A PAIRED CASE STUDY OF CONDITIONAL AND MARGINAL FEATURE-REPLACEMENT SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature-removal explanations measure how a predictor responds when an input feature is replaced. Because the predictor still requires a complete input, every such explanation depends on a replacement distribution. Conditional replacement is commonly motivated by its ability to preserve dependencies among observed features, but observational fidelity does not by itself determine how the predictor will respond to the resulting inputs.

We study this gap with a paired evaluation that holds the predictor, evaluation examples, labels, and scoring rule fixed while changing the replacement distribution. The primary study compares conditional and marginal replacement on Dry Bean and evaluates coordinate prediction, covariance preservation, score magnitude, and feature ordering. Complementary studies examine stochastic draws, deterministic means, and label-informed replacements across several tabular domains.

Conditional replacement more closely reproduces observed coordinate and covariance structure, yet it produces substantially different feature-removal scores and rankings from marginal replacement through the same predictors. The secondary comparisons further show that sensitivity to the replacement procedure depends on the dataset and predictor family. These findings establish the replacement distribution as part of the feature-removal estimand rather than a secondary implementation detail. A reproducible explanation should therefore report the replacement rule together with the predictor and scoring procedure.

1 INTRODUCTION

Feature-removal explanations ask how a predictor changes when part of an input is removed. This idea underlies perturbation explanations, additive feature attribution, and missing-feature value functions (Ribeiro et al., 2016; Fong & Vedaldi, 2017; Covert et al., 2021; Lundberg & Lee, 2017; Covert et al., 2020). In practice, however, a feature cannot simply be deleted: the predictor still requires a complete input. An explanation must therefore specify how the removed feature is replaced. A constant, an unconditional draw, and a conditional draw create different modified inputs and define different prediction-under-replacement quantities (Janzing et al., 2020; Frye et al., 2021; Sundararajan & Najmi, 2020).

Conditional replacement is often preferred when features are dependent because it uses the retained coordinates to generate more plausible replacements (Aas et al., 2021; Frye et al., 2021). This motivation concerns how well the replacement model reproduces the observed data distribution. An explanation score answers a different question: it passes the generated input through a fixed predictor and measures the resulting change in loss, accuracy, or another model response. A replacement model can therefore reproduce observed coordinates closely without producing scores or feature rankings that agree with those obtained under another replacement rule. Existing work establishes that feature-removal quantities depend on their treatment of missing features (Covert et al., 2021; Janzing et al., 2020; Yeh et al., 2022), but the relationship between observational diagnostics and predictor-mediated agreement remains insufficiently measured in a controlled paired setting.

We address this gap by separating three evaluation targets: how a replacement model predicts observed coordinates, how much its samples change the predictor's score, and whether it preserves the resulting feature ordering. Our primary study compares conditional and marginal Gaussian replacement on Dry Bean. Within each comparison, the predictor, evaluation examples, labels, loss, and random

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.

variates are shared, so the replacement rule is the component that changes. The conditional procedure yields lower coordinate-prediction and covariance discrepancies, but the two procedures still produce different score magnitudes and feature rankings. This result shows that observational diagnostics do not determine agreement between feature-removal explanations.

Secondary studies on CelebA, Adult, Spambase, and Bank Marketing examine whether accuracy-based scores also change across stochastic, deterministic, and label-informed conditional replacements.

This paper makes three contributions:

- We introduce a paired protocol separating coordinate diagnostics, score separation, and ranking agreement.
- We show that stronger observational fidelity can coexist with different feature-removal scores and rankings through the same predictor.
- We characterize sensitivity to stochastic, deterministic, and label-informed replacements across four secondary domains.

2 Related Work

Replacement distributions in feature removal. Feature removal requires a rule for completing an input (Ribeiro et al., 2016; Fong & Vedaldi, 2017; Covert et al., 2021); the same choice enters additive and Shapley-style attribution (Lundberg & Lee, 2017; Covert et al., 2020; Sundararajan & Najmi, 2020). Marginal replacement ignores the retained input, whereas conditional replacement uses it to preserve observed dependence (Janzing et al., 2020; Frye et al., 2021). Dependent-feature Shapley approximations explicitly model this dependence (Aas et al., 2021), and conditional permutation importance adjusts the replacement scheme for correlated predictors (Strobl et al., 2008). On- and off-manifold value functions also differ in their treatment of the model and data distribution (Yeh et al., 2022). Causal contribution methods introduce intervention assumptions (Janzing et al., 2024), while Shapley-based feature importance has known interpretive limitations (Kumar et al., 2020). We measure conditional–marginal score and ranking agreement on the same examples and predictors.

Conditional sampling and attribution estimation. Approximate conditional models support feature removal when exact sampling is unavailable (Frye et al., 2021; Covert et al., 2020). A separate line of work reduces attribution cost through learned estimators and additive models (Jethani et al., 2022; Enouen & Liu, 2025). These approaches estimate a chosen removal quantity; our comparison examines how that quantity changes with the replacement distribution. Coordinatewise ridge-Gaussian models make the replacement locations and scales explicit.

Evaluation of perturbation explanations. Perturbation evaluations measure sensitivity to model parameters, agreement with model behavior, and the effects of input changes (Adebayo et al., 2018; Yeh et al., 2019; Madsen et al., 2024; Yin et al., 2022). Our paired protocol complements these tests by holding the predictor, examples, labels, and Monte Carlo variates fixed while changing the replacement distribution.

3 Paired Feature Removal

We evaluate coordinate-prediction diagnostics, score separation, and feature-order agreement. For each held-out example i, feature j, and draw r, the conditional and marginal procedures share $\epsilon_{ijr} \sim \mathcal{N}(0,1)$. Replacement models use training data only; the primary estimator uses $R = 32$ draws per example–feature pair. Figure 1 summarizes the design.

Feature-removal response. For the primary study, X denotes standardized, clipped predictor inputs; P is their joint evaluation distribution with the original labels. Let $f : \mathcal{X} \to \Delta^{C-1}$ be a fixed probabilistic classifier with C classes, and let ℓ be a specified loss. Because the loss is fixed throughout each comparison, we suppress it in the notation. We write $X^{(j \leftarrow \tilde{X}_j)}$ for the input obtained by replacing coordinate j of X with $\tilde{X}_j$. For a replacement kernel $q_j(\cdot \mid X_{-j})$, the population

Held-out $(\mathbf{x}_i, y_i)$; replace j; shared $\epsilon_{ijr} \sim \mathcal{N}(0,1)$	
Conditional Gaussian	**Marginal Gaussian**
$z_{ijr}^{(\mathrm{c})} = \hat{\mu}_j(\mathbf{x}_{i,-j}) + \hat{\sigma}_j \epsilon_{ijr}$	$z_{ijr}^{(\mathrm{m})} = \hat{\mu}_j^{\mathrm{train}} + \hat{\sigma}_j^{\mathrm{train}} \epsilon_{ijr}$
Uses retained coordinates	Uses training moments
$\downarrow \quad \tilde{x}_{ijr}^{(q)} = \mathrm{clip}(z_{ijr}^{(q)}, -5, 5), \quad \tilde{\mathbf{x}}_{ijr}^{(q)} = \mathbf{x}_i^{(j \leftarrow \tilde{x}_{ijr}^{(q)})} \quad \downarrow$	
Coordinate diagnostics	**Predictor response**
Gaussian NLL: $g_j^{(q)}$ evaluated at x_{ij}	Fixed f, original label y_i
Covariance: single-column replacements	$\ell(f(\tilde{\mathbf{x}}_{ijr}^{(q)}), y_i) - \ell(f(\mathbf{x}_i), y_i)$
Average over examples / coordinates / draws as defined	Average over draws, then examples; rank feature scores

Figure 1: Paired protocol. Both kernels retain $\mathbf{x}_{i,-j}$ and share ϵ_{ijr}. The coordinatewise Gaussian NLL evaluates the pre-clipping Gaussian density; covariance and predictor responses use the clipped replacement. The conditional location depends on $\mathbf{x}_{i,-j}$, whereas the marginal location and scale use training-set moments only.

feature-removal response is

$$S_j(q; f, P) = \mathbb{E}_{(X,Y)\sim P} \mathbb{E}_{\tilde{X}_j \sim q_j(\cdot|X_{-j})} \left[\ell\Big(f\Big(X^{(j \leftarrow \tilde{X}_j)} \Big), Y \Big) - \ell(f(X), Y) \right]. \tag{1}$$

Thus, the response is indexed by q, f, ℓ, and P. The kernel q_j denotes the replacement law supplied to the predictor, including clipping. We write g_j for the Gaussian density before clipping, which is used in the Gaussian NLL diagnostic.

For held-out example $(\mathbf{x}_i, y_i)$, let $\tilde{x}_{ijr}^{(q)}$ be the r-th replacement generated by kernel q_j, and let $\tilde{\mathbf{x}}_{ijr}^{(q)} = \mathbf{x}_i^{(j \leftarrow \tilde{x}_{ijr}^{(q)})}$. The finite-draw loss difference, per-example score, and dataset-level score are

$$s_{ijr}(q) = \ell\Big(f\Big(\tilde{\mathbf{x}}_{ijr}^{(q)} \Big), y_i \Big) - \ell(f(\mathbf{x}_i), y_i), \tag{2}$$

$$\hat{s}_{ij}(q) = \frac{1}{R} \sum_{r=1}^{R} s_{ijr}(q), \tag{3}$$

$$\hat{S}_j(q) = \frac{1}{n} \sum_{i=1}^{n} \hat{s}_{ij}(q). \tag{4}$$

The primary Dry Bean analysis uses original-label cross-entropy measured in nats. Features are ranked in descending order of $\hat{S}_j(q)$, so a larger value indicates a larger increase in loss after replacement. No exact score ties occur in the reported rankings. Agreement between the resulting feature orderings is summarized using Kendall's τ_b.

Replacement kernels. For the conditional arm, feature j is predicted from the retained coordinates using ridge regression fitted on the training split. The ridge penalty is fixed at $\lambda = 1$, with an unpenalized intercept. Let $\hat{\mu}_j(\mathbf{x}_{i,-j})$ denote the fitted conditional mean and let $\hat{v}_j$ denote the corresponding training residual variance. We use

$$\hat{\sigma}_j = \sqrt{\max\{\hat{v}_j, v_{\min}\}}, \tag{5}$$

where $v_{\min} = 10^{-6}$ in standardized units and $\hat{v}_j$ is the mean squared training residual. The conditional draw, before the final clipping step, is

$$z_{ijr}^{(\mathrm{c})} = \hat{\mu}_j(\mathbf{x}_{i,-j}) + \hat{\sigma}_j \epsilon_{ijr}, \qquad \epsilon_{ijr} \sim \mathcal{N}(0,1). \tag{6}$$

All coordinates are standardized using training means and population standard deviations, then clipped to $[-5, 5]$ before fitting or evaluation. Conditional locations, residual variances, and generated values remain on this scale.

The marginal arm uses the empirical training-set mean $\hat{\mu}_j^{\text{train}}$ and variance-floored standard deviation $\hat{\sigma}_j^{\text{train}}$:

$$z_{ijr}^{(\text{m})} = \hat{\mu}_j^{\text{train}} + \hat{\sigma}_j^{\text{train}} \epsilon_{ijr}. \tag{7}$$

Before final clipping, the marginal Gaussian uses the first two moments of the clipped standardized training coordinates, subject to the variance floor. Shared variates pair draws at the same Gaussian quantile in the two models.

Both Gaussian models are fitted once on the clipped standardized training data. The predictor receives $\tilde{x}_{ijr}^{(q)} = \text{clip}(z_{ijr}^{(q)}, -5, 5)$, and covariance evaluation uses the same clipped values.

Coordinate-level observational diagnostics. We first evaluate how each fitted kernel predicts the observed held-out coordinate. The conditional and marginal kernels use different predictive information, so their likelihood values are coordinate-prediction diagnostics rather than a model-selection comparison between equivalent estimators. The average coordinatewise pre-clipping Gaussian negative log-likelihood is

$$\widehat{\text{NLL}_{\text{Gauss}}}(q) = -\frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} \log g_j^{(q)}(x_{ij} \mid \mathbf{x}_{i,-j}), \tag{8}$$

where the conditioning argument is omitted for the marginal kernel. This quantity uses the untruncated Gaussian density evaluated at standardized, clipped held-out coordinates.

For each coordinate j and draw r, let $X_{\text{held}}^{(j,q,r)}$ be the held-out matrix with only column j replaced by clipped draws from q_j. We average the covariance discrepancy of these singly replaced matrices:

$$D_{\text{cov}}(q) = \frac{1}{dR} \sum_{j=1}^{d} \sum_{r=1}^{R} \left\| \text{Cov}_0\left(X_{\text{held}}^{(j,q,r)}\right) - \text{Cov}_0(X_{\text{held}}) \right\|_F. \tag{9}$$

Here Cov_0 uses divisor n. We use this discrepancy as a descriptive diagnostic of how closely single-coordinate replacements preserve the held-out covariance structure; it is not a general goodness-of-fit criterion.

Score separation and ordering agreement. For the primary conditional–marginal comparison, the finite-sample per-example, per-feature absolute separation is

$$\hat{A} = \frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} \left| \hat{s}_{ij}(\text{c}) - \hat{s}_{ij}(\text{m}) \right|. \tag{10}$$

We also report the identities of the highest-scoring features and Kendall's τ_b between the complete feature orderings.

For $\hat{A}$, we resample held-out examples jointly across both arms and all features $5{,}000$ times. We report the fifth bootstrap percentile, the lower endpoint of a central 90% percentile interval. These resamples hold fitted models and realized draws fixed. The finite-R absolute contrast depends on the specified coupling.

Interpreting absolute separation. The order of averaging matters. Writing $\hat{\delta}_{ij} = \hat{s}_{ij}(\text{c}) - \hat{s}_{ij}(\text{m})$, the triangle inequality gives

$$\frac{1}{d} \sum_{j=1}^{d} \left| \hat{S}_j(\text{c}) - \hat{S}_j(\text{m}) \right| \le \frac{1}{nd} \sum_{i,j} |\hat{\delta}_{ij}| = \hat{A}. \tag{11}$$

Thus, $\hat{A}$ measures example-level separation before positive and negative differences can cancel across examples. It is not the difference between the two overall mean scores. Both arms subtract the same original-input loss, so that baseline cancels in $\hat{\delta}_{ij}$.

The coupling specifies how the finite-draw estimates are paired. Each arm still samples from its own replacement law; shared Gaussian variates align the quantiles of the two pre-clipping models. Population mean scores depend on the replacement laws, whereas finite-draw absolute separation also depends on the realized coupling. We therefore report $\hat{A}$ together with the draw count and the example-resampling procedure.

Table 1: Datasets, predictors, replacement contrasts, and evaluation endpoints. Dry Bean is the primary conditional–marginal study and uses $R = 32$ paired draws. The remaining datasets report signed accuracy contrasts; a negative contrast indicates lower accuracy under the first-listed procedure.

Dataset	Predictor	Replacement contrast	Endpoint
Dry Bean	16–32–ReLU–7 MLP	Ridge-Gaussian conditional draw versus training-moment marginal Gaussian draw	Cross-entropy response and feature ranking
CelebA	Attribute-wise logistic regression	Conditional draw versus training mean, conditional mean, and label-informed diagnostic	Accuracy contrast
Adult	Logistic regression and MLP	Conditional draw versus conditional mean and label-informed diagnostic	Accuracy contrast
Spambase	MLP and Bernoulli Naive Bayes	Conditional draw versus conditional mean and label-informed diagnostic	Accuracy contrast
Spambase (joint)	MLP	Joint conditional draw versus joint conditional mean and label-informed diagnostic for $K \in \{1, 3, 6\}$	Accuracy contrast
Bank Marketing	MLP	Conditional draw versus conditional mean and label-informed diagnostic	Accuracy contrast

Secondary accuracy contrasts. The secondary studies use binary feature representations and logistic conditional Bernoulli models. Each stochastic replacement uses one seeded draw ($R = 1$). Writing $\hat{p}_j(\mathbf{x}_{i,-j})$ for the fitted Bernoulli probability, deterministic conditional-mean replacement uses

$$\tilde{x}_{ij}^{(\mathrm{mean})} = \hat{p}_j(\mathbf{x}_{i,-j}). \tag{12}$$

For a feature or feature set J, let $\tilde{\mathbf{x}}_{iJ}^{(q)}$ denote the complete input produced by the stated single- or multi-coordinate replacement procedure. Its replacement accuracy is

$$\hat{a}(q; J) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\Big[y_i = \arg\max_c f_c\Big(\tilde{\mathbf{x}}_{iJ}^{(q)}\Big) \Big]. \tag{13}$$

For a collection $\mathcal{J}$ of evaluated units, we report

$$\hat{a}(q; \mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{J \in \mathcal{J}} \hat{a}(q; J), \tag{14}$$

and compare procedures using

$$\Delta_a(q_1, q_2; J) = \hat{a}(q_1; J) - \hat{a}(q_2; J). \tag{15}$$

For an average over $\mathcal{J}$, the reported contrast is $\Delta_a(q_1, q_2; \mathcal{J}) = \hat{a}(q_1; \mathcal{J}) - \hat{a}(q_2; \mathcal{J})$. Negative values indicate lower accuracy for the first procedure. Secondary 90% percentile intervals use $1{,}000$ paired-example resamples, keeping the evaluated replacement units together. Label-informed replacements use the training outcome when fitting the imputer and the original held-out outcome when generating the diagnostic replacement. They are reported only as a diagnostic contrast with additional conditioning information.

Relationship between the endpoints. Cross-entropy scores measure changes in the probability assigned to the original class; accuracy records whether the predicted class matches that label. An input replacement can alter confidence without changing the argmax, so the two endpoints need not respond at the same scale. The primary study compares conditional and marginal stochastic laws through cross-entropy. The secondary studies compare conditional draws, means, and label-informed procedures through accuracy. Their common object is predictor response to replacement, with a different contrast and scoring rule in each study.

4 Experimental Setup

Table 1 summarizes the five datasets. Dry Bean compares conditional and marginal kernels using original-label cross-entropy and full 16-feature rankings. The secondary studies compare stochas-

Table 2: Dry Bean results on the common $10{,}844/2{,}767$ train–held-out split. The upper panel reports the conditional-minus-marginal coordinatewise pre-clipping Gaussian NLL difference and the post-clipping covariance discrepancy. The lower panel reports fixed-predictor performance, mean absolute score separation $\hat{A}$, and the identity of the highest-scoring feature under each replacement kernel. Seeds identify predictor and Monte Carlo runs on the same split.

Sampler diagnostic	1207	3407	7703
$\Delta\mathrm{NLL}_{\mathrm{Gauss}}$ (conditional − marginal)	-3.560	-3.560	-3.560
$D_{\mathrm{cov}}(\mathrm{c})$	0.022	0.022	0.021
$D_{\mathrm{cov}}(\mathrm{m})$	3.156	3.158	3.156

Seed	Macro-F1	Accuracy	$\hat{A}$	Largest conditional	Largest marginal
1207	0.941	0.930	0.256	ShapeFactor4	ShapeFactor1
3407	0.942	0.930	0.248	ShapeFactor4	ShapeFactor1
7703	0.939	0.926	0.237	ShapeFactor4	roundness

tic conditional draws with deterministic replacements and label-informed diagnostics using Equations (13)–(15).

Predictor eligibility uses a fixed rule for each study. Writing b for the held-out majority rate, the accuracy thresholds are $\max(0.60, b{+}0.05)$ for CelebA and Adult, $\max(0.70, b{+}0.05)$ for Spambase, and $\max(0.60, b + 0.01)$ for Bank Marketing. Dry Bean requires macro-F1 of at least 0.85 and at least $b + 0.20$. Eligibility results are reported with the corresponding accuracy comparisons.

Primary study: Dry Bean. The UCI Dry Bean dataset contains $13{,}611$ examples, 16 numerical morphological features, and seven classes (UCI Machine Learning Repository, 2020). We use a deterministic content-hash split constructed from the input features without using the class label, yielding $10{,}844$ training examples and $2{,}767$ held-out examples. The same split is used in all three configurations.

Seeds 1207, 3407, and 7703 train one 16–32–ReLU–7 MLP each. The networks are then fixed. The configurations vary predictor initialization and Monte Carlo draws while sharing the split and replacement-model fitting protocol.

Standardization and both replacement models are fitted on the training split. The two procedures share $R = 32$ variates per example–feature pair and the fixed standardized clipping interval $[-5, 5]$.

Several inputs derive from common geometric measurements. Dry Bean therefore provides a setting with dependent features in which to compare conditional and marginal replacement.

Secondary sensitivity studies. CelebA (Liu et al., 2015) uses the released binary attribute matrix and official partition. Smiling, Wearing_Lipstick, Eyeglasses, and Heavy_Makeup are prediction targets in separate logistic models. Each target–seed configuration evaluates eight attributes, with training capped at $30{,}000$ examples. Adult (Becker & Kohavi, 1996) uses its official train–test partition and 12 engineered binary features, with logistic and 16-hidden-unit ReLU MLP predictors. Spambase (Hopkins et al., 1999) uses 12 binary word, character, and capital-run indicators with MLP and Bernoulli Naive Bayes predictors. Bank Marketing (Moro et al., 2014) uses 12 engineered binary features and an MLP.

Adult, Spambase, and Bank Marketing select six replacement features by absolute training-set correlation with the outcome. The Spambase joint study uses the top $K \in \{1, 3, 6\}$ features under the same selection criterion. Each study uses three seeds. Bank Marketing holds out every fifth row; the Spambase MLP, Naive Bayes, and joint studies hold out row-index residues 0, 1, and 2 modulo five, respectively. Appendix C.2 provides training settings and seed groups.

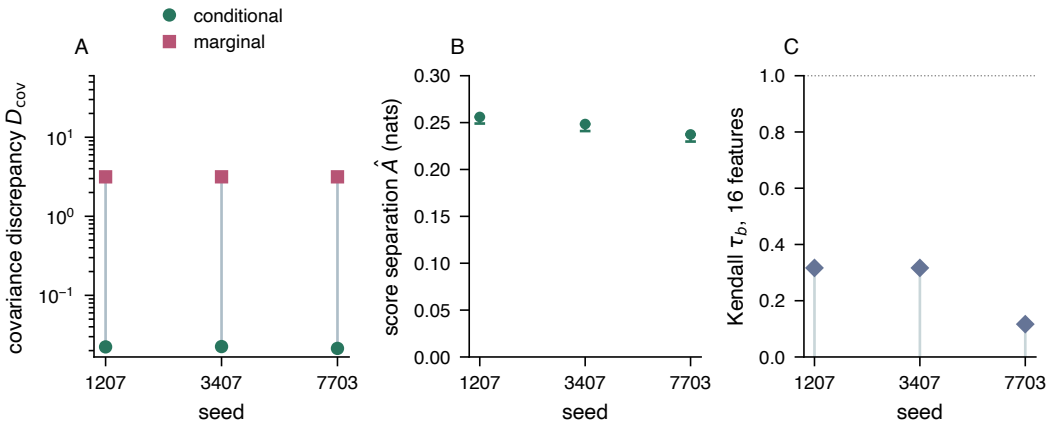


Figure 2: Dry Bean results for three fixed predictors on $n = 2{,}767$ held-out examples, $d = 16$ features, and $R = 32$ coupled draws. Panels report covariance discrepancy, mean absolute score separation $\hat{A}$, and Kendall's τ_b. Panel A uses a logarithmic scale and connects the two kernel values for each configuration; Panel B shows segments from the fifth bootstrap percentile to the point estimate; Panel C shows point-estimate rank agreement.

5 Results

5.1 Primary Dry Bean results

Coordinate diagnostics. Table 2 reports coordinate diagnostics and score results. Conditional Gaussian NLL is -2.143 nats and marginal Gaussian NLL is 1.417 nats, a difference of -3.560 nats in each configuration; the fitted replacement models and evaluation coordinates are shared across runs. The covariance discrepancy is 0.021–0.022 for the conditional kernel and 3.156–3.158 for the marginal kernel.

Score magnitude. The conditional and marginal kernels yield a mean absolute score separation $\hat{A}$ of 0.237–0.256 nats per example–feature pair. Marginal feature scores are substantially larger: their largest point estimate is approximately 0.63 nats, whereas every conditional estimate is below 0.04 nats.

Feature-order agreement. The two replacement kernels select different largest-scoring features by point estimate in all three configurations. Conditional replacement selects ShapeFactor4 in every configuration. Marginal replacement selects ShapeFactor1 for seeds 1207 and 3407 and roundness for seed 7703. Across the complete 16-feature orderings, Kendall's τ_b is 0.12–0.32.

5.2 Secondary replacement comparisons

These analyses examine additional operational choices within conditional replacement. They use different datasets, replacement contrasts, scoring endpoints, and draw counts from the primary Dry Bean study and are not direct replications of its conditional–marginal comparison. The secondary experiments compare stochastic conditional draws with deterministic conditional means and label-informed diagnostics. Appendix Table 3 reports Adult, Spambase, and Bank Marketing; Appendix Table 4 reports CelebA, and Appendix Table 5 reports the joint-removal sweep. Negative contrasts indicate lower accuracy under the conditional draw.

Adult and Spambase. For both logistic-regression and MLP predictors on Adult, conditional draws produce lower replacement accuracy than conditional means in all six configurations. The point estimates range from -0.0094 to -0.0037, and their paired-example 90% intervals exclude zero.

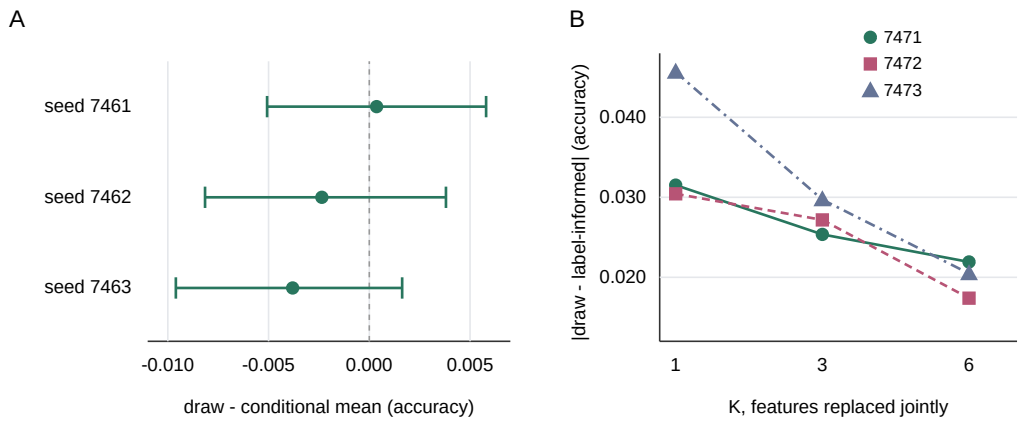


Figure 3: Exploratory Spambase replacement comparisons. Panel A reports signed draw-minus-conditional-mean accuracy contrasts for three Bernoulli Naive Bayes seed configurations; every 90% interval includes zero. Panel B reports the absolute draw-minus-label-informed accuracy contrast for three MLP seed configurations as $K \in \{1, 3, 6\}$ features are jointly replaced. Markers denote predictor configurations, and lines connect values within each seed sequence. Every observed trajectory is lower at $K = 6$ than at $K = 1$. Error bars, where shown, are paired-example 90% percentile intervals from $1{,}000$ bootstrap resamples.

The three Spambase MLP configurations show the same direction, with draw-minus-conditional-mean estimates between -0.0123 and -0.0085. All three Bernoulli Naive Bayes intervals, however, include zero.

Relative to the label-informed diagnostic, conditional draws yield lower replacement accuracy in every reported Adult and Spambase configuration. The absolute point-estimate differences range from 0.0098 to 0.0130 on Adult and from 0.0166 to 0.0253 for the Spambase MLP.

CelebA. Nine CelebA predictor configurations satisfy the fixed eligibility rule: three configurations each for Smiling, Wearing_Lipstick, and Heavy_Makeup. All three Eyeglasses configurations fall below its 0.935 majority-class accuracy and are therefore excluded from the replacement-score interpretation.

Across the nine eligible configurations, draw-minus-conditional-mean point estimates range from -0.012 to -0.002, while draw-minus-label-informed estimates range from -0.027 to -0.006. Contrasts with deterministic training-mean replacement change sign across attributes, indicating that the direction of the draw-versus-training-mean difference is not consistent across the CelebA targets. Full results and paired-example intervals are provided in Appendix Table 4.

Joint replacement and eligibility. In the Spambase joint analysis, the absolute draw-minus-label-informed contrast is lower at $K = 6$ than at $K = 1$ in all three seed sequences. Draw-minus-mean contrasts show no common monotone pattern. Bank Marketing fails the eligibility check in all three runs. Appendix B reports the eligibility details, and Appendix Table 5 reports the joint-removal estimates.

6 Discussion

The Dry Bean comparison separates two questions: how a replacement model predicts observed coordinates, and how its samples change the predictor's loss. The conditional model yields lower coordinatewise Gaussian NLL and covariance discrepancy while producing a different score profile and feature ordering from the marginal model. Replacement choice therefore changes both score magnitude and the feature highlighted by the explanation in this study.

Feature dependence offers an interpretation of the compressed conditional scores. When retained measurements predict the removed coordinate, conditional replacement can leave less variation for the predictor to respond to. Marginal replacement disrupts these relationships and produces larger loss changes in this study. This pattern is consistent with the role of redundancy in conditional feature importance (Janzing et al., 2020; Kumar et al., 2020).

The secondary results show that stochastic-versus-deterministic and label-informed contrasts need not evolve alike under joint replacement: the label-informed gap decreases with K, whereas the draw-versus-mean gap shows no common monotone pattern.

Reporting implications. A replacement score is a response to a specified input modification, not a property of the predictor alone. The operational description should identify the replaced unit, conditioning variables, fitted replacement family, input postprocessing, loss or decision rule, and aggregation across draws and examples. The paired design makes these choices explicit and separates their coordinate-level diagnostics from their effects on the predictor.

The top-ranked feature, τ_b, and $\hat{A}$ summarize different aspects of replacement sensitivity: the leading coordinate, the full ordering, and example-level score separation. Here, all three summaries show disagreement.

7 Limitations

The primary study covers one Dry Bean split, three MLP runs, and Gaussian replacement models. Bootstrap summaries condition on fitted models and realized draws; secondary intervals are unadjusted for multiple comparisons, and eligibility is assessed on held-out data. Spambase model comparisons also differ in split, so their contrasts describe configurations rather than an isolated architecture effect. Ranking results concern point estimates.

8 Conclusion

Conditional replacement yields lower coordinatewise Gaussian NLL and covariance discrepancy on Dry Bean while producing different feature-removal scores and rankings from marginal replacement through the same predictors. Reporting the replacement distribution alongside the predictor and scoring procedure makes clear which feature-removal quantity an explanation measures.

Reproducibility Statement

Sections 3 and 4 define the estimators, replacement models, and eligibility rules. Appendix C records the split construction, numerical settings, and secondary study configurations used for the reported results.

Ethics Statement

The study analyzes existing benchmark datasets and fixed predictors. CelebA uses released attribute annotations rather than images; Adult is evaluated as a benchmark prediction task. The reported analyses concern aggregate replacement sensitivity and do not evaluate deployment decisions.

AI Use Statement

Generative AI tools assisted with manuscript organization, language revision, LaTeX presentation, and figure editing. The authors verified all text, citations, numerical results, and interpretations and take responsibility for the paper.

REFERENCES

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502. URL https://doi.org/10.1016/j.artint.2021.103502.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

Barry Becker and Ronny Kohavi. Adult [dataset]. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20.

Ian Covert, Scott M. Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Ian C. Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *NeurIPS*, 2020.

James Enouen and Yan Liu. InstaSHAP: Interpretable additive models explain Shapley values instantly. In *ICLR*, 2025.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *ICLR*, 2021.

Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase [dataset]. UCI Machine Learning Repository, 1999. URL https://doi.org/10.24432/C53G6X.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *AISTATS*, 2020.

Dominik Janzing, Patrick Blöbaum, Atalanti-Anastasia Mastakouri, Philipp Michael Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions. In *AISTATS*, 2024.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time Shapley value estimation. In *ICLR*, 2022.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *ICML*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Faithfulness measurable masked language models. In *ICML*, 2024.

S. Moro, P. Rita, and P. Cortez. Bank Marketing [dataset]. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5K306.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the predictions of any classifier. In *KDD*, 2016.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008. doi: 10.1186/1471-2105-9-307. URL https://doi.org/10.1186/1471-2105-9-307.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *ICML*, 2020.

UCI Machine Learning Repository. Dry Bean [dataset]. UCI Machine Learning Repository, 2020. URL https://doi.org/10.24432/C50S4B.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, 2019.

Chih-Kuan Yeh, Kuan-Yun Lee, Frederick Liu, and Pradeep Ravikumar. Threading the needle of on and off-manifold value functions for Shapley explanations. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1485–1502, 2022. URL https://proceedings.mlr.press/v151/yeh22a.html.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the sensitivity and stability of model interpretations in NLP. In *ACL*, 2022.

A Secondary Accuracy Results

Table 3 gives the per-seed accuracy contrasts, including the Bank Marketing configurations that fail the eligibility rule.

Table 3: Secondary accuracy contrasts (one seed configuration per row). Full and Majority are held-out accuracies; † marks ineligible Bank Marketing runs, whose accuracies retain six decimals. Contrasts subtract the listed comparator from the conditional draw. Brackets give 90% paired-example percentile intervals from 1,000 resamples.

Dataset	Predictor	Seed	Full	Majority	Draw − cond. mean [90% interval]	Draw − label-informed [90% interval]
Adult	Logistic	7421	0.823	0.764	−0.0094 [−0.0106, −0.0082]	−0.0117 [−0.0130, −0.0105]
Adult	Logistic	7422	0.830	0.764	−0.0037 [−0.0046, −0.0029]	−0.0098 [−0.0107, −0.0088]
Adult	Logistic	7423	0.827	0.764	−0.0050 [−0.0060, −0.0039]	−0.0124 [−0.0135, −0.0112]
Adult	MLP	7431	0.839	0.764	−0.0072 [−0.0082, −0.0061]	−0.0130 [−0.0142, −0.0118]
Adult	MLP	7432	0.839	0.764	−0.0073 [−0.0084, −0.0063]	−0.0114 [−0.0125, −0.0103]
Adult	MLP	7433	0.838	0.764	−0.0056 [−0.0066, −0.0047]	−0.0129 [−0.0139, −0.0119]
Spambase	MLP	7451	0.880	0.606	−0.0087 [−0.0132, −0.0038]	−0.0253 [−0.0300, −0.0203]
Spambase	MLP	7452	0.878	0.606	−0.0123 [−0.0177, −0.0069]	−0.0204 [−0.0252, −0.0156]
Spambase	MLP	7453	0.877	0.606	−0.0085 [−0.0127, −0.0043]	−0.0166 [−0.0212, −0.0121]
Spambase	Bernoulli NB	7461	0.834	0.605	+0.0004 [−0.0051, +0.0058]	−0.0176 [−0.0234, −0.0127]
Spambase	Bernoulli NB	7462	0.834	0.605	−0.0024 [−0.0082, +0.0038]	−0.0121 [−0.0172, −0.0071]
Spambase	Bernoulli NB	7463	0.834	0.605	−0.0038 [−0.0096, +0.0016]	−0.0159 [−0.0210, −0.0112]
Bank Marketing†	MLP	7441	0.892569	0.883335	−0.0003 [−0.0008, +0.0002]	−0.0029 [−0.0036, −0.0022]
Bank Marketing†	MLP	7442	0.892845	0.883335	−0.0010 [−0.0015, −0.0005]	−0.0041 [−0.0047, −0.0034]
Bank Marketing†	MLP	7443	0.892661	0.883335	−0.0007 [−0.0013, −0.0001]	−0.0024 [−0.0031, −0.0018]

B Additional Accuracy Comparisons

Spambase Naive Bayes. All three Spambase Bernoulli Naive Bayes draw-minus-conditional-mean 90% intervals cross zero, while the label-informed contrasts are negative with intervals excluding zero (Table 3 and Figure 3).

Bank Marketing eligibility. The three full accuracies, 0.892569, 0.892845, and 0.892661, are below the 0.893335 threshold set by the majority rate of 0.883335. Table 3 retains these runs, marked †, as failed eligibility cases. Their draw-minus-mean contrasts range from −0.0010 to −0.0003.

Joint replacement. The absolute draw-minus-label-informed contrast decreases from $K = 1$ to $K = 6$ in all three Spambase MLP sequences, by 30%–55%. Draw-minus-conditional-mean contrasts follow different trajectories across seeds; two of nine intervals cross zero, both at $K = 1$. Appendix E reports the complete estimates.

C Implementation Details

C.1 Dry Bean split and training

The split key is the byte string `ce-drybean-v1`, followed by a null byte and the 16 raw features serialized as little-endian IEEE-754 64-bit doubles in source order. SHA-256 digests are interpreted as big-endian integers; residue zero modulo five is held out. Class labels are excluded from the key.

The Dry Bean MLP uses Adam, one 32-unit ReLU hidden layer, batch size 128, initial learning rate 10^{-3}, weight penalty 10^{-4}, a 300-iteration limit, and tolerance 10^{-4}. Training standard deviations use divisor n_{train}. Both observed and generated standardized coordinates are clipped to $[-5, 5]$. Ridge regression uses $\lambda = 1$ and an unpenalized intercept; residual and marginal variances use a 10^{-6} floor. Cross-entropy clips the true-class probability below at 10^{-15}. For seed s, NumPy's `default_rng` uses seed $1000003 + 7919s$ for a $(2767, 16, 32)$ Gaussian array and $91009 + s$ for the $5{,}000$ bootstrap resamples.

C.2 Secondary training settings

Secondary logistic models use 18 training epochs. The secondary MLP has 16 ReLU hidden units and a sigmoid output, trained for 36 epochs with initial learning rate 0.04, per-epoch decay 0.97, and L2 penalty 0.0005. Adult uses 6,000 sampled updates per epoch; the Spambase MLP caps this count at the training size. Bernoulli Naive Bayes uses Laplace smoothing and a posterior threshold of 0.5. Conditional replacement uses logistic Bernoulli sampling with one draw, and secondary bootstrap intervals use $1{,}000$ paired-example resamples.

The seed groups are 7411–7413 (CelebA), 7421–7423 (Adult logistic), 7431–7433 (Adult MLP), 7441–7443 (Bank Marketing), 7451–7453 (Spambase MLP), 7461–7463 (Spambase Naive Bayes), and 7471–7473 (Spambase joint removal).

D CelebA Per-Attribute Results

Table 4: CelebA replacement-accuracy comparisons for nine eligible attribute–run combinations. Signed differences subtract the training mean, conditional mean, or label-informed diagnostic from the conditional draw; negative values indicate lower replacement accuracy under the draw. Brackets give paired-example 90% percentile intervals from 1,000 resamples. Eyeglasses predictors attain 0.919–0.926 accuracy against a 0.935 majority rate.

Attribute	Seed	Full	Draw − train. mean [90%]	Draw − cond. mean [90%]	Draw − label-informed [90%]
Smiling	7411	0.822	-0.0061 $[-0.0071, -0.0052]$	-0.0116 $[-0.0124, -0.0108]$	-0.0266 $[-0.0277, -0.0257]$
Smiling	7412	0.824	-0.0002 $[-0.0011, +0.0008]$	-0.0074 $[-0.0082, -0.0066]$	-0.0272 $[-0.0282, -0.0263]$
Smiling	7413	0.815	-0.0002 $[-0.0011, +0.0008]$	-0.0056 $[-0.0063, -0.0049]$	-0.0202 $[-0.0210, -0.0192]$
Wearing_Lipstick	7411	0.918	$+0.0017$ $[+0.0009, +0.0024]$	-0.0029 $[-0.0035, -0.0024]$	-0.0075 $[-0.0081, -0.0069]$
Wearing_Lipstick	7412	0.922	-0.0006 $[-0.0013, +0.0001]$	-0.0031 $[-0.0036, -0.0026]$	-0.0073 $[-0.0078, -0.0067]$
Wearing_Lipstick	7413	0.919	$+0.0008$ $[+0.0002, +0.0015]$	-0.0039 $[-0.0044, -0.0035]$	-0.0094 $[-0.0100, -0.0088]$
Heavy_Makeup	7411	0.876	$+0.0068$ $[+0.0059, +0.0078]$	-0.0021 $[-0.0026, -0.0017]$	-0.0063 $[-0.0068, -0.0057]$
Heavy_Makeup	7412	0.875	$+0.0063$ $[+0.0054, +0.0072]$	-0.0029 $[-0.0034, -0.0024]$	-0.0065 $[-0.0071, -0.0060]$
Heavy_Makeup	7413	0.867	$+0.0071$ $[+0.0061, +0.0081]$	-0.0023 $[-0.0028, -0.0017]$	-0.0063 $[-0.0069, -0.0056]$

E Joint-Removal Sweep on Spambase

Table 5: Spambase secondary comparisons for three MLP seeds and $K \in \{1, 3, 6\}$ jointly replaced features. Signed differences are conditional-draw accuracy minus conditional-mean or label-informed accuracy, with negative values indicating lower accuracy under the draw; brackets give paired-example 90% percentile intervals from 1,000 resamples, and every seed has a smaller label-informed contrast magnitude at $K = 6$ than at $K = 1$.

K	Seed	Draw − cond. mean [90%]	Draw − label-informed [90%]
1	7471	-0.0087 $[-0.0207, +0.0033]$	-0.0315 $[-0.0467, -0.0174]$
1	7472	-0.0098 $[-0.0239, +0.0022]$	-0.0304 $[-0.0446, -0.0174]$
1	7473	-0.0250 $[-0.0370, -0.0120]$	-0.0457 $[-0.0609, -0.0326]$
3	7471	-0.0123 $[-0.0192, -0.0058]$	-0.0254 $[-0.0333, -0.0170]$
3	7472	-0.0127 $[-0.0210, -0.0062]$	-0.0272 $[-0.0355, -0.0199]$
3	7473	-0.0185 $[-0.0272, -0.0101]$	-0.0297 $[-0.0377, -0.0225]$
6	7471	-0.0123 $[-0.0170, -0.0078]$	-0.0219 $[-0.0275, -0.0161]$
6	7472	-0.0076 $[-0.0130, -0.0029]$	-0.0174 $[-0.0219, -0.0127]$
6	7473	-0.0107 $[-0.0154, -0.0063]$	-0.0205 $[-0.0254, -0.0159]$