DECISION EQUIVALENCE DOES NOT IDENTIFY THE EXPLANATION: SURROGATE MULTIPLICITY IN CLUSTER EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A common way to explain a clustering is to train a classifier on its labels and explain that surrogate with SHAP. How much does the resulting explanation depend on the classifier? Across six datasets and six model families, estimated attributions were less similar across families than across bootstrap refits of the same family. Under a shared probability-space KernelSHAP protocol, the median paired similarity gap was 0.130; its sign persisted across tested estimator settings, while its magnitude changed. We show that this ambiguity can remain even when decisions agree at every input. An exact construction rescales all class scores by the same positive, input-dependent factor, preserving every decision while making a decision-irrelevant feature uniquely most important. The result depends on what is explained: centering and normalizing scores gives exact attribution invariance within the common positive affine family. Decision fidelity alone therefore does not identify a score-based explanation. We recommend reporting the surrogate set and fidelity criterion, explained target, value function and background, and explanation variation across the selected models.

1 INTRODUCTION

Which features explain a clustering's assignments? A common approach trains a classifier to reproduce the cluster labels, then explains this *surrogate* with SHAP (Morichetta et al., 2019; Lundberg and Lee, 2017). High accuracy establishes agreement on assignments. The feature attributions explain the surrogate's numerical outputs, such as cluster probabilities. How much of this explanation depends on the chosen surrogate?

An assignment specifies which class score is largest, but leaves the scores themselves undetermined. Similarly accurate models are already known to assign different feature importance (Breiman, 2001; Fisher et al., 2019; Laberge et al., 2023). We ask whether disagreement remains under a stronger condition: *do identical decisions at every input determine the same score-based feature attributions?*

Our key idea is to vary scores while holding every decision fixed. Multiplying all class scores by one positive, input-dependent factor preserves their ordering but can change feature importance. Together with common additive shifts, these transformations define the *affine decision orbit*. We use it to establish what decision agreement leaves undetermined, then examine variation among fitted surrogates (Figure 1).

Our contributions are:

1. **An exact ranking reversal with the decision function fixed.** The construction preserves the decision at every input, a stronger constraint than similar predictive accuracy or agreement on observed instances. With the explanation method and background fixed, the winning-class Shapley vector changes from $(2, 1, 0)$ to $(8, 4, 9)$. A decision-irrelevant feature becomes uniquely most important. The reversal also holds for probability outputs and a nondegenerate background, without fitting or attribution estimation error (§4).
2. **An invariant attribution target.** Centering and normalizing scores gives identical exact attributions throughout the affine orbit. This invariance depends on the target and does not cover all decision-preserving transformations (§4.3).

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.

3. **Evidence from fitted surrogates.** Across six datasets and six model families, estimated attributions differ more across families than across same-family bootstrap refits to resampled training data. The median paired similarity gap is 0.130 under a common probability-space KernelSHAP protocol. The gap remains positive across tested estimator settings, with varying magnitudes (§6).

Supporting diagnostics relate disagreement to affine rescaling without assigning causes (§7). We recommend reporting the surrogate set, attribution target, value function and background, and variation across the selected models. Background and related work follow; the theory, protocol, and results appear in §4–7.

2 Background: labels, scores, and attributions

An input $x \in X \subseteq \mathbb{R}^d$ has d features. A fixed clustering assigns it a label $C(x) \in \{1, \dots, K\}$, where K is the number of clusters. A *surrogate* is a classifier trained to predict these labels. Its *fidelity* is its accuracy against C, not against external class labels. A score map $s : X \to \mathbb{R}^K$ gives one score $s_k(x)$ per class; its decision is the class with largest score. A probability surrogate $f : X \to \Delta^{K-1}$ is a special case: Δ^{K-1} is the set of nonnegative vectors summing to one, and $P_f(x) = \arg\max_k f_k(x)$ is its predicted partition.

What SHAP explains. The *attribution target* is a scalar function h, such as one class score s_k. A random vector B drawn from a *background* distribution supplies reference values for features that are not retained. For a set of retained features $S \subseteq [d] = \{1, \dots, d\}$, called a *coalition*, interventional SHAP uses

$$v_h(S; x) = \mathbb{E}_B\big[h(x_S, B_{-S})\big]. \tag{1}$$

The hybrid (x_S, B_{-S}) keeps the coordinates in S from x and takes the rest from one background draw. Equivalently it is $z = x \odot m + B \odot (1 - m)$, where m is the binary mask of S and $\odot$ denotes coordinatewise multiplication. The exact Shapley attribution of feature j is

$$\varphi_j[h](x) = \sum_{S \subseteq [d] \setminus \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \big[v_h(S \cup \{j\}; x) - v_h(S; x)\big]. \tag{2}$$

It averages how much adding feature j changes the target over all coalitions. The attribution vector $\varphi[h](x)$ contains these d contributions; absolute values measure importance. A *contrast* subtracts two outputs' attributions, $\varphi_j[s_k] - \varphi_j[s_l]$. KernelSHAP estimates these values using sampled coalitions (Lundberg and Lee, 2017; Chen et al., 2020). We use exact values for the theory and estimated values for fitted models, with the output coordinate, background, and value function fixed within comparisons.

3 Related work

Multiplicity and attribution uncertainty. The Rashomon effect and model class reliance show that similarly accurate models can assign different variable importance (Breiman, 2001; Fisher et al., 2019; Marx et al., 2020). Explanation variability can also arise within a fixed model architecture: Watson et al. (2022) show that changes in training setup, including random initialization and training-data order, can yield substantially different explanations. For feature attributions, Laberge et al. (2023) derive consensus partial orders across a Rashomon set, and Gunasekaran et al. (2024) aggregate importance ranges across models and methods. Hwang et al. (2026) distinguish variability from model selection and training from SHAP randomness using a common model-agnostic explainer. Methods can also disagree for a fixed predictor (Krishna et al., 2024). For surrogates of one fixed clustering, we ask a stronger question: even if decisions agree exactly, is the score explanation uniquely determined?

Decision agreement and score transformations. Thackshanaramana B (2026) reports structured cross-family disagreement on prediction-agreeing instances in binary classification. We extend this question to multiclass clustering. Our exact construction fixes the entire decision function, including at every feature/background hybrid, rather than selecting prediction-agreeing observed instances. The fitted-model experiments use a common estimator and bootstrap refits; they do not assume

global decision equivalence. For calibration, Decker et al. (2025) change perturbation-based explanations through class-preserving temperature scaling tailored to perturbation intensity. We establish a ranking reversal and an orbit-invariant target, then evaluate affine-orbit reconstructions between trained surrogates.

Attribution semantics and invariance. Choosing an interventional or conditional value function changes what an attribution explains (Chen et al., 2020; Sundararajan and Najmi, 2020). Model behavior outside the data support can also affect interventional attributions (Frye et al., 2021; Slack et al., 2020). Our construction requires neither correlated inputs nor disagreement outside the support. Implementation invariance concerns models that compute identical real-valued functions (Sundararajan et al., 2017). Our transformed functions make the same decisions but have different scores. Separately, Huang & Marques-Silva (2024) prove that SHAP scores can misrank features relative to logic-based relevancy for a fixed model; our reversal instead holds the attribution definition fixed and varies the decision-equivalent target, locating the failure in target non-identification. The construction also complements general limits on what complete, linear attributions can reveal (Bilodeau et al., 2024).

Cluster explanations. Surrogate-based approaches transfer supervised explanation tools to cluster assignments (Morichetta et al., 2019); Kauffmann et al. (2024) instead neuralize the clustering model itself and explain assignments without a surrogate, and explainable clustering constrains the partition to admit a compact tree representation (Moshkovitz et al., 2020; Frost et al., 2020). We keep the clustering fixed and ask how much its reported feature attribution depends on the fitted surrogate; a direct explanation of the K-means affinity (Appendix G) avoids surrogate choice but explains affinity rather than membership.

4 THEORY: THE DECISION ORBIT

Can attribution differences persist even when decisions agree everywhere? We show they can, through an exact ranking reversal — separating the conceptual problem from surrogate fitting and estimation error. The exact results below assume that the domain contains every feature/background hybrid and that coalition expectations exist. We hold the background, value function, and output coordinates fixed throughout.

Definition 1 (Decision equivalence)**.** Two score maps $s, s' : X \to \mathbb{R}^K$ are **decision-equivalent** when they select the same winning class at every input: $\arg\max_k s_k(x) = \arg\max_k s'_k(x)$ as sets for all $x \in X$, with ties resolved by the same deterministic rule.

Definition 2 (Decision orbit)**.** The **orbit** of s is

$$\mathcal{O}(s) \;=\; \big\{\, s' : s'_k(x) = a(x)\, s_k(x) + b(x),\; a(x) > 0,\; b(x) \in \mathbb{R} \,\big\},$$

where a and b are functions of x **common to all** K **coordinates** $s_1, \ldots, s_K$ (one score per class).

Proposition 1 (Closure)**.** *$\mathcal{O}(s)$ is contained in the decision-equivalence class of s.*

Proof. Fix x. Adding the same $b(x)$ to all K coordinates preserves their order. Multiplying them by $a(x) > 0$ also preserves their order. Hence $\arg\max_k [a(x)s_k(x) + b(x)] = \arg\max_k s_k(x)$. ■

The orbit generally covers only part of decision equivalence; the affine subfamily suffices for the results below.

4.1 ADDITIVE COMMON TERMS LEAVE CONTRASTS INVARIANT

Proposition 2 (Common additive shifts preserve contrasts)**.** *If $s'_k = s_k + g$ with $g : X \to \mathbb{R}$ common to all k, then for every pair k, l and every feature j,*

$$\varphi_j[s'_k] - \varphi_j[s'_l] \;=\; \varphi_j[s_k] - \varphi_j[s_l].$$

The level *of each attribution may change: $\varphi_j[s'_k] = \varphi_j[s_k] + \varphi_j[g]$.*

Proof sketch. The value function is linear in the explained function, and the Shapley value is a fixed linear combination of the coalition values $v(S)$. Therefore, $\varphi_j[s_k + g] = \varphi_j[s_k] + \varphi_j[g]$. Since g is shared by every output, its attribution cancels when we subtract two outputs' attributions. The full proof is in Appendix A. ■

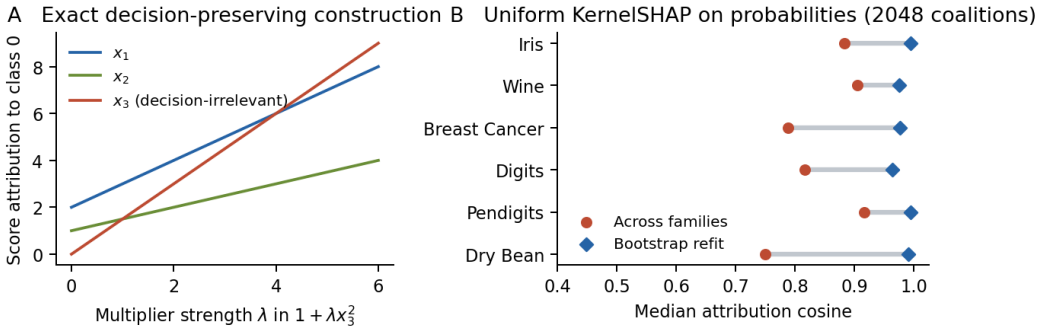

Figure 1: (A) The analytic construction of §4: exact interventional attributions to the winning cluster at x^*, as the multiplier strength λ in $a = 1 + \lambda x_3^2$ grows. The decision-irrelevant feature x_3 (absent from the scores) overtakes both decision-relevant features. **(B)** The uniform-protocol experiment of Table 3: cross-family similarity against the bootstrap-refit baseline, per dataset.

Thus, differences between outputs' attributions remain fixed within the additive family, even though the individual attributions can change. Appendix A gives a numerical illustration.

4.2 Input-dependent scaling can reverse feature rankings

Positive input-dependent rescaling preserves every decision while allowing both attribution levels and contrasts to change. The following construction gives an exact ranking reversal within a single decision orbit.

Proposition 3 (Analytic non-identifiability). *On $D = [-2, 2]^3$ take $s_0 = x_1 + x_2$, $s_1 = -x_1$, $s_2 = -x_2$, and $t = a \cdot s$ with $a(x) = 1 + 6{x_3}^2$. Then s and t induce the same partition at every point of D; the feature x_3 is absent from s and cannot affect either* $\arg\max$ *anywhere. Yet with a background at the origin, the exact interventional Shapley vectors at $x^* = (2, 1, 1)$ for the winning cluster move from $(2, 1, 0)$ under s to $(8, 4, 9)$ under t: x_3 goes from uniquely last to uniquely first. The across-output mean absolute importances move from $(4/3, 2/3, 0)$ to $(16/3, 8/3, 6)$ — the same strict reversal — and the cluster-0/1 contrast moves from $(4, 1, 0)$ to $(16, 4, 15)$.*

Proof. With a zero background, each linear monomial contributes its value to its feature's attribution. Each interaction term $6x_j x_3^2$ involves two active features and is zero unless both enter the coalition, so its value is split equally between them. For the winning cluster at x^*, the two interactions have values 12 and 6. Adding their contributions to those of $s_0 = x_1 + x_2$ gives

$$\varphi[t_0](x^*) = (2, 1, 0) + (6, 0, 6) + (0, 3, 3) = (8, 4, 9).$$

The other vectors follow by the same calculation and Shapley linearity. ■

The reversal also holds for valid probabilities and independent, nondegenerate backgrounds; neither feature correlation nor off-support disagreement is required. Appendix A gives these variants and the Gaussian illustration.

Remark 1. Removing a common additive term is not enough to make attributions invariant under decision equivalence. Input-dependent multiplication can change even the contrasts: $\varphi[a \cdot s_k]$ need not equal $a \cdot \varphi[s_k]$ when a varies with x. This is a failure of invariance in general; equality can still hold in special cases.

4.3 An orbit-invariant attribution target

Whether attributions are uniquely determined depends on the *estimand*: the quantity being explained. It is not an inherent limitation of Shapley values. Let $\bar{s} = K^{-1} \sum_k s_k$ be the mean class score, $\mathbf{1}$ the all-ones vector, and $s_c = s - \bar{s}\,\mathbf{1}$. Define the centered-normalized target $H(s)(x) = s_c(x)/\|s_c(x)\|_2$ where $s_c(x) \neq 0$, and 0 otherwise. For any orbit member $t = a\,s + b\,\mathbf{1}$,

Table 1: Datasets: n rows, d features, K clusters, and the range over the six families of five-fold accuracy in recovering the fixed partition (reduced-feature variants excluded). News embeddings is used only for the multiclass diagnostics, where it replaces Breast Cancer.

dataset	n	d	K	accuracy range
Main experiment (six datasets)				
Iris	150	4	3	0.933–0.973
Wine	178	13	3	0.972–0.989
Breast Cancer	569	30	2	0.977–0.995
Digits	1,797	64	10	0.940–0.972
Pendigits	10,992	16	10	0.973–0.991
Dry Bean	13,611	16	7	0.973–0.995
Orbit diagnostics only				
News embeddings	4,200	32	6	0.971–0.984

centering gives $t_c = a\, s_c$ with $a > 0$. Thus, for every output coordinate k, $H_k(t) = H_k(s)$ pointwise and

$$\varphi[H_k(t)](x) = \varphi[H_k(s)](x)$$

for every x, with the value function and background fixed. Centering removes the common shift, and normalization removes the positive scale.

This invariance has a specific scope. Non-affine transformations that preserve the $\arg\max$ can still change H, so H is not invariant under every form of decision equivalence. Explaining the hard-label indicator gives that broader invariance, but introduces a discontinuous target. An analyst can therefore obtain affine-orbit invariance by changing the quantity being explained. Fidelity alone does not uniquely determine the original *score-valued* explanation. Appendix C compares H and probability attributions on fixed real surrogate pairs, for which orbit membership is not assumed.

5 Experimental protocol

We tested whether attribution variation across model families exceeded variation from refitting the same family. For each dataset, we fixed the clustering, trained six surrogate families, and explained the same points against the same background. We used K-means with a fixed seed, except in controls varying the clustering algorithm.

Models and datasets. We used random forests (`RF`), extremely randomized trees (`ExtraTrees`), histogram gradient boosting (`HistGB`), extreme gradient boosting (`XGB`), logistic regression (`LogReg`), and a multilayer perceptron (`MLP`). Their off-the-shelf settings are listed in Appendix B. Some default-cap MLP fits did not converge; refitting all MLPs to convergence changes the median gap by less than 0.001 and leaves its sign unchanged in every dataset (Appendix C). We grouped them as forests, boosting, and linear/network models: $\binom{6}{2} = 15$ pairs, 3 within groups and 12 across groups. Table 1 reports fidelity. Scaling and PCA were fit on the full data before cross-validation: these accuracies recover a fixed transductive partition, not a nested clustering pipeline. Family accuracies differed by up to four points on Iris.

The main comparison used six tabular datasets. Orbit diagnostics replaced `breast_cancer` ($K = 2$), whose unconstrained centered-output residual is uninformative where defined, with `news_embeddings`: MiniLM embeddings of 20 Newsgroups posts reduced to 32 PCA components (§7).

Shuffling avoided row-order artifacts; capping points kept the cost of evaluating points $\times$ coalitions $\times$ background size feasible. We reset the estimator seed before each call. SHAP selected up to ten features when `l1_reg` was unset. The earlier 100-coalition comparison and sensitivity checks appear in Appendix C.

Similarity and baseline. We explained the output for the assigned cluster $C(x_i)$, the *own-cluster convention*. Let $u_{f,i}$ be its estimated d-feature attribution vector at explained point x_i, with N points

Table 2: Attribution settings. A fit seed controls model fitting; an estimator seed controls coalition sampling. Within each experiment, every model and refit shared the background, explained rows, and coalition samples.

Setting	Main comparison	Orbit diagnostics
Explained points	First 150 shuffled rows, or all if fewer	100 points
Requested coalitions	2048	512
Feature selection	None	SHAP default: up to ten features
Fit seeds	Three	Same three
Shared across both settings		
Background	Weighted 10-center K-means background per dataset	
Estimator	Interventional KernelSHAP (SHAP 0.48.0)	

in total. We computed

$$\mathrm{sim}(f, g) = \mathrm{median}_{i=1}^{N} \frac{u_{f,i}^{\top} u_{g,i}}{\|u_{f,i}\|_2 \|u_{g,i}\|_2 + 10^{-12}}. \tag{3}$$

This cosine measures alignment: values near one indicate similar directions. For each fit seed r, let U_r be the median similarity over the 15 family pairs and B_r the median over within-family bootstrap refits (models retrained on rows sampled with replacement). The reported paired gap is $G = \mathrm{median}_r(B_r - U_r)$. A positive gap means cross-family explanations are less similar than refits. The group contrast is within-group minus across-group median similarity. All comparisons held preprocessing, labels, background, rows, and estimator seed fixed.

Scope. The baseline measures conditional refit stability, not uncertainty from SHAP, the background, or clustering. Shared settings need not give equal error; finite budgets do not establish convergence, and hybrid points were not checked for data-manifold membership. Each family occurs in five pairs, so pair correlations are descriptive and datasets are the replication units. Archived controls use explicitly marked mixed protocols (Appendix B).

6 Results: variation across faithful surrogates

6.1 Cross-family similarity against the refit baseline

Does the variation that decision equivalence permits actually appear across fitted surrogates? Table 3 reports the main comparison. With the partition, points, background, and estimator held fixed, median cross-family similarity (0.750–0.916) was below the bootstrap-refit baseline (0.964–0.994) in every dataset. The paired gap within seeds ranged from 0.064–0.242, with a median of 0.130 across datasets (positive in all six; one-sided sign test $p = 2^{-6} \approx 0.016$). Each gap subtracts similarities within a seed before taking medians, so it need not equal the difference of the displayed medians.

On `Dry Bean` (n = 13,611, $d = 16$, $K = 7$), five-fold accuracy was 0.973–0.995, yet cross-family similarity was 0.750 against a refit baseline of 0.990: high fidelity coexists with different estimated explanations.

Repeating the full 150-point protocol with three estimator seeds kept the gap positive in all 54 dataset/fit-seed/estimator-seed combinations. The largest within-fit seed spread was 0.012 (Appendix C.2). Exhaustive Shapley values on `Wine` give a gap of 0.0631, versus 0.0619–0.0644 for the sampled estimates on the same models (Appendix C.3).

Gap size still depends on the budget: the earlier 100-coalition experiment gave a median of 0.205 and a different dataset ordering (Appendix C). The supported claim is a persistent positive gap, not a fixed magnitude or ordering. Figure 1 places this comparison alongside the exact construction of §4.

Exploratory grouping. Within-group similarity exceeded across-group similarity in all six datasets (contrasts 0.072–0.230; Table 3). The grouping is coarse, especially LogReg/MLP, and the main cross-family gap does not depend on it. Appendix J gives the supporting analysis.

Table 3: The uniform-protocol experiment (2048 requested coalitions per point, no ℓ_1 selection, shared estimator seed): bootstrap-refit baseline, cross-family similarity, within-seed paired gap (median and range over the three fit seeds), and the same/different functional-group contrast (medians per dataset).

dataset	baseline	cross-family	paired gap	gap range	group contrast
Iris	0.994	0.883	0.111	[0.105, 0.112]	0.112
Wine	0.974	0.904	0.064	[0.058, 0.091]	0.107
Breast Cancer	0.977	0.787	0.185	[0.174, 0.221]	0.215
Digits	0.964	0.816	0.149	[0.148, 0.150]	0.167
Pendigits	0.994	0.916	0.078	[0.072, 0.078]	0.072
Dry Bean	0.990	0.750	0.242	[0.205, 0.251]	0.230

6.2 Additional controls and their scope

The supplementary controls varied, one at a time, factors that could explain the gap: surrogate accuracy (families retrained to a matched cross-validation target), identical predictions on every observed point, correlated features, the attribution method (LIME (Ribeiro et al., 2016) and permutation importance), the clustering algorithm, and the SHAP attribution target (Appendix D, Table 11). The gap persisted under each control. These mixed-protocol controls constrain particular alternatives without isolating every factor; the main experiment uses probability-space KernelSHAP throughout, and the §4 construction preserves decisions everywhere, including at hybrids.

6.3 Feature-importance rankings and stability

We also compared feature-importance rankings under the uniform protocol. For each model, we averaged the absolute attribution of each feature over points, then compared the resulting vectors with Spearman correlation. All families emit probabilities, so their attribution magnitudes share a unit, although approximation error remains. The median gap from the bootstrap-refit baseline was 0.149 across the six datasets. On `dry_bean`, cross-family correlation was 0.379 and the baseline was 0.978. Appendix D reports the earlier experiment.

A feature passes our stability rule if its rank range across families is no wider than its bootstrap-refit range, which must itself span at most three positions — otherwise an unstable refit reference could certify a rank. The rule describes observed stability without a statistical identification or confidence guarantee. Very few features pass: 4 of 134 eligible features under the headline setup, and 2 with converged MLPs. The exact count and feature identities depend on the protocol, so we treat scarcity of stable rankings as the finding. Appendix D gives the counts, passing features, and cutoff sensitivity; Appendix C assesses estimator repeatability.

7 Diagnostics: relating the orbit to fitted models

The construction gives an exact mechanism, but trained surrogates need not lie in one orbit. We therefore asked two questions: how well can one model's outputs be reconstructed by rescaling the other, and does the multiplier's input dependence track attribution disagreement?

A centered-output residual is nearly degenerate for confident models predicting the same class and is uninformative for binary outputs where defined. Simple multiplier variability also has a weak association with explanation similarity (Appendix E). We therefore examine the feature dependence of the extracted multiplier and compare it with disagreement at SHAP's query points.

7.1 Multiplier attributions and query-point disagreement

Definition 3 ($\Phi[a]$)**.** Given two surrogates f, g fit to the same labels, let a be the least-squares coefficient fitting g's centered outputs to f's ((4), Appendix E). We can evaluate it by calling both models. We explain it using the matched background and 100 points from the orbit analyses, where N is the number of explained points. Define $\bar{\varphi}_j[a] = \frac{1}{N}\sum_{n=1}^{N}|\varphi_j[a](x_n)|$ and $\Phi[a] = \frac{1}{d}\,\|\bar{\varphi}[a]\|_1$, the absolute value taken before any averaging. Thus, $\Phi[a]$ averages absolute attribution over points and

Table 4: Spearman association of each diagnostic with explanation similarity under the orbit-diagnostic protocol (Table 2), joined by family pair and fit seed, pair medians over seeds, 15 pairs per dataset. Negative values mean that larger diagnostic values accompany lower similarity. "TV" is a mean total-variation distance between the two models' output probabilities, at hybrid (coalition) points and at observed points. Generated from the archived artifacts.

dataset	$\Phi[a]$	$\mathrm{cv}(a)$	hybrid-point TV	observed-point TV
Iris	-0.421	-0.107	-0.836	-0.125
Wine	-0.664	-0.429	-0.539	+0.025
Digits	-0.400	-0.014	-0.832	-0.014
Pendigits	-0.582	-0.086	-0.650	-0.068
Dry Bean	-0.779	+0.150	-0.168	+0.071
News embeddings	-0.821	+0.107	-0.939	+0.100
Median	-0.623	-0.050	-0.741	+0.005

features. It explains the *raw* coefficient, a well-defined diagnostic function of x. Orbit membership is claimed only for the positive reconstruction in §7.2.

For comparison, total variation (TV) measures probability-output disagreement: $\mathrm{TV}(f(z), g(z)) = \frac{1}{2}\sum_k |f_k(z) - g_k(z)|$. We averaged it separately over observed points and hybrid query points. Table 4 reports Spearman rank correlations ρ over family pairs, after taking each pair's median over fit seeds. The simpler comparator $\mathrm{cv}(a)$ is the multiplier's standard deviation divided by its absolute mean.

Empirical hypothesis. Larger $\Phi[a]$ is associated with lower explanation similarity between f and g. We assess this directional hypothesis through descriptive correlations within the measured pairs.

Table 4 compares the multiplier diagnostics with output-space distances. $\Phi[a]$ is negatively associated with similarity in all six datasets (median $\rho = -0.623$). The coefficient of variation has a weak median association (-0.050), as does observed-point total variation ($+0.005$). Hybrid-point total variation has the strongest median association (-0.741), exceeding that of $\Phi[a]$ in four of six datasets. This suggests that agreement on observed outputs is a poor guide to agreement on interventional explanations: disagreement at the explainer's hybrid queries tracks attribution disagreement more closely. The result connects to work on off-support behavior (Frye et al., 2021; Slack et al., 2020), but we did not test whether these hybrids lie off the data manifold. This is an empirical diagnostic, not an explanation of the exact theorem or a causal attribution of the observed gap.

7.2 Projecting real pairs onto the orbit

We fit a positive affine transformation of each model f to a second model g and explain the reconstructed score map $\tilde{g}$. This guarantees that $\tilde{g}$ lies in f's orbit, although it need not be a probability vector. Appendix F gives the projection and numerical guards.

The projection changes attributions substantially, but reconstruction remains imperfect relative to the bootstrap baseline. Similarity between explanations of $\tilde{g}$ and g falls below that baseline by 0.003–0.104, with the largest gap on `dry_bean` (0.880 versus 0.984). Thus, affine rescaling provides a useful diagnostic without accounting for all observed disagreement. Reference swaps and constant-multiplier controls qualify this interpretation (Appendices F and F.3).

A directional refinement of the multiplier diagnostic does not consistently improve its association with disagreement (Appendix I).

8 Discussion: attribution targets and reporting

Do the attributions describe the clustering or the chosen classifier? Fidelity alone cannot answer: a partition fixes decisions, but score attributions also depend on their numerical representation.

Choosing the target. Table 5 distinguishes four questions an attribution can answer. All invariance statements hold with the value function and background fixed. The direct K-means affinity $a_k(x) =$

Table 5: Attribution targets and the scope of their interpretation. Invariance refers to exact attributions for each scalar output coordinate.

Target	Explains	Invariance	Main limitation
Probability f_k	Surrogate's predicted probability	Not ensured by decision agreement	Depends on score representation
$H_k(s)$	Centered, normalized scores	Positive common affine changes	Not all decision-preserving changes
Label indicator	Membership in a predicted class	All globally decision-equivalent scores	Discontinuous; discards score variation
Direct affinity a_k	Cluster geometry	Avoids surrogate choice	Explains affinity, not membership

$-\sum_j (x_j - \mu_{kj})^2$, where μ_k is cluster k's center, has exact additive Shapley values. Its comparisons with surrogates fall within the surrogate-pair spread (Appendix G); this selects no canonical target.

Defining the surrogate set. To make "faithful" operational, predeclare a finite candidate set $\mathcal{F}$, a fidelity evaluation protocol, and a tolerance $\epsilon \geq 0$. With $q(f)$ denoting accuracy against the fixed clustering under that protocol, define

$$q^\star = \max_{f \in \mathcal{F}} q(f), \qquad \mathcal{F}_\epsilon = \{f \in \mathcal{F} : q(f) \geq q^\star - \epsilon\}.$$

For example, $\epsilon = 0.01$ allows one percentage point below the best candidate. This reporting prescription does not retrospectively filter our six families; we compare them at their reported fidelities.

Reporting variation. Report the candidate and retained sets, fidelity protocol and tolerance, scalar target, value function and background, and ranges over $\mathcal{F}_\epsilon$ for signed attributions, absolute importance, or ranks. This follows the Rashomon and consensus literature (Fisher et al., 2019; Laberge et al., 2023; Gunasekaran et al., 2024; Hwang et al., 2026). Our rank-stability results illustrate such reporting (§6.3).

9 Limitations

Scope. The main experiments explain K-means; alternative clusterings and attribution methods appear only in the archived mixed-protocol controls (Appendix D). Cross-family comparisons cover six tabular datasets; orbit analyses replace the binary dataset with embeddings. Accuracies recover a fixed transductive partition and differ across families.

Estimation. The full-protocol replication varies three estimator seeds on fixed models across all six datasets; budget and feature-selection sensitivity covers only two datasets and 12 points each (Appendix C). Exhaustive Wine attributions validate one dataset and fit seed relative to the finite background (Appendix C.3). These checks do not establish exact convergence elsewhere or robustness of every archived result. Only new runs record runtime package versions.

Interpretation. The orbit diagnostics do not identify a causal contribution to disagreement. Reconstruction similarity falls below the refit baseline by up to 0.104. Hybrid-point total variation has a stronger median association with disagreement than $\Phi[a]$. We evaluate the invariant target H on the same small fixed samples, leaving its usefulness across datasets and refits open.

10 Conclusion

Identical clustering decisions can coexist with reversed feature rankings. Centering and normalizing scores gives exact invariance within the affine orbit. Across fitted surrogates, attribution differences exceeded a conditional refit baseline, with estimator-dependent magnitudes. Reports should specify the surrogate set and fidelity criterion, scalar target, value function and background, and explanation variation. Broader evaluation of the invariant target across datasets and refits remains open.

REPRODUCIBILITY STATEMENT

All datasets are public: scikit-learn built-ins, UCI `dry_bean` and `pendigits`, and embeddings of 20 Newsgroups posts. The embedding configuration uses `all-MiniLM-L6-v2` on six categories with up to 700 posts each. We remove headers, footers, and quotes, drop posts shorter than 50 characters, and reduce the 384-dimensional embeddings to 32 PCA components. The protocol fixes and reports the clustering seed, the six families and their hyperparameters, the estimator per experiment arm (uniform probability KernelSHAP for the main similarity and orbit analyses and headline rank-stability count; mixed estimators for marked legacy controls and the archived rank-stability analysis), the shared background, the explained points, the fit seeds, the coalition budgets, the denominator guard, the positive-part constraint of the projection, and the aggregation order of every table.

The pinned SHAP version defaults to ℓ_1 selection of ten features per point when `l1_reg` is unset. Archived artifacts contain experiment signatures but no complete software/hardware manifest for each run. Regenerating the embeddings requires the optional `sentence_transformers` dependency and the pretrained model revision; the included embedding table supports all analyses without regeneration. Two standalone scripts audit the principal archived summaries: `audit_results.py` (standard library only) recomputes the main numerical comparisons, while `build_assets.py` checks the analytic construction and regenerates the main tables, the uniform directional comparison and the overview figure. The remaining appendix results refer to their archived experiment outputs; these scripts do not rerun training or establish estimator convergence. The new `36_kernel_validation.py` runs the targeted sensitivity and invariant-target study, `37_headline_2048.py` supplies the arm reported in Table 3, and `38_mlp_convergence.py` refits every MLP to convergence; `audit_validation.py` audits the saved attribution tensors of all three, recomputes the rank-stability counts, and generates the appendix tables. `39_headline_estimator_seeds.py` saves reconstructed models and varies estimator seeds under the full protocol; `40_wine_exhaustive.py` enumerates all Wine coalitions. `audit_estimator_seeds.py` checks the saved models and tensors and generates their validation tables. Anonymized code, data, and archived artifacts are available at `https://anonymous.4open.science/r/decision-equivalence-repro`. The repository README pins the software environment and maps each table and figure to the script that produces it; the four audit scripts regenerate the paper's tables from the saved attribution tensors in about one minute, and rerunning the main comparison on Iris from scratch reproduces the archived tensors exactly.

ETHICS STATEMENT

This work studies the reliability of post-hoc explanations on public datasets and involves no human subjects, personal data, or deployment. Its main societal implication concerns the use of explanations in decisions: feature importance reported from a single clustering surrogate can depend on an undisclosed model choice. Reporting ranges alongside a refit baseline helps address this risk. The recommendation mitigates an existing problem without introducing a new capability, and we see no plausible dual-use concern.

AI USE STATEMENT

We used generative AI tools, including Anthropic's Claude and OpenAI's Codex, for tasks requiring disclosure. These included feedback on experimental methodology; design of the later orbit-interpolation, orbit-projection, reference-swap, embedding-replication, and targeted estimator and invariant-target experiments; implementation of measurement code under the authors' specifications; and review of mathematical arguments. We also used these tools for tasks with recommended disclosure: drafting, restructuring, and editing the manuscript, and locating and checking related work. Empirical summaries come from archived experimental outputs. We take responsibility for the final text, claims, and artifacts, including those produced with generative AI assistance.

REFERENCES

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2), 2024.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Thomas Decker, Volker Tresp, and Florian Buettner. Improving perturbation-based explanations by understanding the role of uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. ExKMC: Expanding explainable k-means clustering. *arXiv preprint arXiv:2006.02399*, 2020.

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

Abirami Gunasekaran, Pritesh Mistry, and Minsi Chen. Which explanation should be selected: A method agnostic model class reliance explanation for model and explanation multiplicity. *SN Computer Science*, 5(5):503, 2024.

Xuanxiang Huang and Joao Marques-Silva. On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*, 171:109112, 2024.

Hyunseung Hwang, Seungeun Lee, Lucas Rosenblatt, Steven Euijong Whang, and Julia Stoyanovich. Explanation multiplicity in SHAP: Characterization and assessment. *arXiv preprint arXiv:2601.12654*, 2026.

Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1926–1940, 2024.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the Rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Andrea Morichetta, Pedro Casas, and Marco Mellia. EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pp. 22–28, 2019.

Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7055–7065, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017.

Thackshanaramana B. Hypothesis class determines explanation: Why accurate models disagree on feature attribution. *arXiv preprint arXiv:2603.15821*, 2026.

Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 875–884, 2022.

Guide to the appendices. The exact construction and protocol appear first (Appendices A–B), followed by estimator validation (Appendix C) and controls (Appendix D). Orbit diagnostics and the direct clustering target are in Appendices E–G. Further diagnostic comparisons and archived pairwise and regional analyses follow. Legacy calibration and normalized comparisons are collected at the end (Appendix L). These archived analyses support the stated controls and scope checks; they use mixed estimators where explicitly marked. Their "floor" denotes a refit or seed-change baseline, not a bound on estimator error. Class I denotes within-group pairs and class II across-group pairs. The headline rank-stability analysis uses the uniform probability protocol.

A Construction details and the additive/cube controls

A.1 Worked example: fixed decisions, changed attributions

Figure 2 illustrates Proposition 3: the partition stays fixed, but the output values and their attributions change. The tables below explain why, using every coalition at $x^* = (2, 1, 1)$.

Coalition values. Recall $s = (x_1 + x_2, -x_1, -x_2)$ and $t = a\, s$, where $a(x) = 1 + 6x_3^2$. With the background at the origin, the coalition point z_S retains the coordinates of x^* in S and sets all others to zero. Thus the value function for a scalar score h is simply $v_h(S) = h(z_S)$. Table 6 lists these values for output 0, the winning cluster at x^*. The argmax sets agree in every row, including the ties. However, adding feature 3 can change t_0 because it changes the common multiplier.

Table 6: Coalition values for output 0 at $x^* = (2, 1, 1)$, with background at the origin. Bold entries mark score values changed by the multiplier. The two score maps have identical argmax sets at every coalition point.

S	z_S	$a(z_S)$	$s_0(z_S)$	$t_0(z_S)$	$\arg\max_k s_k(z_S)$	$\arg\max_k t_k(z_S)$
$\emptyset$	$(0,0,0)$	1	0	0	$\{0,1,2\}$	$\{0,1,2\}$
$\{1\}$	$(2,0,0)$	1	2	2	$\{0\}$	$\{0\}$
$\{2\}$	$(0,1,0)$	1	1	1	$\{0\}$	$\{0\}$
$\{3\}$	$(0,0,1)$	7	0	0	$\{0,1,2\}$	$\{0,1,2\}$
$\{1,2\}$	$(2,1,0)$	1	3	3	$\{0\}$	$\{0\}$
$\{1,3\}$	$(2,0,1)$	7	2	**14**	$\{0\}$	$\{0\}$
$\{2,3\}$	$(0,1,1)$	7	1	**7**	$\{0\}$	$\{0\}$
$\{1,2,3\}$	$(2,1,1)$	7	3	**21**	$\{0\}$	$\{0\}$

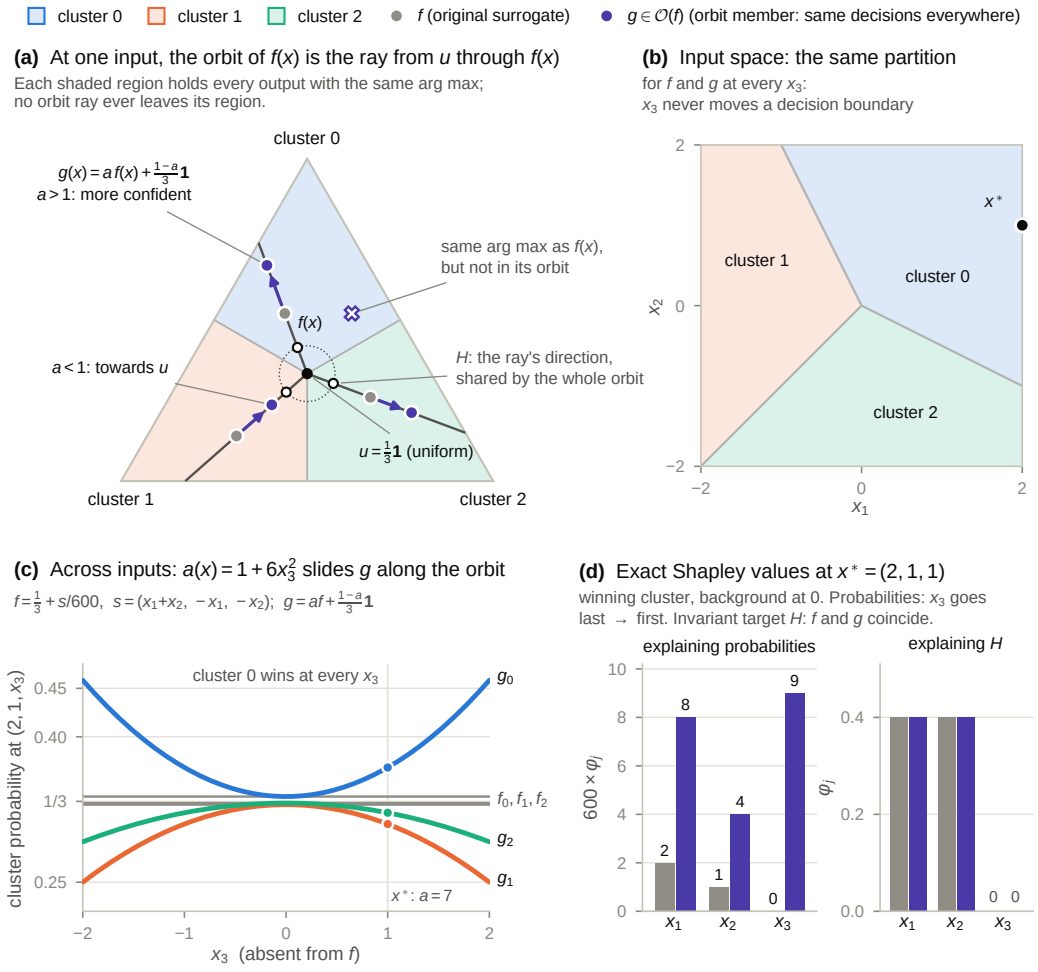

Figure 2: Decision preservation and attribution change. **(a)** For probability outputs, positive common affine changes follow the portion of a ray from $u = \mathbf{1}/3$ that remains inside the simplex. For a nonuniform starting vector, u is excluded because $a > 0$; the uniform vector itself stays fixed and has $H = 0$. The marked off-ray output has the same argmax but is outside this affine family. **(b)** The construction's partition is independent of x_3. **(c)** Along $(2, 1, x_3)$, probabilities change while cluster 0 remains the winner. **(d)** At x^*, probability attributions reverse their ranking, whereas attributions to H_0 agree. The probability bars are multiplied by 600 to match the score attributions in the tables.

From coalition values to attributions. For $S \subseteq \{1, 2\}$, write $\Delta_3 h(S) = h(z_{S\cup\{3\}}) - h(z_S)$. The Shapley value averages these changes with weights $w(S) = |S|!\,(2 - |S|)!/3!$ (Table 7). Under s_0, every change is zero. Under t_0, adding x_3 changes the value whenever x_1 or x_2 is already present:

$$\varphi_3[t_0](x^*) = \tfrac{1}{3}(0) + \tfrac{1}{6}(12) + \tfrac{1}{6}(6) + \tfrac{1}{3}(18) = 9.$$

Thus x_3 receives attribution for changing the explained score, even though it never changes a decision.

Table 7: Marginal contributions of x_3 and their Shapley weights. Each change is the difference between two entries in Table 6.

x_3 joins S	$w(S)$	$\Delta_3 s_0(S)$	$\Delta_3 t_0(S)$
$\emptyset$	$1/3$	$0 - 0 = 0$	$0 - 0 = 0$
$\{1\}$	$1/6$	$2 - 2 = 0$	$14 - 2 = 12$
$\{2\}$	$1/6$	$1 - 1 = 0$	$7 - 1 = 6$
$\{1, 2\}$	$1/3$	$3 - 3 = 0$	$21 - 3 = 18$
Weighted sum		0	9

The same calculation for x_1 and x_2 gives $\varphi[s_0](x^*) = (2, 1, 0)$ and $\varphi[t_0](x^*) = (8, 4, 9)$. Their sums are $3 = s_0(x^*) - s_0(0)$ and $21 = t_0(x^*) - t_0(0)$, respectively, as efficiency requires.

Scores, probabilities, and the invariant target. The figure uses the probability maps $f = \mathbf{1}/3 + s/600$ and $g = \mathbf{1}/3 + t/600$. The constant term has zero attribution, so $\varphi[f_0] = \varphi[s_0]/600$ and $\varphi[g_0] = \varphi[t_0]/600$. These maps retain the same ranking reversal. By contrast, $H_0(f) = H_0(g)$ at every input. Their exact attribution vectors at x^* both equal $(3/(2\sqrt{14}),\, 3/(2\sqrt{14}),\, 0)$, using $H = 0$ at the uniform background. The right-hand bars in panel **(d)** show this common vector without rescaling.

A.2 Extensions and supplementary controls

The same construction also applies to probabilities and nondegenerate backgrounds. *Probabilities*: $f = 1/3 + s/600$ and $g = 1/3 + t/600$ are valid probability vectors on D (the scores sum to zero; all entries lie in $[1/6, 1/2]$) with $g = af + \frac{1-a}{3}\mathbf{1}$ — the same positive affine orbit, the same decision everywhere, and all Shapley vectors divided by 600. *Nondegenerate backgrounds*: with independent, symmetric, mean-zero background coordinates supported in D with $\mathrm{Var}(B_3) = v$, the transformed class-0 vector at x^* is $(8+6v,\, 4+3v,\, 9-9v)$, so the reversal holds for every $v < 1/15$ (uniform product background on $[-0.1, 0.1]^3$: $(8.02, 4.01, 8.97)$). *Independent inputs*: the construction requires no feature correlation and no appeal to off-support model behavior. Below we give an additional Gaussian illustration, with exact attributions relative to its sampled background. We also establish that the cubic transformation s^3 belongs to the affine orbit in the zero-common-term, three-class variant, by Vieta's identity; that transformation therefore does not provide an outside-orbit control in this setting.

Full proof of Proposition 2. The interventional value function is linear in the explained function: $v_{s_k+g}(S) = v_{s_k}(S) + v_g(S)$ for every coalition S, because the expectation over coalition points is linear. The Shapley value of feature j is a fixed linear combination of the $v(S)$ over coalitions, hence $\varphi_j[s_k + g] = \varphi_j[s_k] + \varphi_j[g]$. Since g does not depend on k, the added term $\varphi_j[g]$ is identical across outputs and cancels in every contrast $\varphi_j[s'_k] - \varphi_j[s'_l]$. ■

Archived Gaussian illustration. This experiment is distinct from the analytic construction of Proposition 3. It uses a sample of 400 independent standard Gaussian inputs and scores $s_k = w_k^\top x$, $w_k = (\cos(2\pi k/3), \sin(2\pi k/3), c)$, $k = 0, 1, 2$, with $c = 0.5$ and multiplier $a = 1 + 5x_3^2$. The following values enumerate interventional coalitions exactly relative to the sampled background and average $|\varphi|$ over the sampled points and three outputs; they are not population identities.

transformation	in orbit?	x_1	x_2	x_3	ranking
s (reference)	—	0.552	0.433	0.382	$x_1 > x_2 > x_3$
$a \cdot s$	**yes**	3.452	2.727	4.943	$\mathbf{x_3 > x_1 > x_2}$
s^3	no	1.956	1.489	1.488	$x_1 > x_2 > x_3$
$s + g$, $g = 2x_3^2$	additive	0.552	0.433	1.919	$x_3 > x_1 > x_2$

Both the multiplicative and additive transformations in the orbit reverse the level ranking in this sample. The additive transformation leaves contrasts invariant to 10^{-14}. Contrasts between outputs 0 and 1 (mean over points; the other two class pairs behave identically):

transformation	$\lvert\varphi_j[s_0] - \varphi_j[s_1]\rvert$ at x_1	x_2	x_3
s (reference)	1.242	0.650	$\mathbf{3.7 \times 10^{-16}}$
$s + g$ (additive)	1.242	0.650	$\mathbf{8.1 \times 10^{-16}}$
$a \cdot s$ (orbit)	7.767	4.090	$\mathbf{3.548}$
s^3 (outside orbit)	4.169	2.273	0.926

Because c is shared by all three classes, the attribution contrast for x_3 is exactly zero under both s and $s + g$, numerically satisfying Proposition 2 at 10^{-16}. Under $a \cdot s$, it rises to 3.548: input-dependent multiplication can change contrasts as well as attribution levels (Remark 1). The cube transformation also changes the contrast (0.926), but does not reverse the ranking. These examples illustrate different behaviors; they do not show that only affine-orbit transformations can produce both a ranking reversal and a contrast change.

The $c = 0$ variant. With $c = 0$, x_3 is absent from s and its attribution is exactly 0. Under $a \cdot s$ it moves to 2.239, below x_1 (3.452) and x_2 (2.727) in this particular sample and parametrization. This does not make the common term necessary for reversal: Proposition 3 reverses the ranking with x_3 absent from all original scores and no additive common term.

Orbit membership of the cube. With $c = 0$, the three scores sum to zero. Vieta's identity gives

$$s_k^3 = \frac{s_0^2 + s_1^2 + s_2^2}{2}\, s_k + s_0 s_1 s_2.$$

The multiplier is positive away from the all-zero vector; at that vector any positive multiplier with zero shift works. Thus the cube lies inside the orbit in this variant. For $c = 0.5$ it is not generally in the orbit: at $x = (1, 1, 1)$ the three scores are distinct and sum to $3/2$. If all three satisfied $s_k^3 = As_k + B$, they would be the three distinct roots of $u^3 - Au - B$, whose roots sum to zero, a contradiction. Class-dependent expressions for a multiplier alone would not prove non-membership.

B Instrument details

Estimator scope and protocol differences. SHAP 0.48.0 selects ten features when `l1_reg` is unset. Shared settings need not give equal approximation error. Appendix C varies budget, regularization, and estimator seed on fixed models and 12 points each from Digits and Dry Bean. Digits same-model repeatability rises from 0.362 at the archived budget to 0.92–0.94 at 2048 coalitions. This does not establish exact values: the empirical results concern estimated attributions, while the theory concerns exact Shapley values.

Archived controls and the archived rank-stability record mix path-dependent TreeSHAP with interventional probability KernelSHAP and differ in output units for some families. These results are marked in the appendices and when cited in the main text. Main attribution comparisons use the uniform protocol; the residual analysis (§7) uses model outputs directly.

Family hyperparameters. `RF` and `ExtraTrees`: scikit-learn defaults; `HistGB`: 120 iterations; `XGB`: 100 trees; `LogReg`: L_2 regularization; `MLP`: one hidden layer of 64 units, with scikit-learn's default cap of 200 iterations (Appendix C refits every MLP to convergence). The three same-class pairs are `RF`×`ExtraTrees`, `HistGB`×`XGB`, and `LogReg`×`MLP`.

B.1 THE FOUR SIMILARITY MEASURES

Each explanation is an attribution tensor φ with dimensions (points $\times$ features $\times$ clusters). For each point, we select the output corresponding to its assigned cluster, following the own-cluster convention. We compare the resulting attributions using four measures:

measure	definition
per-instance	for each point, the cosine between the two attribution vectors of that point; the reported number is the **median over points**. The canonical measure: do the two explain *the same point* the same way?
magnitude	**Spearman** correlation between the two per-feature mean-$\lvert\varphi\rvert$ vectors; only the feature *ordering* matters
Pearson	**Pearson** correlation over the same two vectors; sensitive to the size of differences, not only the order
direction	cosine between the two *signed* mean attribution vectors: do the two push each feature the same way?

The archived calibration ladder and normalized comparisons are collected in Appendix L.

C ESTIMATOR VALIDATION AND INVARIANT-TARGET EVALUATION

We first vary the estimator budget and feature selection on small fixed samples, then check the full protocol and MLP convergence. Finally, we vary estimator seeds on all headline points (§C.2) and compare Wine estimates with exhaustive Shapley values (§C.3).

C.1 TARGETED SENSITIVITY AND IMPLEMENTATION CHECKS

This study examines how estimator settings affect the results on `digits`, the highest-dimensional dataset in Table 3, and `dry_bean`, our running example. It is separate from the archived results with three model-fitting seeds. For each dataset, we hold fixed the partition, all six fitted models (fit seed 42), one bootstrap refit per family (resampling seed 42), the first 12 shuffled observations, and the weighted 10-center background. The small sample and single fitted instance per family limit generalization. This study supplements the main experiment.

We evaluate every combination of requested coalition budget (100, 512, or 2048), feature selection (`l1_reg="num_features(10)"` or 0 for no selection), and estimator seed (42, 43, or 44). The fitted models remain unchanged across settings. Every comparison uses the output for the point's assigned cluster. We concatenate probability outputs and H in a single KernelSHAP call so that both targets use identical coalition samples. A small check that enumerates all coalitions confirms agreement with separate calls.

The new runs record SHAP 0.48.0, NumPy 1.26.4, scikit-learn 1.7.2, SciPy 1.11.4, and XGBoost 2.1.4. They also save data and script hashes, background weights, explained rows, fitted models, and full attribution tensors. These runtime records describe this study only; they do not establish the versions used in the older experiments. The MLP keeps the archived 200-iteration limit and issued convergence warnings on Dry Bean. We did not tune or refit it between estimator settings; the converged-MLP paragraph below addresses the main comparison.

Sensitivity on Digits. With ten-feature selection, increasing the budget from 100 to 2048 reduces the probability-attribution gap from 0.325 to 0.172. At the same time, median same-model similarity across estimator seeds rises from 0.362 to 0.920. Without selection, 100 coalitions give repeatability of only 0.040, so the larger gap in that setting is not more reliable evidence. At 2048 coalitions without selection, the gap is 0.150 (range 0.142–0.153 across estimator seeds) and repeatability is 0.941. The corresponding gap for H is 0.098. The positive gap therefore persists on these fixed observations, but its size and the estimated feature rankings depend on the estimator settings. This does not establish convergence of the main archived measurement.

Table 8: Estimator sensitivity on 12 fixed points with one fixed fit/refit per family. For baseline and cross-family cosine, we take medians over families or pairs within each estimator seed, then over seeds. The gap is the median of the differences within seeds. The H column reports the gap for centered-normalized scores. Repeatability is the median cosine similarity between probability attributions for the same fitted model across estimator-seed pairs. When few features vary, the requested budget can exceed the number of available coalitions.

dataset	budget	selection	baseline	cross	gap	H gap	repeatability
Digits	100	none	0.782	0.091	0.692	0.149	0.040
Digits	100	10 features	0.858	0.497	0.325	0.177	0.362
Digits	512	none	0.956	0.707	0.249	0.124	0.713
Digits	512	10 features	0.909	0.669	0.249	0.176	0.778
Digits	2048	none	0.958	0.810	0.150	0.098	0.941
Digits	2048	10 features	0.907	0.738	0.172	0.135	0.920
Dry Bean	100	none	0.989	0.657	0.332	0.274	0.898
Dry Bean	100	10 features	0.984	0.665	0.318	0.289	0.919
Dry Bean	512	none	0.990	0.697	0.292	0.298	0.992
Dry Bean	512	10 features	0.987	0.690	0.296	0.293	0.990
Dry Bean	2048	none	0.990	0.698	0.292	0.287	0.998
Dry Bean	2048	10 features	0.988	0.699	0.289	0.292	0.997

Dry Bean and the scope of the invariant target. With ten-feature selection, the gap falls from 0.318 at 100 coalitions to 0.289 at 2048, while same-model repeatability rises from 0.919 to 0.997. At 2048 coalitions without selection, the probability gap is 0.292 (range 0.291–0.301), and repeatability is 0.998. The gap for H is 0.287, a small change. With ten-feature selection, the H gap is slightly larger than the probability gap (0.292 versus 0.289). Explaining H therefore does not reduce disagreement in every empirical setting. Its exact invariance applies to transformations within an affine orbit, not to arbitrary trained surrogates.

The probability gap is positive within every estimator seed in all 36 dataset–budget–selection–seed combinations. These combinations reuse fixed models and points, so they are not 36 independent replications.

Full-protocol rerun at 2048 coalitions. Table 3 reports a rerun of the full protocol on all six datasets: three fit seeds, bootstrap refits, the first 150 observations, and shared coalition masks, with 2048 requested coalitions and no feature selection. The script `37_headline_2048.py` saves the attribution tensors for each model. The earlier 100-coalition experiment gave gaps 0.111 (Iris), 0.085 (Wine), 0.368 (Breast Cancer), 0.375 (Digits), 0.130 (Pendigits), 0.281 (Dry Bean), with median 0.205. Every gap is positive in both experiments. For datasets above $d = 4$, the magnitudes change with the budget (Wine $0.085 \to 0.064$, Pendigits $0.130 \to 0.078$, Breast Cancer $0.368 \to 0.185$, Digits $0.375 \to 0.149$, Dry Bean $0.281 \to 0.242$), as does the dataset ordering. Iris ($d = 4$) gives identical results because both budgets enumerate all 14 coalitions. The group contrast in §6.1 remains positive in $6/6$ datasets under both budgets.

Converged MLP. The main comparison keeps scikit-learn's default MLP cap of 200 iterations, which 20 of the 36 MLP fits and bootstrap refits reach before meeting the convergence tolerance: all six on Iris, Wine, and Dry Bean, and two on Pendigits. The script `38_mlp_convergence.py` refits every MLP with a cap of 5000 (all converge, using 133–558 iterations) and re-explains the 20 fits that change, with the same background, points, and coalition samples. A fit that stops before 200 iterations is identical under both caps and keeps its tensor, and the other five families are unchanged. Only the Iris gap moves ($0.111 \to 0.110$); the other five gaps are unchanged to six decimals, and the median gap goes from 0.1302 to 0.1295. A deterministic LogReg tensor recomputed for each dataset reproduces the archived arm exactly, except on Digits, where the k-means background summary is not bitwise reproducible across runs (maximum attribution difference 2×10^{-4}); no Digits MLP needed re-explanation. With converged MLPs, 2 of 134 features pass the rank-stability rule (Appendix D).

Implementation of the invariant target. We compute H in double precision. We first subtract one output coordinate from all coordinates, then center and normalize the vector. This is alge-

braically equivalent to centering the original outputs and maps exactly uniform floating-point vectors to zero. At every queried point, including coalition hybrids, we check the equality of $H(p)$ and $H((1 + 0.1x_1^2)p + \sin(x_2)\mathbf{1})$. This checks the implementation at the queried points; the proof of exact orbit invariance is in §4.3. Different fitted models need not belong to the same orbit, so their H attributions can differ. Normalization can also remain numerically sensitive near uniform outputs. The maximum observed pointwise invariance error was 1.7×10^{-15} or less in both datasets. This numerical check supplements the proof.

Denominator-guard sensitivity. For the 15 family pairs, we also explain the extracted multiplier using denominator thresholds of 10^{-4}, 10^{-3}, and 10^{-2}. Below each threshold, we replace the multiplier with 1. All thresholds and pairs use the same 512-coalition run, ten-feature selection, and estimator seed 42. We record how often the squared norm of the centered reference output falls below each threshold at synthetic hybrid points. We exclude background evaluations during initialization and singleton evaluations of explained points.

This frequency counts repeated queries without weighting. It is neither a probability under the data distribution nor a Shapley-weighted measure. We save the frequencies and full multiplier attributions so that behavior at hybrid points can be compared with the observed-point result in §7. Each dataset supplied 61,440 hybrid evaluations in this guard audit. At thresholds 10^{-4} and 10^{-3}, no reference model in the compared pairs activated the guard. At 10^{-2}, the maximum frequency over pairs was 0.0082% in Digits and zero in Dry Bean. Relative to the default threshold, the median feature-rank correlation of multiplier attributions was 1.000 in both datasets at all tested thresholds; the Spearman correlation of the 15 pairwise $\Phi[a]$ magnitudes was 0.996 in Digits at 10^{-2} and 1.000 otherwise. The guard has little effect on these measured queries. These checks do not show that it is inactive for every possible query.

The scripts `36_kernel_validation.py`, `37_headline_2048.py`, `38_mlp_convergence.py`, and `audit_validation.py` reproduce and audit these studies. From the saved attribution tensors, the auditor recomputes pair similarities, summary gaps, guard magnitudes, the full-protocol and converged-MLP summaries, and the rank-stability counts, and generates Table 8 and the rank-stability table of Appendix D. The optional `run_kernel_validation.py` entry point parallelizes HistGB prediction while preserving the fitted models and attribution protocol; it records the thread setting separately. These studies measure sensitivity on fixed models and points. Finite budgets and high repeatability do not certify exact Shapley values.

C.2 Estimator seeds under the full headline protocol

We repeated the headline probability-attribution experiment on all six datasets, using all 150 explained observations, the archived weighted 10-center background, and all three fit seeds. Each fit seed supplied six model families and one bootstrap refit per family. We reconstructed these models once from the recorded fitting recipe, saved them, and held them fixed while varying only the estimator seed (42, 43, 44). Every call requested 2048 coalitions with no feature selection. Within each estimator seed, models shared coalition samples. The original 200-iteration MLP cap was retained to match the headline setup; the convergence control above addresses this separate source of variation.

The experiment saved 648 attribution tensors: six datasets, three fit seeds, twelve fitted models per fit seed, and three estimator seeds. We also recomputed seed 42 and compared every attribution with the archived tensor, rather than assuming that reconstructing a model reproduced the original result. Saved models, design hashes, software versions, and tensor-level checks support this comparison. Table 3 retains the original headline results; this replication measures estimator variation conditional on the reconstructed, saved models. The maximum seed-42 attribution difference from the archive was 0.00144 on Digits and below 1.4×10^{-8} elsewhere. The largest change in a paired similarity gap was 1.7×10^{-5} or less. Reconstruction therefore does not reproduce every attribution bit for bit, but these differences have little effect on the reported gap.

For estimator seed e and fit seed r, let $G_{r,e} = B_{r,e} - U_{r,e}$ be the paired refit-minus-cross-family similarity gap. Table 9 reports $G_e = \mathrm{median}_r\, G_{r,e}$ and its range over the three estimator seeds. It separately reports the largest within-fit seed spread, $\max_r(\max_e G_{r,e} - \min_e G_{r,e})$. Repeatability

is the median same-model attribution cosine over the 36 fitted models and three estimator-seed pairs per dataset, including bootstrap refits. This isolates sampling variation conditional on the fitted models and background; the repeated fits, pairs, and seeds are not independent dataset replications.

All 54 paired gaps remain positive; the smallest is 0.0553. The largest within-fit seed spread is 0.0116, on Dry Bean, and each dataset's largest spread is smaller than its smallest measured gap. Median same-model repeatability ranges from 0.960 on Digits to 1.000 on Iris. These are dataset medians: the least-repeatable individual model/seed-pair comparison on Digits is 0.866. Thus, the cross-family gap persists under estimator-seed variation, while some individual attribution estimates remain sensitive to coalition sampling.

Table 9: Full-protocol estimator-seed replication. Gap is the median of G_e; the range varies estimator seed after taking paired gaps within fit seeds. Within-fit spread varies only the estimator seed at a fixed fitted-model set. Repeatability compares the same fitted function across estimator seeds.

Dataset	Gap	Seed range	Max. seed spread	Repeatability
Iris	0.111	[0.111, 0.111]	0.000	1.000
Wine	0.063	[0.062, 0.064]	0.003	1.000
Breast Cancer	0.185	[0.185, 0.187]	0.004	0.994
Digits	0.149	[0.147, 0.150]	0.009	0.960
Pendigits	0.078	[0.075, 0.079]	0.003	0.998
Dry Bean	0.245	[0.242, 0.245]	0.012	0.998

C.3 Exhaustive interventional Shapley reference on Wine

To distinguish repeatability from approximation accuracy, we independently enumerated all $2^{13} = 8192$ coalitions on Wine. This check used all 150 headline observations, all six families and their bootstrap refits at fit seed 42, and the same weighted background. For every coalition, we evaluated all ten background hybrids and took their weighted mean. We then applied the exact Shapley sum in (2), without KernelSHAP regression or coalition sampling. These are exhaustive attributions relative to the finite background, subject to floating-point arithmetic.

We checked the efficiency identity at every explained point and output. The enumeration routine also recovered the analytic ranking-reversal example and weighted-background linear attributions, and agreed with enumerated KernelSHAP on Iris to numerical precision. Table 10 compares the Wine reference with the 2048-coalition estimates on exactly these same models and rows. Agreement is the median, over the twelve models, of each model's median per-point cosine with its exhaustive attribution vector. The single dataset and fit seed make this a numerical anchor, not an exact-value certificate for the full six-dataset experiment.

The exhaustive refit baseline is 0.9742, cross-family similarity is 0.9111, and their paired gap is 0.0631. The sampled gaps range from 0.0619 to 0.0644, within 0.0013 of the exhaustive result. Median agreement with exhaustive attributions exceeds 0.9997 for every estimator seed; even the least-agreeing model has a median cosine above 0.9993. Mean absolute error over all points, features, models, and output coordinates is 0.0010, with a maximum of 0.0146 or less. Thus, the positive gap on this fixed Wine model set persists under exhaustive evaluation, although individual estimated attributions still have error.

Table 10: Wine: exhaustive attributions and sampled estimates on the same 150 points and fixed models. The last column compares estimated with exhaustive attributions for the same model.

Attributions	Refit baseline	Cross-family	Gap	Exact agreement
Exhaustive	0.974	0.911	0.063	—
Seed 42	0.974	0.910	0.064	0.9997
Seed 43	0.974	0.911	0.063	0.9997
Seed 44	0.975	0.913	0.062	0.9997

`39_headline_estimator_seeds.py` reconstructs and saves the fixed models and runs the seed comparison; `40_wine_exhaustive.py` computes the full Wine reference. Both save attri-

bution tensors. The separate `audit_estimator_seeds.py` verifies checkpoint completeness and model hashes, recomputes every similarity and gap from tensors, and generates both tables.

D CONTROL EXPERIMENTS

Table 11: Six alternative accounts of the divergence, each tested under control (§6.2). "Gap" is floor minus cross.

account	control	result
residual fidelity	capacity matched to a common cv target (0.97)	spread $0.0220 \rightarrow 0.0153$, cross *falls* ($0.6966 \rightarrow 0.6279$); gap > 0 in $12/12$ dataset×target combinations
same partition, different angles	pairs with *identical* predicted partition, every point, every seed	gap > 0 in $6/6$ datasets, median $+0.2361$ (on `dry_bean`: floor 0.9902 vs. cross 0.7103)
collinearity	remove correlated vs. random features; gap vs. d across datasets	gap tracks d ($\rho = +0.986$; budget-confounded, see below), not correlated fraction ($+0.143$); floor stays 0.95–0.99 at every d incl. 64
the explainer (SHAP)	replicate with LIME and permutation importance	gap > 0 in all 30 cells; SHAP reads 1.7–5.0× LIME, no method inverts sign
the algorithm (K-means)	K-medians, full GMM, spherical GMM as targets (ARI 0.70–0.85 apart)	gap > 0 in $9/9$; switching target moves the cross 0.0551, switching dataset 0.2502
the SHAP estimand	all six families under one uniform interventional KernelSHAP, floors recomputed	median gap shrinks $\approx 12\%$ but stays > 0 in $6/6$; the same-group regularity holds in $6/6$

D.1 MATCHED FIDELITY

We tune each family's capacity toward a common cross-validation accuracy. The 0.95 target fails the stability criterion: the MLP with hidden size 2 has degenerate attributions, and its refit baseline drops to 0.6539. At the 0.97 target, the refit baselines remain high:

condition	cv range	floor (bootstrap)	cross	magnitude
free	0.0220	0.9902	0.6966	0.3559
target 0.97	**0.0153**	0.9901	0.6279	0.2153

(Magnitude cells are medians of per-pair medians, the convention of Table 13; the observation-level aggregation gives 0.3382 and 0.1888.)

Closer matching does not increase cross-family similarity, and the gap remains positive in all 12 dataset×target combinations. The difference between the two targets does not support a finer conclusion: its sign changes with the order of aggregation ($+0.0039$ versus -0.0034), while the between-seed standard deviation is 0.0187–0.0488. Matching remains approximate, and reaching the targets deliberately reduces model capacity.

D.2 COLLINEARITY

Removing correlated features does not reduce the gap more consistently than removing the same number of features at random. Across datasets, $\rho(\text{gap}, d) = +0.986$ versus $\rho(\text{gap}, \text{correlated fraction}) = +0.143$; the apparent association with d also depends on the estimator budget. At 2048 coalitions (Table 3), the association falls to $+0.49$ (Digits: $0.375 \rightarrow 0.149$). It therefore does not separate dimension from approximation error.

Random feature removal changes d and is associated with the gap ratio ($\rho = -0.829$ between the removed fraction and the gap ratio). However, it also changes information content, partition

difficulty, and support. We did not run the experiment of adding pure noise features to isolate dimension. The floor stays at 0.95–0.99 at every d (including 64), and the relative rank-stability criterion of §6.3 admits wider ranges when the refit range grows (from 1.0 position on `pendigits` to 9.0 on `digits`), yet only 3 of 55 `digits` features pass the relative criterion and none pass after requiring a narrow floor. These observations do not isolate the effect of dimension on cosine similarity, particularly under the legacy mixed estimators.

D.3 Attribution method

We repeat the comparison with different attribution methods, each evaluated against its own refit baseline because estimator variances differ. SHAP and LIME provide both magnitude and per-instance comparisons. Permutation feature importance (PFI) provides only magnitude, with no signed or pointwise attribution. The gap is positive in all 30 comparisons. SHAP gaps are 1.7–$5.0\times$ larger than LIME gaps across the six datasets; both methods give the same sign. PFI has insufficient refit stability in 4 of 6 datasets (baseline 0.57–0.90 with 5 repetitions). We therefore use PFI only to count gap signs, not to compare gap sizes.

D.4 Clustering algorithm

We use K-medians, full-covariance GMM, and spherical GMM as alternative clustering targets in 9 dataset$\times$algorithm combinations. Agreement between targets is ARI 0.70–0.85, and the gap is positive in all 9 of 9 combinations. Within a dataset, changing the target changes cross-family similarity by a median amplitude of 0.0551 (0.0043 on `wine`). Changing the dataset gives an amplitude of 0.2502.

D.5 SHAP estimand

The older comparison uses path-dependent TreeSHAP for tree families and interventional KernelSHAP for `LogReg`/`MLP`. Eight of the fifteen pairs therefore compare different attribution targets. As a control, we explain all six families with the same interventional KernelSHAP protocol and recompute the refit baselines. The uniform arm below is the earlier, lower-budget experiment:

	mixed arm (legacy)		uniform arm	
dataset	gap	I − II	gap	I − II
`iris`	0.0912	+0.1313	0.1111	+0.1125
`wine`	0.1194	+0.1563	0.0851	+0.1161
`breast_cancer`	0.4188	+0.4490	0.3682	+0.4022
`digits`	0.5205	+0.5389	0.3752	+0.4833
`pendigits`	0.1813	+0.1674	0.1300	+0.1231
`dry_bean`	0.2844	+0.2560	0.2808	+0.2390

Using one attribution target reduces the median gap from 0.233 to 0.205 ($\approx 12\%$), but the gap remains positive in $6/6$ datasets. Within-group similarity also remains higher than across-group similarity (I $>$ II) in $6/6$. In the uniform arm, all families emit probabilities, so attribution magnitudes share a unit. The median magnitude gap is 0.2155 across datasets (on `dry_bean`: cross 0.4206 versus baseline 0.9794). At 2048 coalitions, the corresponding median is 0.149 (§6.3).

The mixed arm instead compares margins with probabilities. On `wine`, raw φ values reach 2.2 for boosting and 0.1 for forests. The unnormalized ratio of $18.7\times$ falls to $1.58\times$ after normalization at each point. The uniform comparison avoids these differences in output units.

D.6 Feature-level dispersion and greedy removal

We compare per-point normalized attribution shares $\tilde{\varphi}$. Both the cross-family comparison and the refit baseline use the same statistic: the standard deviation over six values.

dataset	comparable	above floor	median ratio	sign flips
dry_bean	16	**16/16**	6.58×	0
pendigits	16	16/16	4.00×	0
breast_cancer	30	28/30	3.24×	3
digits	55	55/55	2.92×	1
iris	4	3/4	2.64×	0
wine	13	12/13	1.58×	0
total	**134**	**130/134**		4

(digits has 64 features; nine coordinates with zero bootstrap dispersion are excluded from the denominator rather than counted as success or failure.)

We remove correlated features greedily. While any remaining pair has $|r| \geq 0.7$, we drop the member with the largest mean absolute correlation with the other features. On dry_bean, this removes 12 of 16 features. The table compares same-family similarity before and after removal with cross-family similarity, restricting both comparisons to the surviving features:

family	per-instance	magnitude†	Pearson†
HistGB	**0.3583**	**−0.40**	**−0.4212**
XGB	**0.3716**	**−1.00**	**−0.8864**
LogReg	0.8969	0.80	0.9713
MLP	0.9501	0.80	0.9782
RF	0.9507	0.80	0.9825
ExtraTrees	0.9840	1.00	0.9914
median	0.9235		
family swap (restricted)	**0.5330**		

† over the 4 surviving features; with $n = 4$ Spearman takes only 11 values, so the column supports reading the *sign*, not the value.

The aggregate comparison holds in 6/6 datasets (on digits, 0.9688 vs. 0.4312). On dry_bean, switching family changes attribution similarity more than removing features for four of the six families; the two exceptions are the boosting families (Figure 3).

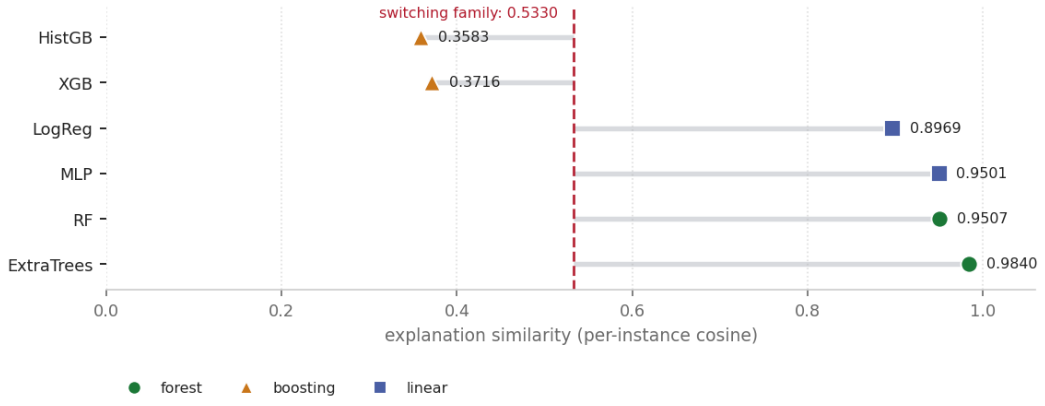

Figure 3: Removing 75% of the features versus switching family, per family. Each marker is the similarity of the same family before/after greedy removal; the dashed line is the cross-family similarity under the same restriction (0.5330). The two families left of the line are exactly the boosting class.

D.7 Rank-stability rule: counts and cutoff sensitivity

The rule is evaluated on the attribution tensors of Table 3 (2048 coalitions, no ℓ_1 selection). A feature's rank orders the mean absolute attribution over explained points on the assigned-cluster

output. The family range spans the 18 family×fit-seed fits; the refit range is the median, over families, of the range across each family's three fits and three bootstrap refits. Nine Digits pixels take a single value across the explained points and the background, so no model can attribute to them; they are not eligible, leaving 134 features. The archived row recounts the mixed-protocol record under the same rule.

refit-range cutoff	≤ 2.0	≤ 2.5	$\leq \mathbf{3.0}$	≤ 3.5	≤ 4.0
uniform, 2048 coalitions	3	3	**4**	4	5
archived mixed protocol	3	4	**5**	6	6

Under the uniform protocol, the count ranges from 3 to 5 as the cutoff varies from 2 to 4 positions. None pass in Iris, Breast Cancer, Digits, or Dry Bean. The Digits zero occurs at a budget where same-model repeatability on fixed fits is 0.94 (Appendix C). Four features pass at the chosen cutoff: `alcalinity_of_ash`, `nonflavanoid_phenols`, and `od280/od315_of_diluted_wines` (`wine`); `input15` (`pendigits`). With every MLP refit to convergence (Appendix C), only `od280/od315_of_diluted_wines` and `input15` pass. The five that pass in the archived record — `sepal length` (`iris`); `color_intensity`, `alcohol`, `magnesium` (`wine`); `mean concavity` (`breast_cancer`) — include none of them: the count is similar across protocols, but which features appear stable is not. The feature `worst fractal dimension` (`breast_cancer`) illustrates why an absolute cutoff is needed. Its rank varies by 10 positions across families and 10 under refitting. It passes the relative criterion, but its refit variation is too large to count as stable. All 16 features fail on `dry_bean`. For example, `Eccentricity` spans ranks 2–16, a family range of 14.0 compared with a refit range of 1.0 (Figure 4 shows the archived record).

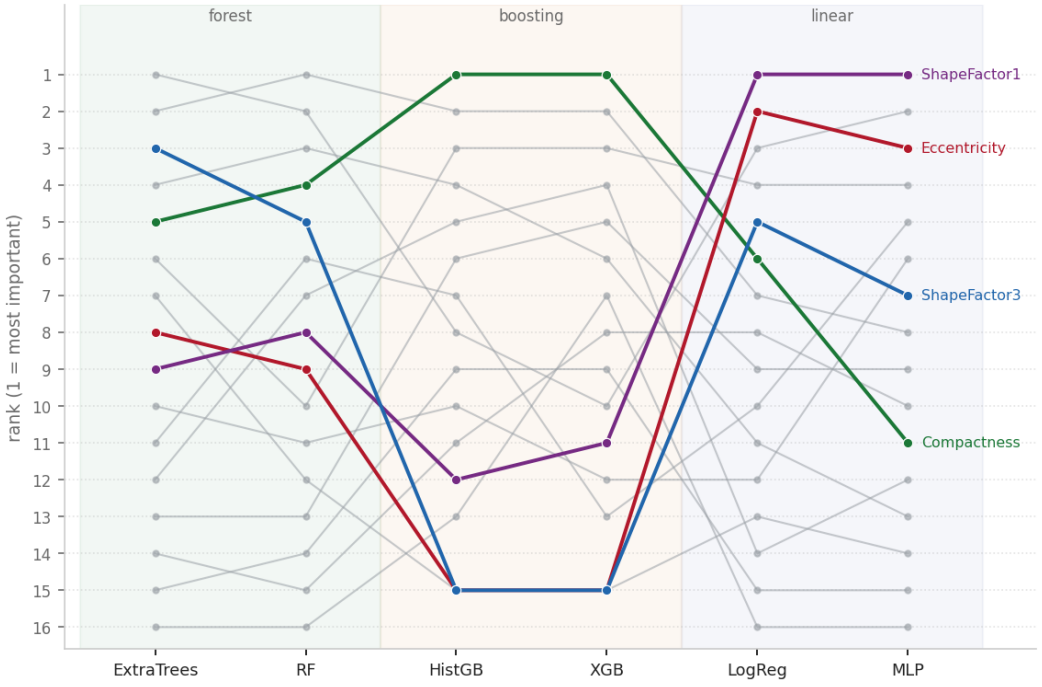

Figure 4: Archived mixed-protocol record. Rank of each of the 16 `dry_bean` features in each family, columns grouped by functional class. Within a block the lines run nearly flat — same-class families agree on the order; between blocks they cross. `Eccentricity` is ranked 8 by ExtraTrees and 9 by RF, falls to 15 in boosting (which attributes exactly zero), and rises to 2 in the linear models.

E RESIDUAL AND MULTIPLIER DIAGNOSTICS: SUPPLEMENTARY RESULTS

Lemma 4 (Multiplier extraction). *Let $f_c(x) = f(x) - \bar{f}(x)\mathbf{1}$, with $\bar{f}(x) = K^{-1}\sum_k f_k(x)$, be the class-centered output, and likewise g_c. Then (i) centering absorbs b exactly: if $g = af + b\mathbf{1}$ then $g_c = af_c$; (ii) wherever $\langle f_c, f_c\rangle > 0$, the multiplier is identified in closed form by*

$$a(x) \;=\; \langle g_c(x), f_c(x)\rangle \,/\, \langle f_c(x), f_c(x)\rangle \tag{4}$$

— the pointwise least-squares fit of g_c on f_c; (iii) where both centered vectors are nonzero, the relative squared fit residual is $r(x) = \|g_c - af_c\|_2^2/\|g_c\|_2^2 = 1-\cos^2(f_c, g_c)$. It is zero exactly when those vectors are collinear; positive orbit membership additionally requires $a > 0$. If both centered vectors vanish, the maps are related by a common shift at that point; if exactly one vanishes, no positive affine transformation relates them there.

Proposition 5 (Binary degeneracy). *For $K = 2$ write $f = (p, 1-p)$ and $g = (q, 1-q)$. Wherever both centered outputs are nonzero they are parallel to $(1, -1)$, so the residual $r(x)$ is identically zero and carries no information; the unguarded coefficient is $\hat{a}(x) = (q(x) - \frac{1}{2})/(p(x) - \frac{1}{2})$, which need not lie in $\{\pm 1\}$ and is undefined at $p = \frac{1}{2}$. If $p \neq \frac{1}{2}$ and $q = \frac{1}{2}$, the coefficient is zero, the normalized residual is undefined, and no positive affine map relates the outputs at that point. Binary problems make the* residual *uninformative — not the multiplier or its feature dependence.*

Confidence dependence of the residual. In the archived five-seed run, the median residual is ≈ 0 for all 15 pairs in every dataset, including `digits`, with 9 centered-output dimensions. This uses model outputs directly. Confident models predicting the same class have centered outputs $\approx e_j - K^{-1}\mathbf{1}$ for the same j, where e_j is the vector with one in coordinate j and zero elsewhere, making them nearly parallel. Separating points by confidence changes the residual by one to two orders of magnitude (table below). Joint confidence can therefore explain small residuals without implying similar attributions.

Weak association of multiplier variability. The coefficient of variation, $\mathrm{cv}(a) = \mathrm{sd}(a)/|\bar{a}|$ (standard deviation divided by absolute mean over points), has a weak and inconsistent association with the uniform-protocol similarity: median $\rho = -0.050$ across the six datasets, with per-dataset values from -0.429 to $+0.150$ (Table 4; the archived mixed-outcome analysis below shows a similar pattern).

Numerical guard and positivity. To avoid division by a small denominator, the implementation sets a to 1, the orbit identity, wherever $\langle f_c, f_c\rangle \leq 10^{-3}$. This guard was never activated on the explained observations. Appendix C separately measures its activation at coalition points for fixed fits.

A small residual does not by itself establish positive orbit membership. At observed *data* points, $\hat{a}(x) \leq 0$ never occurs in five of six datasets and occurs at $\leq 1\%$ of points in one `iris` pair. At *coalition* points, where interventional SHAP evaluates the models, the unconstrained coefficient is non-positive at a median 0.3–4.8% of points per dataset. The reconstruction in §7.2 therefore constrains the multiplier to be positive.

Residual by confidence stratum (median orbit residual r):

dataset	r, confidence > 0.95	r, medium confidence
`wine`	0.00003	0.00209 (70×)
`digits`	0.00018	0.00435 (24×)
`dry_bean`	0.00000	0.00916

$\mathrm{cv}(a)$ **against similarity** (prediction: negative):

dataset	$\rho(\mathrm{cv}(a), \mathrm{sim})$	$\mathrm{cv}(a)$, class I	class II	separates?
`iris`	$+0.225$	0.069	0.163	yes
`wine`	$\mathbf{-0.421}$	0.047	0.100	yes
`digits`	$+0.082$	0.035	0.050	yes
`pendigits`	$+0.096$	0.074	0.064	**no**
`dry_bean`	$+0.171$	0.057	0.049	**no**
median	$\mathbf{+0.096}$			3 of 5

This archived table uses the legacy similarity outcome. Table 4 instead joins multiplier diagnostics to uniform probability-space similarities by pair and fit seed (three seeds, 100 points), obtaining median $\rho = -0.050$ across the six multiclass datasets. Both analyses give weak, inconsistent associations; their outcomes and dataset coverage differ.

$\Phi[a]$ **vs.** $\mathrm{cv}(a)$**:** $\rho(\Phi[a], \mathrm{cv}(a))$ per dataset: `iris` $+0.054$, `wine` -0.036, `pendigits` -0.179, `dry_bean` $+0.089$, `digits` -0.236; median -0.036.

F Orbit interpolation and orbit projection

F.1 Controlled interpolation (dose–response)

We use the same sample, background, and seed as the Gaussian illustration in Appendix A ($X \sim \mathcal{N}(0, I_3)$, $n = 400$, $c = 0.5$). The multiplier is $a_\lambda(x) = 1 + \lambda\, 5x_j^2$. All interventional Shapley values, including those used in $\Phi[a_\lambda]$, are exact relative to this background. Both arms preserve the partition of s at every tested λ, as checked pointwise. In the table, "sim" is the median per-instance cosine for the output of each point's own class. "Rank" is the Spearman correlation between the mean-$|\varphi|$ vectors of s and $a_\lambda \cdot s$. With $d = 3$, this correlation takes few values and indicates whether the ranking changes rather than how much it changes.

arm	λ	$\Phi[a_\lambda]$	$\mathrm{cv}(a_\lambda)$	sim	rank	ranking
x_3	0	0.000	0.000	1.0000	$+1.0$	$x_1 > x_2 > x_3$
x_3	0.05	0.081	0.259	0.9980	$+0.5$	$x_1 > x_3 > x_2$
x_3	0.25	0.404	0.743	0.9851	-0.5	$\mathbf{x_3 > x_1 > x_2}$
x_3	1	1.615	1.142	0.9688	-0.5	$\mathbf{x_3 > x_1 > x_2}$
x_3	2	3.230	1.254	0.9633	-0.5	$\mathbf{x_3 > x_1 > x_2}$
x_1	0	0.000	0.000	1.0000	$+1.0$	$x_1 > x_2 > x_3$
x_1	0.05	0.094	0.289	0.9971	$+1.0$	$x_1 > x_2 > x_3$
x_1	0.25	0.468	0.785	0.9808	$+1.0$	$x_1 > x_2 > x_3$
x_1	1	1.871	1.160	0.9629	$+1.0$	$x_1 > x_2 > x_3$
x_1	2	3.743	1.260	0.9592	$+1.0$	$x_1 > x_2 > x_3$

In both arms, $\rho(\Phi[a_\lambda], \mathrm{sim}) = -1.000$ over the six nonzero λ values. The grid includes $\lambda = 0.1$ and 0.5, omitted from the table but included in the correlation. At $\lambda = 0$, the negative-control divergence is 1.1×10^{-12}.

As discussed in Appendix I, this correlation is not independent validation of the diagnostic. Within each arm, $\Phi[a_\lambda]/\lambda$ is constant: 1.615 for x_3 and 1.871 for x_1. Exact linearity gives $\Phi[a_\lambda] = \lambda\, \Phi[a_1]$, so the correlation of -1.000 expresses monotonicity with the dose. Across arms, isotropy gives identical cv in distribution; sampled values differ by at most 1.6% at $\lambda = 1$. The ranking reversal occurs only in the x_3 arm, starting at $\lambda = 0.25$. Per-instance divergence tracks Φ across arms: sampling gives the x_1 arm $\approx 16\%$ larger Φ, with larger divergence at every matched λ (divergence ratios $x_3/x_1 = 0.68$–0.90).

F.2 Orbit projection: protocol and detail

We compare observed similarity $\mathrm{sim}(f, g)$, within-orbit similarity $\mathrm{sim}(f, \tilde{g})$, and reconstruction similarity $\mathrm{sim}(\tilde{g}, g)$. With $\Theta(f, g) = \arccos(\mathrm{sim}(f, g))$, the ratio $\Theta(f, \tilde{g})/\Theta(f, g)$ is a descriptive normalized angular displacement. Its median is 53%, with a range of 48–65% across datasets. This

is neither an additive nor a causal share of disagreement. Swapping the reference gives aggregate medians 0.53 and 0.54, while individual pairs depend on the reference (Appendix F.3).

For each across-group pair (f, g), we construct $\tilde{g} = \hat{a}_+ f_c + \bar{g}\mathbf{1}$ with $\hat{a}$ from (4) under the denominator guard of Appendix E ($\hat{a} = 1$ where $\langle f_c, f_c \rangle \leq 10^{-3}$) and $\hat{a}_+ = \max(\hat{a}, 10^{-6})$, so $\tilde{g} \in \mathcal{O}(f)$ at every evaluated point. If the guard is inactive, a zero residual and $\hat{a} \geq 10^{-6}$ give $\tilde{g} = g$. Positivity alone is insufficient because clipping can still change the coefficient. Where the guard is active, setting $\hat{a} = 1$ is a numerical convention. All three explanations (f, g, $\tilde{g}$) use the same uniform interventional KernelSHAP protocol (`nsamples` $= 512$, 100 points, 3 fit seeds, shared background). Bootstrap refits are explained under this same protocol to provide the baselines. For every pair and seed, we save per-point cosines, joint confidence $\min(\max_k f_k, \max_k g_k)$, and residuals $r(x)$. These records allow the confidence-stratified analysis below. High confidence means joint confidence above 0.95; medium confidence means at most 0.95. We computed a stratum's ratio only when it contained at least five points. The medium/high-confidence summary divides the medium-confidence normalized displacement by the high-confidence normalized displacement, both on the $1 - \cos$ scale.

Table 12 reports displacement ratios on both scales, using medians over across-group pairs. It also compares reconstruction similarity with the refit baseline and reports confidence-stratified ratios and pair-level correlations.

Table 12: Projection of real cross-class pairs onto the orbit, via $\tilde{g} = \hat{a}_+ f_c + \bar{g}\mathbf{1}$ (medians over the 12 cross-class pairs, 3 seeds, uniform interventional KernelSHAP). Ratios compare orbit-induced with total displacement on the indicated scale; they are not additive fractions. The stratified ratios use $1 - \cos$, and their persistence constrains a saturation account without ruling out confidence effects.

	ratio		recon		ratio by stratum		pair-level ρ (15 pairs)		
dataset	angle	$1-\cos$	$\mathrm{sim}(\tilde{g}, g)$	floor	high	med	(orb, tot)	(Φ, tot)	(Φ, ratio)
`iris`	0.53	0.28	0.970	0.994	0.28	0.27	$+0.69$	$+0.42$	-0.62
`wine`	0.53	0.29	0.969	0.972	0.28	0.39	$+0.85$	$+0.66$	-0.61
`digits`	0.65	0.45	0.881	0.908	0.45	0.60	$+0.84$	$+0.40$	-0.34
`pendigits`	0.49	0.25	0.952	0.989	0.25	0.40	$+0.58$	$+0.58$	-0.29
`dry_bean`	0.48	0.24	0.880	0.984	0.24	0.21	$+0.87$	$+0.78$	$+0.20$
`news_emb.`	0.57	0.33	0.993	0.996	0.33	0.58	$+0.85$	$+0.82$	-0.16
median	**0.53**	0.29					$+0.85$	$+0.62$	-0.32

The $1 - \cos$ and angular ratios use different nonlinear scales and need not agree. The angular ratio uses the arccosine of a median cosine; separately aggregated angles need not satisfy a triangle inequality. The ratio can exceed one and is not a geodesic decomposition. Neither ratio is an additive variance allocation or a causal contribution. The medium/high-confidence ratio on the $1 - \cos$ scale ranges from 0.86 to 1.77 and exceeds one in four of six datasets. This persistence weakens a simple saturation account; confidence at an explained point does not determine behavior at its hybrids. The pointwise reconstruction is exact for matching one-hot outputs, not for arbitrary one-hot outputs with different winning classes.

$\rho(\mathrm{orb}, \mathrm{tot})$ compares orbit-induced and total $1-\cos$ displacement across pairs. The positive $\rho(\Phi, \mathrm{tot})$ values correspond to the negative associations with similarity in Table 4. Associations with the ratio itself have a different interpretation and are negative in five of six datasets. Reconstruction similarity is below the bootstrap baseline by 0.003–0.104. On `dry_bean`, some `ExtraTrees` pairings have reconstruction-to-total $1 - \cos$ ratios of 0.4–0.5. These discrepancies preclude treating the reconstruction as an exhaustive account of the observed disagreement.

The constant-multiplier null is exact. Probability outputs have class mean $1/K$. Replacing $\hat{a}(x)$ with a constant $\bar{a}$ therefore reduces the reconstruction to $\tilde{g}_{\text{const}} = \bar{a} f_c + \frac{1}{K}\mathbf{1} = \bar{a} f + \frac{1-\bar{a}}{K}\mathbf{1}$: a positive rescaling plus a constant shift of f. By linearity of the Shapley value, $\varphi[\tilde{g}_{\text{const}}] = \bar{a}\,\varphi[f]$ exactly, because the constant contributes zero. Wherever the attribution vectors are nonzero, their cosine is one and the angular displacement is zero. The implementation adds a 10^{-12} cosine regularizer. We checked this numerically with script 35, setting $\bar{a}$ to the mean of $\hat{a}$ over the explained points. Similarity was $\mathrm{sim}(f, \tilde{g}_{\text{const}}) = 1.000$ to machine precision for all 15 pairs $\times$ 3 seeds of `iris` and `wine`. We did not run the remaining datasets after this check. Constant multipliers thus provide an exact

zero-displacement control. The measured 48–65% angular ratios use input-dependent multipliers, but this control does not assign a causal fraction of the disagreement to input dependence.

The saved per-point cosines also measure total cross-family divergence within each confidence stratum. Pooling points over all 15 family pairs and three fit seeds, the ratio of median $1 - \cos(f, g)$ at medium confidence to its median at high confidence ranges from 0.69 to 1.59. It is below one in Iris and Dry Bean and above one in the other four datasets. This is a ratio of *raw total divergences*, distinct from the 0.86–1.77 ratio above and in Figure 5C: that comparison uses *orbit-to-total displacement ratios*, aggregated over across-group pairs within each stratum. The raw comparison shows no consistent concentration of disagreement at medium confidence.

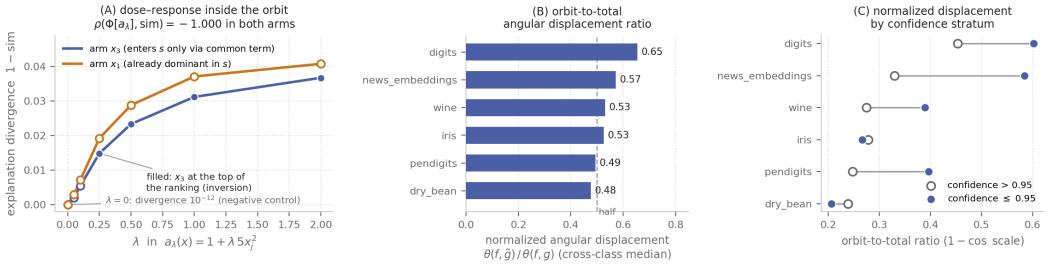

Figure 5: The two closing experiments. **(A)** Dose–response inside the orbit: sweeping only the feature dependence of the multiplier, divergence rises monotonically in both arms ($\rho = -1.000$), the negative control at $\lambda = 0$ is exact, and the ranking inversion in this sampled Gaussian illustration (filled markers) occurs only in the x_3 arm. This is distinct from the analytic construction of Proposition 3. **(B)** The ratio of orbit-induced to total angular displacement ranges from 48–65% (median 53%) across the six datasets. **(C)** The confidence-stratified comparison: the normalized $1-\cos$ displacement at medium confidence (filled) against high confidence (open), per dataset. The medium/high ratio ranges from 0.86 to 1.77 and exceeds one in four of six datasets.

F.3 The reference swap

Equation (4) projects g onto the orbit of f. By default, we use the alphabetically first family as f. To assess this choice, we rerun the multiplier attribution and constrained projection with the model roles swapped in all six datasets, using the same $\hat{a}_+$ construction in both directions. The angular ratios below use the uniform projection experiment. The Φ–similarity correlations use the earlier mixed-estimator outcome and therefore differ from those in Table 4.

	angular ratio					
dataset	orig	swap	$q + q'$	$\rho(q, q')$	$\rho(\Phi[a'], \text{sim})$ [orig]	II/I (swap)
iris	0.53	0.54	1.06 ± 0.06	-0.81	-0.16 [-0.51]	2.08
wine	0.53	0.51	1.12 ± 0.08	-0.75	-0.29 [-0.64]	1.71
digits	0.65	0.65	1.38 ± 0.13	-0.50	-0.59 [-0.54]	1.65
pendigits	0.49	0.57	1.14 ± 0.11	-0.68	-0.63 [-0.46]	1.76
dry_bean	0.48	0.52	0.98 ± 0.10	-0.42	-0.75 [-0.77]	1.93
news_emb.	0.57	0.54	1.24 ± 0.17	$+0.13$	-0.26 [-0.78]	1.57
median	0.53	0.54			-0.44 [-0.59]	

The aggregate angular ratio is similar in the two directions (medians 0.53 and 0.54; largest dataset difference 0.08). Pair-level ratios vary and are negatively associated across directions in five of six datasets; their near-constant sums are empirical observations, not a complementarity identity. For nonzero centered vectors, the unguarded coefficients satisfy $\hat{a}_{f\to g}\hat{a}_{g\to f} = \cos^2(f_c, g_c)$, and are reciprocal only at exact collinearity. Clipping and denominator guards add exceptions. Shapley values have no general reciprocal chain rule, so this identity does not determine the ordering of $\Phi[a]$ and $\Phi[a']$ across pairs. Under the legacy outcome, the correlations with similarity remain negative in all six datasets in both directions, but their strength changes with the reference (median -0.59 versus -0.44).

G THE DIRECT EXPLANATION OF THE CLUSTERING

For K-means, the affinity $a_k(x) = -\sum_j (x_j - \mu_{kj})^2$ is additive over features, so its Shapley values can be computed exactly:

$$\varphi_j^k \;=\; \underbrace{2\mu_{kj}\,(x_j - \mathbb{E}[X_j])}_{\text{discriminative}} \;-\; \underbrace{(x_j^2 - \mathbb{E}[X_j^2])}_{\text{common to all } k},$$

The efficiency axiom holds to 10^{-14} in all six datasets, without an estimator. The common term cancels in the $\arg\max$, so it cannot be recovered from the labels. We remove it for the comparison.

The reference distribution is the 15 surrogate×surrogate similarities recomputed in this experiment (range on `dry_bean`: 0.4333–0.9328). After removing the common term, the absolute difference between the median direct-to-surrogate similarity and the median surrogate-to-surrogate similarity is 0.0147 (median over six datasets). On `dry_bean`, the surrogate-pair median is 0.7058, the direct-to-surrogate median is 0.7370 before the quotient and 0.7694 after it. These are descriptive comparisons. The 15 pairs share families, so they do not supply 15 independent replications for a standard error of the median. Proximity to the pairwise distribution neither establishes equivalence nor selects a canonical attribution target.

Direct explanations are most often closest to the tree group. Across three additive, well-conditioned targets (K-means, K-medians, spherical GMM), trees rank closest in 15 of 18 target×dataset pairs. The recurring exception is `iris` ($d = 4$). For individual families, the pattern is weaker and depends on whether the common term is removed: `ExtraTrees` is closest in 5 of 6 datasets after removal and 3 of 6 with raw attributions. This does not identify a uniquely correct surrogate. The surrogate explains a score for *membership*, while the direct explanation explains *affinity*; neither target is canonical.

H $\Phi[a]$: SEPARATION AND FLOOR TABLES

Coordinated feature permutations leave the scalar $\Phi[a]$ unchanged. Feature identities are retained in the vector $\bar{\varphi}[a]$ of Definition 3, but not in its scalar average.

dataset	$\Phi[a]$, class I	class II	rank positions of the 3 class-I pairs
`iris`	0.0547	0.1091	$[1, 3, 5]$
`wine`	0.0170	0.0327	$[1, 2, 5]$
`pendigits`	0.0122	0.0257	$[\mathbf{1}, \mathbf{2}, \mathbf{3}]$
`dry_bean`	0.0144	0.0245	$[\mathbf{1}, \mathbf{2}, \mathbf{3}]$
`digits`	0.0028	0.0053	$[1, 2, 4]$
`news_embeddings`	0.0059	0.0087	$[\mathbf{1}, \mathbf{2}, \mathbf{3}]$

dataset	floor	class I / floor	class II / floor	ratio II/I	pairs above floor
`iris`	0.041	1.33×	2.66×	2.00	15 of 15
`wine`	0.019	**0.88**×	1.70×	1.93	12 of 15
`pendigits`	0.009	1.39×	2.93×	2.11	14 of 15
`dry_bean`	0.008	1.80×	3.06×	1.70	15 of 15
`digits`	0.003	**0.92**×	1.72×	1.87	13 of 15
`news_embeddings`	0.005	1.27×	1.88×	1.48	14 of 15

I THE DIRECTIONAL DIAGNOSTIC AND ITS NULL

The archived interpolation illustrates how attributions can change within the orbit (Appendix F). By linearity, $\Phi[a_\lambda]$ is proportional to the interpolation dose, so its monotone association does not independently validate it as a predictor. We also test a directional refinement: the component of $\bar{\varphi}[a]$ perpendicular to the surrogate's mean absolute attribution vector. Its median correlation with uniform-protocol similarity is $\rho = -0.559$, weaker than that of $\Phi[a]$ (-0.623). It improves the

association in only two of six datasets. The tables below distinguish this reanalysis from the earlier mixed-protocol result.

Uniform-protocol outcome. Joining the archived diagnostic to the uniform probability-space similarities by family pair and seed gives the following correlations. Pair medians precede the Spearman correlation over the 15 pairs; the directional magnitude improves the association in two of six datasets.

dataset	$\rho(\Phi[a])$	$\rho(\varphi[a]^{\perp})$
Iris	-0.421	-0.400
Wine	-0.664	-0.721
Digits	-0.400	-0.511
Pendigits	-0.582	-0.386
Dry Bean	-0.779	-0.693
News embeddings	-0.821	-0.607
Median	-0.623	-0.559

Legacy-outcome diagnostics. The remaining tables in this appendix use the archived mixed-estimator similarity outcome and are retained as exploratory results. Their four-of-six improvement is not the result of the uniform reanalysis above.

Here $\bar{\varphi}[f]$ averages absolute probability-space attributions over points and outputs (unlike the own-cluster similarity convention). With $\cos\theta = \langle\bar{\varphi}[a], \bar{\varphi}[f]\rangle/(\|\bar{\varphi}[a]\|_2\|\bar{\varphi}[f]\|_2)$, $\varphi[a]^{\perp} = \|\bar{\varphi}[a]\|_2 \sin\theta$, and concentration $= \|\bar{\varphi}[a]\|_2/\|\bar{\varphi}[a]\|_1$:

dataset	d	$\rho(\Phi[a])$	$\rho(\varphi[a]^{\perp})$	$\rho(\sin\theta)$	ρ(conc.)
`iris`	4	$\mathbf{-0.514}$	-0.329	$+0.196$	$+0.300$
`wine`	13	-0.636	$\mathbf{-0.696}$	-0.246	-0.157
`pendigits`	16	-0.461	$\mathbf{-0.675}$	-0.611	-0.371
`dry_bean`	16	$\mathbf{-0.768}$	-0.611	$+0.018$	$+0.754$
`news_emb.`	32	-0.782	$\mathbf{-0.793}$	$+0.136$	$+0.043$
`digits`	64	-0.536	$\mathbf{-0.682}$	-0.868	-0.089
median		-0.586	$\mathbf{-0.679}$	-0.114	-0.023
negative in		$6/6$	$\mathbf{6/6}$	$3/6$	$3/6$

(The factor-decomposition and permutation-null analyses below predate the embedding dataset and cover the original five.)

In the three datasets where $\sin\theta$ correlates negatively (`wine`, `pendigits`, `digits`), $\varphi[a]^{\perp}$ has a stronger association than $\Phi[a]$. It has a weaker association in `iris` ($+0.196$ for $\sin\theta$) and `dry_bean` ($+0.018$). The exception is `news_embeddings` ($\sin\theta = +0.136$), where the refinement improves the association by 0.011. This correspondence holds in 5 of 6 datasets: the 4–2 comparison and the 3-of-6 result for $\sin\theta$ reflect largely overlapping patterns. On the original five datasets, $\|\bar{\varphi}[a]\|_2$ alone has median correlation -0.525, negative in $5/5$. The directional refinement also includes $\sin\theta$, which has the hypothesized negative association in $3/6$ datasets; concentration reaches $+0.754$ on `dry_bean`, opposite to that hypothesized direction.

Both $\bar{\varphi}[a]$ and $\bar{\varphi}[f]$ are nonnegative, so $\cos\theta \geq 0$. We assess their alignment against a permutation null that shuffles feature indices of one vector, preserving its sparsity profile and dimension while removing feature correspondence:

dataset	d	$\cos\theta$ obs.	null (median)	null (p95)	excess
`iris`	4	0.970	0.951	$\mathbf{1.000}$	$\mathbf{+0.019}$
`wine`	13	0.952	0.771	0.913	$+0.181$
`pendigits`	16	0.943	0.862	0.938	$+0.081$
`dry_bean`	16	0.912	0.765	0.892	$+0.147$
`digits`	64	0.873	0.517	0.733	$\mathbf{+0.357}$

The permutation null changes the interpretation of the raw cosine. As d increases, raw $\cos\theta$ falls from $0.970 \to 0.873$, which might suggest less alignment. However, the null median falls faster, from $0.951 \to 0.517$. The excess above the null is smallest on `iris` ($+0.019$, within the null's p95 $= 1.000$) and largest on `digits` ($+0.357$). The intermediate datasets show no monotone pattern ($d \in \{4, 13, 16, 16, 64\}$: four distinct dimensions with one tie). Relative to chance, the multiplier is more aligned with $\varphi[f]$ in the highest-dimensional dataset. The reason remains unclear.

The fixed Digits fits in Appendix C rule out one candidate explanation for those fits. Recomputing both attributions without ℓ_1 selection (12 points, 512 coalitions, shared estimator seed) raises the median pairwise $\cos(\bar{\varphi}[a], \bar{\varphi}[f])$ from 0.73 to 0.92. It also scales the fifteen $\Phi[a]$ magnitudes by $\approx 2.8\times$ while largely preserving their ordering (Spearman 0.93). Thus, the ten-feature default reduces the measured alignment. A common support induced by feature selection does not explain the excess in this check.

J PAIRWISE TABLE AND FAMILY MATRIX (DRY_BEAN)

Table 13: `dry_bean`: medians across the 15 cross-family pairs versus the seed and bootstrap refit-stability baselines ("floors"), for the four similarity measures. Summary rows are medians of the per-pair medians shown in the table below; the observation-level aggregation (over pair × seed) differs for two cells — magnitude 0.3382, direction 0.8086 — and the aggregation order is declared wherever either number is used.

	per-instance (cosine)	magnitude (Spearman)	Pearson	direction (signed cos.)
seed floor	0.9879	0.9587	0.9802	0.9953
bootstrap floor	0.9902	0.9735	0.9847	0.9965
cross-family	**0.6966**	**0.3559**	**0.3146**	**0.8018**

pair	per-instance	magnitude	Pearson	direction
HistGB×MLP	**0.3942**	−0.0855	−0.2018	0.6319
MLP×XGB	0.4655	−0.2773	−0.3305	0.6839
HistGB×LogReg	0.4761	−0.0147	−0.0264	0.7045
LogReg×XGB	0.5182	−0.2124	−0.1941	0.7590
ExtraTrees×XGB	0.6374	0.0737	0.0275	0.7823
MLP×RF	0.6386	0.3559	0.2292	0.8419
ExtraTrees×HistGB	0.6907	0.3687	0.3760	0.7685
LogReg×RF	0.6966	0.4853	0.3146	0.8801
ExtraTrees×MLP	0.7068	0.3382	0.4291	0.9213
RF×XGB	0.7103	0.1091	0.2162	0.7886
ExtraTrees×LogReg	0.7686	0.5176	0.5570	0.9561
HistGB×RF	0.7813	0.3658	0.4786	0.8018
HistGB×XGB	0.8571	0.8432	0.8762	0.9597
ExtraTrees×RF	0.9265	0.8588	0.7933	0.9516
LogReg×MLP	**0.9310**	0.8618	0.9039	0.9842

Within-group pairs (LogReg×MLP, ExtraTrees×RF, HistGB×XGB) occupy the bottom three rows. Four across-group pairs—HistGB×MLP, MLP×XGB, HistGB×LogReg, LogReg×XGB—compare boosting with linear/network models and have negative correlations in both magnitude measures. These pairs also cross margin↔probability outputs and TreeSHAP↔KernelSHAP estimators. Under the uniform attribution target (Appendix D), their rank correlations range from -0.07 to $+0.08$. They therefore show little agreement in ranking, without a systematic reversal. Figure 6 plots the baseline and cross-family distributions; Figure 7 orders the pairwise matrix by functional group.

Uniform-protocol group comparison. For the analyst-defined forest, boosting, and linear/network groups, the within-minus-across-group contrast is 0.072–0.230 under the headline protocol and positive in all six datasets (one-sided sign test across datasets, $p = 0.016$). The grouping,

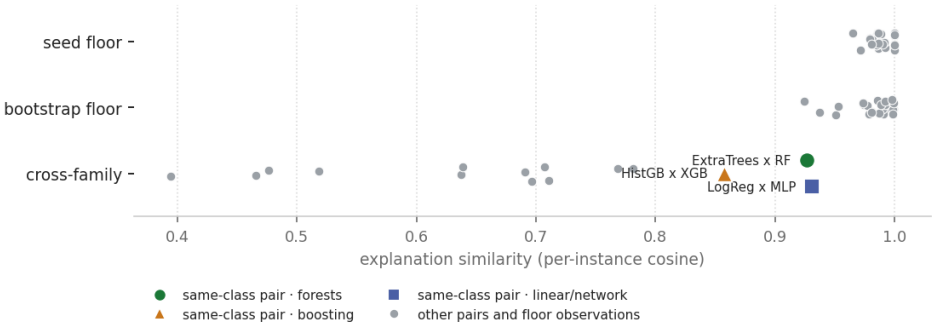

Figure 6: The earlier mixed-estimator result on `dry_bean`. Gray points represent cross-family pairs or refit-baseline observations. The three within-group pairs are colored and labeled. The baseline and cross-family distributions overlap little; the overlap comes from within-group pairs, consistent with the pattern in §6.1.

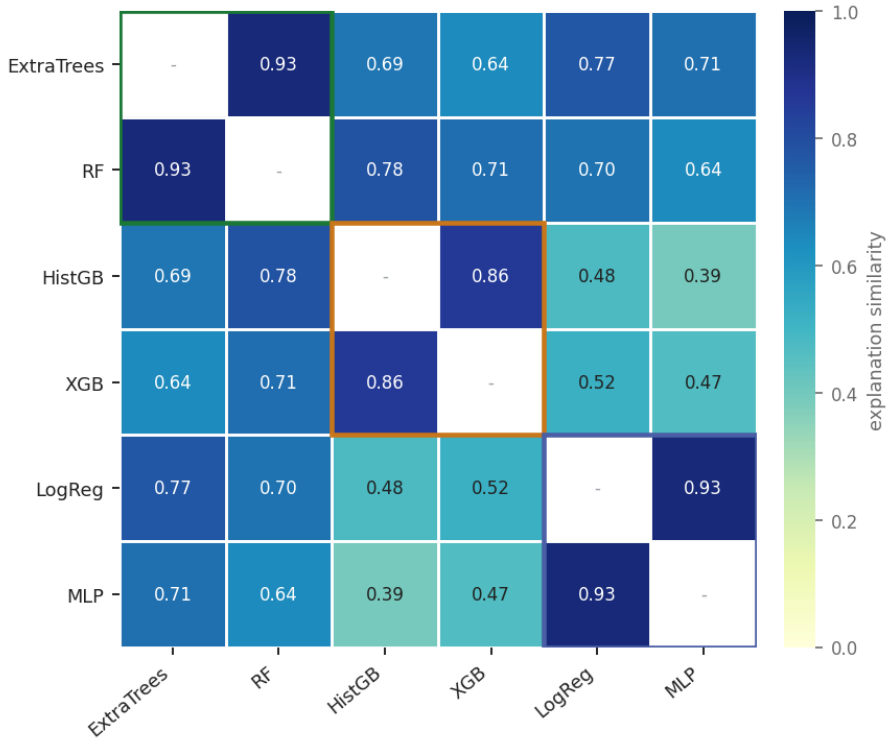

Figure 7: Cross-family similarity with families ordered by functional group. The three outlined diagonal blocks contain within-group pairs. Their median similarity is 0.9265, compared with 0.6646 for the remaining pairs, illustrating the group contrast in §6.1.

especially LogReg/MLP, is exploratory. The main comparison pools all 15 family pairs and does not require these categories.

K Archived regional analysis and the agreement stratum

Interior/boundary. We define two regions without a numerical threshold. First, the Hungarian algorithm matches each cluster to a class. A point is *interior* if its cluster is matched to its true class, and *boundary* otherwise. These names describe label agreement rather than geometric distance. The normalized discrepancy d_{norm} used below is defined in Appendix L. On `dry_bean`, 10,706 points

are interior (78.7%) and 2,905 are boundary (21.3%). Of the 16 features, 15 separate the regions at FDR < 0.05.

		interior		boundary	
comparison (per-instance)	pairing	simil.	d_{norm}	simil.	d_{norm}
surrogate × clustering	map	0.7997	+0.032	0.8405	+0.003
classifier × clustering	map	0.8457	−0.001	0.1749	+0.479
classifier × clustering	class	0.8457	−0.001	0.7503	+0.067

The surrogate's similarity does not decrease at the boundary. For the classifier, the large drop under *map* pairing depends on which output is read: at boundary points, this pairing selects a class to which the point does not belong. As a control, *class* pairing selects the true class. The decrease then becomes $0.8457 \rightarrow 0.7503$. Four of nine comparisons show significant regional instability after Benjamini–Hochberg correction. All four involve the classifier, and three use map pairing. Switching to class pairing removes 86–99% of the difference. Only the classifier's per-instance comparison remains significant ($\delta = 0.0954$, $p_{\text{BH}} = 0.0112$). No surrogate comparison is significant ($p_{\text{BH}} = 0.0522$). Figure 8 shows the comparison.

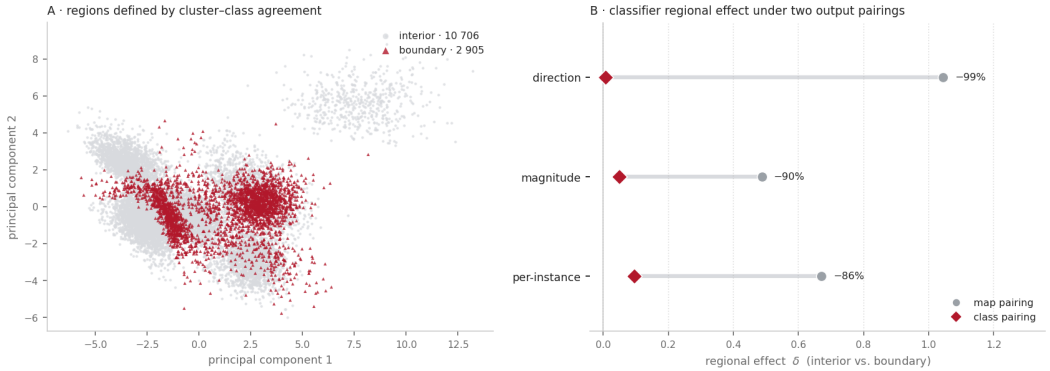

Figure 8: (A) The two regions of `dry_bean` in the first two principal components; the regions are defined by agreement between the matched cluster label and the true class, rather than by geometric distance. Points with label disagreement concentrate where clusters overlap. **(B)** The classifier's regional effect under map and true-class pairing: changing the pairing reduces the measured effect by 86–99%.

The agreement stratum. We next compare the refit baseline with cross-family similarity separately in two strata. A point belongs to stratum **A** ("everything agrees") if it is interior, all six surrogates predict its assigned cluster, and a held-out classifier ($X \rightarrow y$, 5-fold `cross_val_predict`) predicts its true class. Stratum **D** contains the remaining points. We define the strata over the explained sample of up to 150 points. In all six datasets, A is contained in the interior (median Jaccard 0.9856). We compute a separate refit baseline for each stratum because easier points can also be more stable under refitting.

	stratum A			stratum D			
dataset	floor	cross	gap	floor	cross	gap	n in D
`iris`	0.9949	0.8882	0.1067	0.9846	0.8773	0.1026	30
`wine`	0.9680	0.8546	0.1134	0.9291	0.7805	**0.1498**	**8**
`breast_cancer`	0.9717	0.5388	0.4215	0.9485	0.6102	0.3383	17
`digits`	0.9521	0.4092	0.5424	0.9493	0.4482	0.4974	48
`pendigits`	0.9936	0.7911	0.1996	0.9898	0.8337	0.1552	53
`dry_bean`	0.9917	0.6966	**0.2951**	0.9892	0.7201	0.2696	39

Each gap is the median of differences within seeds, so it need not equal the difference of the displayed medians. The gap is larger in stratum A in 5 of 6 datasets. The refit baseline rises in A in every dataset, while cross-family similarity falls in 4 of 6. Thus, stronger agreement on labels

can coexist with weaker agreement between families' explanations. The exception to the larger gap has only $n = 8$ points in D, limiting resolution. The gap persists where predictions agree, but this descriptive comparison does not isolate every effect of ambiguity or estimation noise.

L ARCHIVED CALIBRATION AND NORMALIZED COMPARISONS

These mixed-protocol analyses compare three attribution targets: a surrogate trained on cluster labels, a classifier trained on the dataset's true classes, and direct K-means affinity. In the calibration ladder, "proxy versus classifier" compares the two fitted random forests; "across paradigms" compares the supervised classifier with direct clustering affinity. The ladder is a set of reference comparisons for the similarity measure, separate from the main cross-family experiment on cluster labels.

For observed similarity s, a matched ceiling c, and a shuffled-label baseline b, the archived normalized discrepancy is

$$d_{\text{norm}} = \frac{c - s}{c - b}.$$

The ceiling maps to zero and the shuffled-label baseline to one. Values are not clipped to this interval; the checks below determine whether this normalization is used for a given measure and method.

L.1 GATES, CALIBRATION LADDER, AND GUARDS

Before comparing explanations on `dry_bean`, we check cluster-to-class agreement and surrogate fidelity. The Hungarian cluster→class matching has accuracy 0.7866, compared with a maximum of 0.1858 under shuffled labels. The surrogate's out-of-bag accuracy against the partition is 0.9771. The clustering has Silhouette 0.3094 and adjusted Rand index (ARI) 0.6688 against the classes.

We assess the resolution of the similarity measure with a calibration ladder (SHAP, per-instance, `dry_bean`). Similarity is 1.000 for a model compared with itself; 0.985 for the same family trained with another seed; 0.866 for proxy versus classifier; 0.826 across paradigms; 0.599 **across families** in the supervised task; and -0.529 for shuffled labels. The separation between repeated fits (0.985) and shuffled labels (-0.529) shows that the measure can distinguish these cases. It does not determine the value of the main comparison (Figure 9).

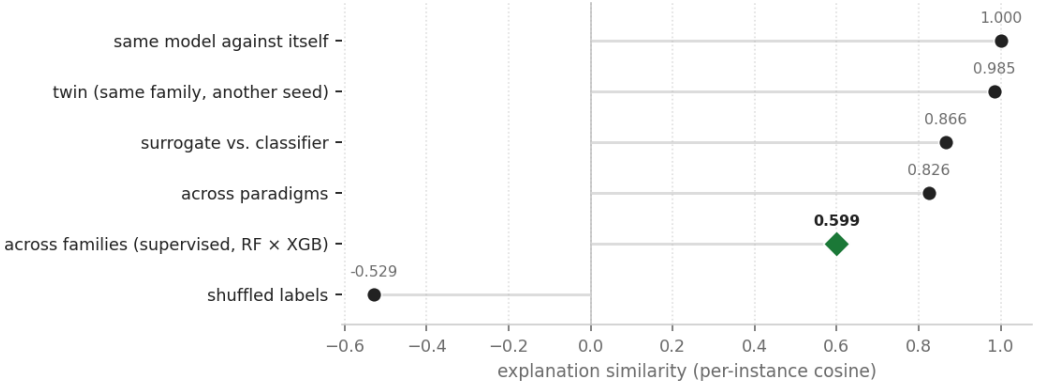

Figure 9: The calibration ladder on `dry_bean` (SHAP, per-instance): from its top rung — the same model against itself — to the shuffled-label floor. Each rung compares the supervised random-forest reference with the listed counterpart. The cross-family comparison of §6.1 (fifteen pairs on cluster labels) uses a different target and does not appear here.

For per-instance SHAP on `dry_bean`, the matched ceiling is 0.8445, the shuffled baseline is -0.5547, and their span is 1.3992. An observed cosine of 0.80 therefore corresponds to $d_{\text{norm}} = 0.032$. We require a positive ceiling and a sufficiently large span in every seed. Six of the nine measure×method combinations fail these checks:

measure	method	anchor holds	reason
per-instance	SHAP	**yes**	—
per-instance	LIME	**yes**	—
direction	LIME	**yes**	—
direction	SHAP	no	insufficient span
magnitude	SHAP	no	non-positive ceiling
magnitude	LIME	no	insufficient span
magnitude	PFI	no	insufficient span
direction	PFI	no	dimension not covered
per-instance	PFI	no	dimension not covered

None of the three methods supports a calibrated magnitude comparison on this dataset. Magnitude claims in the main text therefore compare the refit baseline with cross-family similarity within a single method; they do not use d_{norm}.

L.2 The three views under the valid comparisons

candidate	measure / method	d_{norm}	
surrogate	per-instance / SHAP	-0.003	*above the ceiling*
surrogate	direction / LIME	$+0.036$	
surrogate	per-instance / LIME	$+0.081$	
classifier	direction / LIME	$+0.006$	
classifier	per-instance / SHAP	$+0.022$	
classifier	per-instance / LIME	$+0.046$	

A value of $d_{\text{norm}} = -0.003$ means that the observed difference is smaller than the measure's resolution. It does not indicate negative divergence. We leave two quantities outside this normalized scale. Interaction strength is 0.4587 for the surrogate, 0.5166 for the classifier, and 0.0000 for direct clustering. It describes each model individually; the K-means affinity is additive by construction. CKA/RSA values (0.39–0.46) lack an anchor for normalization.