slopes define the distances used in sequential training–test allocation.

4 ALGORITHM OF HILBERT SPLIT

We select one test observation from each stratum in Hilbert order and assign the remaining observations to training. The local distances from section 3 guide each choice by accounting for its effect on the separation between the training and test observations assigned so far.

4.1 JOINT HILBERT STRATA

Scale each selected predictor and the response to $[0, 1]$ using the development-sample minimum and maximum. For a nonconstant coordinate v, write $s_v(v_i) = (v_i - \min_k v_k)/(\max_k v_k - \min_k v_k)$ and form

$$Z_i^H = \left\{ s_{X_j}(X_{ij}) : j \in \widehat{A};\ s_Y(Y_i) \right\} \in [0, 1]^{a+1}. \tag{9}$$

Sorting the joint observations by their Hilbert indices gives a permutation π. For the requested test size m, use the boundaries $s_b = \lfloor bn/m \rfloor$ from section 2. Stratum C_b contains sorted ranks $s_{b-1} + 1, \ldots, s_b$ and has size $p_b = s_b - s_{b-1}$; it contributes one test point. The sizes may differ by one. The ordering uses both selected predictors and response, so its dimension is $a + 1$. To evaluate the geometry at a fixed location in each stratum, take its central rank:

$$i_b^\circ = \pi\{s_{b-1} + \lceil p_b/2 \rceil\}, \qquad x_b^\circ = \widetilde{X}_{i_b^\circ, \widehat{A}}. \tag{10}$$

4.2 SEQUENTIAL SELECTION

Set $D_b = D(x_b^\circ)$ and $\Phi_b = \Phi_{x_b^\circ}$; for $a \geq 2$, write $\widehat{g}_b = \widehat{g}(x_b^\circ)$ for the evaluated slope. All candidates in the stratum share this diagonal, which is held fixed during allocation. Distances are

$$d_b(i,j) = \left\| \Phi_b(\widetilde{X}_{i,\widehat{A}}) - \Phi_b(\widetilde{X}_{j,\widehat{A}}) \right\|_2 . \tag{11}$$

Each test point is compared with training observations using the local geometry of its stratum. We select test points sequentially in Hilbert order, keeping earlier assignments fixed. Within each stratum, we choose the candidate that maximizes the minimum local distance over all newly formed training–test pairs.

Before stage b, let $T_{b-1} = \{t_1, \ldots, t_{b-1}\}$ and $R_{b-1} = \bigcup_{k<b}(C_k \setminus \{t_k\})$. Selecting $t \in C_b$ for test assigns $U_b(t) = C_b \setminus \{t\}$ to training and introduces three types of training–test relations:

$$A_b(t) = \min_{j \in R_{b-1}} d_b(t,j), \tag{12}$$

$$B_b(t) = \min_{j \in U_b(t)} d_b(t,j), \tag{13}$$

$$G_b(t) = \min_{k<b} \min_{j \in U_b(t)} d_k(t_k, j). \tag{14}$$

The first two compare the new test candidate with previous training points and with the remaining points in its stratum. The third compares those remaining points with earlier test points, using the geometry of each earlier test point. At the first stratum, set $A_1(t) = G_1(t) = +\infty$, so only $B_1(t)$ enters the score. Select

$$S_b(t) = \min\{A_b(t), B_b(t), G_b(t)\}, \qquad t_b \in \arg\max_{t \in C_b} S_b(t). \tag{15}$$

Numerical score tolerances and tie-breaking are specified in appendix B.2.

Algorithm 1 Hilbert Split

Require: Development sample $\mathcal{D}$ of size n, test size $1 \leq m \leq \lfloor n/2 \rfloor$

1: Estimate $\widehat{A}$ using the local-linear risk contrasts
2: Min–max scale selected predictors and response as in (9)
3: Hilbert-sort; form C_b from ranks $\lfloor (b-1)n/m \rfloor + 1$ through $\lfloor bn/m \rfloor$
4: Store $D_b = [1]$ if $a = 1$; otherwise evaluate anchor slopes and form D_b
5: $T_0 \leftarrow \varnothing$; $R_0 \leftarrow \varnothing$
6: **for** $b = 1, \ldots, m$ **do**
7: Evaluate $S_b(t)$ for every $t \in C_b$ using (12)–(15)
8: Select a maximizer t_b, breaking ties by central rank and then row key
9: $T_b \leftarrow T_{b-1} \cup \{t_b\}$; $R_b \leftarrow R_{b-1} \cup (C_b \setminus \{t_b\})$
10: **end for**
11: **return** (T_m, R_m)

5 EMPIRICAL EVALUATION

We compare Random, SPlit, Twinning, MaxTwin, and Hilbert Split using a 20% test fraction. All splitters receive the observed predictors and response, and every predictive model receives the full predictor vector. Known active coordinates are used only for evaluation diagnostics. We first examine distributional coverage with irrelevant predictors, then contrast model assessment on real data. A paired-data illustration makes the allocation mechanism visible, and the final study measures the independent generalization error of models selected using each split. Appendix B gives implementation and experimental details.

5.1 REPRESENTATION WITH IRRELEVANT PREDICTORS

We use the eight-input Borehole function (Surjanovic & Bingham, 2013), drawing its inputs independently and uniformly over their physical ranges. We add Gaussian response noise with standard

deviation 9.34093 and append $q \in \{0, 5, 10, 20, 50, 100\}$ independent uniform irrelevant predictors. Each condition has $n = 2000$ and 100 replications. Energy distance (Székely & Rizzo, 2013) is evaluated in all eight function inputs and the response, standardized over the pooled sample. Within each replication, the active sample is shared across q. stochastic allocations are redrawn in every condition.

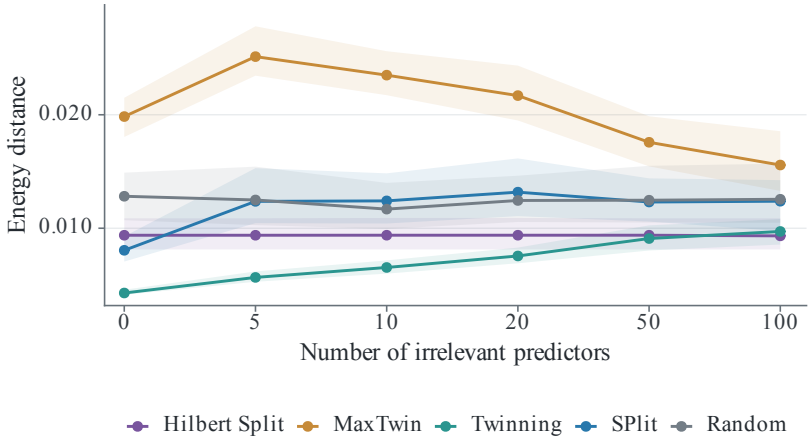

Figure 2: Borehole: energy distance under added irrelevant predictors. Lines and bands show the median and interquartile range over 100 replications.

As irrelevant predictors are added, Energy increases for SPlit and Twinning in figure 2. Their splits become less representative in the coordinates relevant to the response, even though the full input space is used to construct them. MaxTwin has the largest discrepancy, illustrating the loss of coverage that can accompany separation in that space. Hilbert Split remains stable because screening largely excludes the added predictors before the Hilbert strata and local distances are constructed. Random splitting also changes little across dimensions, but gives less representative subsets than Hilbert Split.

5.2 Assessment on real regression data

We compare a model that closely fits individual training observations with a model that fits a linear trend. Nearby training observations can make the first model appear accurate on the test set even when its predictions away from those observations are less reliable. We examine whether Hilbert Split makes this dependence more visible without producing a similar increase in the training–test gap for the linear model.

Tecator predicts fat content from 100 near-infrared measurements (Tecator, n.d.). Airfoil predicts scaled sound-pressure level from frequency and four aerodynamic variables (Brooks et al., 1989). We draw 20 subsets of size 200 from Tecator's 215 eligible records, and 20 of size 800 from Airfoil's 1503 records, using the same subset for all five splitting methods in each repetition. On each training set we fit Gaussian kernel ridge regression (KRR) with a small ridge penalty as a quasi-interpolating model, together with linear ridge regression. The sampled subsets overlap within each source dataset. For each fitted model, we report $|\,\mathrm{MSE}_{\mathrm{test}} - \mathrm{MSE}_{\mathrm{train}}\,|$, with matched axes for the two learners within each dataset.

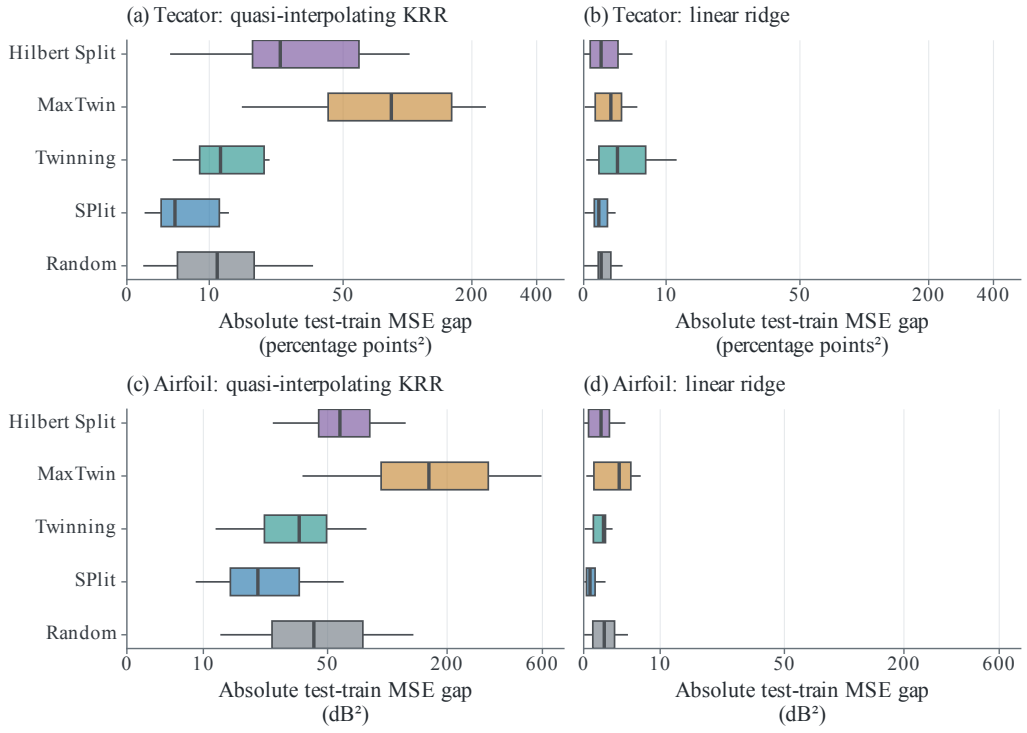

Figure 3: Absolute training–test MSE gaps for Gaussian KRR and linear ridge regression.

Hilbert Split produces a wider gap for the quasi-interpolating KRR than Random, SPlit, and Twinning on both datasets (figure 3). For linear ridge regression, the gaps remain small on the same axes. The contrast is consistent with the intended role of separation: evaluation becomes more demanding for a fit that depends on nearby observations, while a linear fit is less affected by their allocation. The separation-oriented MaxTwin also produces a pronounced kernel gap. The paired-data illustration below shows how the allocations differ. The model-selection experiment in section 5.4 evaluates the generalization of models chosen using each split.

5.3 Separation of near-repeated observations

We next visualize how the methods allocate near-repeated observations when irrelevant coordinates obscure their similarity. Consider the Friedman function (Friedman, 1991)

$$f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5. \tag{16}$$

Figure 4 uses fifteen pairs whose active coordinates $X_{1:5}$ are close, while the five irrelevant coordinates $X_{6:10}$ vary independently between records. Here $Y = f(X)$. All methods receive all ten predictors and the response. The horizontal coordinate is a common PC1 of standardized active predictors and the vertical coordinate is Y. This reveals the pair structure relevant to prediction while keeping the displayed coordinates fixed across methods.

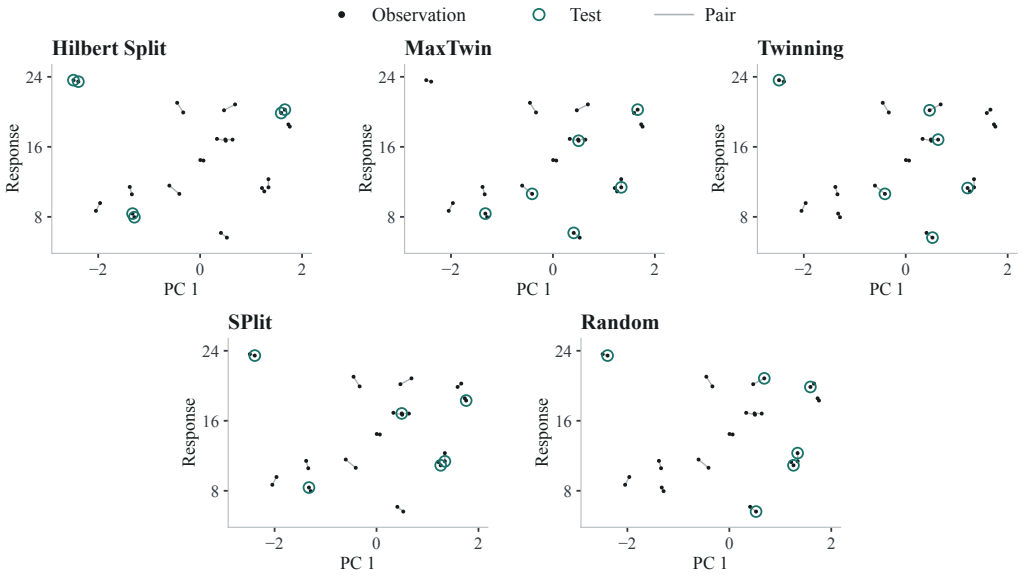

Figure 4: Visualization of data splitting for paired Friedman data.

In this example Hilbert Split assigns complete pairs to the test set. Under each comparator, the test observations have their near mate in the training set. Selecting a pair together removes this immediate training analogue. Hilbert stratification spreads the test points across the joint predictor–response distribution, and the local geometry guides selection away from such cross-set redundancies, even when independent irrelevant coordinates obscure the pairs in the full input space.

5.4 Generalization after test-based selection

We assess generalization by selecting kernel parameters on the test set and evaluating the full-data refits on independent observations. The target population has 10 independent $\mathrm{Unif}(0,1)$ predictors and regression mean (16). To combine broad coverage with local redundancy, we construct 100 support-point centers $c_h \in [0.005, 0.995]^5$ (Mak & Joseph, 2018) and create two observations per center, giving $n = 200$:

$$\begin{aligned} &A_{h\ell} = c_h + U_{h\ell}, \quad X_{h\ell} = (A_{h\ell}, Z_{h\ell}), \\ &Y_{h\ell} = f(A_{h\ell}) + 0.1 s_f \varepsilon_{h\ell}, \qquad h = 1, \ldots, 100, \quad \ell = 1, 2. \end{aligned} \tag{17}$$

The components of $U_{h\ell}$ and $Z_{h\ell}$ are independent $\mathrm{Unif}[-0.005, 0.005]$ and $\mathrm{Unif}(0,1)$, respectively, and $\varepsilon_{h\ell} \sim N(0,1)$ independently. Here s_f is a fixed reference standard deviation of f. Its calibration and the support-point construction are given in appendix B.5. The pairs are close in the five active coordinates but can be far apart in the five irrelevant coordinates.

Following (Anonymous Authors, 2026b), each method fits the same anisotropic Gaussian KRR family on the training set using all 10 predictors. The 10 length scales and ridge penalty are selected by minimizing test MSE through a common candidate search and local refinement. These parameters are then fixed for refitting on the full sample. The resulting predictor is evaluated against the true f at 160,000 independent uniform reference points, excluded from parameter selection.

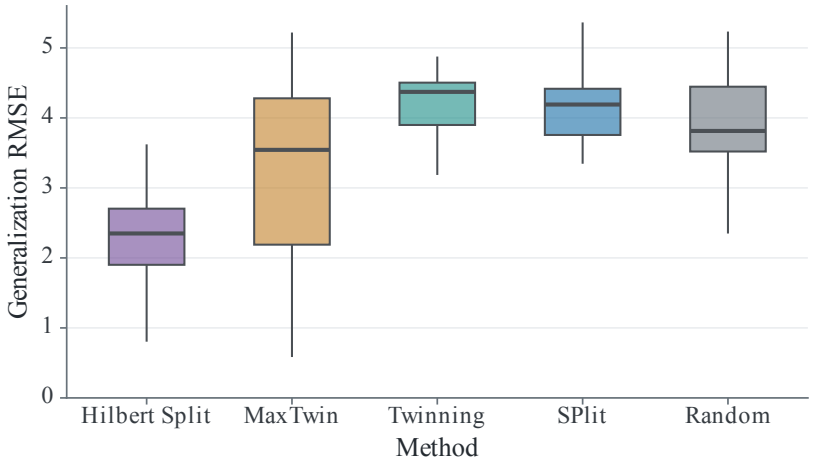

Figure 5: Generalization error of KRR models selected using each splitting method.

We can see from figure 5 that the errors from the model based on Hilbert Split are much smaller than those of the other competing methods. Since the selected models are refitted on the same full sample, this improvement reflects the parameters chosen using each split rather than a difference in the final training data. Hilbert Split identifies the active coordinates and usually places both members of a near pair in the same subset. Its test set therefore relies less on immediate training analogues, while Hilbert stratification maintains coverage of the relevant space. In this design, these allocations lead to parameter choices that generalize better to independent observations. The 30 repetitions measure split/search variation on the fixed observed design. Hilbert Split's allocation is deterministic, so its variability comes from the parameter search. The search settings and numerical status are given in appendix B.5.

6 Conclusion

We introduced Hilbert Split to preserve representative training and test sets while reducing cross-set similarities that can make generalization assessment overly optimistic. Sensitivity analysis identifies the predictors for low-dimensional Hilbert stratification, and the same local regression fits define the geometry for sequential separation. This combines distributional coverage with test-point selection guided by the prediction problem. The theory quantifies the distributional control supplied by Hilbert stratification, while the experiments demonstrate robustness to irrelevant predictors and benefits for assessing and selecting models under local redundancy.

AI USE STATEMENT

A generative AI tool assisted with manuscript polishing, code implementation, and experiments.

REPRODUCIBILITY STATEMENT

The randomized results are proved in appendix A. The algorithm, screening settings, numerical conventions, and experiment specifications are given in section 3 and appendix B. The reproduction materials contain recorded seeds, allocations, numerical outcomes, and source fingerprints. Saved-model replay checks the independent generalization errors, and reporting scripts regenerate the experimental summaries. Synthetic data and reference points accompany the records. Measured-data acquisition instructions retain the original usage terms.

REFERENCES

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pp. 420–434. Springer, 2001. doi: 10.1007/3-540-44503-X_27.

Anonymous Authors. Measuring variable importance using proportional inclusive marginal effects. Unpublished working manuscript, 2026a.

Anonymous Authors. A robust deterministic data splitting method for model validation. Unpublished manuscript, 2026b.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999. doi: 10.1007/3-540-49257-7_15.

Thomas F. Brooks, D. Stuart Pope, and Michael A. Marcolini. Airfoil self-noise. UCI Machine Learning Repository dataset, `https://doi.org/10.24432/C5VW2C`, 1989. URL `https://archive.ics.uci.edu/dataset/291/airfoil+self+noise`.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G. Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6:1512–1524, 2024. doi: 10.1038/s42256-024-00931-6.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991. doi: 10.1214/aos/1176347963.

Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260, 2019. doi: 10.1214/18-AOS1746.

Zhijian He and Art B. Owen. Extensible grids: Uniform sampling on a space-filling curve. *Journal of the Royal Statistical Society: Series B*, 78(4):917–931, 2016. doi: 10.1111/rssb.12132.

David Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38(3):459–460, 1891. doi: 10.1007/BF01199431. `https://doi.org/10.1007/BF01199431`.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16:3337, 2025. doi: 10.1038/s41467-025-58606-8.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Addendum: Data splitting against information leakage with DataSAIL. *Nature Communications*, 17:1597, 2026. doi: 10.1038/s41467-025-67495-w.

V. Roshan Joseph and Akhil Vakayil. SPlit: An optimal method for data splitting. *Technometrics*, 64 (2):166–176, 2022. doi: 10.1080/00401706.2021.1921037.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Yichao Li, Wenshuo Wang, Ke Deng, and Jun S. Liu. Stratification and optimal resampling for sequential Monte Carlo. *Biometrika*, 109(1):181–194, 2022. doi: 10.1093/biomet/asab004.

Simon Mak and V. Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018. doi: 10.1214/17-AOS1629.

Bongki Moon, H. V. Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):124–141, 2001. doi: 10.1109/69.908985.

Richard R. Picard and Kenneth N. Berk. Data splitting. *The American Statistician*, 44(2):140–147, 1990. doi: 10.1080/00031305.1990.10475704.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9: 129–145, 1935.

Assaf Rabinowicz and Saharon Rosset. Cross-validation for correlated data. *Journal of the American Statistical Association*, 117(538):718–731, 2022. doi: 10.1080/01621459.2020.1801451.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Ronald D. Snee. Validation of regression models: Methods and examples. *Technometrics*, 19(4): 415–428, 1977. doi: 10.1080/00401706.1977.10489581.

I. M. Sobol' and S. Kucherenko. Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79(10):3009–3017, 2009. doi: 10.1016/j.matcom.2009.01.023.

Sonja Surjanovic and Derek Bingham. Virtual library of simulation experiments: Test functions and datasets. Simon Fraser University, 2013. URL `https://www.sfu.ca/~ssurjano/`.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Tecator. Tecator meat spectroscopy data. StatLib dataset, n.d. URL `https://lib.stat.cmu.edu/datasets/tecator`. Original 215 calibration, monitoring, and test observations.

Akhil Vakayil and V. Roshan Joseph. Data twinning. *Statistical Analysis and Data Mining*, 15(5): 598–610, 2022. doi: 10.1002/sam.11574.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Chenbin Zhang, Zhiqiang Hu, Chuchu Jiang, Wen Chen, Jie Xu, and Shaoting Zhang. Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=j7cyANIAxV`.

A PROOFS AND FURTHER STATISTICAL PROPERTIES

A.1 RANDOM SPLITTING

Every observation enters T with probability m/n, and hence

$$\mathbb{E}(\bar{\phi}_T) = \frac{1}{m}\sum_{i=1}^{n}\phi(D_i)\frac{m}{n} = \bar{\phi}_D.$$

The same calculation applies to R. The finite-population variance is

$$\mathrm{Var}(\bar{\phi}_T) = \frac{\sigma_D^2}{m}\frac{n-m}{n-1}. \tag{18}$$

Because

$$n\bar{\phi}_D = m\bar{\phi}_T + (n-m)\bar{\phi}_R,$$

we have

$$\bar{\phi}_T - \bar{\phi}_R = \frac{n}{n-m}(\bar{\phi}_T - \bar{\phi}_D)$$

and therefore

$$\mathrm{Var}(\bar{\phi}_T - \bar{\phi}_R) = \left(\frac{n}{n-m}\right)^2 \frac{\sigma_D^2}{m}\frac{n-m}{n-1}.$$

Chebyshev's inequality gives (2). □

A.2 CONDITIONAL UNBIASEDNESS

Uniform selection in C_b gives

$$\mathbb{E}\{p_b\phi(D_{J_b})\} = \sum_{i\in C_b}\phi(D_i).$$

Also, each index in C_b remains in training with probability $(p_b-1)/p_b$, so

$$\mathbb{E}\left\{\frac{p_b}{p_b-1}\sum_{i\in C_b\setminus\{J_b\}}\phi(D_i)\right\} = \sum_{i\in C_b}\phi(D_i).$$

Summing over the disjoint strata and dividing by n proves both identities. □

A.3 VARIANCE AND CONCENTRATION

Write $u_i = h_d(D_i)$. For stratum C_b, let

$$a_b = \min_{i\in C_b} u_i, \qquad b_b = \max_{i\in C_b} u_i, \qquad \Delta_b = b_b - a_b.$$

The strata are ordered and disjoint, so $\sum_{b=1}^{m}\Delta_b \le 1$. Since $d \ge 2$, concavity gives

$$\sum_{b=1}^{m}\Delta_b^{2/d} \le m^{1-2/d}\left(\sum_{b=1}^{m}\Delta_b\right)^{2/d} \le m^{1-2/d}. \tag{19}$$

For $i, j \in C_b$, (3) and Lipschitz continuity give

$$|\phi(D_i) - \phi(D_j)| \le 2\sqrt{d+3}L_\phi\Delta_b^{1/d}.$$

Popoviciu's inequality (Popoviciu, 1935) therefore implies

$$\mathrm{Var}\{\phi(D_{J_b})\} \le (d+3)L_\phi^2\Delta_b^{2/d}. \tag{20}$$

The stratum draws are independent. Since $p_b \le n/m + 2 = \rho^{-1} + 2$, equations (19)–(20) yield

$$\begin{aligned}\mathrm{Var}(\widehat{\phi}_T) &= \frac{1}{n^2}\sum_{b=1}^{m} p_b^2 \,\mathrm{Var}\{\phi(D_{J_b})\} \\ &\le \frac{(\rho^{-1}+2)^2}{n^2}(d+3)L_\phi^2 m^{1-2/d} \\ &= (1+2\rho)^2(d+3)L_\phi^2 m^{-1-2/d}.\end{aligned}$$

For the training estimator, the random part of the contribution from block b is

$$-\frac{1}{n}\frac{p_b}{p_b-1}\phi(D_{J_b}).$$

Because $p_b \ge 2$ and $p_b/(p_b-1) \le p_b$, the same calculation proves the second variance bound.

For concentration, define the centered block contributions

$$X_b = p_b[\phi(D_{J_b}) - \mathbb{E}\{\phi(D_{J_b})\}], \qquad Y_b = -\frac{p_b}{p_b-1}[\phi(D_{J_b}) - \mathbb{E}\{\phi(D_{J_b})\}].$$

Each family is independent across strata. Since $\phi \in [0,1]$, $|X_b|, |Y_b| \le p_b \le B := \rho^{-1} + 2$. The variance calculation gives

$$\sum_b \mathrm{Var}(X_b),\ \sum_b \mathrm{Var}(Y_b) \le V := B^2(d+3)L_\phi^2 m^{1-2/d}.$$

Applying the two-sided Bernstein inequality (Vershynin, 2018) to each sum at threshold $nt/2$ and using the union bound gives

$$\begin{aligned}\mathbb{P}\{|\widehat{\phi}_T - \widehat{\phi}_R| \ge t\} &\le \mathbb{P}\{|\widehat{\phi}_T - \bar{\phi}_D| \ge t/2\} + \mathbb{P}\{|\widehat{\phi}_R - \bar{\phi}_D| \ge t/2\} \\ &\le 4\exp\left\{-\frac{n^2t^2/8}{V + ntB/6}\right\},\end{aligned}$$

which is (6). □

Proof of corollary 2.4. Under the Hölder condition, the oscillation within C_b is at most

$$L_\phi\{2\sqrt{d+3}\}^\alpha \Delta_b^{\alpha/d}.$$

Repeating the preceding proof replaces $2/d$ by $2\alpha/d$ in the powers of m. The same displayed constant remains valid because

$$\frac{\{2\sqrt{d+3}\}^{2\alpha}}{4} = 2^{2\alpha-2}(d+3)^\alpha \le d+3 \qquad (d \ge 2,\ 0 < \alpha \le 1).$$

This proves corollary 2.4. □

A.4 Ordering-rate optimality

Proposition A.1 (Ordering-rate optimality). *Let $n = 2m$, $d \ge 2$, and let Φ_d be the set of one-Lipschitz maps $\phi : [0,1]^d \to [0,1]$. Form consecutive pairs under an ordering and select one member of every pair independently and uniformly. Under Hilbert ordering, as $n \to \infty$ through even integers,*

$$\limsup_{n\to\infty} n^{1+2/d} \sup_{\substack{D\subset[0,1]^d \\ |D|=n}} \sup_{\phi\in\Phi_d} \mathrm{Var}(\widehat{\phi}_T - \widehat{\phi}_R) \le (d+3)2^{3+2/d}. \tag{21}$$

For any one-to-one map $o : [0,1]^d \to [0,1]$, with the same independent uniform selection from consecutive pairs under the order induced by o,

$$\limsup_{n\to\infty} n^{1+2/d} \sup_{\substack{D\subset[0,1]^d \\ |D|=n}} \sup_{\phi\in\Phi_d} \mathrm{Var}(\widehat{\phi}_T - \widehat{\phi}_R) \ge \frac{2}{d}. \tag{22}$$

Hence no scalar ordering improves the exponent $1 + 2/d$.

When $n = 2m$, each stratum has size $p_b = 2$ and $\widehat{\phi}_T + \widehat{\phi}_R = 2\bar{\phi}_D$. Therefore,

$$\mathrm{Var}(\widehat{\phi}_T - \widehat{\phi}_R) = 4\,\mathrm{Var}(\widehat{\phi}_T).$$

Using $p_b = 2$ in the variance calculation above gives

$$\begin{aligned}\mathrm{Var}(\widehat{\phi}_T - \widehat{\phi}_R) &= \frac{16}{n^2}\sum_{b=1}^{m}\mathrm{Var}\{\phi(D_{J_b})\} \\ &\le \frac{16}{n^2}(d+3)m^{1-2/d} \\ &= (d+3)2^{3+2/d}n^{-1-2/d},\end{aligned}$$

which proves (21).

For the lower bound, take $n = 2^{cd}$ and the grid

$$D_c = \{0, 2^{-c}, \ldots, (2^c - 1)2^{-c}\}^d.$$

Any injective scalar order pairs distinct grid points. If $\Delta_b \in \mathbb{R}^d$ is the coordinate difference in pair b, then $\|\Delta_b\|_2^2 \ge 2^{-2c} = n^{-2/d}$. For $\phi_\ell(x) = x_\ell$, $\ell = 1, \ldots, d$,

$$\begin{aligned}\sum_{\ell=1}^{d}\mathrm{Var}(\widehat{\phi}_{\ell,T} - \widehat{\phi}_{\ell,R}) &= \frac{1}{m^2}\sum_{b=1}^{m}\|\Delta_b\|_2^2 \\ &\ge \frac{1}{m}n^{-2/d} = 2n^{-1-2/d}.\end{aligned}$$

At least one coordinate projection therefore has variance at least $(2/d)n^{-1-2/d}$, proving (22) along this sequence and hence the stated limsup bound. □

A.5 Wasserstein control

Theorem A.2 (Wasserstein control)**.** *Let $1 \le m \le \lfloor n/2 \rfloor$, and define $\mu_T = m^{-1}\sum_{i\in T}\delta_{D_i}$ and $\mu_R = (n-m)^{-1}\sum_{i\in R}\delta_{D_i}$. For every $p \ge 1$, with Euclidean ground distance, the randomized Hilbert design satisfies*

$$W_p(\mu_T, \mu_R) \le 4\sqrt{d+3}\,m^{-1/\max\{p,d\}} \qquad \textit{almost surely}. \tag{23}$$

Proof. Let $\tau = \frac{1}{n}\sum_{t=1}^{n}\delta_{D_{(t)}}$ be the empirical measure of the full dataset. $D_{(1)}, \ldots, D_{(n)}$ are the Hilbert-sorted observations in section 2, and T_i is the point selected from stratum C_i. We define a coupling between $X_T \sim \mu_T$, $X_D \sim \tau$ by letting $(X_T, X_D) = (T_I, D_{(J)})$ where, for $1 \le i \le m$ and $1 \le j \le n$,

$$P(I = i, J = j) = \begin{cases} \frac{1}{n}(\lceil \frac{i-1}{m}n \rceil - \frac{i-1}{m}n) & if \quad j = \lceil \frac{i-1}{m}n \rceil \\ \frac{1}{n} & if \quad \lceil \frac{i-1}{m}n \rceil < j \le \lfloor \frac{i}{m}n \rfloor \\ \frac{1}{n}(\frac{i}{m}n - \lfloor \frac{i}{m}n \rfloor) & if \quad j = \lfloor \frac{i}{m}n \rfloor + 1 \end{cases}$$

and all unlisted probabilities are zero.

Let

$$\alpha_i = \begin{cases} h_d\big(D_{(\lceil (i-1)\frac{n}{m} \rceil)}\big), & m \nmid (i-1)n, \\ h_d\big(D_{((i-1)\frac{n}{m}+1)}\big), & m \mid (i-1)n \end{cases},$$

$$\beta_i = \begin{cases} h_d\big(D_{(\lfloor (i)\frac{n}{m} \rfloor + 1)}\big), & m \nmid in, \\ h_d\big(D_{(i\frac{n}{m})}\big), & m \mid in \end{cases}.$$

Then $0 \le \alpha_1 \le \beta_1 \le \alpha_2 \le \cdots \le \alpha_m \le \beta_m \le 1$. Under the coupling above,

$$\begin{aligned}\mathbb{E}\big[\|X_T - X_D\|^p\big] &\le \frac{1}{m}\sum_{i=1}^{m} \max_{z,z' \in [\alpha_i,\beta_i]} \|H_d(z) - H_d(z')\|^p \\ &\le \frac{1}{m}\sum_{i=1}^{m} \big(2\sqrt{d+3}\,(\beta_i - \alpha_i)^{1/d}\big)^p \\ &\le \begin{cases} 2^p (d+3)^{p/2}\, m^{-p/d}, & p \le d, \text{ (by Hölder inequality)} \\ \frac{2^p (d+3)^{p/2}}{m}, & p > d, \end{cases}\end{aligned}$$

In the same way, let $X_R \sim \mu_R$, by coupling with Hilbert index,

$$\mathbb{E}\big[\|X_R - X_D\|^p\big] \le \begin{cases} 2^p (d+3)^{p/2}\, m^{-p/d}, & p \le d, \\ \frac{2^p (d+3)^{p/2}}{m}, & p > d, \end{cases}$$

Thus, by triangle inequality,

$$W_p(\mu_T, \mu_R) \le W_p(\mu_T, \tau) + W_p(\tau, \mu_R) \le 4\sqrt{d+3}\, m^{-1/\max\{p,d\}}, \quad a.s.$$

□

B Implementation and experimental protocols

B.1 Neighborhoods and screening

The local-linear residual-risk construction developed for PRIME (Anonymous Authors, 2026a) uses the following dimension-dependent neighborhood rule shared by screening and local geometry:

$$K_n(d) = \left\lceil \max\{n^{4/(d+4)}, 3d\} \right\rceil, \qquad d+1 < K_n(d) \le \lfloor n/2 \rfloor. \tag{24}$$

Here d is a predictor-subset dimension. For a nested comparison $S \subset T$, both fits use $K_n(|T|)$. Neighbors are found separately in the standardized coordinates of S and T. The stratum geometry uses $K_n(a)$, where $a = |\widehat{A}|$. The joint Hilbert dimension is $a+1$, because the response is included in the ordering but not in neighborhood search.

All development observations serve as screening queries. Predictor coordinates are centered and scaled by development-sample means and sample standard deviations. Write the standardized predictors as $\widetilde{X}$. A query's own row is excluded by its row identifier. Other rows with the same coordinates remain eligible. The same query-exclusion rule applies at stratum anchors.

Let $\mathcal{N}_{S,K}(x)$ contain the K nearest eligible rows in standardized coordinates S. At query x, the shared local fit is

$$(\widehat{\alpha}_{S,K}(x), \widehat{g}_{S,K}(x)) = \arg\min_{\alpha, g} \sum_{j \in \mathcal{N}_{S,K}(x)} \{Y_j - \alpha - g^\top (\widetilde{X}_{j,S} - x)\}^2 + 10^{-6}\, \|g\|_2^2 . \tag{25}$$

The intercept is unpenalized, and the residual sum of squares excludes the slope penalty. For nonempty S, the local residual-risk contribution is $g_{i,K}(S) = \mathrm{SSE}_{i,S,K} / (K - |S| - 1)$. The empty-subset contribution is the sample variance of Y. Define

$$H_i(S,T) = g_{i,K}(S) - g_{i,K}(T), \qquad \widehat{\Gamma}_d(S,T) = \left[\overline{H} - \sqrt{2\log\{\max(2,d)\}}\, \frac{\mathrm{sd}(H_i)}{\sqrt{n}} \right]_+ . \tag{26}$$

Here $d = |T|$ for each nested comparison. Both sides use the same query rows. Their neighborhoods may overlap.

Squared standardized ridge coefficients with penalty 10^{-6} rank the initial coordinates. BIC selects an OLS prefix, including the empty prefix, among at most $\lceil 1.5 \log\{\max(2,p)\} \rceil$ coordinates subject to local-fit feasibility. Forward search ranks additions by the raw mean gain $\overline{H}$, accepting the leader if its calibrated gain is positive and stopping otherwise. A subsequent backward phase removes coordinates with zero calibrated deletion loss one at a time, recomputing comparisons after each removal. An empty retained set is recorded as an allocation failure.

B.2 Ordering, local fits, and sequential distances

Selected raw predictors and Y are min–max scaled separately. The general scaling routine maps constants to $1/2$. The screening interface requires nonconstant inputs. A 16-bit Hilbert traversal sorts the quantized coordinates $\lfloor (2^{16}-1)z \rfloor$, with response weight one. The continuous ordering in section 2 is the theoretical construction. The finite traversal is the implemented approximation.

At each anchor in (10), coefficients are evaluated afresh using (25) and $K_n(a)$. The solver uses the symmetric positive-definite system when available and a pseudoinverse fallback. An exactly zero slope uses $D_b = I_a$. For $a = 1$, $D_b = [1]$ and coefficient evaluation is unnecessary.

Neighborhood search uses exact projected nearest neighbors through nanoflann KD trees, with leaf size 16 and approximation tolerance zero. Equal distances are ordered by original row identifier. This search is distinct from sequential allocation: the three distances in (12)–(14) use direct weighted coordinate differences, without retaining a full $n \times n$ distance matrix. Score ties use tolerance $10^{-14}+10^{-12}\max(1, |S_{\max}|)$, then proximity to the stratum's central rank and original row identifier.

Comparison settings. SPlit uses 500 iterations, tolerance 10^{-10}, and one worker. Twinning uses leaf size 8. In the Borehole and real-data studies and the paired illustration, it starts farthest from the standardized joint centroid, and MaxTwin uses its default centroid initialization. In the final selection study, Twinning uses its random default start and MaxTwin uses `initialization=FALSE`, following the original MaxTwin repeated-split experiment. MaxTwin standardizes the joint data in both cases. Random samples test rows without replacement. Package versions and software dependencies accompany the reproduction code.

B.3 Borehole design and representation metric

For two sets of transformed observations U and V, empirical Energy is

$$\mathcal{E}(U,V) = \frac{2}{|U||V|}\sum_{u\in U, v\in V}\|u-v\|_2 - \frac{1}{|U|^2}\sum_{u,u'\in U}\|u-u'\|_2 - \frac{1}{|V|^2}\sum_{v,v'\in V}\|v-v'\|_2. \quad (27)$$

In the irrelevant-predictor study its evaluation coordinates are the true active predictors and response, standardized over the pooled sample. Those coordinates are used to assess the split, not supplied to the splitters.

The Borehole function is

$$f = \frac{2\pi T_u(H_u - H_l)}{\log(r/r_w)\{1 + 2LT_u/[\log(r/r_w)r_w^2 K_w] + T_u/T_l\}}.$$

For inputs $(r_w, r, T_u, H_u, T_l, H_l, L, K_w)$, the lower and upper endpoints of the independent uniform ranges are

$$\ell = (0.05, 100, 63070, 990, 63.1, 700, 1120, 9855),$$
$$u = (0.15, 50000, 115600, 1110, 116, 820, 1680, 12045),$$

respectively. Inputs are recorded on $[0,1]^8$ before mapping to these ranges to generate the response. Independent numerical calibration sets the signal-to-noise variance ratio to 23.8265. Irrelevant coordinates are independent uniforms on $[0,1]$. Their prefixes are shared across q within each replication. The six q settings in figure 2 are equally spaced for display.

B.4 Real-data regression models

The real-data study uses Gaussian KRR with kernel $\exp\{-\|z-z'\|_2^2/(2h^2)\}$, bandwidth h equal to half the median positive training distance, and mean-loss penalty 10^{-8}. Linear ridge uses mean-loss penalty 0.01. Both learners have an unpenalized intercept, and their system matrices use the shift $n_{\text{train}}\lambda$. Predictors are standardized using the training set only. Both learners use all 100 Tecator measurements or all five Airfoil predictors, with identical fitting rules across the datasets. The KRR fits have small, nonzero training residuals and are described as quasi-interpolating.

Airfoil is split at the record level, so aerodynamic settings may be shared between training and test sets. For figure 3, quantiles are computed in the original units and displayed on $\operatorname{asinh}(x/10)$ axes

with original-unit ticks. The two learners have matched axes within each dataset. In figures 3 and 5, boxes show medians and interquartile ranges. Whiskers extend to the most extreme values within 1.5 interquartile ranges of the hinges. All observations enter these summaries and the axis limits, including those beyond the whiskers whose individual markers are omitted.

B.5 Paired Friedman experiments

Paired illustration. The illustration in figure 4 uses fifteen centers drawn uniformly on $[0.005, 0.995]^5$. Two independent $\mathrm{Unif}[-0.04, 0.04]^5$ jitters are added to each center. Coordinates outside $[0, 1]$ are reflected at the boundary. Each row receives five independent uniform irrelevant coordinates and the noiseless response (16). All methods use the same 30 rows and a 20% test fraction. Hilbert Split retains X_1, X_2, X_4, X_5 in this small sample. Its six test observations form three complete pairs. The other four methods each place six test observations opposite their training mate.

The horizontal coordinate is PC1 of the pooled sample-standardized $X_{1:5}$, explaining 33.15% of their variation. The vertical coordinate is Y in its original units. The PCA uses only the active predictors. All methods share the same coordinates and limits, without point displacement. Pair identities and the known active set are used only for display and diagnostics.

Observed designs and target population. The reported design follows (17). The $n/2$ centers are constructed in $[0.005, 0.995]^5$ by `support::sp`, using 8,192 fixed Owen-scrambled Sobol integration candidates, no resampling during optimization, at most 500 iterations, at least 50 iterations, and tolerance 10^{-5}. The noise scale uses $s_f = 4.8750735541$, the sample standard deviation of f on the common independent reference bank. Row order is randomized once per observed design. These are designed samples, not iid draws from the uniform target.

Candidate predictor and test objective. We use the anisotropic Gaussian kernel

$$K_\ell(x, z) = \exp\left\{ -\frac{1}{2} \sum_{j=1}^{10} \frac{(x_j - z_j)^2}{\ell_j^2} \right\}. \tag{28}$$

At a candidate (ℓ, λ), set $A_R = K_\ell(X_R, X_R) + \lambda I$ and compute

$$\widehat{\mu}_R = \frac{\mathbf{1}^\top A_R^{-1} Y_R}{\mathbf{1}^\top A_R^{-1} \mathbf{1}}, \quad \alpha_R = A_R^{-1}(Y_R - \widehat{\mu}_R \mathbf{1}), \quad \widehat{f}_R(x) = \widehat{\mu}_R + K_\ell(x, X_R)\alpha_R. \tag{29}$$

The candidate loss is the mean of $(Y_i - \widehat{f}_R(X_i))^2$ over $i \in T$. The penalty is the matrix shift λ, not a mean-loss penalty multiplied by training size. Cholesky solves form A_R directly. The full observed ten-coordinate vector enters every KRR fit in its original $[0, 1]$ units, with the unstandardized response and a fitted GLS intercept.

Common search and repetitions. Each repetition uses 110 maximin Latin-hypercube candidates, stratified in $\log_{10} \ell_j \in [-2, 2]$ and $\log_{10} \lambda \in [-4, 0]$. The candidate bank is shared across all five methods at each repetition. From the candidate with the smallest test MSE, one L-BFGS-B run optimizes in the original parameter coordinates, with bounds $\ell_j \in [\ell_{0j}/2, 2\ell_{0j}]$ and $\lambda \in [10^{-6}, 1]$. All methods use an analytic gradient, including the derivative of the GLS intercept, and a common 500-iteration limit. The returned optimizer parameters define the fitted model. Termination flags are retained. Search never uses the independent risk.

The observed design is held fixed for 30 split/search repetitions, with 160 training and 40 test observations. Randomized splitters and search initializations vary by repetition. Hilbert Split's

allocation is deterministic on each design, so its variation reflects the randomized parameter search. Pair identities and the true active set are available only to the generator and evaluation diagnostics.

Full-data refitting and independent generalization error. After test-based selection, $(\widehat{\ell}, \widehat{\lambda})$ is fixed and (29) is recomputed on all n observations. For the resulting $\widehat{f}_{\text{full}}$, the target and its plotted estimate are

$$\begin{aligned} &\mathcal{R}_f(\widehat{f}_{\text{full}}) = \mathbb{E}_{X^*}\{\widehat{f}_{\text{full}}(X^*) - f(X^*)\}^2, \\ &\widehat{\text{RMSE}}_f = \left[\frac{1}{M}\sum_{i=1}^{M}\{\widehat{f}_{\text{full}}(X_i^*) - f(X_i^*)\}^2\right]^{1/2}, \quad M = 160000. \end{aligned} \tag{30}$$

All 150 reported fitted models use the same independent uniform reference bank, which is held outside the selection procedure. Risk is relative to the true regression function. Squared risk for a future noisy response adds the common noise variance $0.01 s_f^2$. All searches converge except Random repetition 18, which reaches the common 500-iteration limit. Its returned model is included in every summary. The saved models, search histories, and termination codes accompany the reproduction code.