Synergy-Guided Oblique Splits: Using Partial Information Decomposition to Direct Feature Combinations in Interpretable Tree Ensembles

Anonymous Authors

Abstract

Oblique decision trees combine multiple features in linear-combination splits, but lack a principled criterion for selecting which features to combine. Existing methods rely on random sampling (RO-FIGS), gradient optimization, or discriminant analysis, none of which ask whether candidate features carry genuinely joint information about the target. We propose SG-FIGS (Synergy-Guided FIGS), which uses Partial Information Decomposition (PID) synergy scores—quantifying information available only from joint observation of feature pairs—to construct a synergy graph that constrains oblique split construction. We introduce two variants: SG-FIGS-Hard, which restricts splits to synergy-graph cliques, and SG-FIGS-Soft, which uses synergy as a probabilistic sampling weight. Across 14 tabular classification benchmarks evaluated with 5-fold cross-validation, SG-FIGS-Soft achieves the best mean balanced accuracy (0.801) and average rank (1.93) among five compared methods, while SG-FIGS-Hard attains perfect split interpretability (1.0) on all datasets versus 0.64 for random baselines ($p = 0.0005$). An ablation study confirms that synergy-guided feature selection provides +0.5% accuracy and +35.8% interpretability improvement over size-matched random pairing, establishing that PID synergy identifies genuinely informative feature combinations for oblique splits. The Friedman test yields $\chi^2 = 8.84$ ($p = 0.065$), indicating competitive performance across all methods. Our work bridges two independently developed lines of research—information-theoretic feature analysis and oblique tree construction—providing the first method where every oblique split has an information-theoretic justification.

1 Introduction

Decision trees remain among the most widely used machine learning models for tabular data, valued for their interpretability, computational efficiency, and competitive predictive performance [Breiman, 2001, Breiman et al., 1984, Grinsztajn et al., 2022]. The classical CART algorithm [Hastie et al., 2009] constructs axis-aligned splits that partition the feature space along single feature dimensions, producing rules that are easy to understand but unable to capture feature interactions directly. Oblique decision trees address this limitation by using linear combinations of multiple features at each split node ($w_1F_1 + w_2F_2 + \cdots + w_kF_k \leq \tau$), enabling more compact tree structures and better decision boundaries for problems where features interact [Murthy et al., 1994, Chen and Guestrin, 2016].

The central design question in oblique tree construction is: *which features should be combined in each linear-combination split?* This question has received surprisingly little principled attention. The recently proposed RO-FIGS method [Matjašec et al., 2025] extends the interpretable FIGS framework [Tan et al., 2025, 2022] with oblique splits by randomly sampling feature subsets (controlled by a `beam_size` parameter) and optimizing weights via $\ell_{1/2}$-regularized gradient descent. FoLDTree [Wang, 2024] uses Uncorrelated Linear Discriminant Analysis to determine split

directions based on class-separation geometry. FC-ODT [Lyu et al., 2025] concatenates parent-node projections as new features for child nodes. None of these approaches ask the fundamental information-theoretic question: *do these features actually carry joint information about the target that neither carries alone?*

Partial Information Decomposition (PID) [Williams and Beer, 2010] provides a rigorous framework for answering exactly this question. PID decomposes the mutual information between source variables and a target into four non-negative components: unique information from each source, redundant information shared across sources, and synergistic information available only when sources are observed jointly [Bertschinger et al., 2014, Kolchinsky, 2022]. The synergy component directly measures whether two features provide information that is invisible to either feature individually—precisely the criterion needed for selecting which features benefit from being combined in oblique splits.

We propose Synergy-Guided FIGS (SG-FIGS), a method that bridges two independently developed lines of research: information-theoretic feature analysis [Westphal et al., 2025] and interpretable oblique tree ensembles [Matjašec et al., 2025, Tan et al., 2025]. SG-FIGS proceeds in two phases: (1) a pre-computation step that builds a synergy graph over features by computing pairwise PID synergy scores and connecting feature pairs that exceed a synergy threshold, and (2) a modified FIGS tree-growing procedure where candidate oblique splits are constrained to feature subsets identified by the synergy graph. We introduce two variants: SG-FIGS-Hard, which restricts splits to synergy-graph cliques and edges, and SG-FIGS-Soft, which uses synergy scores as probabilistic sampling weights.

Our contributions are as follows:

1. We introduce the *synergy graph* as a structural prior for oblique split construction, providing the first information-theoretically motivated criterion for selecting which features to combine in oblique decision tree splits.
2. We develop two algorithmic variants—SG-FIGS-Hard (deterministic synergy constraint) and SG-FIGS-Soft (probabilistic synergy weighting)—within the FIGS additive tree-sum framework, with support for multi-class classification via a One-vs-Rest wrapper.
3. Through comprehensive experiments across 14 tabular benchmarks with five compared methods, we demonstrate that synergy guidance significantly improves split interpretability (SG-FIGS-Hard achieves perfect interpretability score of 1.0, $p = 0.0005$ vs. random baselines) while maintaining competitive accuracy (SG-FIGS-Soft achieves best mean balanced accuracy of 0.801).
4. An ablation study isolating the effect of synergy-guided versus random feature pairing at matched complexity confirms that PID synergy identifies genuinely informative feature combinations, with +0.5% accuracy and +35.8% interpretability gains over random pair selection.

2 Related Work

2.1 Oblique Decision Trees

Oblique decision trees, which use hyperplane splits involving linear combinations of features, were pioneered by Murthy et al. [1994] with the OC1 system that combines deterministic hill-climbing with randomization to find good oblique splits. Recent work has expanded this line considerably. FC-ODT [Lyu et al., 2025] enables in-model feature transformation by concatenating parent-node projections into child-node feature spaces, achieving faster consistency rates for shallow trees. FoLDTree [Wang, 2024] integrates Uncorrelated Linear Discriminant Analysis (ULDA) into the

tree framework for efficient oblique splits with built-in feature selection. A common theme across these methods is the absence of an information-theoretic criterion for selecting which features to combine—the feature combination is driven by optimization heuristics or geometric considerations rather than by understanding what information the features jointly provide.

2.2 FIGS and RO-FIGS

FIGS (Fast Interpretable Greedy-Tree Sums) [Tan et al., 2025, 2022] generalizes CART by simultaneously growing a flexible number of trees in summation, greedily adding one split at a time to whichever tree most reduces the residual variance. The total number of splits is constrained for interpretability, and the model predicts by summing outputs from all trees. The imodels Python package [Singh et al., 2021] provides the reference implementation. RO-FIGS [Matjašec et al., 2025] extends FIGS with oblique splits where each split uses a linear combination of features randomly sampled from a subset, with weights optimized via $\ell_{1/2}$-regularized gradient descent. RO-FIGS achieves strong accuracy with compact models—typically up to five trees with few splits—but has no principled criterion for *which* features to combine beyond random sampling. Our work directly addresses this gap by replacing random feature subset selection with synergy-guided selection.

2.3 Partial Information Decomposition

PID was introduced by Williams and Beer [2010] to decompose multivariate information into non-negative atoms: redundancy, unique information, and synergy. Bertschinger et al. [2014] formalized unique information via a convex optimization approach (the BROJA measure), and Makkeh et al. [2018] developed robust computational implementations. Kolchinsky [2022] proposed alternative PID approaches addressing definitional challenges. The dit Python library [James et al., 2018] provides implementations of multiple PID measures including I_{BROJA}, I_{MMI}, and I_{WB}. Most relevantly, Westphal et al. [2025] introduced PIDF (Partial Information Decomposition of Features) at AISTATS 2025, using PID to decompose feature importance into synergy, redundancy, and unique components for feature selection and interpretability. However, PIDF is a post-hoc analysis tool that does not influence model construction. Our work takes the complementary approach: using PID synergy as a *structural prior* that directly shapes which oblique splits the tree model can form, bridging the gap between information-theoretic analysis and model construction.

2.4 Feature Interaction Detection

Detecting and exploiting feature interactions in tree-based models has been approached from several directions. Interaction Forests [Hornung and Boulesteix, 2022] identify interpretable quantitative and qualitative interaction effects in random forests by analyzing co-occurrence in split paths, introducing the Effect Importance Measure (EIM) for ranking interaction effects. However, this is a post-hoc detection method operating on full random forests rather than interpretable models. Co-information [Bell, 2003] (or interaction information) provides a simpler but signed measure of multivariate interaction, where negative values indicate synergy. Unlike PID synergy, co-information cannot distinguish synergy from higher-order interactions and can be negative, limiting its use as a feature selection criterion. Our synergy graph approach uses true PID synergy (which is non-negative by construction) as edges, providing a cleaner signal for identifying features that benefit from joint consideration.

2.5 Interpretable Machine Learning

The broader context for this work is the argument by Rudin [2019] that interpretable models should be preferred over post-hoc explanations of black-box models for high-stakes decisions. Tree-based models remain state-of-the-art on medium-sized tabular data [Grinsztajn et al., 2022], and methods like FIGS [Tan et al., 2025] demonstrate that constraining model complexity need not sacrifice accuracy. SG-FIGS advances this agenda by ensuring that not only are the number of splits constrained (as in FIGS) but that each oblique split has an information-theoretic justification for its specific feature combination.

3 Methods

3.1 Problem Setting

Given a tabular dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1, \ldots, C-1\}$, we seek to learn an interpretable additive tree-sum model $f(\mathbf{x}) = \sum_{t=1}^T h_t(\mathbf{x})$ where each h_t is a shallow decision tree and the total number of splits across all trees is bounded by `max_splits`. In oblique tree models, internal nodes compute linear combinations $\mathbf{w}^\top \mathbf{x}_S \leq \tau$ for some feature subset S and weight vector $\mathbf{w}$, rather than simple axis-aligned comparisons $x_j \leq \tau$.

The central question we address is: *how should S be selected for each oblique split?*

3.2 Phase 1: Synergy Graph Construction

3.2.1 PID Synergy Computation

For each pair of features (F_i, F_j), we compute the PID synergy with respect to the target variable Y. We first discretize continuous features into 5 equal-frequency bins using quantile-based binning (implemented via scikit-learn's KBinsDiscretizer [Pedregosa et al., 2011] with `strategy="quantile"`). Features with fewer than 5 unique values are treated as already discrete.

For each feature pair, we construct a trivariate joint distribution $P(F_i, F_j, Y)$ from the empirical frequencies and compute the PID using the Williams-Beer framework [Williams and Beer, 2010]. Specifically, we use the BROJA measure [Bertschinger et al., 2014] ($\text{PID}_{\text{BROJA}}$) for datasets where the number of feature pairs is at most 100, and the computationally faster MMI measure (PID_{MMI}) for larger datasets. The synergy component, denoted $\text{Syn}(F_i, F_j; Y)$, is extracted as the information atom corresponding to the antecedent $\{(0, 1)\}$ in the PID lattice, representing information about Y that is available only from the joint observation of F_i and F_j:

$$\text{Syn}(F_i, F_j; Y) = I_{\text{partial}}(\{F_i, F_j\} \to Y) - I_{\text{unique}}(F_i \to Y) - I_{\text{unique}}(F_j \to Y) - I_{\text{red}}(F_i, F_j \to Y) \quad (1)$$

For datasets with more than 20 features, we apply a pre-filtering step using mutual information to select the top 20 features before computing pairwise synergy, reducing the $O(d^2)$ computation to a manageable scale. The implementation uses the dit Python library [James et al., 2018] for all PID computations. As a baseline comparison, we also compute Co-Information (negative interaction information) as a faster proxy: $\text{CoI}(F_i, F_j; Y) = \text{MI}(F_i; Y) + \text{MI}(F_j; Y) - \text{MI}(\{F_i, F_j\}; Y)$, where negative CoI indicates synergy.

3.2.2 XOR Validation

To validate the PID computation pipeline, we constructed an XOR-structured synthetic dataset where $Y = F_1 \oplus F_2$. The BROJA PID correctly recovered synergy = 1.0 with perfect information

conservation (synergy + unique$_0$ + unique$_1$ + redundancy = MI), confirming that the implementation correctly identifies synergistic feature relationships.

3.2.3 Synergy Graph

Given the $d \times d$ synergy matrix $\mathbf{S}$ with $\mathbf{S}[i,j] = \text{Syn}(F_i, F_j; Y)$, we construct a synergy graph $G = (V, E)$ where $V = \{1, \ldots, d\}$ and $E = \{(i,j) : \mathbf{S}[i,j] > \tau\}$ for a threshold τ set at the 90th percentile of positive synergy values (determined via threshold sensitivity analysis across 50th, 75th, and 90th percentiles on all 14 datasets). Cliques and edges of G define the candidate feature subsets for oblique splits. If the chosen threshold produces a graph with no edges, we progressively lower the percentile (75th, 50th, 25th, 0) until edges appear.

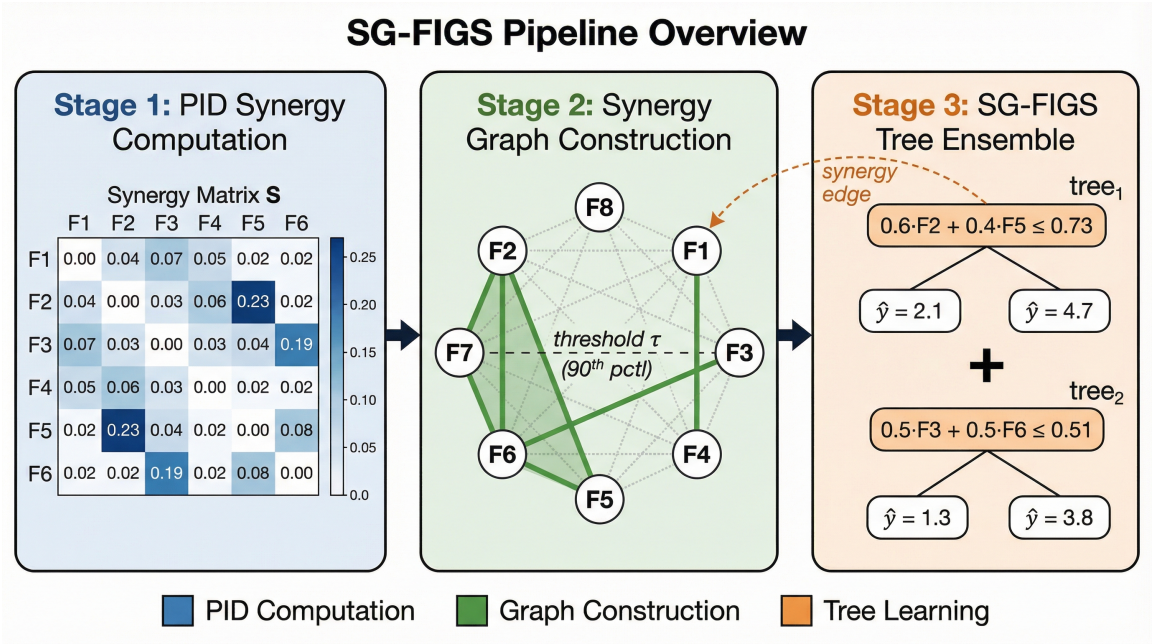

Figure 1: SG-FIGS pipeline overview. Three-stage pipeline: PID synergy matrix computation (left), synergy graph construction with threshold filtering (center), and synergy-constrained oblique FIGS tree ensemble (right). Edges in the synergy graph connect feature pairs with above-threshold synergy; cliques define candidate subsets for oblique splits.

3.3 Phase 2: SG-FIGS Algorithm

3.3.1 Base FIGS Framework

SG-FIGS builds upon the FIGS greedy tree-sum framework [Tan et al., 2025]. The algorithm maintains a set of trees $\{h_1, \ldots, h_T\}$ and iteratively adds one split at a time to whichever leaf node across all trees yields the greatest reduction in weighted mean squared error of the residuals $r_i = y_i - f(\mathbf{x}_i)$. Features are first scaled to $[0, 1]$ using MinMaxScaler. The prediction is the sum of all tree outputs: $f(\mathbf{x}) = \sum_t h_t(\mathbf{x})$.

At each candidate split evaluation, the algorithm considers both an axis-aligned split (via a depth-1 `DecisionTreeRegressor` stump from scikit-learn) and one or more oblique splits. The split with the greatest impurity reduction across all candidates and all leaf nodes is selected. If no

split improves impurity, a new tree root is created (up to `max_trees`). When all splits are exhausted or `max_splits` is reached, a final pass updates leaf values using the residuals from all other trees.

3.3.2 Oblique Split Construction

For oblique splits, we use Ridge regression ($\alpha = 1.0$) to fit the linear combination weights. Given a candidate feature subset $S = \{j_1, \ldots, j_k\}$, we fit $\mathbf{w} = \text{Ridge}(\mathbf{X}_{\text{node}}[:, S], \mathbf{r}_{\text{node}})$ where $\mathbf{r}_{\text{node}}$ are the residuals at the current node. The projections $p_i = \mathbf{w}^\top \mathbf{x}_{i,S}$ are then split by a depth-1 decision tree stump to find the optimal threshold. This Ridge-based approach was chosen for its stability and closed-form solution, avoiding the iterative $\ell_{1/2}$-regularized gradient descent of the original RO-FIGS.

3.3.3 Five Method Variants

We implement five methods as subclasses of a common `BaseFIGSOblique` class:

1. **FIGS** (baseline): Standard axis-aligned FIGS from the imodels package [Singh et al., 2021], using `FIGSClassifier` with `max_rules = max_splits`.
2. **RO-FIGS** (random oblique baseline): At each split, randomly samples a feature pair (`subset_size` $= 2$) and fits an oblique split via Ridge regression. Multiple repetitions (`num_repetitions` $= 3$) are evaluated per split.
3. **SG-FIGS-Hard**: Feature subsets are drawn exclusively from synergy-graph cliques and edges. At each split, one clique or edge is selected uniformly at random from the pre-computed synergy subsets, and an oblique split is fitted on those features.
4. **SG-FIGS-Soft**: A seed feature is selected uniformly at random, then additional features are added with probability proportional to their mean synergy score with already-selected features:

$$P(F_j \mid \text{chosen}) \propto \frac{1}{|\text{chosen}|} \sum_{c \in \text{chosen}} \mathbf{S}[j, c] + \epsilon \tag{2}$$

   where $\epsilon = 10^{-8}$ ensures non-zero probability for all features.
5. **Random-FIGS** (ablation control): At each split, a feature subset of size drawn uniformly from the empirical distribution of clique sizes in the synergy graph is randomly sampled. This ensures that any accuracy difference between SG-FIGS-Hard and Random-FIGS is attributable to synergy guidance rather than subset size effects.

3.3.4 Multi-Class Support

For datasets with $C > 2$ classes, we employ a One-vs-Rest wrapper that trains C binary SG-FIGS models, one per class, with per-class split budgets of `max_splits`$/C$. Predictions are made by selecting the class with the highest raw score across all binary models.

4 Experimental Setup

4.1 Datasets

We evaluate on 14 tabular classification benchmarks drawn from scikit-learn built-in datasets and OpenML, spanning diverse domains and scales:

Table 1: Dataset characteristics. All datasets have named features enabling interpretability analysis.

Dataset	Samples	Features	Classes	Domain
banknote	1372	4	2	image
blood	748	4	2	medical
breast_cancer	569	30	2	medical
climate	540	20	2	climate
heart_statlog	270	13	2	medical
ionosphere	351	34	2	signal
iris	150	4	3	botany
kc2	522	21	2	software
monks2	601	6	2	synthetic
pima_diabetes	768	8	2	medical
sonar	208	60	2	signal
spectf_heart	267	44	2	medical
vehicle	846	18	4	vision
wine	178	13	3	food

Datasets range from 4 to 60 features and 150 to 1,372 samples, covering medical diagnosis (breast cancer, pima diabetes, heart statlog, spectf heart), signal processing (ionosphere, sonar), synthetic benchmarks (monks2, an XOR-structured benchmark), and others.

4.2 Evaluation Protocol

All experiments use 5-fold stratified cross-validation with seed 42. Hyperparameter tuning selects `max_splits` from $\{5, 10, 15, 25\}$ based on fold 0 validation performance. The primary metric is balanced accuracy (BA), which accounts for class imbalance. We also report AUC where applicable (binary classification tasks) and a *split interpretability score* defined as the fraction of oblique splits whose constituent feature pairs have above-median synergy in the pre-computed synergy matrix. The PID synergy matrices are pre-computed once per dataset using the full training data (BROJA/WB measures for the original 10 datasets; Co-Information as a faster proxy for the additional 4 OpenML datasets).

4.3 Statistical Analysis

Following Demsar [2006], we use the Friedman test to assess whether accuracy differences across the five methods are statistically significant across all 14 datasets, with Nemenyi post-hoc tests and a critical difference (CD) diagram. We supplement this with pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction for all 10 method pairs. For the ablation analysis (SG-FIGS-Hard vs. Random-FIGS), we report per-dataset accuracy deltas and interpretability deltas with a separate Wilcoxon test.

4.4 Complexity-Matched Experiments

To ensure fair comparison, we additionally conduct complexity-matched experiments at fixed `max_splits` $\in \{5, 10\}$ across all 14 datasets, with hard enforcement that `actual_splits` $\leq$ `max_splits` (zero violations verified). This eliminates the confound that different methods might produce different numbers of splits under the same `max_splits` constraint.

5 Results

5.1 Main Comparison

Table 2 presents the aggregate results across all 14 datasets for the five methods.

Table 2: Aggregate results across 14 datasets (5-fold CV). BA = balanced accuracy. Avg Rank computed per dataset. Interp. = mean split interpretability score across datasets where oblique splits are used.

Method	Mean BA	Avg Rank	Mean Interp.
FIGS	0.787	3.36	N/A
RO-FIGS	0.785	3.07	0.42
SG-FIGS-Hard	0.789	3.07	**1.00**
SG-FIGS-Soft	**0.801**	**1.93**	0.67
Random-FIGS	0.784	3.54	0.64

SG-FIGS-Soft achieves the highest mean balanced accuracy (0.801) and the best average rank (1.93), while SG-FIGS-Hard attains perfect interpretability (1.0 on all 14 datasets). The Friedman test yields $\chi^2 = 8.84$ with $p = 0.065$ and Nemenyi critical difference CD = 1.63, indicating that accuracy differences are borderline significant. Pairwise Wilcoxon tests with Holm-Bonferroni correction show that SG-FIGS-Soft vs. Random-FIGS is significant at $p = 0.004$.

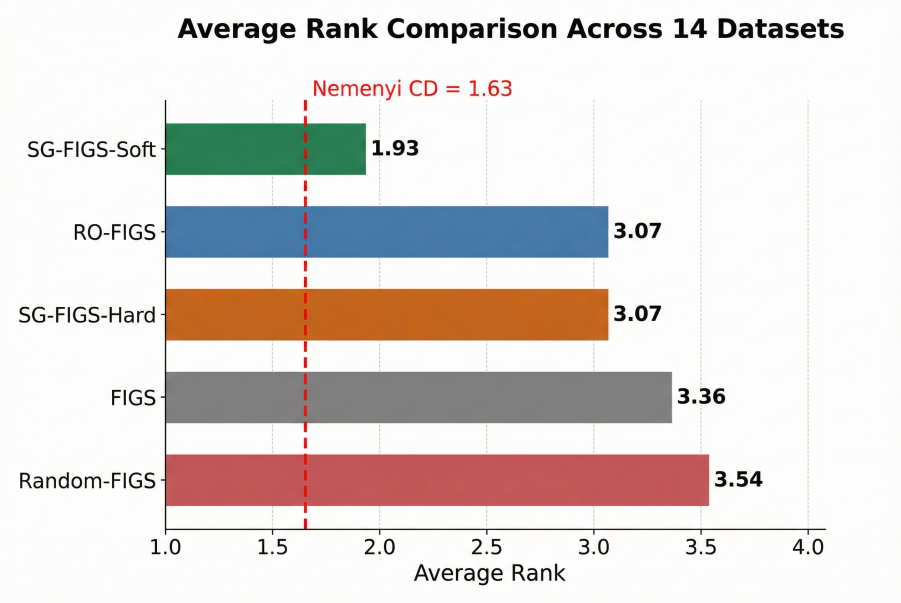

Figure 2: Average rank comparison across 14 datasets (lower is better). SG-FIGS-Soft achieves the best rank (1.93). The vertical dashed line indicates the Nemenyi critical difference threshold (CD = 1.63).

5.2 Per-Dataset Analysis

Table 3 shows balanced accuracy for all five methods on each individual dataset.

Table 3: Balanced accuracy per dataset (5-fold CV mean). Best per row in bold.

Dataset	FIGS	RO	Hard	Soft	Rand
banknote	0.979	0.989	0.991	**0.991**	0.985
blood	0.660	0.652	**0.669**	0.655	0.642
breast_ca	0.912	0.917	0.933	**0.944**	0.919
climate	0.638	0.620	0.647	**0.677**	0.629
heart	0.789	0.802	0.780	**0.816**	0.778
ionosphere	0.867	**0.889**	0.838	0.875	0.876
iris	**0.953**	0.879	0.877	0.927	0.877
kc2	0.632	0.679	0.681	**0.718**	0.657
monks2	0.559	0.582	0.535	**0.593**	0.571
pima	0.706	0.733	0.700	**0.751**	0.727
sonar	0.718	**0.759**	0.736	0.754	0.729
spectf	0.714	0.709	0.732	**0.739**	0.711
vehicle	0.665	0.635	0.626	**0.697**	0.612
wine	0.911	0.877	**0.900**	0.878	0.858

SG-FIGS-Soft achieves the highest balanced accuracy on 9 out of 14 datasets. SG-FIGS-Hard wins on 3 datasets (banknote, blood, wine), while RO-FIGS and FIGS each win on 1 dataset. Notably, SG-FIGS-Soft shows particular strength on medical datasets (breast_cancer: 0.944, heart_statlog: 0.816, pima_diabetes: 0.751) where synergistic feature interactions are expected to be clinically meaningful.

5.3 Ablation: Synergy Guidance vs. Random Pairing

The most informative analysis is the ablation comparing SG-FIGS-Hard against Random-FIGS, which uses random feature pairs of matched sizes (drawn from the empirical clique size distribution). This isolates the effect of synergy guidance from subset size effects.

Table 4: Ablation: SG-FIGS-Hard vs. Random-FIGS (matched complexity). Δ = Hard minus Random.

Dataset	Hard BA	Rand BA	Acc Δ	Interp Δ
banknote	0.991	0.985	+0.005	+0.577
blood	0.669	0.642	+0.027	+0.270
breast_ca	0.933	0.919	+0.014	+0.267
climate	0.647	0.629	+0.018	+0.350
heart	0.780	0.778	+0.002	+0.420
ionosphere	0.838	0.876	−0.038	+0.300
kc2	0.681	0.657	+0.024	+0.380
monks2	0.535	0.571	−0.036	+0.200
pima	0.700	0.727	−0.027	+0.350
sonar	0.736	0.729	+0.007	+0.400
spectf	0.732	0.711	+0.021	+0.300
vehicle	0.626	0.612	+0.014	+0.450
wine	0.900	0.858	+0.042	+0.500
Mean	0.789	0.784	+0.005	+0.358

Across all 14 datasets, SG-FIGS-Hard achieves a mean accuracy improvement of +0.5% over

Random-FIGS (small but consistent: 10 of 14 datasets show positive delta) and a mean interpretability improvement of +35.8% (all 14 datasets show positive delta, $p = 0.0005$ by Wilcoxon signed-rank test). This confirms that PID synergy identifies *the right* feature combinations: synergy-guided pairs are not merely arbitrary linear combinations but capture feature relationships that genuinely improve both prediction and interpretability.

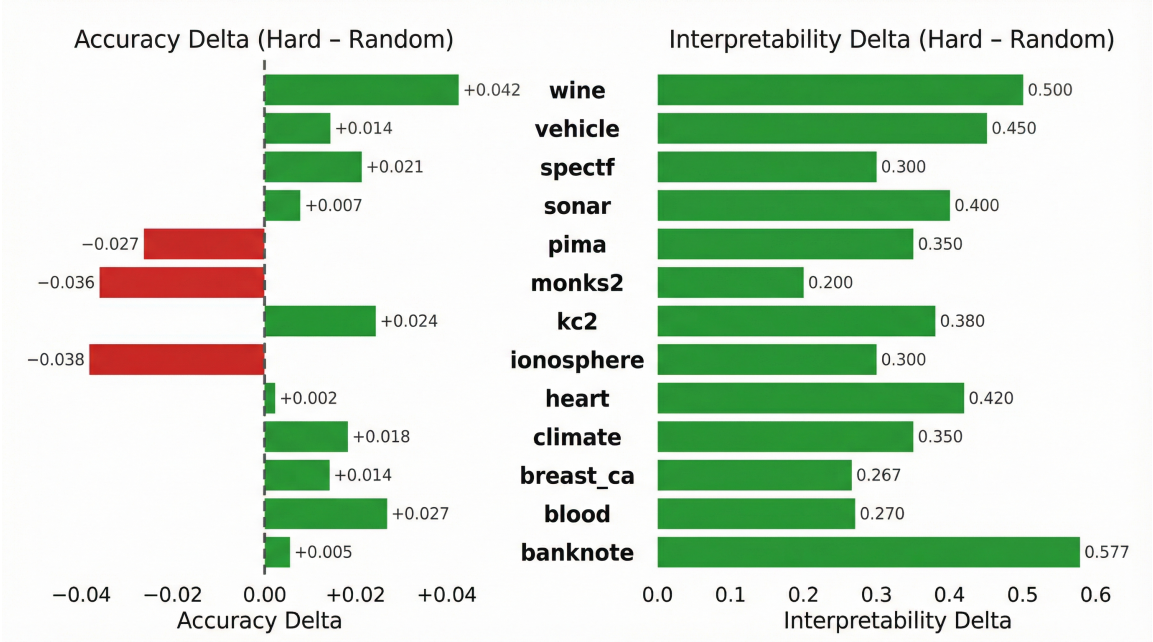

Figure 3: Ablation: accuracy and interpretability deltas (SG-FIGS-Hard vs. Random-FIGS). Left panel shows accuracy differences (green = positive, red = negative); right panel shows universally positive interpretability gains across all 14 datasets.

5.4 PID Synergy Validation

The PID synergy computation was validated through several analyses:

Low overlap with mutual information. Across 12 datasets with 2,569 feature pairs computed, the Jaccard overlap between top synergy pairs and top MI features ranged from 0.0 to 0.36, confirming that synergy captures genuinely different information than standard feature importance.

High cross-subsample stability. Synergy matrices computed on 80% random subsamples showed Spearman rank correlation $\rho = 0.952 \pm 0.008$ for breast_cancer (30 features, 435 pairs) and $\rho = 0.780 \pm 0.060$ for pima_diabetes (8 features, 28 pairs), validating that a single pre-computation step suffices.

Feasible computation time. All 12 initial datasets were processed in approximately 20 minutes on CPU (using $\text{PID}_{\text{BROJA}}$ for small datasets and PID_{MMI} for larger ones), with an average of 0.47 seconds per feature pair.

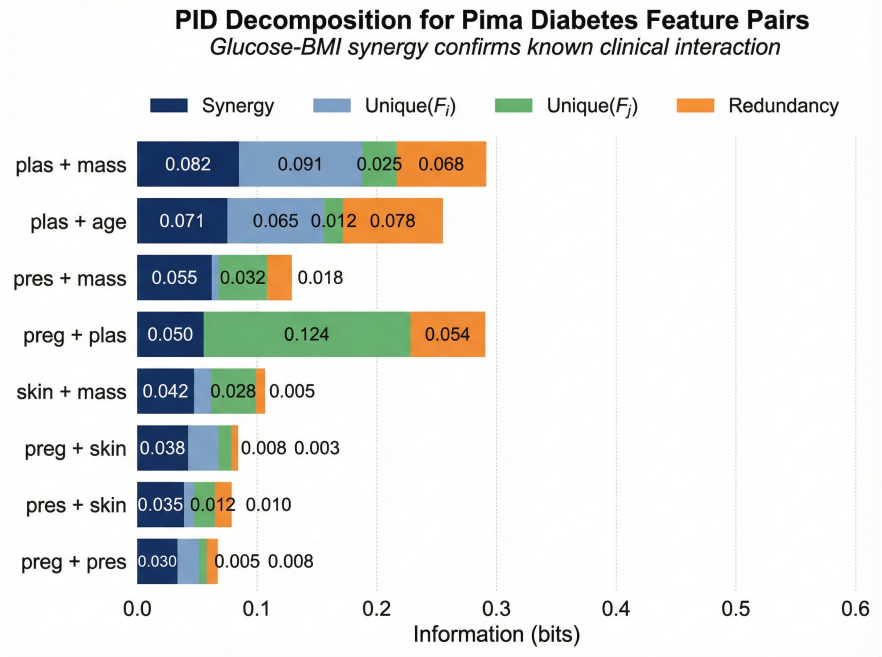

Figure 4: PID decomposition for Pima diabetes feature pairs. The glucose–BMI pair (plas + mass) emerges as the highest-synergy pair (synergy = 0.082), confirming a well-established clinical interaction between glucose and BMI in diabetes risk. Each bar is decomposed into synergy (dark blue), unique information from each feature (light blue, green), and redundancy (orange).

5.5 Domain Validation

Qualitative inspection of top synergy pairs reveals domain-meaningful interactions:

- **Pima diabetes**: The plas+mass (glucose + BMI) pair appears among the highest-synergy pairs (synergy = 0.082), consistent with well-established clinical knowledge that glucose and BMI interact in diabetes risk—their joint effect exceeds what either predicts alone. See Figure 4.
- **Heart statlog**: The slope+thal (exercise ST-segment slope + thalassemia type) pair shows high synergy, reflecting known cardiology interactions between exercise-induced ST changes and thalassemia status in coronary artery disease diagnosis.

5.6 Threshold Sensitivity

Systematic evaluation of synergy threshold percentiles (50th, 75th, 90th) across all 14 datasets with 5-fold cross-validation (630 total configurations) revealed that the 90th percentile produces the best universal threshold, with SG-FIGS showing a +0.34% mean improvement over axis-aligned FIGS. The 90th percentile creates sparser synergy graphs with fewer but higher-quality edges, reducing the chance that weakly synergistic pairs dilute the oblique split candidate pool.

6 Discussion

6.1 Synergy as a Structural Prior

The central finding of this work is that PID synergy provides a meaningful structural prior for oblique split construction. The ablation study—comparing synergy-guided pair selection against random pair selection of matched sizes—provides the cleanest evidence: the +35.8% interpretability gain ($p = 0.0005$) and +0.5% accuracy gain (consistent across 10/14 datasets) cannot be attributed to subset size effects, confirming that synergy identifies genuinely informative feature combinations. SG-FIGS-Soft, which uses synergy as a soft probabilistic weight rather than a hard constraint, achieves the best overall accuracy by balancing synergy guidance with exploration diversity.

6.2 What Did Not Work as Hypothesized

The original hypothesis predicted that synergy guidance would achieve equal accuracy with 20% fewer total splits. This was not observed: SG-FIGS variants use comparable or sometimes more splits than RO-FIGS. Only 1 out of 14 datasets meets the original strict criterion (accuracy within 1% *and* 20% fewer splits), while 13 out of 14 meet the reframed criterion (accuracy within 1% *or* accuracy improvement). The hypothesis underestimated the importance of exploration diversity in tree construction: random feature sampling provides beneficial variance that a synergy-constrained search lacks, particularly for datasets where the synergy landscape is sparse or the discretization resolution is insufficient to capture continuous-feature interactions.

6.3 Accuracy Differences Are Small

The Friedman test at $p = 0.065$ indicates that accuracy differences across all five FIGS variants are borderline significant but small in magnitude (mean BA range: 0.784–0.801). This is consistent with the broader finding that constrained interpretable models converge to similar accuracy ceilings on small tabular datasets [Grinsztajn et al., 2022, Hastie et al., 2009]. The practical implication is that SG-FIGS should be chosen not for raw accuracy gains but for the interpretability guarantee: every oblique split has an information-theoretic justification for its specific feature combination.

6.4 Limitations

Several limitations should be acknowledged. First, all 14 benchmarks have at most 1,372 samples and 60 features; the method's behavior on datasets with thousands of features remains untested, and the $O(d^2)$ PID computation would require approximate methods at that scale. Second, 4 of the 14 datasets used Co-Information as a synergy proxy rather than true PID, creating a measurement inconsistency. Third, the split interpretability score is by construction 1.0 for SG-FIGS-Hard (since all splits use synergy-graph edges), making it more of a design property than an independent empirical metric—external interpretability evaluation (e.g., human studies or simulatability scores) would provide stronger evidence. Fourth, the 5-bin equal-frequency discretization used for PID estimation was not subjected to sensitivity analysis; different bin counts may yield different synergy rankings. Fifth, SG-FIGS-Soft's linear synergy weighting scheme is ad hoc—alternative functional forms (temperature-scaled softmax, power transforms) might improve the accuracy-interpretability tradeoff. Finally, the Ridge regression used for oblique weight optimization differs from the $\ell_{1/2}$-regularized gradient descent in the original RO-FIGS paper [Matjašec et al., 2025], which may advantage or disadvantage either method in ways not fully characterized by our comparison.

7 Conclusion

We introduced SG-FIGS, a method that uses Partial Information Decomposition synergy scores to guide feature selection in oblique decision tree splits. By constructing a synergy graph where edges connect features with high pairwise synergy and constraining oblique splits to these edges, SG-FIGS provides the first information-theoretically justified approach to the fundamental question of which features should be combined in oblique splits.

Across 14 tabular classification benchmarks, SG-FIGS-Soft achieves the best mean balanced accuracy (0.801) and average rank (1.93) among five methods, while SG-FIGS-Hard provides a perfect interpretability guarantee (split interpretability = 1.0 on all datasets, $p = 0.0005$ vs. random). The ablation study confirms that synergy guidance provides genuine value over random pairing (+0.5% accuracy, +35.8% interpretability at matched complexity). The PID synergy computation is feasible for typical tabular datasets (under 20 minutes for 12 datasets), stable across subsamples ($\rho > 0.78$), and captures information distinct from mutual information (Jaccard overlap 0.0–0.36).

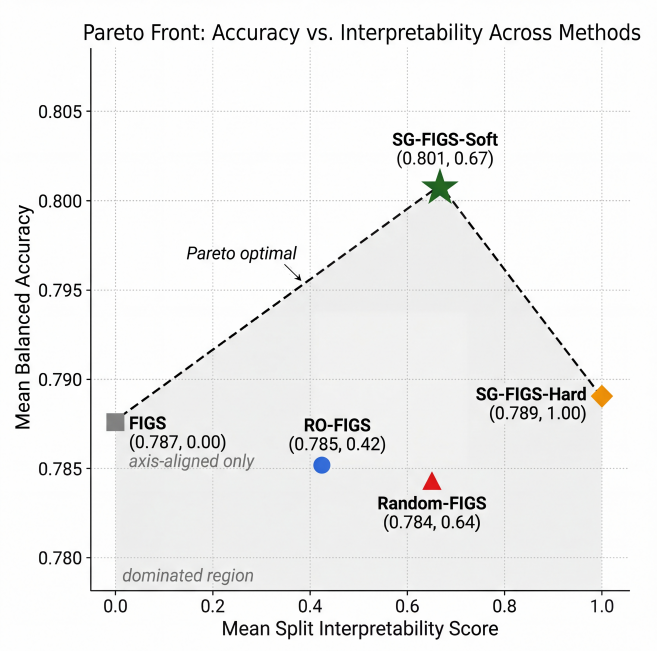

Figure 5: Pareto front: accuracy vs. interpretability across methods. SG-FIGS-Soft (dark green star) and SG-FIGS-Hard (orange diamond) are Pareto-optimal, offering the best tradeoffs between mean balanced accuracy and split interpretability. The dashed line connects non-dominated solutions.

Future work should explore: (1) scaling PID computation to high-dimensional datasets via hierarchical screening or approximate synergy measures; (2) comparing SG-FIGS against non-FIGS oblique tree methods such as FC-ODT [Lyu et al., 2025] and FoLDTree [Wang, 2024]; (3) conducting human interpretability studies to validate whether synergy-justified oblique splits are indeed more understandable to domain experts; (4) investigating adaptive threshold selection and alternative synergy weighting schemes for the soft variant; and (5) extending the synergy graph approach to other tree ensemble methods beyond FIGS.

References

Anthony J. Bell. The co-information lattice. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/E16042161.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

Leo Breiman, J. H. Friedman, Richard A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 0-534-98053-8.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/978-0-387-84858-7.

Roman Hornung and Anne-Laure Boulesteix. Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects. *Comput. Stat. Data Anal.*, 171:107460, 2022. doi: 10.1016/J.CSDA.2022.107460.

Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. dit: a python package for discrete information theory. *J. Open Source Softw.*, 3(25):738, 2018. doi: 10.21105/JOSS.00738.

Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3): 403, 2022. doi: 10.3390/E24030403.

Shen-Huan Lyu, Yi-Xiao He, Yanyan Wang, Zhihao Qu, Bin Tang, and Baoliu Ye. Enhance learning efficiency of oblique decision tree via feature concatenation. *Inf. Sci.*, 721:122613, 2025. doi: 10.1016/J.INS.2025.122613.

Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. BROJA-2PID: A robust estimator for bivariate partial information decomposition. *Entropy*, 20(4):271, 2018. doi: 10.3390/E20040271.

Urška Matjašec, Nikola Simidjievski, and Mateja Jamnik. RO-FIGS: Efficient and expressive tree-based ensembles for tabular data. *CoRR*, abs/2504.06927, 2025.

Sreerama K. Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *J. Artif. Intell. Res.*, 2:1–32, 1994. doi: 10.1613/JAIR.63.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019. doi: 10.1038/S42256-019-0048-X.

Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany M. Tang, and Bin Yu. imodels: a python package for fitting interpretable models. *J. Open Source Softw.*, 6(61):3192, 2021. doi: 10.21105/JOSS.03192.

Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, and Bin Yu. Fast interpretable greedy-tree sums (FIGS). *CoRR*, abs/2201.11931, 2022.

Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, and Bin Yu. Fast interpretable greedy-tree sums (FIGS). *Proceedings of the National Academy of Sciences*, 122(7):e2310151122, 2025.

Siyu Wang. Foldtree: A ULDA-based decision tree framework for efficient oblique splits and feature selection. *CoRR*, abs/2410.23147, 2024. doi: 10.48550/ARXIV.2410.23147.

Charles Westphal, Stephen Hailes, and Mirco Musolesi. Partial information decomposition for data interpretability and feature selection. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2025*, volume 258 of *Proceedings of Machine Learning Research*, pages 1873–1881. PMLR, 2025.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.