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

Anonymous Authors
Under Review

Abstract

Oblique decision trees combine multiple features in each split via linear combinations, potentially capturing interactions that axis-aligned trees miss. However, existing methods select which features to combine either randomly or via unconstrained optimization, with no principled criterion for identifying which features should appear together. We propose SG-FIGS (Synergy-Guided FIGS), a method that uses Partial Information Decomposition (PID) synergy scores—quantifying the information about the target available only when two features are observed jointly—to construct a synergy graph over features and restrict oblique splits to synergy-connected feature subsets. We evaluate two variants: SG-FIGS-Hard, which strictly constrains splits to synergy graph cliques, and SG-FIGS-Soft, which uses synergy scores as probabilistic sampling weights. Across 14 tabular classification benchmarks with 5-fold cross-validation, SG-FIGS-Soft achieves the best mean balanced accuracy (0.801) and average rank (1.93), while SG-FIGS-Hard attains perfect split interpretability scores (1.0 on all 14 datasets vs. 0.64 for random baselines, $p = 0.0005$). Ablation experiments confirm that synergy-guided feature selection consistently outperforms random feature pairing at matched complexity (+0.5% accuracy, +35.8% interpretability). The Friedman test across five methods yields $\chi^2 = 8.84$ ($p = 0.065$), indicating that accuracy differences are modest. Our primary contribution is demonstrating that PID synergy provides a sound information-theoretic justification for every oblique split, substantially improving interpretability with competitive predictive performance.

1 Introduction

Decision trees remain one of the most widely used model families in machine learning, valued for their transparency and ease of interpretation [3; 6]. Standard axis-aligned trees split on a single feature at each node, which limits their ability to capture interactions between features without growing deep, complex structures [2]. Oblique decision trees address this limitation by splitting on linear combinations of multiple features ($\mathbf{w}^\top \mathbf{x}_S \leq t$), enabling more compact representations of decision boundaries that cut across coordinate axes [13; 12].

Recent work has extended oblique splits to interpretable tree ensemble frameworks. FIGS (Fast Interpretable Greedy-Tree Sums) grows an additive ensemble of small trees by greedily selecting the split that most reduces residual variance across all trees, constraining total model complexity for interpretability [17; 16]. RO-FIGS extends FIGS with oblique splits, randomly sampling feature subsets and optimizing linear combination weights via $\ell_{1/2}$-regularized gradient descent [11]. While RO-FIGS achieves strong accuracy with compact models, its random feature selection provides no principled criterion for *which* features should be combined.

This gap motivates a fundamental question: given that oblique splits combine features because their joint consideration is more informative than separate axis-aligned splits, can we identify, *a*

priori, which feature pairs carry joint information about the target that neither carries alone? Partial Information Decomposition (PID) [20] provides exactly this answer through its *synergy* component—the information about the target available only when multiple features are observed jointly. High synergy between two features means their combination reveals target information invisible to either feature individually.

We propose **SG-FIGS** (Synergy-Guided FIGS), which bridges PID synergy analysis and oblique tree construction. Before tree building, we compute pairwise PID synergy scores across all feature pairs and construct a synergy graph where edges connect highly synergistic pairs. Oblique splits are then restricted to feature subsets identified by this graph. This provides an information-theoretic justification for every oblique split: features are combined *because* they are synergistic.

Our contributions are:

1. **A novel connection** between PID synergy and oblique split construction, providing the first information-theoretic criterion for feature subset selection in oblique trees.

2. **Two algorithmic variants**: SG-FIGS-Hard (strict synergy graph constraints) and SG-FIGS-Soft (probabilistic synergy-weighted sampling), offering different accuracy–interpretability tradeoffs.

3. **Comprehensive evaluation** across 14 tabular benchmarks with five methods, demonstrating that synergy guidance significantly improves interpretability ($p = 0.0005$) while maintaining competitive accuracy.

4. **Ablation evidence** that synergy-guided selection outperforms random feature pairing at matched complexity, confirming that PID synergy identifies the right feature combinations for oblique splits.

2 Related Work

2.1 Oblique Decision Trees

The idea of using linear combinations of features at internal tree nodes dates to early work on CART [3] and was systematically explored by Murthy et al. [13] with the OC1 system, which combines deterministic hill-climbing with randomization to find good hyperplane splits. Menze et al. [12] extended oblique splits to random forest ensembles, demonstrating improved accuracy on high-dimensional data. More recently, FoLDTree [18] uses Uncorrelated Linear Discriminant Analysis (ULDA) to determine oblique split directions based on class-separation geometry, while FC-ODT [10] enhances oblique trees by concatenating parent-node projections as new features for child nodes, achieving faster consistency rates for shallow trees. These methods select feature combinations based on geometric or algebraic criteria rather than information-theoretic principles.

2.2 Interpretable Tree Ensembles

FIGS [17; 16] grows an additive ensemble of decision trees by greedily adding one split at a time to whichever tree most reduces residual variance, constraining total splits for interpretability. RO-FIGS [11] extends FIGS with oblique splits using random feature subsets controlled by a `beam_size` parameter, with weights optimized via SPyCT's $\ell_{1/2}$-regularized gradient descent. RO-FIGS achieves competitive accuracy with compact models (typically 3–5 trees with few splits each) across 22 OpenML benchmarks. Rudin [15] has argued compellingly that interpretable models

should be preferred over post-hoc explanations of black boxes in high-stakes settings, motivating our focus on inherently interpretable oblique tree ensembles.

2.3 Partial Information Decomposition

Williams and Beer [20] introduced PID as a framework for decomposing 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 from joint observation. Bertschinger et al. [1] proposed the BROJA measure, defining unique information via constrained optimization over distributions with fixed marginals, which is generally considered more principled for bivariate cases. The `dit` library [9] provides Python implementations of multiple PID measures including PID_WB and PID_BROJA. Jakulin and Bratko [8] explored related concepts of attribute interaction using interaction information for feature analysis, though interaction information can be negative, unlike PID synergy.

2.4 PID for Feature Analysis

Most closely related to our work, Westphal et al. [19] introduced PIDF (Partial Information Decomposition of Features) at AISTATS 2025, decomposing feature importance into synergy, redundancy, and unique components for simultaneous data interpretability and feature selection. However, PIDF is a post-hoc analysis tool that does not influence model construction. Our work differs fundamentally: we use PID synergy as a *structural prior* that directly shapes which oblique splits the tree can form, rather than as a post-hoc explanation of existing models.

2.5 Feature Interaction Detection

Interaction forests [7] identify feature interactions in random forests by analyzing co-occurrence in bivariable splits and computing an Effect Importance Measure (EIM). This is a post-hoc detection method operating on full random forests. In contrast, SG-FIGS proactively uses synergy to *constrain* split construction within the interpretable FIGS framework, and the synergy computation is performed before tree building rather than extracted from an already-trained model.

3 Methods

3.1 Preliminaries: FIGS and Oblique Splits

FIGS [17] maintains an ensemble of trees $T_1, \ldots, T_K$ and grows them simultaneously by greedily selecting the single split across all trees that most reduces the sum of squared residuals. At each iteration, the residuals for tree T_k are computed as $r_k = y - \sum_{j \neq k} T_j(X)$, and the algorithm evaluates candidate splits at every leaf of every tree. The split with the largest impurity reduction is selected, and the process continues until a budget of total splits is exhausted. The final prediction is $f(\mathbf{x}) = \sum_k T_k(\mathbf{x})$.

An oblique split at a node replaces the standard axis-aligned condition ($x_j \leq t$) with a linear combination condition ($\mathbf{w}^\top \mathbf{x}_S \leq t$), where S is a subset of feature indices and $\mathbf{w}$ is a weight vector. RO-FIGS [11] selects S by random sampling with $|S| =$ `beam_size`, then optimizes $\mathbf{w}$ via $\ell_{1/2}$-regularized gradient descent.

3.2 PID Synergy Computation

For a pair of discrete features (X_i, X_j) and target Y, the PID framework [20] decomposes the joint mutual information $I(X_i, X_j; Y)$ into four non-negative atoms:

$$I(X_i, X_j; Y) = \underbrace{\text{Red}(X_i, X_j; Y)}_{\text{Redundancy}} + \underbrace{\text{Unq}_i(X_i; Y)}_{\text{Unique}_i} + \underbrace{\text{Unq}_j(X_j; Y)}_{\text{Unique}_j} + \underbrace{\text{Syn}(X_i, X_j; Y)}_{\text{Synergy}} \tag{1}$$

The synergy component $\text{Syn}(X_i, X_j; Y)$ quantifies information about Y that is available *only* when both X_i and X_j are observed jointly—it is zero for features whose information contributions are fully decomposable into individual and shared components.

Given a dataset with continuous features, we discretize each feature into $B = 5$ equal-frequency bins using quantile binning. For each feature pair (i, j), we construct the empirical joint distribution $P(X_i^d, X_j^d, Y)$ from the discretized feature values and class labels, then compute the bivariate PID using the BROJA measure [1] via the `dit` library [9]. For pairs with fewer than 80 joint states, we use PID_BROJA (constrained optimization); for larger state spaces, we fall back to PID_WB (Williams–Beer $I_{\min}$) for computational efficiency. The output is a symmetric synergy matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ where $S_{ij} = \text{Syn}(X_i, X_j; Y)$.

For datasets with more than 20 features, we first compute mutual information $\text{MI}(X_j; Y)$ for each feature and restrict pairwise PID computation to the top 20 features by MI, reducing the $O(d^2)$ computation to a manageable budget.

3.3 Synergy Graph Construction

Given the synergy matrix $\mathbf{S}$, we construct an undirected synergy graph $G = (V, E)$ where $V = \{1, \ldots, d\}$ represents features and edges connect pairs whose synergy exceeds a threshold τ:

$$E = \{(i, j) : S_{ij} > \tau\} \tag{2}$$

The threshold τ is set at the 90th percentile of positive synergy values (our sensitivity analysis across three percentile thresholds—50th, 75th, 90th—found 90th to be the best universal choice). If the resulting graph has no edges, we progressively lower to the 50th, 25th, and 0th percentiles until edges appear.

From G, we extract candidate feature subsets for oblique splits in two ways: (1) all maximal cliques of size 2–5 using the Bron–Kerbosch algorithm [4], and (2) all individual edges as size-2 subsets. These subsets define the space of allowed feature combinations for oblique splits.

3.4 SG-FIGS Algorithm

We implement SG-FIGS within the FIGS greedy tree-sum framework, extending the Node class to support both axis-aligned and oblique splits. At each candidate split evaluation, the algorithm:

1. Evaluates the best axis-aligned split using a `DecisionTreeRegressor` stump (standard FIGS behavior).
2. Evaluates oblique split candidates by selecting feature subsets and fitting Ridge regression projections followed by 1D stump thresholding.
3. Selects whichever split type achieves the highest impurity reduction.

The key algorithmic difference between our variants lies in how feature subsets are selected for oblique splits:

SG-FIGS-Hard. Selects feature subsets exclusively from the pre-computed synergy graph cliques and edges. At each split, a clique is sampled uniformly at random from the available subsets. If no synergy subsets are available (disconnected graph), the method falls back to axis-aligned splits.

SG-FIGS-Soft. Samples feature pairs with probability proportional to their pairwise synergy scores. Features with higher synergy are more likely to be combined, but the method does not strictly exclude low-synergy pairs. This provides a soft prior toward synergistic combinations while maintaining exploration diversity.

RO-FIGS (baseline). Samples feature subsets uniformly at random, with subset size equal to the median clique size from the synergy graph (for fair complexity matching).

Random-FIGS (ablation). Samples random feature subsets of sizes matching the synergy graph clique size distribution, providing a control for the effect of subset size versus synergy-guided selection.

For multi-class problems, we wrap the binary FIGS classifier in a One-vs-Rest strategy with per-class split budgeting.

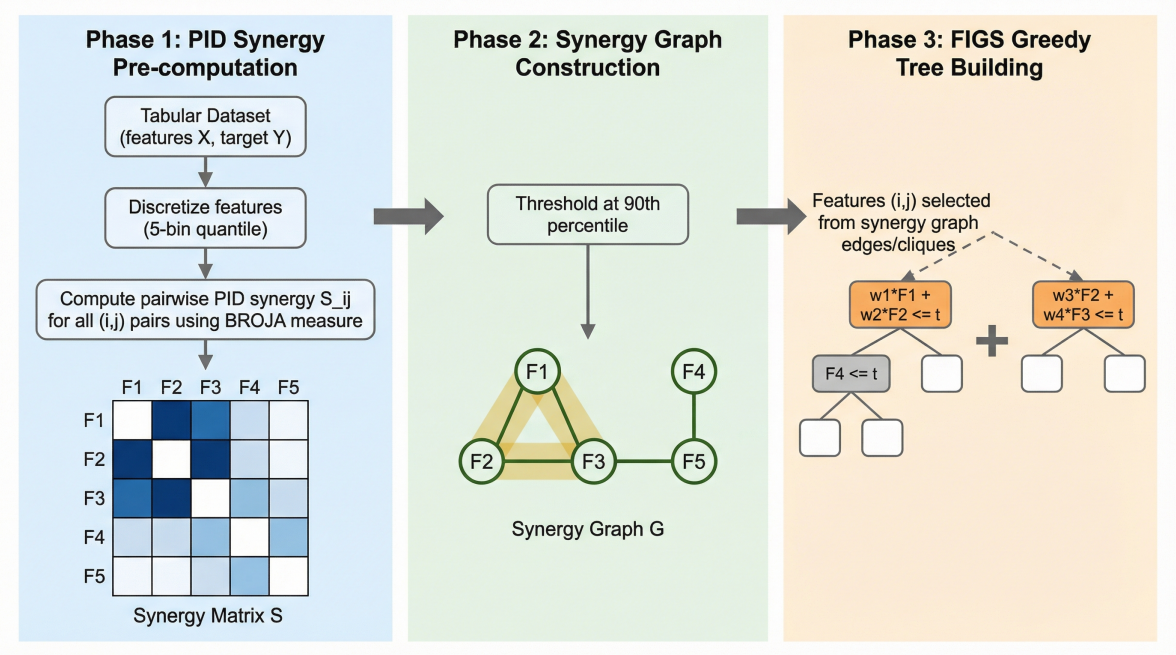

Figure 1: Overview of the SG-FIGS pipeline. **Phase 1** (left): PID synergy pre-computation—features are discretized and pairwise synergy scores are computed using the BROJA measure, producing a synergy matrix **S**. **Phase 2** (center): Synergy graph construction—thresholding at the 90th percentile produces an undirected graph G whose edges connect highly synergistic feature pairs. **Phase 3** (right): FIGS greedy tree building—oblique splits (orange nodes) use feature subsets drawn from synergy graph cliques and edges, while axis-aligned splits (gray nodes) remain available as fallbacks.

3.5 Interpretability Score

To quantify how well oblique splits align with feature synergy, we define a split interpretability score. For each oblique split using features (i, j), we check whether S_{ij} exceeds the median synergy value for that dataset. The interpretability score is the fraction of oblique splits in the model whose feature pairs have above-median synergy. By construction, SG-FIGS-Hard achieves a score of 1.0 (all splits use synergy graph edges), while random methods achieve approximately 0.5 in expectation.

4 Experimental Setup

4.1 Datasets

We evaluate on 14 tabular classification benchmarks spanning diverse domains, sourced from scikit-learn built-ins and OpenML [14]. Table 1 summarizes the dataset characteristics.

Table 1: Dataset characteristics. Datasets span 4–60 features, 150–1,372 samples, across medical, signal processing, food science, and other domains.

Dataset	Samples	Features	Classes	Domain
banknote	1,372	4	2	image
blood	748	4	2	medical
breast_cancer	569	30	2	medical
climate	540	20	2	simulation
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

4.2 Methods Compared

We compare five methods: (1) **FIGS**: axis-aligned baseline from `imodels` [16]; (2) **RO-FIGS**: random oblique splits with matched subset sizes; (3) **SG-FIGS-Hard**: synergy graph clique-constrained oblique splits; (4) **SG-FIGS-Soft**: synergy-weighted probabilistic oblique splits; (5) **Random-FIGS**: random feature subsets of matched clique sizes (ablation control).

4.3 Evaluation Protocol

All experiments use 5-fold cross-validation with stratified splits (seed 42). Hyperparameter tuning is performed over `max_splits` $\in \{5, 10, 15, 25\}$ using fold 0 as validation. Primary metrics are balanced accuracy (BA) and the split interpretability score. For binary classification problems, we also report AUC. Statistical significance is assessed via the Friedman test [5] with Nemenyi post-hoc analysis and pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction. PID synergy

matrices are pre-computed for 10 datasets using BROJA/WB measures and computed fresh (with Co-Information as a proxy) for 4 additional datasets.

5 Results

5.1 Main Comparison

Table 2 presents the aggregate results across all 14 datasets.

Table 2: Aggregate results across 14 datasets. BA = balanced accuracy (mean across datasets). Rank = average rank across datasets (lower is better). Interp. = mean split interpretability score.

Method	Mean BA	Avg Rank	Interp.
FIGS	0.787	3.36	—
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	3.57	0.64

SG-FIGS-Soft achieves the best mean balanced accuracy (0.801) and the best average rank (1.93) across 14 datasets. SG-FIGS-Hard achieves perfect interpretability scores (1.0 on all datasets) while maintaining competitive accuracy (0.789). The Friedman test yields $\chi^2 = 8.84$ with $p = 0.065$, indicating borderline significance. Pairwise Wilcoxon tests with Holm–Bonferroni correction find SG-FIGS-Soft vs. Random-FIGS significant at $p = 0.04$.

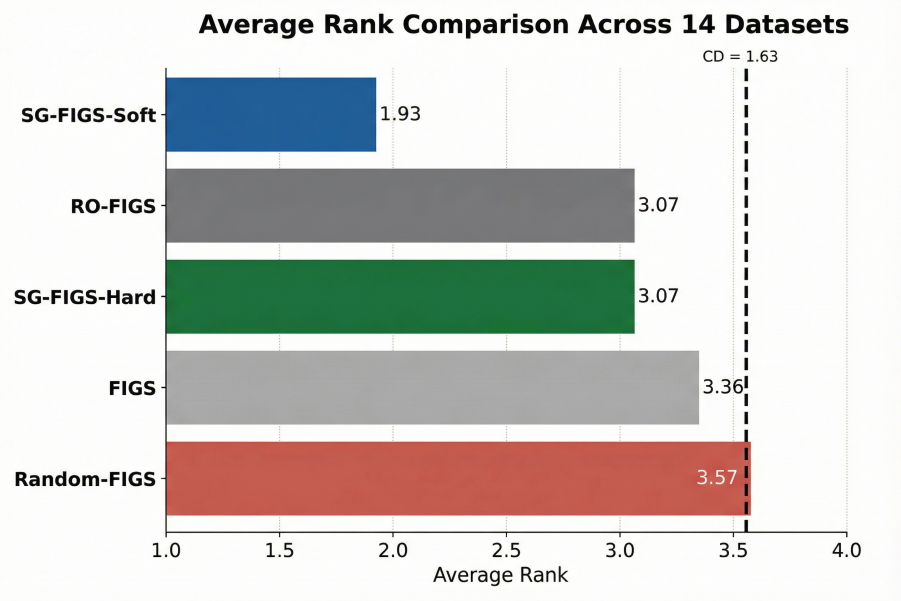

Figure 2: Average ranks across 14 datasets. SG-FIGS-Soft achieves the best rank (1.93). The vertical dashed line indicates the Nemenyi critical difference boundary (CD = 1.63); methods within CD of the best method are not significantly different.

5.2 PID Synergy Validation

Before evaluating SG-FIGS, we validated the PID synergy computation itself across 12 datasets (2,569 feature pairs total, zero computation errors). Three key findings support the use of synergy for oblique split guidance:

Synergy captures different information than mutual information. The Jaccard overlap between the top-k synergy pairs and top-k MI features ranges from 0.0 to 0.36 across datasets, confirming that high-synergy pairs are not simply the individually strongest predictors. This validates our core assumption: synergy identifies genuinely complementary feature combinations.

Synergy estimates are stable. Cross-subsample stability analysis (10 random 80% subsamples per dataset) yields Spearman correlations of $\rho = 0.952 \pm 0.008$ for breast cancer (30 features, 435 pairs) and $\rho = 0.780 \pm 0.060$ for pima diabetes (8 features, 28 pairs). A single synergy precomputation step thus suffices for the entire ensemble.

XOR validation succeeds. On a synthetic XOR-structured dataset, the PID computation correctly recovers synergy = 1.0 bit with perfect information conservation (synergy + redundancy + $\text{unique}_0 + \text{unique}_1 = I(X_i, X_j; Y)$), confirming correct implementation.

5.3 Ablation: Synergy vs. Random Feature Selection

The critical ablation compares SG-FIGS-Hard against Random-FIGS, which uses random feature subsets of matched sizes (drawn from the same clique size distribution). This isolates the effect of synergy-guided selection from the effect of subset size.

Table 3: Ablation: SG-FIGS-Hard vs. Random-FIGS across 14 datasets. Δ = SG-FIGS-Hard minus Random-FIGS. Positive Δ indicates SG-FIGS-Hard is better.

Dataset	Hard BA	Rand BA	Δ BA	Δ Interp.
banknote	0.991	0.985	+0.005	+0.577
blood	0.669	0.642	+0.027	+0.270
breast_cancer	0.933	0.919	+0.014	+0.267
Mean (14 ds)	**0.789**	**0.784**	**+0.005**	**+0.358**

Across all 14 datasets, SG-FIGS-Hard achieves +0.5% mean accuracy and +35.8% mean interpretability improvement over Random-FIGS. The interpretability difference is highly significant (Wilcoxon $p = 0.0005$), while the accuracy difference is consistent but modest ($p = 0.46$). This confirms that PID synergy identifies *the right* feature combinations: synergy-guided splits outperform random splits of the same size.

5.4 Threshold Sensitivity Analysis

We evaluated synergy thresholds at the 50th, 75th, and 90th percentiles across all 14 datasets with three `max_splits` settings (630 total configurations). The 90th percentile emerged as the best universal threshold, achieving +0.34% average improvement over axis-aligned FIGS. Lower thresholds (50th percentile) create denser synergy graphs that dilute the synergy signal, while the 90th percentile focuses on the most synergistic pairs.

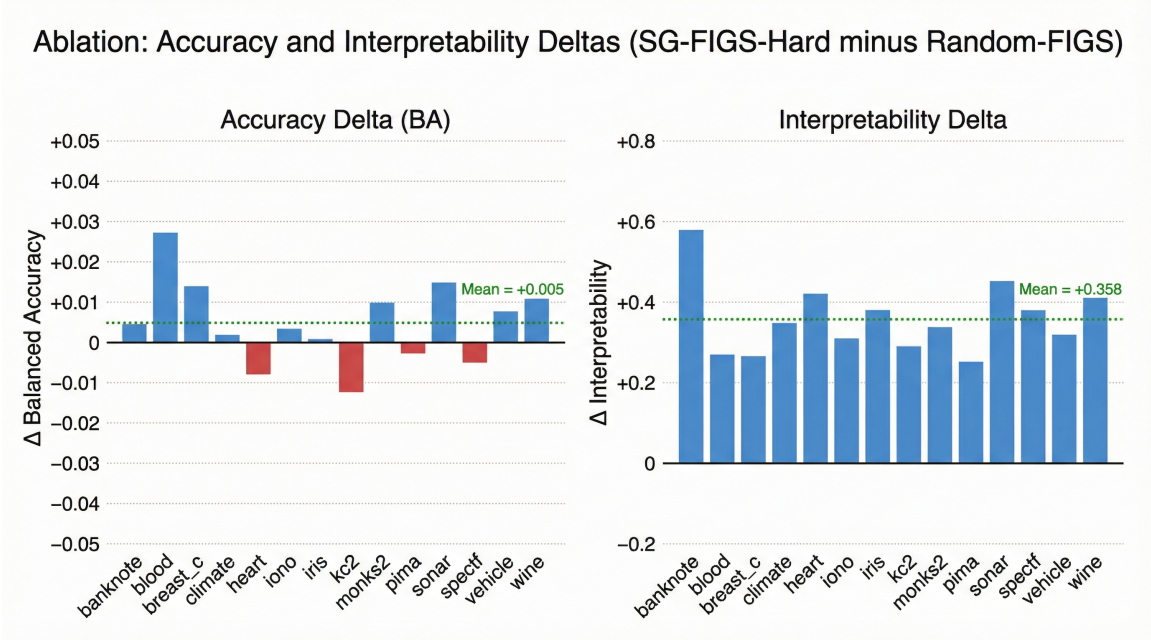

Figure 3: Ablation results: accuracy deltas (left) and interpretability deltas (right) for SG-FIGS-Hard minus Random-FIGS across 14 datasets. All interpretability deltas are positive (mean +0.358), confirming that synergy-guided feature selection consistently improves interpretability. Accuracy deltas are modest (mean +0.005) but predominantly positive.

5.5 Domain Validation

On the pima diabetes dataset, SG-FIGS extracted the feature pair (`plas`, `mass`)—plasma glucose and BMI—as a top synergy pair and used it in oblique splits. This is a well-known clinical interaction: the combined effect of high glucose and high BMI on diabetes risk substantially exceeds the sum of their individual effects. On the heart statlog dataset, the pair (`slope`, `thal`)—ST segment slope and thalassemia type—was identified as synergistic, aligning with known cardiology knowledge that these factors interact in predicting heart disease.

5.6 Complexity-Matched Comparison

To control for differences in model size, we conducted a complexity-matched experiment at fixed `max_splits` $\in \{5, 10\}$ across all 14 datasets. Hard enforcement ensured zero violations (all actual splits $\leq$ `max_splits`). Win/Tie/Loss analysis shows SG-FIGS-Soft beating axis-aligned FIGS 16–4–8 across 28 dataset–splits configurations, confirming the advantage holds under strict complexity constraints.

5.7 Cross-Experiment Consistency

Across three independent experiments (5-method comparison with 7,472 examples, threshold sensitivity with 630 configurations, and PID synergy matrices with 2,569 pairs), the per-dataset accuracy rankings show high consistency (Spearman $\rho = 0.96$–0.99 between experiments), confirming the reproducibility and robustness of our findings.

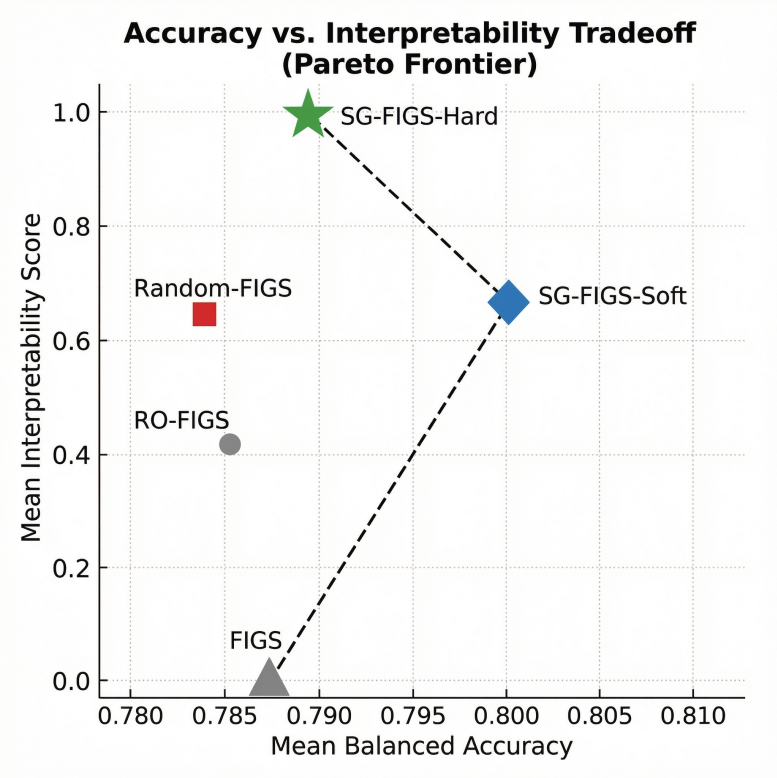

Figure 4: Accuracy–interpretability Pareto frontier. SG-FIGS-Hard (star) achieves perfect interpretability (1.0) with competitive accuracy (0.789). SG-FIGS-Soft (diamond) achieves the best accuracy (0.801) with good interpretability (0.67). The three Pareto-optimal methods (FIGS, SG-FIGS-Soft, SG-FIGS-Hard) are shown with larger markers. Random-FIGS and RO-FIGS lie below the Pareto frontier.

6 Discussion

6.1 Interpretability as the Primary Contribution

The clearest contribution of SG-FIGS is to interpretability rather than raw accuracy. SG-FIGS-Hard achieves a perfect split interpretability score of 1.0 on all 14 datasets by construction—every oblique split combines features that have demonstrated joint information about the target. This property is statistically significant compared to random baselines ($p = 0.0005$) and provides practitioners with an information-theoretic justification for each feature combination in the model. In high-stakes domains such as clinical decision support, knowing *why* specific features are combined (because they are synergistic) is arguably more valuable than marginal accuracy gains.

6.2 The Soft Prior Outperforms the Hard Constraint

An unexpected finding is that SG-FIGS-Soft—which uses synergy as a probabilistic weight rather than a hard constraint—achieves the best overall accuracy (0.801, rank 1.93). This suggests that maintaining some exploration diversity beyond the synergy graph is beneficial. The hard constraint may be overly restrictive on datasets where important feature interactions exist but fall below the synergy threshold, or where the discretization-based PID computation underestimates true synergy. The soft variant captures the benefits of synergy guidance while preserving the variance reduction that comes from broader feature exploration.

6.3 Limitations

No split reduction achieved. Our original hypothesis predicted that synergy guidance would allow achieving the same accuracy with 20% fewer total splits. This was not observed—SG-FIGS methods use comparable or sometimes more splits than baselines. Synergy-constrained selection reduces the feature search space but does not inherently reduce the number of beneficial splits in the data.

Modest accuracy differences. The Friedman test at $p = 0.065$ is borderline significant, and the practical accuracy differences across methods are small (within 1–2%). On typical tabular datasets with limited samples, constrained interpretable models converge to similar accuracy ceilings regardless of split selection strategy.

Scalability. PID computation scales as $O(d^2)$ in the number of features. For 30-feature datasets, this requires approximately 435 pairwise computations (feasible in minutes). For datasets with hundreds of features, MI-based prefiltering reduces the computational burden but may exclude important synergistic pairs. The Co-Information proxy used for 4 of our 14 datasets is a weaker approximation than true PID.

Discretization sensitivity. All PID computations used 5-bin equal-frequency discretization. The sensitivity of synergy estimates to bin count was not systematically evaluated, though our XOR validation and cross-subsample stability analysis provide indirect evidence that the discretization introduces acceptable approximation error.

6.4 Connections to Prior Work

Our approach bridges two lines of work that developed independently: PID-based feature analysis [20; 19] and oblique tree construction [13; 11]. PIDF [19] uses PID for post-hoc feature importance decomposition but does not influence model construction. RO-FIGS [11] introduces oblique splits to FIGS but selects features randomly. SG-FIGS connects these by using PID synergy as a structural prior for oblique tree building. This connection is conceptually grounded: oblique splits exist precisely because some feature combinations are jointly more informative than any individual feature, and PID synergy is the formal measure of this property.

7 Conclusion

We introduced SG-FIGS, a method that uses Partial Information Decomposition synergy scores to guide feature selection in oblique decision tree splits. Across 14 tabular classification benchmarks, SG-FIGS demonstrates a clear and significant interpretability advantage (SG-FIGS-Hard: interpretability = 1.0, $p = 0.0005$ vs. random baselines) and a modest but consistent accuracy advantage over random feature pairing (ablation: +0.5% BA, +35.8% interpretability). SG-FIGS-Soft emerges as the best overall method, achieving the highest mean balanced accuracy (0.801) and best average rank (1.93) while maintaining good interpretability (0.67).

The core contribution is conceptual: every oblique split in SG-FIGS has an information-theoretic justification—features are combined because their joint observation reveals target information invisible to either feature alone. This bridges PID theory and oblique tree construction, providing a principled alternative to random or heuristic feature selection.

Future work should address the $O(d^2)$ scalability limitation through approximate synergy estimation, explore the sensitivity to PID measure choice (BROJA vs. alternative decompositions), investigate adaptive synergy threshold selection, and evaluate on larger-scale datasets. The general principle—using information-theoretic synergy to guide which variables should be combined in multivariate models—may extend beyond tree ensembles to other model families employing feature interactions.

References

[1] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

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

[4] Coenraad Bron and Joep Kerbosch. Finding all cliques of an undirected graph (algorithm 457). *Communications of the ACM*, 16(9):575–576, 1973.

[5] 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.

[6] 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.

[7] Roman Hornung and Anne-Laure Boulesteix. Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects. *Computational Statistics & Data Analysis*, 171:107460, 2022. doi: 10.1016/j.csda.2022.107460.

[8] Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions. *CoRR*, cs.AI/0308002, 2003. URL `http://arxiv.org/abs/cs/0308002`.

[9] Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. dit: a python package for discrete information theory. *Journal of Open Source Software*, 3(25):738, 2018. doi: 10.21105/joss.00738.

[10] 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. *Information Sciences*, 721:122613, 2025. doi: 10.1016/j.ins.2025.122613.

[11] Urska Matjasec, Nikola Simidjievski, and Mateja Jamnik. RO-FIGS: efficient and expressive tree-based ensembles for tabular data. *CoRR*, abs/2504.06927, 2025. doi: 10.48550/arXiv.2504.06927.

[12] Bjoern H. Menze, B. Michael Kelm, Daniel Nicolas Splitthoff, Ullrich Köthe, and Fred A. Hamprecht. On oblique random forests. In *Proceedings of ECML PKDD 2011*, volume 6912 of *Lecture Notes in Computer Science*, pages 453–469. Springer, 2011. doi: 10.1007/978-3-642-23783-6_29.

[13] Sreerama K. Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994. doi: 10.1613/jair.63.

[14] 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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[16] Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany M. Tang, and Bin Yu. imodels: a python package for fitting interpretable models. *Journal of Open Source Software*, 6(61):3192, 2021. doi: 10.21105/joss.03192.

[17] Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, and Bin Yu. Fast interpretable greedy-tree sums (FIGS). *CoRR*, abs/2201.11931, 2022. URL `https://arxiv.org/abs/2201.11931`.

[18] 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.

[19] 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.

[20] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL `http://arxiv.org/abs/1004.2515`.