
Pareto-Optimal Probabilistic Explanations: Balancing Cognitive Constraints and User Preferences

Louenas Bounia¹

¹LIPN-UMR CNRS 7030., Université Sorbonne Paris Nord, Villetaneuse, France

Abstract

Machine learning classifiers deployed in critical domains require comprehensible explanations. Abductive explanations identify minimal feature sets guaranteeing a decision but suffer from two limitations: exceeding human cognitive limits and ignoring user preferences (actionability, fairness, cost). Probabilistic explanations reduce size via controlled error, while preferred explanations integrate preferences; however, these approaches remain disjoint. We introduce the first framework unifying these paradigms: preferred probabilistic explanations balancing cognitive constraints, probabilistic accuracy, and user preferences. We formulate the problem via weighted scalarization (WPPE) and Pareto optimization (PPPE), prove NP-hardness for decision trees, and exploit supermodularity to establish approximation guarantees. We propose three complementary algorithms: weighted greedy descent (WGD) with ratio $(e^{p_w} - 1)/p_w$, Pareto frontier enumeration (PFE) for interactive exploration of non-dominated trade-offs, and lexicographic stratified algorithm (LSA) for strict ordinal preferences. Our framework enables rigorous navigation of the Pareto-optimal trade-off space between cognitive limits and user preferences.

1 INTRODUCTION

1.1 MOTIVATION AND CONTEXT

Machine learning classifiers are increasingly deployed in high-stakes domains such as healthcare [Esteva et al., 2017], criminal justice [Dressel and Farid, 2018], loan approval, and employment decisions [Raghavan et al., 2020]. These models—including decision trees, random forests, neural networks—have demonstrated remarkable predictive perfor-

mance across diverse applications.

However, with deployment in critical sectors, the ability to explain model decisions has become vital for transparency, trust, and regulatory compliance (GDPR’s right to explanation, EU AI Act). This need has driven the development of Explainable AI (XAI), with popular post-hoc agnostic methods such as LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017], and counterfactual explanations [Wachter et al., 2018] gaining widespread adoption. These methods rely on local perturbations or game theory without considering model structure.

Despite their popularity, agnostic methods suffer from fundamental limitations: they can generate identical explanations for opposite predictions [Slack et al., 2020, Ignatiev et al., 2019, Huang and Marques-Silva, 2024], lack rigorous theoretical foundations [Kumar et al., 2019], and exhibit sensitivity to input perturbations [Alvarez-Melis and Jaakkola, 2018]. These weaknesses undermine reliability in critical contexts where decisions have significant consequences.

Our work lies within Formal XAI [Marques-Silva, 2023], providing mathematically validated explanations with guaranteed faithfulness and robustness. The most prominent form are *abductive explanations*, which identify minimal feature sets where fixing their values guarantees the same prediction regardless of other features. Each feature is necessary, and collectively sufficient to justify the decision.

However, two interdependent challenges emerge:

Challenge 1: Cognitive Overload. Even minimal explanations can contain too many features to be cognitively processed. Miller [1956] established that humans can simultaneously process only 7 ± 2 items, a limit later refined to as low as 4 items for complex reasoning tasks [Cowan, 2001].

Challenge 2: Preference Heterogeneity. Not all features are equally relevant, and the same explanation may suit one user but not another. Explanation utility depends on user profile, expertise, and decision context, particularly regarding

interpretability, actionability, and ethical constraints [Miller, 2019]. Ignoring this heterogeneity can lead to explanations that are formally correct but uninformative for end users, as illustrated by the following scenarios:

- **Scenario A (Medical diagnosis):** A non-expert patient prefers explanations with common medical terms, while a physician is comfortable with this terminology.
- **Scenario B (Loan application):** Certain features, such as *age* > 50, are non-actionable, while others, like *debt-to-income ratio* > 0.6, can be modified.
- **Scenario C (Cost-sensitive applications):** In industrial / medical contexts, certain variables rely on expensive sensors or tests, limiting their use in explanations.

1.2 THE GAP IN CURRENT RESEARCH

Existing work addresses conciseness and comprehensibility disjointly. **Probabilistic explanations** [Izza et al., 2022a, Wäldchen, 2022, Izza et al., 2024b, Koriche et al., 2025, Wang et al., 2021, Bounia and Koriche, 2023, Bounia and Setitra, 2025b, Arenas et al., 2022] relax exactness via controlled error ϵ , reducing size to respect cognitive limits, but treat all features uniformly without integrating user preferences. Conversely, **preferred explanations** [Audemard et al., 2022b] integrate preferences (cost, actionability, fairness) [Miller, 2019, Sokol and Flach, 2020, Langer et al., 2021] but require $\epsilon = 0$, yielding long explanations poorly adapted to diverse stakeholder needs [Langer et al., 2021].

1.3 CONTRIBUTIONS

We propose the first principled framework unifying probabilistic explanations and user preferences. Our approach computes explanations that are simultaneously **short** (via controlled error ϵ), **precise** (with explicit probabilistic guarantees), and **preference-aligned** (optimizing utility).

Algorithmic contributions. We develop three complementary algorithms exploiting supermodularity: **Weighted Greedy Descent (WGD)** optimizes a scalarized objective with approximation ratio $(e^{p_w} - 1)/p_w$; **Pareto Frontier Enumeration (PFE)** efficiently explores non-dominated trade-offs without exhaustive search; and **Lexicographic Stratified Algorithm (LSA)** handles ordinal preferences via priority levels with polynomial-time guarantees.

Theoretical contributions. We establish: (i) NP-hardness of WPPE for decision trees, (ii) preservation of supermodularity under preference weighting—crucial for our approximation guarantees, and (iii) polynomial-time tractability for decision trees despite general hardness, enabling practical implementation.

2 PRELIMINARIES

Classification problems. We consider Boolean binary classification with d features. Let $[d] = \{1, \dots, d\}$ denote feature indices and $X_d = \{x_1, \dots, x_d\}$ the feature set. Each instance $\mathbf{x} = (x_1, \dots, x_d) \in \{0, 1\}^d$ is classified by $h : \{0, 1\}^d \rightarrow \{0, 1\}$. We assume the reader’s familiarity with supervised learning, decision trees.

Abductive and Probabilistic Explanations. An *abductive explanation*¹ for $\mathbf{x}$ is a feature subset $S \subseteq [d]$ such that any instance $\mathbf{y}$ agreeing with $\mathbf{x}$ on S satisfies $h(\mathbf{y}) = h(\mathbf{x})$. A *sufficient reason* is a minimal abductive explanation. Since these can be long, *probabilistic explanations* [Arenas et al., 2022, Wang et al., 2021] relax correctness via controlled error. The *error function* $\mu_{h,\mathbf{x}}(S)$ counts instances $\mathbf{y}$ where $y_S = x_S$ but $h(\mathbf{y}) \neq h(\mathbf{x})$. The *normalized error*

$$\epsilon_{h,\mathbf{x}}(S) = \frac{\mu_{h,\mathbf{x}}(S)}{2^{d-|S|}}$$

measures the proportion of incompatible completions. A set S is a *probabilistic explanation of level ϵ* if $\epsilon_{h,\mathbf{x}}(S) \leq \epsilon$.

Preferred Explanations. Not all explanations are equally useful. *Preferred explanations* [Audemard et al., 2022b] account for user preferences (actionability, cost, constraints) [Miller, 2019, Ustun et al., 2019] via feature weightings or ordinal preferences. We use additive costs: x_i receives weight $w(x_i) \in \mathbb{R}^+$, yielding $\text{cost}(S) = \sum_{i \in S} w(x_i)$.

Submodular Functions. A function $f : 2^{[d]} \rightarrow \mathbb{R}$ is *monotone* if $f(S) \leq f(T)$ (increasing) or $f(S) \geq f(T)$ (decreasing) for all $S \subseteq T$. It is *supermodular* if the marginal loss $L_f(i | S) = f(S) - f(S \setminus \{i\})$ satisfies $L_f(i | S) \geq L_f(i | T)$ for all $S \subseteq T$ and $i \in S$ —intuitively, removing an element impacts smaller sets more. A function is *submodular* if the reverse holds; *modular* if contributions are additive. The *curvature* measures deviation from modularity and is crucial for approximation guarantees. For non-negative f and non-empty $I \subseteq [d]$: $c_f(I) = 1 - \min_{i \in I} \frac{f(I) - f(I \setminus \{i\})}{f(\{i\}) - f(\emptyset)}$. We have $c_f(I) \in [0, 1]$, where $c_f(I) = 0$ indicates modularity and $c_f(I) = 1$ maximal sub/supermodularity. For decreasing supermodular f , $c_f(I) < 1$ guarantees constant-factor approximation under cardinality constraints [Iyer and Bilmes, 2013, Sviridenko et al., 2017].

2.1 RELATED WORK

Our work is part of a line of research known as *formal explainable AI* (formal XAI) Marques-Silva [2023], Darwiche

¹Unlike Ignatiev et al. [2019], we do not restrict to minimal abductive explanations under set inclusion.

[2023], Darwiche and Hirth [2020], which investigates explanations equipped with provable guarantees Shih et al. [2018], Audemard et al. [2021], Azzolin et al. [2025], Bassan and Katz [2022], Bassan et al. [2026b].

Abductive and Probabilistic Explanations. Darwiche and Hirth [2020] characterized sufficient reasons for various model classes, with efficient algorithms for decision trees [Izza et al., 2020, Arenas et al., 2022, Bounia and Setitra, 2025b], random forests [Izza and Marques-Silva, 2021, Izza et al., 2024a], linear models [Marques-Silva et al., 2020, Subercaseaux et al., 2025], tree ensembles [Audemard et al., 2024], and neural networks [Bassan and Katz, 2022, Wu et al., 2023]. On the probabilistic side, Wäldchen et al. [2021] proved NP^{PP} -completeness for Boolean circuits, Arenas et al. [2022] showed NP-completeness for decision trees, and Bounia and Koriche [2023] exploited submodularity for approximation with formal guarantees.

Preferred and Multi-Objective Explanations. Miller [2019], Ramon et al. [2021] emphasize personalized interpretability. Audemard et al. [2022b] integrated cost, actionability, and fairness constraints [Ustun et al., 2019, Karimi et al., 2023] but restrict to exact explanations ($\epsilon = 0$). Our work generalizes this by allowing $\epsilon > 0$ via a λ -parameterized trade-off, handled by WGD for quantitative and LSA for ordinal preferences. Mothilal et al. [2020] considered multi-objective optimization for counterfactuals; our PFE extends this to probabilistic explanations over the (ϵ, cost) Pareto space, building on submodular optimization [Nemhauser and Wolsey, 1988, Sviridenko et al., 2017] for efficient approximation of NP-hard problems.

3 PROBLEM FORMULATIONS

3.1 MAIN PROBLEM

The number of abductive explanations can grow exponentially with features, even for simple models like decision trees and tree ensembles [Molnar, 2020, Izza et al., 2020, Audemard et al., 2022a, Izza et al., 2025], making exhaustive analysis impractical. Some are long and difficult to interpret Izza et al. [2022b], Audemard et al. [2022c], while others may not reflect user preferences. To resolve this trade-off, we consider *preferred probabilistic explanations*, which allow controlled probabilistic error ϵ to reduce size while explicitly integrating user preferences over features.

Definition 1 (Preferred Probabilistic Explanations). *Given classifier $h : \{0, 1\}^d \rightarrow \{0, 1\}$, instance $x \in \{0, 1\}^d$, and weight function $w : X_d \rightarrow \mathbb{R}^+$ modeling user preferences, a preferred probabilistic explanation is a subset $S \subseteq [d]$ minimizing probabilistic error $\epsilon_{h,x}(S)$ while optimizing cost $\text{cost}(S) = \sum_{i \in S} w(x_i)$ under constraint $|S| \leq k$.*

The central problem computes subsets of size at most k

achieving optimal trade-offs between precision and preference satisfaction, leading to weighted scalarization and Pareto formulations below.

3.2 WEIGHTED SCALARIZATION PROBLEM (WPPE)

The first approach aggregates both objectives—probabilistic error and preference cost—into a single objective function via a trade-off parameter λ .

Problem 1 (WPPE – Weighted Preferred Probabilistic Explanations). *Input:*

- A classifier $h : \{0, 1\}^d \rightarrow \{0, 1\}$
- An instance $x \in \{0, 1\}^d$ to explain
- A candidate feature set $I \subseteq [d]$
- A size limit $k \leq |I|$ reflecting cognitive constraints
- A weight function $w : X_d \rightarrow \mathbb{R}^+$ encoding preferences
- A trade-off parameter $\lambda \geq 0$ controlling relative importance of cost

Output: A feature subset S^* solving

$$S^* \in \underset{S \subseteq I, |S| \leq k}{\operatorname{argmin}} [\epsilon_{h,x}(S) + \lambda \cdot \text{cost}(S)]$$

Interpretation: Parameter λ acts as an exchange rate between objectives: $\lambda = 0$ exclusively prioritizes error minimization (ignoring preferences), while high λ favors low-cost explanations at the expense of precision. Since we minimize preference cost, non-preferred features receive high weights while preferred features receive low weights. This formulation is particularly suited when users can explicitly express preferences as a quantitative trade-off, as often occurs in medical contexts.

Remark 1 (Special Cases and Complexity). WPPE generalizes classical probabilistic explanations: when $\lambda = 0$, or when preferences are uniform ($w(x_i) = c$ for all i) and the size $|S|$ is fixed to exactly k , it reduces to $\min_{S \subseteq I, |S| \leq k} \epsilon_{h,x}(S)$. Since this problem is NP^{PP} -complete for CNF classifiers [Wäldchen et al., 2021]—strictly above NP—WPPE inherits this hardness.

Proposition 1 (Complexity). *The WPPE problem is NP-hard when classifier h is represented by a decision tree and I is a sufficient reason.*

3.3 PARETO OPTIMIZATION PROBLEM (PPPE)

The second formulation avoids choosing λ a priori by considering the complete set of optimal trade-offs, called the Pareto frontier [Ehrgott, 2005]. Unlike weighted scalarization aggregating both objectives into a scalar function, the Pareto approach simultaneously minimizes probabilistic error $\epsilon_{h,x}(\cdot)$ and preference cost $\text{cost}(\cdot)$ without privileging either.

Definition 2 (Pareto Dominance). *Let two explanations $S, S' \subseteq I$ satisfy $|S|, |S'| \leq k$. We say S Pareto-dominates S' , denoted $S \prec_P S'$, if and only if:*

$$\epsilon_{h,x}(S) \leq \epsilon_{h,x}(S') \quad \text{and} \quad \text{cost}(S) \leq \text{cost}(S'),$$

with at least one inequality being strict.

Definition 3 (Pareto Optimality). *An explanation $S \subseteq I$ with $|S| \leq k$ is Pareto-optimal if no other explanation $S' \subseteq I$ of size at most k dominates it, i.e., there exists no S' such that $S' \prec_P S$. The set of all Pareto-optimal explanations constitutes the Pareto frontier, denoted $\mathcal{P}_k$.*

Problem 2 (PPPE – Pareto Preferred Probabilistic Explanations). *Input:*

- A classifier $h : \{0, 1\}^d \rightarrow \{0, 1\}$
- An instance $x \in \{0, 1\}^d$ to explain
- A candidate feature set $I \subseteq [d]$
- A size limit $k \leq |I|$ reflecting cognitive constraints
- A weight function $w : X_d \rightarrow \mathbb{R}^+$ encoding preferences

Output: *The Pareto frontier $\mathcal{P}_k$, or an approximation thereof, i.e., the set (or superset) of all Pareto-optimal explanations of size at most k with respect to minimization objectives $\epsilon_{h,x}(\cdot)$ and $\text{cost}(\cdot)$.*

Advantages of the Pareto approach. This formulation offers several advantages. First, it requires no a priori trade-off specification, suited for ill-defined or evolving preferences [Miettinen, 1999]. Second, it enables interactive exploration: users examine frontier $\mathcal{P}_k$ and select their preferred trade-off. Third, it reveals *knee points* [Das and Dennis, 1997] where error-cost substitution rates change abruptly. Finally, unlike weighted scalarization capturing one point per λ , it provides complete characterization of non-dominated solutions.

4 APPROXIMATION ALGORITHMS

We now present three complementary algorithms to efficiently solve WPPE and PPPE problems, systematically exploiting underlying mathematical structure.

4.1 STRUCTURAL PROPERTIES

Our algorithmic approach fundamentally relies on a relaxed supermodularity property of $\mu_{h,x}$ established in [Bounia and Koriche, 2023], noting that $\epsilon_{h,x}$ is neither supermodular nor monotone in general.

Lemma 1 (Supermodularity Preservation). *Let $\mu_{h,x} : 2^{[d]} \rightarrow \mathbb{R}^+$ be a supermodular and monotone non-increasing function. Let $w : X_d \rightarrow \mathbb{R}^+$ be a weight function. Define the weighted objective function:*

$$f_\lambda(S) = \mu_{h,x}(S) + \lambda \sum_{i \in S} w(x_i)$$

Then $f_\lambda(\cdot)$ is supermodular, non-negative, and monotone non-increasing for all $\lambda \geq 0$ satisfying:

$$\lambda \leq \lambda^* = \min_{S \subseteq [d], i \notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S \cup \{i\})}{w(x_i)}$$

4.2 ALGORITHM 1 (WGD)

Our first algorithm solves WPPE via a greedy descent strategy, iteratively removing features that minimize increase in weighted objective function.

Algorithm 1: Weighted Greedy Descent (WGD)

Require: Classifier h , instance x , initial set I , limit k , weights w , parameter λ

Ensure: Explanation $S_{\text{WGD}} \subseteq I$ with $|S_{\text{WGD}}| \leq k$

- 1: $n \leftarrow |I|, S_n \leftarrow I$
 - 2: **for** $j = n$ **down to** 1 **do**
 - 3: **for** each feature $i \in S_j$ **do**
 - 4: Compute marginal: $\Delta_i = f_\lambda(S_j \setminus \{i\}) - f_\lambda(S_j)$
 - 5: **end for**
 - 6: $i^* \leftarrow \text{argmin}_{i \in S_j} \Delta_i$
 - 7: $S_{j-1} \leftarrow S_j \setminus \{i^*\}$
 - 8: Store pair $(S_j, \epsilon_{h,x}(S_j))$ for later evaluation
 - 9: **end for**
 - 10: **return** $S_{\text{WGD}} \in \text{argmin}_{S_j : |S_j| \leq k} [\epsilon_{h,x}(S_j) + \lambda \cdot \text{cost}(S_j)]$
-

Algorithmic intuition. The algorithm starts with complete set I (which could be a sufficient reason) and iteratively removes the feature whose deletion causes smallest degradation in weighted objective. This process generates a nested sequence $I = S_n \supset S_{n-1} \supset \dots \supset S_1 \supset S_0 = \emptyset$, and the algorithm finally returns the best solution among those.

Proposition 2 (Approximation Guarantee). *Let $\lambda \leq \lambda^*$ as defined in Lemma 1, so that f_λ is supermodular, monotone non-increasing, and non-negative on 2^I . Let $S^* \in \text{argmin}_{S \subseteq I, |S| \leq k} f_\lambda(S)$ be an optimal solution to WPPE, and let c_w be the curvature of function f_λ on 2^I . Then algorithm WGD returns a solution S_{WGD} satisfying:*

$$f_\lambda(S_{\text{WGD}}) = \epsilon_{h,x}(S_{\text{WGD}}) + \lambda \cdot \text{cost}(S_{\text{WGD}}) \leq \alpha_w \cdot f_\lambda(S^*)$$

where $\alpha_w = \frac{e^{p_w} - 1}{p_w}$ with $p_w = \frac{c_w}{1 - c_w}$.

Remark 2. *In practice, observed approximation ratios are significantly better than theoretical bounds, as shown by our experiments (Section 5). This is explained by theoretical bounds being established for worst case, while real instances generally present more favorable structure.*

4.3 ALGORITHM 2: PARETO FRONTIER ENUMERATION (PFE)

To solve PPPE and construct an approximation of the Pareto frontier, we propose a systematic sampling strategy of parameter space λ .

Algorithm 2: Pareto Frontier Enumeration (PFE)

Require: Classifier h , instance x , set I , limit k , weights w , number of samples L
Ensure: Approximate Pareto frontier $\tilde{\mathcal{P}}$

- 1: Define $\lambda_{\min} = 10^{-3}$
- 2: Compute $\lambda^* = \min_{S \subseteq [d], i \notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S \cup \{i\})}{w(x_i)}$
- 3: Compute geometric factor: $\gamma = \left(\frac{\lambda^*}{\lambda_{\min}}\right)^{1/(L-1)}$
- 4: Generate sequence $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$ with $\lambda_1 = \lambda_{\min}$ and $\lambda_{\ell+1} = \gamma \cdot \lambda_\ell$
- 5: Solutions $\leftarrow \emptyset$
- 6: **for** each $\lambda \in \{\lambda_1, \dots, \lambda_L\}$ **do**
- 7: $S_\lambda \leftarrow \text{WGD}(h, x, I, k, w, \lambda)$
- 8: Compute $\epsilon_\lambda = \epsilon_{h,x}(S_\lambda)$ and $c_\lambda = \text{cost}(S_\lambda)$
- 9: Solutions $\leftarrow$ Solutions $\cup \{(S_\lambda, \epsilon_\lambda, c_\lambda)\}$
- 10: **end for**
- 11: Remove duplicate solutions (same pair (ϵ, c))
- 12: $\tilde{\mathcal{P}} \leftarrow \text{FilterDominance}(\text{Solutions})$
- 13: **return** $\tilde{\mathcal{P}}$

FilterDominance procedure. This procedure iterates through solution set and eliminates those dominated in Pareto sense. For each solution pair (S, S') , it checks if S dominates S' (in which case S' is removed) or vice versa. This can be implemented efficiently in time $O(L^2)$ by first sorting solutions by one dimension.

Choice of geometric progression. Using a geometric sequence for λ (rather than arithmetic) is motivated by two considerations. First, it ensures uniform exploration of trade-off space on logarithmic scale over $[\lambda_{\min}, \lambda^*]$, which is natural since cost ratios are often more significant than their absolute differences. Second, it guarantees complete coverage of Pareto frontier within the validity range of Lemma 1, as formalized by the following theorem.

Definition 4 ((α, β) -Pareto Approximation). *An approximate Pareto frontier $\tilde{\mathcal{P}}$ is an (α, β) approximation of true Pareto frontier $\mathcal{P}$ if, for any Pareto-optimal solution $S^* \in \mathcal{P}$, there exists a solution $\tilde{S} \in \tilde{\mathcal{P}}$ satisfying:*

$$\epsilon_{h,x}(\tilde{S}) \leq \alpha \cdot \epsilon_{h,x}(S^*) \quad \text{and} \quad \text{cost}(\tilde{S}) \leq \beta \cdot \text{cost}(S^*)$$

Theorem 1 (Guarantee for PFE). *Let α_w be WGD's approximation ratio as defined in Proposition 2, under the condition $\lambda \leq \lambda^*$ as defined in Lemma 1. Then algorithm PFE with L samples drawn from $[\lambda_{\min}, \lambda^*]$ returns a frontier $\tilde{\mathcal{P}}$ that is an (α_w, α_w) -Pareto approximation of $\mathcal{P}$.*

4.4 ALGORITHM 3: LEXICOGRAPHIC STRATIFIED APPROACH (LSA)

In certain application contexts, user preferences are not quantitative but ordinal: users can rank features by priority levels, wishing to use priority level 1 features first, then level 2 features only if necessary, and so on. This situation

is frequent in medicine (invasive vs non-invasive tests) or finance (public vs private data).

Lexicographic preference model. Suppose features are partitioned into m strata $I = I_1 \cup I_2 \cup \dots \cup I_m$ with $I_i \cap I_j = \emptyset$ for $i \neq j$, where I_1 contains most preferred features and I_m least preferred. An explanation S is lexicographically preferred to S' if, denoting $n_i(S) = |S \cap I_i|$ the number of features from S in stratum i , there exists level ℓ such that:

$$n_i(S) = n_i(S') \quad \forall i < \ell \quad \text{and} \quad n_\ell(S) < n_\ell(S')$$

Algorithm 3: Lexicographic Stratified Algorithm (LSA)

Require: Classifier h , instance x , strata $\{I_1, \dots, I_m\}$, limit k , error threshold $\epsilon_{\max}$
Ensure: Lexicographically optimal explanation S_{LSA}

- 1: $S \leftarrow I_1 \cup I_2 \cup \dots \cup I_m$
- 2: **if** $\epsilon_{h,x}(S) > \epsilon_{\max}$ **then**
- 3: **return** S {Infeasible}
- 4: **end if**
- 5: **for** level $\ell = m$ **down to** 1 **do**
- 6: **while** $S \cap I_\ell \neq \emptyset$ **do**
- 7: $i^* \leftarrow \text{argmin}_{i \in S \cap I_\ell} [\mu_{h,x}(S \setminus \{i\}) - \mu_{h,x}(S)]$
- 8: $S_{\text{temp}} \leftarrow S \setminus \{i^*\}$
- 9: **if** $\epsilon_{h,x}(S_{\text{temp}}) \leq \epsilon_{\max}$ **and** $|S_{\text{temp}}| \geq 1$ **then**
- 10: $S \leftarrow S_{\text{temp}}$
- 11: **else**
- 12: **break**
- 13: **end if**
- 14: **end while**
- 15: **end for**
- 16: **return** S

Operating principle. The LSA algorithm proceeds by priority levels, starting with the least preferred features (level m) and progressively moving to the most preferred (level 1). At each level ℓ , it iteratively removes features from I_ℓ :

1. Size limit k is not reached
2. The probabilistic error remains below $\epsilon_{\max}$
3. Features still need to be removed in the current level

This process guarantees that low-priority features are eliminated first, preserving high-priority ones much as possible.

Theorem 2 (Guarantee for LSA). *For each priority level ℓ , algorithm LSA produces a solution that is α_w -approximate among all solutions using at most the same number of features from levels 1 to $\ell - 1$.*

4.5 APPLICATION TO DECISION TREES

Approximation guarantees in Proposition 2 and Theorems 1 and 2 apply to any Boolean hypothesis class. However, for computational efficiency, each call to oracle $\mu_{h,x}(S)$ must

execute in polynomial time. Evaluating $\mu_{h,x}$ in the general case is #P-hard [Crama and Hammer, 2011]. For decision trees, Izza et al. [2020], Koriche et al. [2013], Barceló et al. [2020] showed it can be computed in polynomial time, while Bounia and Koriche [2023] demonstrated linear-time evaluation in $O(|T| \cdot |S|)$, where $|T|$ is the number of tree nodes.

Resulting complexities. Our algorithms inherit the following worst-case complexities for decision trees. WGD makes $O(n^2)$ oracle calls (where $n = |I|$), each requiring $O(|T| \cdot d)$ operations, yielding $O(n^2 \cdot |T| \cdot d)$. PFE performs L calls to WGD over $[\lambda_{\min}, \lambda^*]$, yielding $O(L \cdot n^2 \cdot |T| \cdot d)$ with typically $L \in [20, 50]$. LSA performs at most n iterations each requiring $O(n)$ evaluations of $\mu_{h,x}(\cdot)$, yielding the same $O(n^2 \cdot |T| \cdot d)$ as WGD. This tractability is particularly relevant as decision trees are widely used in critical domains (healthcare, finance, justice) due to their interpretability [Rudin, 2018], and our algorithms extend this by enabling more concise preference-aligned explanations while preserving computational efficiency.

4.6 COMPARATIVE ANALYSIS

Having established polynomial-time tractability for decision trees, we now compare the three algorithms:

Table 1: Comparison of the three proposed algorithms

Criterion	WGD	PFE	LSA
Preference type	Quantitative	Quantitative	Ordinal
Output	One solution	Complete frontier	One solution
Choice of λ required	Yes ($\leq \lambda^*$)	No	No
Complexity	$O(n^2 T d)$	$O(Ln^2 T d)$	$O(n^2 T d)$
Approximation ratio	α_w	(α_w, α_w)	α_w per level
Interactive exploration	No	Yes	No

Usage recommendations:

- **WGD:** Prefer when desired trade-off is well-defined ($\lambda \leq \lambda^*$) and single solution is sought. Ideal for integration in automated systems.
- **PFE:** Recommended for interactive exploration, when user preferences are not fully specified a priori, or to present multiple options to end user.
- **LSA:** Use in contexts where preferences are naturally hierarchical (medicine, regulated finance) and where quantitative trade-offs are difficult to express.

5 EXPERIMENTAL EVALUATION

To validate the effectiveness and scalability of our three proposed algorithms (WGD, PFE, LSA), we conducted extensive experiments on decision tree classifiers across diverse benchmark datasets from the UCI Machine Learning Repository. All algorithms were implemented in Python, leveraging scikit-learn for model training and NumPy for

numerical computations. Experiments were performed on a workstation equipped with a 3.1 GHz Intel Core i9-9900 CPU and 64 GiB of RAM, running Ubuntu 22.04 LTS.

Datasets. In our experiments, we considered $B = 25$ binary classification datasets (10 to 10^5 features) from Kaggle, OpenML, and UCI, including MNIST38 and MNIST49 two binary subsets of MNIST opposing digits 3 vs. 8 and 4 vs. 9 respectively. Multi-class tasks were converted to binary by considering the dominant class versus all others. Due to space constraints, we report results on $B = 10$.

User Preferences. We considered four preference models: (i) **uniform preferences**, where all features receive equal weight, serving as a baseline; (ii) **SHAP-based preferences**, using Shapley values as local importance weights; (iii) **LIME-based preferences**, derived from LIME’s local explanations; and (iv) **feature importance**, using scikit-learn’s feature importance for decision trees. Due to space constraints, we report results only for SHAP-based preferences in the main paper, as they represent the most widely adopted approach in practice. Results for other preference scenarios are provided in the supplementary material.

5.1 EXPERIMENTAL SETUP

Explanation Tasks. For each benchmark b , an explanation task is (T, x, I, k, w) . The classifier h is represented by a decision tree T trained via scikit-learn’s CART implementation Pedregosa et al. [2011]. Instance x is randomly selected from the test set; candidate set I is the root-to-leaf path consistent with x (path explanation [Izza et al., 2020]). The weight function w is derived from preference scenarios via a monotone decreasing transformation: $w(x_i) = \max\{s\} - s_i + \delta$ where s_i is the raw importance score and $\delta > 0$ ensures positivity. This maps higher importance to lower weights, which is consistent with our cost minimization objective. We set cognitive limit $k = 7 \pm 2$ Miller [1956] in general; when $|I|$ is small, we adapt and set $k = 4 \pm 2$ to avoid trivial cases where $k \geq |I|$. We impose a time limit of 60 minutes per instance. Performance is evaluated by averaging $\epsilon_{h,x}(S)$, $\text{cost}(S)$, and $|S|$ over $m = \min\{s, 150\}$ test instances.

Scalability and Approximation Quality. Outside the case $\epsilon = 0$, where exact preferred abductive explanations can be computed via MAXSAT [Audemard et al., 2022b] and the case $\lambda = 0$ where exact probabilistic explanations are accessible via SAT encoding [Arenas et al., 2022], no exact method is known for the general preferred probabilistic explanation problem. This motivates our approximation approach. To empirically assess its quality, we compare Algorithm 1 (WGD) against two exact baselines. First, for $\lambda = 0$, we use the SAT-based method of Arenas et al. [2022], extended to the decision version of Problem 1 by adding

the clause $\bigvee \{x_i : (x_I)_i = 1\} \vee \bigvee \{\neg x_i : (x_I)_i = 0\}$, and solved via binary search over $(0, 1]$ with precision 10^{-3} , requiring at most 10 calls to GLUCOSE 4 (PySAT, timeout 60 min). Second, for $\lambda > 0$, we compare against an exhaustive enumeration baseline, computing $\arg\min_{S \subseteq I, |S| \leq k} f_\lambda(S)$ over all possible subsets, on small-dimensional datasets where this is tractable, as *placement*, *tic-tac-toe* and others. These comparisons jointly quantify the practical cost of approximation across both the $\lambda = 0$ and $\lambda > 0$ regimes.

Impact of λ on WGD. We evaluate Algorithm 1 (WGD) across $\lambda \in [10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$, where d is the number of features and k the size limit, using SHAP-based normalized preferences with $\sum_{i \in I} w(x_i) = 1$ to ensure a meaningful exchange rate between $\epsilon_{h,x}(\cdot)$ and $\text{cost}(\cdot)$. We report a table summarizing $\epsilon_{h,x}(S_\lambda)$, $\text{cost}(S_\lambda)$, and $|S_\lambda|$ for representative values of λ , as well as a figure illustrating the evolution of both objectives as λ varies.

Pareto Frontier Exploration via PFE. A key limitation of WGD is that it requires specifying λ a priori, which may be difficult in practice when user preferences are not fully quantified. Algorithm 2 (PFE) addresses this by computing a set of Pareto-optimal solutions covering the full trade-off space between $\epsilon_{h,x}(\cdot)$ and $\text{cost}(\cdot)$, without requiring any explicit choice of λ . We evaluate PFE with $L = 30$ samples drawn from $\lambda \in [10^{-3}, 10^3]$ with a geometric step, on our benchmark datasets using SHAP-based normalized preferences, and report the resulting approximate Pareto frontiers as 2D plots (ϵ vs. cost), highlighting knee points where the marginal substitution rate between both objectives changes abruptly. These visualizations allow users to interactively select their preferred trade-off without committing to λ .

Ordinal Preferences and LSA. While WGD and PFE handle quantitative preferences via λ , certain domains naturally express preferences in ordinal terms, without any meaningful numerical trade-off. Algorithm 3 (LSA) addresses this by partitioning features into priority strata. We evaluate LSA on COMPAS, where actionability constraints are well-documented [Ustun et al., 2019], and compare it against WGD in terms of $\epsilon_{h,x}(S)$, stratum composition $n_\ell(S)$, $|S|$.

Benchmark				$\epsilon_{h,x}(S)$		$ S $		Time (s)
name	acc	d	$ I $	WGD	SAT	WGD	SAT	SAT
<i>student perf.</i>	92.04	30	5.37	0.27 (± 0.10)	0.27 (± 0.10)	2.03	2.03	2.11
<i>hungarian</i>	63.71	13	6.69	0.12 (± 0.11)	0.12 (± 0.09)	3.57	3.57	1.79
<i>horse colic</i>	75.94	40	6.74	0.14 (± 0.06)	0.14 (± 0.06)	4.09	4.09	11.71
<i>loan eligibility</i>	74.08	68	8.49	0.18 (± 0.12)	0.22 (± 0.13)	5.73	6.81	43.96
<i>wine</i>	70.11	11	9.05	0.09 (± 0.09)	0.10 (± 0.11)	5.66	5.64	35.48
<i>employee attr.</i>	83.37	63	10.58	0.07 (± 0.08)	0.21 (± 0.10)	6.44	7.04	1008.42
<i>compas</i>	68.02	40	10.97	0.05 (± 0.08)	0.06 (± 0.09)	5.86	6.82	1091.88
<i>mnist49</i>	96.28	784	15.60	0.38 (± 0.13)	—	6.92	—	—
<i>gisetete</i>	94.37	5000	21.39	0.33 (± 0.10)	—	6.92	—	—
<i>farm ads</i>	81.12	54877	23.18	0.14 (± 0.16)	—	6.34	—	—

Table 2: Experimental results for $\lambda = 0$ (Problem 1) on 10 benchmarks when h is a decision tree, with $k \in \{5, 6, 7\}$.

λ	$\epsilon_{h,x}(S)$		$ S $		$f_\lambda(S)$ (norm.)	
	WGD	Exact	WGD	Exact	WGD	Exact
0.0010	0.1105 (± 0.0689)	0.1105 (± 0.0689)	4.7387 (± 0.6536)	4.7387 (± 0.6536)	0.0587	0.0587
0.2507	0.1105 (± 0.0689)	0.2268 (± 0.0789)	4.7387 (± 0.6536)	3.0270 (± 1.3046)	0.0587	0.1205
0.5005	0.1122 (± 0.0684)	0.2401 (± 0.0719)	4.7297 (± 0.6837)	2.8919 (± 1.3244)	0.0596	0.1275
0.7502	0.1122 (± 0.0684)	0.2626 (± 0.0904)	4.7297 (± 0.6837)	2.7658 (± 1.2445)	0.0596	0.1395
10^0	0.1291 (± 0.0614)	0.2672 (± 0.0882)	4.6396 (± 0.9280)	2.7387 (± 1.2499)	0.0686	0.1419
2.15×10^6	0.2409 (± 0.1115)	0.3825 (± 0.1694)	3.1171 (± 1.4316)	1.7748 (± 0.9743)	0.1280	0.2032
6.98×10^6	0.2490 (± 0.1174)	0.3905 (± 0.1730)	3.0180 (± 1.4205)	1.7027 (± 0.9260)	0.1323	0.2074
1.18×10^7	0.2566 (± 0.1157)	0.3944 (± 0.1743)	2.9009 (± 1.4203)	1.6847 (± 0.9201)	0.1363	0.2095
1.66×10^7	0.2636 (± 0.1214)	0.3949 (± 0.1740)	2.8198 (± 1.4155)	1.6847 (± 0.9201)	0.1395	0.2098
2.15×10^7	0.2632 (± 0.1230)	0.3949 (± 0.1740)	2.8108 (± 1.4111)	1.6847 (± 0.9201)	0.1398	0.2098
2.15×10^{11}	0.3388 (± 0.1526)	0.4342 (± 0.1916)	2.0991 (± 1.2003)	1.2432 (± 0.5061)	0.2766	0.2306
3.82×10^{11}	0.3549 (± 0.1562)	0.4351 (± 0.1911)	2.0270 (± 1.1505)	1.2252 (± 0.4966)	0.3320	0.2311
6.79×10^{11}	0.3549 (± 0.1562)	0.4400 (± 0.1934)	2.0270 (± 1.1505)	1.1982 (± 0.4615)	0.4436	0.2337
1.21×10^{12}	0.3555 (± 0.1561)	0.4402 (± 0.1933)	2.0180 (± 1.1546)	1.1802 (± 0.4492)	0.6425	0.2338
2.15×10^{12}	0.3639 (± 0.1581)	0.4428 (± 0.1927)	1.9730 (± 1.1347)	1.1622 (± 0.3923)	1.0000	0.2352

Table 3: WGD vs. exact method on Horse Colic ($d = 36$, $k = 5$, $\text{acc.} = 77.48\%$, SHAP preferences, $|R_d| \approx 6$, $\lambda \in [10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$). $f_{\hat{\lambda}}(S)$ (with $\hat{\lambda} = \lambda/2^{d-k}$).

5.2 EXPERIMENTAL RESULTS

Table 2 reports results on 10 of the 33 benchmarks for decision tree classifiers with $k = 7$ and $\lambda = 0$. Columns *acc* and *d* give accuracy and number of binary features. Rows are sorted by average path explanation size $|I|$ [Izza et al., 2020]. We report average error $\epsilon_{h,x}(S)$ and explanation size $|S|$ for both WGD and the exact SAT-based method, along with SAT runtime. Datasets in **blue** indicate partial timeout; **magenta** indicates complete timeout. WGD always terminates in under 1 second. For the general $\lambda > 0$ regime, Table 3 (supplementary material for full results) treats the *horse colic* dataset across 15 representative values of $\lambda \in [10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$, comparing WGD against exhaustive enumeration; results confirm that WGD achieves consistently lower normalized $f_\lambda(S)$ than the exact method for small and moderate λ , with a practical approximation bound well below the theoretical guarantee of Proposition 2. However, for large λ (beyond λ^*), the theoretical guarantees of Proposition 2 no longer hold, and the approximation ratio may degrade significantly. This degradation is further amplified when I is a path explanation rather than a sufficient reason, since f_λ may lose its monotonicity structure in this regime, explaining the occasional bound explosion observed in Table 3. Overall, WGD demonstrates remarkable scalability, remaining stable in milliseconds on large datasets such as *gisetete* and *farm ads* where SAT times out—while the gap $\epsilon_{h,x}(S_{\text{WGD}}) - \epsilon_{h,x}(S^*)$ remains negligible for $\lambda = 0$ and $|S_{\text{WGD}}|$ is on average smaller than $|S^*|$, confirming the effectiveness of our algorithm in practice.

Impact of λ on WGD. Figure 1 illustrates the effect of λ on the *placement* dataset: as λ increases, $\text{cost}(S)$ and $|S|$ decrease monotonically while $\epsilon_{h,x}(S)$ increases, confirming that λ effectively controls the trade-off between preference satisfaction and probabilistic precision. Conversely, small λ favors precision at the expense of preferences. For $\lambda \leq \lambda^*$, the empirical approximation ratio remains well below the theoretical bound of Proposition 2, confirming near-optimality of WGD. Beyond λ^* , approximation guarantees no longer hold and solution quality may degrade,

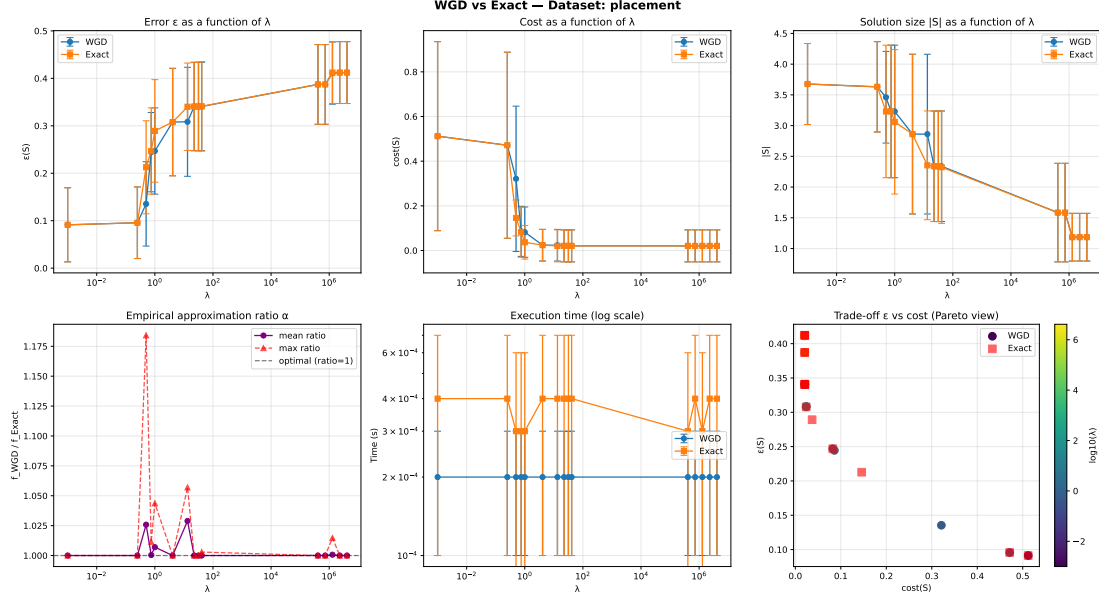


Figure 1: WGD vs. exact method on placement dataset (SHAP preferences) and $k = 4$.

particularly when I is not a sufficient reason. This highlights the necessity of exploring the full trade-off space via PFE rather than committing to a single λ a priori, especially when preferences are not fully quantified.

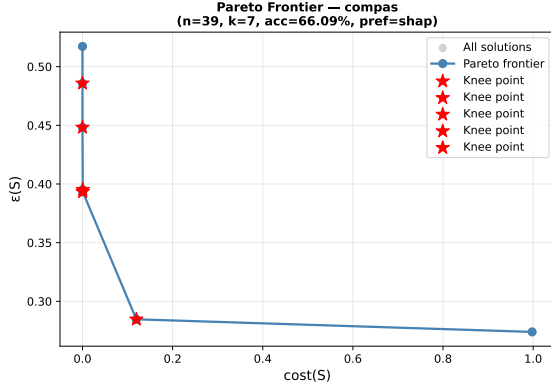


Figure 2: Pareto frontier obtained by PFE on the COMPAS.

Pareto Frontier Exploration via PFE. Figure 2 reports the approximate Pareto frontier returned by PFE on COMPAS (SHAP preferences, $k = 7$, $|I| = 11$), comprising 7 non-dominated solutions ranging from high fidelity ($\epsilon = 0.274$, $\text{cost} = 0.998$, $|S| = 7$) to maximal preference alignment ($\epsilon = 0.517$, $\text{cost} \approx 0$, $|S| = 1$). PFE avoids specifying λ a priori by exposing the full trade-off space. A particularly attractive knee point is the second Pareto solution ($\epsilon = 0.285$, $\text{cost} = 0.119$, $|S| = 7$), which achieves an 88% reduction in preference cost at the expense of only a marginal error increase of 0.011, illustrating how PFE empowers users to identify strongly preference-aligned so-

lutions without explicit knowledge of λ .

Ordinal Preferences and LSA. While WGD and PFE handle quantitative preferences via λ , certain domains naturally express preferences in ordinal terms. Algorithm 3 (LSA) addresses this by partitioning features into priority strata. We evaluate LSA on COMPAS [Ustun et al., 2019] with $\epsilon_{\max} = 0.1$, $k = 7$, and three strata: I_1 (actionable: *Number_of_Priors*, *score_factor*, *Age_**), I_2 (partial: *Misdemeanor*), and I_3 (non-actionable demographic attributes). For the selected instance ($|I| = 10$), Figure 3 compares LSA against WGD in terms of $\epsilon_{h,x}(S)$, stratum composition $n_\ell(S)$, and $|S|$. LSA lexicographically eliminates three non-actionable features from I_3 while preserving $\epsilon_{h,x}(S) = 0$, yielding $n_1 = 4$, $n_2 = 1$, $n_3 = 2$ with $|S| = 7$. This is meaningful in the COMPAS context: the explanation relies primarily on actionable features (*Number_of_Priors*, *score_factor*) while minimizing sensitive demographic attributes. By contrast, WGD ($\lambda = 0$) returns $|S| = 5$ with only $n_1 = 3$ and $n_2 = 0$, ignoring the ordinal structure entirely. As shown in the third subplot, LSA achieves a strictly lower non-actionable ratio at equal error, confirming that it produces explanations that are both probabilistically valid and maximally aligned with user priorities.

6 CONCLUSION AND FUTURE WORK

We introduced the first unified framework for *preferred probabilistic explanations*, bridging probabilistic accuracy, cognitive constraints, and user preferences via three complementary algorithms: WGD for quantitative preferences with formal approximation ratio $(e^{p_w} - 1)/p_w$, PFE for interac-

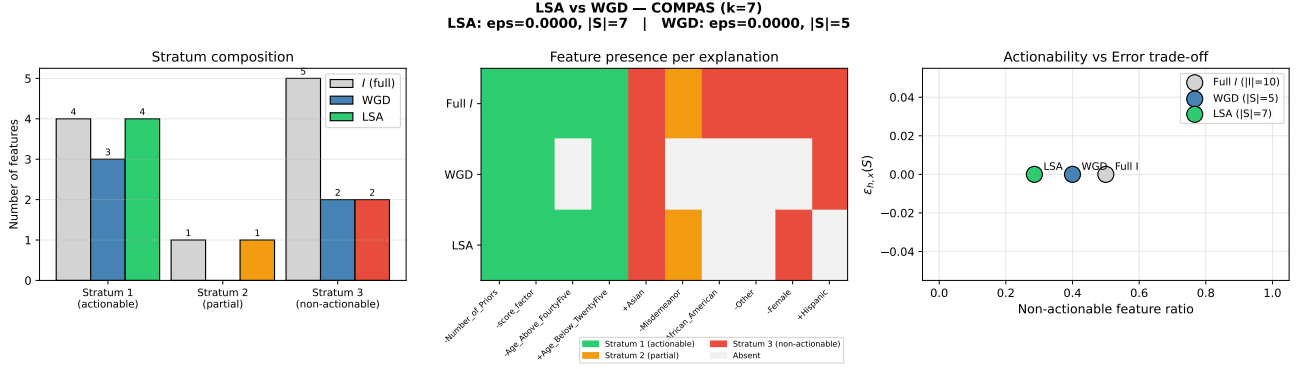


Figure 3: LSA vs. WGD on compas ($\epsilon_{max} = 0.1$, $k = 7$).

tive Pareto frontier exploration without committing to λ a priori, and LSA for ordinal lexicographic preferences. All three exploit the supermodular structure of $\mu_{h,x}$ to provide polynomial-time tractability and formal quality certificates for decision trees. Experiments on 25 benchmarks confirm that WGD scales to datasets with up to 54,877 features in milliseconds where exact SAT-based methods time out, PFE reveals meaningful knee points allowing users to interactively select their preferred precision–cost trade-off, and LSA consistently produces explanations with fewer non-actionable features than WGD at equal probabilistic error, as demonstrated on COMPAS.

Several directions naturally extend this work. **(i)** Extending efficient oracles for $\mu_{h,x}$ to random forests, tree ensembles, and neural networks would significantly broaden applicability. **(ii)** Instance-adaptive bounds exploiting dataset or tree topology structure could yield guarantees tighter than the worst-case curvature ratio. **(iii)** Replacing exhaustive λ -sampling in PFE with Bayesian strategies would identify knee points more efficiently. **(iv)** Integrating group fairness metrics as hard constraints into WPPE would strengthen applicability in regulated domains such as criminal justice and credit scoring. **(v)** Finally, real user studies validating preference alignment beyond synthetic proxies (SHAP, LIME, feature importance) remain an essential step toward deployment in human-facing decision systems.

ACKNOWLEDGEMENTS

We would like to thank the reviewers for their valuable comments and constructive feedback, which helped improve the quality and clarity of this work. This work is funded by the ANR JCJC project FORM-XAI (ANR-25-CE23-0146).

References

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Ma-*

chine Learning (WHI 2018), Stockholm, Sweden, 2018. arXiv:1806.08049.

Guy Amir, Shahaf Bassan, and Guy Katz. Hard to explain: On the computational hardness of in-distribution model interpretation. In *ECAI*, 2024.

Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems*, 35:28695–28707, 2022.

Marcelo Arenas, Pablo Barceló, Diego Bustamante, Jose Caraball, and Bernardo Subercaseaux. A uniform language to explain decision trees. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2024.

G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis. Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI’22*, 2022a.

G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis. On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI’22*, 2022b.

Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In *Proc. of KR’21*, 2021.

Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of boolean decision trees. *Data Knowledge Engineering*, 142, 2022c. ISSN 0169-023X. URL <https://www.sciencedirect.com/science/article/pii/S0169023X22000799>.

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. Computing abductive explanations for boosted trees. In *AISTATS*, 2023.

- Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. On the computation of example-based abductive explanations for random forests. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3679–3687, 2024.
- Steve Azzolin, SAGAR MALHOTRA, Andrea Passerini, and Stefano Teso. Beyond topological self-explainable GNNs: A formal explainability perspective. In *Forty-second International Conference on Machine Learning*, 2025.
- Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15487–15498, 2020.
- Shahaf Bassan and Guy Katz. Towards formal xai: Formally approximate minimal explanations of neural networks. In *Proceedings of the 29th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2022.
- Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability: A computational complexity perspective. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Shahaf Bassan, Guy Amir, Meirav Zehavi, and Guy Katz. What makes an ensemble (un) interpretable? In *Proceedings of the International Conference on Machine Learning*, 2025.
- Shahaf Bassan, Yizhak Yisrael Elboher, Tobias Ladner, Volkan Şahin, Jan Kretinsky, Matthias Althoff, and Guy Katz. Provably explaining neural additive models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026a. Poster.
- Shahaf Bassan, Xuanxiang Huang, and Guy Katz. Unifying formal explanations: A complexity-theoretic perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026b. Poster.
- Guy Blanc, Jane Lange, and Li-Yang Tan. Provably efficient, succinct, and precise explanations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=9UjRw5bqURS>.
- Ryma Boumazouza, Fatiha Cheikh-Alili, Bernard Mazure, and Karim Tabia. Asteryx: A model-agnostic sat-based approach for symbolic and score-based explanations. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2021.
- Louenas Bounia. *Modèles formels pour l’IA explicable : des explications pour les arbres de décision*. PhD thesis, Université d’Artois, Lens, France, December 2023. PhD thesis.
- Louenas Bounia. Computing an approximating version of a minimum-size explanation for boolean decision tree classifiers. In *Proceedings of ICCART 2025*, 2025.
- Louenas Bounia. k -Pseudo-Probable Path Explanations for Weighted Tree-Based Ensembles: Bridging Interpretability and Computational Efficiency. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2026)*, Naples, Italy, 2026.
- Louenas Bounia and Frederic Koriche. Approximating probabilistic explanations via supermodular minimization. In *Uncertainty in Artificial Intelligence (UAI 2023)*, 2023.
- Louenas Bounia and Insaf Setitra. Computing improved explanations for random forests: k -majoritary reasons. In *Proceedings of the ICCART 2025*, 2025a.
- Louenas Bounia and Insaf Setitra. Enhancing the intelligibility of decision trees with concise and reliable probabilistic explanations. *Data & Knowledge Engineering*, 2025b.
- Louenas Bounia, Juba Agoun, and Soundouss Messoudi. Approximating user-preferred abductive explanations via supermodular optimization: Application to tree-based classifiers. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2026)*, Naples, Italy, 2026.
- Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1433–1452. SIAM, 2014.
- Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001. doi: 10.1017/S0140525X01003922. URL <https://doi.org/10.1017/S0140525X01003922>.
- Yves Crama and Peter L Hammer. *Boolean functions: Theory, algorithms, and applications*. Cambridge University Press, 2011.
- A. Darwiche and A. Hirth. On the reasons behind decisions. In *Proc. of ECAI’20*, 2020.
- Adnan Darwiche. Logic for explainable AI. In *Proceedings of the 38th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–11. IEEE, 2023.

- Indraneel Das and J. E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14(1):63–69, 1997. doi: 10.1007/BF01197559.
- Alessandro De Palma, Greta Dolcetti, and Caterina Urban. Faster verified explanations for neural networks. *arXiv preprint arXiv:2512.00164*, 2025.
- Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. In *Science Advances*, volume 4, page eaao5580, 2018. doi: 10.1126/sciadv.aao5580.
- Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2005. ISBN 978-3-540-21398-7. doi: 10.1007/3-540-27659-9.
- Ouns El Harzli, Bernardo Cuenca Grau, and Ian Horrocks. Cardinality-minimal explanations for monotonic neural networks. In *Proceedings of the International Conference on Joint Artificial Intelligence (ICJAI)*, 2023.
- Andre Esteva, Brett Kuperl, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. doi: 10.1038/nature21056.
- Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. Poster.
- Xuanxiang Huang and João Marques-Silva. On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, 2024.
- A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI’19*, pages 1511–1519, 2019.
- Victor P. Il’ev. An approximation guarantee of the greedy descent algorithm for minimizing a supermodular set function. *Discrete Applied Mathematics*, 114(1):131–146, 2001. ISSN 0166-218X. doi: [https://doi.org/10.1016/S0166-218X\(00\)00366-8](https://doi.org/10.1016/S0166-218X(00)00366-8). URL <https://www.sciencedirect.com/science/article/pii/S0166218X00003668>. discrete analysis operations research.
- Rishabh Iyer and Jeff Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. *NeurIPS*, 2013.
- Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In *Proc. of IJCAI’21*, pages 2584–2591, 2021.
- Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *ArXiv*, abs/2010.11034, 2020. URL <https://api.semanticscholar.org/CorpusID:224814214>.
- Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 2022a.
- Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75:261–321, 2022b.
- Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.
- Yacine Izza, Kuldeep S. Meel, and João Marques-Silva. Locally-minimal probabilistic explanations. *ECAI*, 2024b.
- Yacine Izza, Alexey Ignatiev, Sasha Rubin, Joao Marques-Silva, and Peter J. Stuckey. Most general explanations of tree ensembles (extended version). *IJCAI 2025*, 2025.
- Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Hungarian heart disease dataset. UCI Machine Learning Repository, 1988. URL <https://archive.ics.uci.edu/ml/datasets/heart+Disease>.
- Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2023. doi: 10.1145/3527848.
- G. Katz, D. A. Huang, D. Ibeling, K. Julian, Ch. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljic, D. L. Dill, M. J. Kochenderfer, and C. W. Barrett. The marabou framework for verification and analysis of deep neural networks. In *Proc. of CAV’19*, pages 443–452, 2019.
- F. Koriche, J.-M. Lagniez, P. Marquis, and S. Thomas. Knowledge compilation for model counting: Affine decision trees. In *Proc. of IJCAI’13*, pages 947–953, 2013.
- Frédéric Koriche, Jean-Marie Lagniez, and Chi Tran. Probabilistic explanations for regression models. In *Proceedings on Uncertainty in Artificial Intelligence*, 2025.
- M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, and E. Aswini. Credit card fraud detection using random forest algorithm. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, 2019. doi: 10.1109/ICCCT2.2019.8824930.

- F. Labbaf, T. Kolárik, M. Blich, G. Fedyukovich, M. Wand, et al. Space explanations of neural network classification. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, 2025.
- Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021. doi: 10.1016/j.artint.2021.103473.
- Bounia Louenas. Enhancing the Intelligibility of Boolean Decision Trees with Concise and Reliable Probabilistic Explanations. In *IPMU 2024*, Lisboa, Portugal, 2024.
- S. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. In *Proc. of NIPS'17*, 2017.
- J Marques-Silva and A Ignatiev. Efficient & correct predictive equivalence for decision trees. *arXiv preprint arXiv:2509.17774*, 2025.
- Joao Marques-Silva. Logic-based explainability in machine learning. *ArXiv*, abs/2211.00541, 2023.
- Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. *Neural Information Processing Systems*, 2020.
- Ramy Marzouk, Shahaf Bassan, Guy Katz, and Colin de la Higuera. On the computational tractability of the (many) shapley values. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, 2025.
- Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, 1999. ISBN 978-0-7923-8278-2.
- G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, 1956.
- T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019.
- C. Molnar. *Interpretable Machine Learning*. Leanpub, 2020.
- Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.
- George Nemhauser and Laurence Wolsey. *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1988. Advisory Editors: Ronald L. Graham, Jan Karel Lenstra, Robert E. Tarjan.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020. doi: 10.1145/3351095.3372828.
- Yanou Ramon, Tom Vermeire, Olivier Toubia, David Martens, and Theodoros Evgeniou. Understanding consumer preferences for explanations generated by xai algorithms. *arXiv*, 2021.
- M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pages 1135–1144, 2016.
- Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *ArXiv*, abs/1811.10154, 2018. URL <https://api.semanticscholar.org/CorpusID:53747370>.
- A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proc. of IJCAI'18*, 2018.
- 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 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 180–186. ACM, 2020.
- Kacper Sokol and Peter A. Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020. doi: 10.1007/s13218-020-00637-y.
- Bernardo Subercaseaux, Marcelo Arenas, and Kuldeep S. Meel. Probabilistic explanations for linear models. In *Proceedings of the Thirty-Ninth AAAI Conference*, 2025.
- Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4):1197–1218, 2017. doi: 10.1287/moor.2016.0842.

- Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019. doi: 10.1145/3287560.3287566.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.
- Stephan Wäldchen. *Towards explainable artificial intelligence: Interpreting neural network classifiers with probabilistic prime implicants*. PhD thesis, Technische Universität Berlin, 2022. URL <https://depositonce.tu-berlin.de/items/e25aff2c-bf11-4bff-9cb6-57a1802cdda2>. Ph.D. thesis.
- Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3082–3088, 2021. doi: 10.24963/ijcai.2021/424. URL <https://doi.org/10.24963/ijcai.2021/424>.
- Huan Wu, Ovidiu Isac, Aleksandar Zeljic, Tomoki Tagomori, Matthew Daggitt, Wouter Kokke, Itay Refaeli, Guy Amir, Kyle Julian, Shahar Bassan, et al. Marabou 2.0: A versatile formal analyzer of neural networks. In *Proceedings of the 36th International Conference on Computer Aided Verification (CAV 2024)*, pages 249–264, 2024.
- Min Wu, Haoze Wu, and Clark Barrett. Verix: Towards verified explainability of deep neural networks. In *NeurIPS*, 2023.
- Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70:351–387, 2021.
- J. Yu, A. Ignatiev, P. J. Stuckey, N. Narodytska, et al. Eliminating the impossible, whatever remains must be true: On extracting and applying background knowledge in the context of formal explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. Focuses on applying background knowledge to improve the succinctness and accuracy of formal post-hoc explanations in XAI.

APPENDIX

We apologize for condensing Figures 1 and 3, as well as the two tables in the main paper, due to space limitations. We opted for this presentation to fit the current format. In the final version, we will have an additional page, which will easily resolve this issue.

The following appendix is organized as follows:

- **Appendix A:** Extended Background.
- **Appendix B:** Provides additional preliminary details to help for understand the key concepts and notions used in the paper.
- **Appendix C:** Contains the proofs of Lemma 1, Propositions 1 and 2, as well as Theorems 1 and 2.
- **Appendix D:** Provides additional details on the evaluation of the function $\mu_{h,x}(S)$ and on Algorithms 1, 2, and 3 when h is a decision tree.
- **Appendix E:** Additional experiments on the full set of 33 benchmarks with alternative preference models (LIME-based, SHAP-based, and global feature importance weights), reporting $\epsilon_{h,x}(S)$, $\text{cost}(S)$, and $|S|$ for each configuration. These results confirm that the conclusions drawn in the main paper are robust across preference scenarios and datasets of varying size and dimensionality.
- **Appendix F:** To further ground our framework in realistic deployment scenarios, we design synthetic user studies simulating diverse stakeholder profiles (e.g., medical practitioners, loan officers, data scientists) with heterogeneous preference structures across actionability, fairness, and cost constraints. Grounded in well-documented stakeholder needs from the literature Miller [2019], Ustun et al. [2019], Langer et al. [2021], each profile is associated with explicit preference weights, and we evaluate the utility of WGD, PFE, and LSA explanations in terms of comprehensibility, actionability, and alignment with stated preferences. These scenarios demonstrate that our three algorithms naturally complement each other across stakeholder types: WGD for well-specified quantitative trade-offs, PFE for interactive exploration when preferences are ill-defined, and LSA for hierarchical ordinal constraints common in regulated domains. While a rigorous real-world user study — comparing preference-aligned and non-aligned explanations, assessing Pareto frontier visualizations for decision support, and measuring task completion time and accuracy — remains an important direction for future work, our synthetic evaluation provides a principled first step toward validating preference alignment beyond purely quantitative metrics.

A EXTENDED BACKGROUND

Abductive and Formal Explanations. Formal explainable AI [Marques-Silva, 2023] investigates explanations equipped with provable guarantees [Yu et al., 2023, Darwiche, 2023, Darwiche and Hirth, 2020, Shih et al., 2018, Audemard et al., 2021, Azzolin et al., 2025]. Darwiche and Hirth [2020] characterized sufficient reasons for various model classes, establishing the theoretical foundations of abductive explanations. Efficient algorithms have since been developed for decision trees [Izza et al., 2020, Arenas et al., 2024, 2022, Bounia and Setitra, 2025b, Bounia and Koriche, 2023, Louenas, 2024, Bounia, 2025, 2023], random forests [Izza and Marques-Silva, 2021, Izza et al., 2024a], linear models [Marques-Silva et al., 2020, Subercaseaux et al., 2025], tree ensembles [Audemard et al., 2024, 2023, Marques-Silva and Ignatiev, 2025, Izza et al., 2025, Bassan et al., 2025, Boumazouza et al., 2021, Bounia et al., 2026, Bounia, 2026, Bounia and Setitra, 2025a], and neural networks [Bassan and Katz, 2022, Wu et al., 2023, Hadad et al., 2026, Bassan et al., 2026a, De Palma et al., 2025, Labbaf et al., 2025], with connections to SAT solving [Ignatiev et al., 2019]. Because producing such guaranteed explanations is often computationally intractable [Barceló et al., 2020, Bassan et al., 2024, 2026b, Blanc et al., 2021, Amir et al., 2024, Marzouk et al., 2025], much of the literature concentrates on restricted model classes [Marques-Silva, 2023], including monotonic models [El Harzli et al., 2023] and Boolean circuits [Darwiche, 2023, Darwiche and Hirth, 2020]. The complexity of generating explanations for additive and linear models has also been studied theoretically [Subercaseaux et al., 2025]. Our work focuses on decision trees [Izza et al., 2020, Audemard et al., 2022c], for which tractable algorithms exist, and extends them to the preference-aware probabilistic setting, a direction not previously explored in the formal XAI literature.

Probabilistic Explanations. Probabilistic explanations [Arenas et al., 2022, Koriche et al., 2025] relax the deterministic sufficiency requirement by allowing a bounded error $\epsilon > 0$, asking for a subset $S \subseteq I$ such that $\epsilon_{h,x}(S) \leq \epsilon$. Wäldchen

et al. [2021] proved NP^{PP} -completeness for Boolean circuits, while Arenas et al. [2022] showed NP-completeness for decision trees and proposed a SAT-based exact method. Barceló et al. [2020] further investigated the model counting complexity of explanation problems, and Bassan et al. [2024] studied the computational hardness of local explanations for neural networks. Bounia and Koriche [2023] exploited the submodular structure of $\mu_{h,x}$ to derive approximation algorithms with formal guarantees, establishing a $(1 - 1/e)$ -approximation ratio via greedy descent — a result central to our WGD algorithm. Despite verification tools [Wu et al., 2024, Katz et al., 2019], computing provable explanations remains costly, motivating approximation schemes [Koriche et al., 2025, Bounia and Koriche, 2023]. Our WGD algorithm builds directly on Bounia and Koriche [2023], extending it to the weighted setting where user preferences modulate the trade-off between $\epsilon_{h,x}(S)$ and $\text{cost}(S)$ via a parameter λ . The relaxed supermodularity condition is central to our approximation guarantees in Proposition 2, and our empirical results confirm that the practical approximation ratio remains well below the theoretical bound across all tested benchmarks.

Preferred Explanations and Multi-Objective XAI. Miller [2019] and Ramon et al. [2021] emphasize that explanations must be personalized to be practically useful. Audemard et al. [2022b] incorporated user preferences — including cost, actionability [Ustun et al., 2019], and fairness [Karimi et al., 2023] — into abductive explanation frameworks via MAXSAT-based methods, but restrict to exact explanations ($\epsilon = 0$), limiting applicability on complex or high-dimensional models. Audemard et al. [2024] further extended this line to tree ensembles, while Izza et al. [2020] and Izza et al. [2024a] studied preference-aware explanations for decision trees and random forests respectively. Darwiche [2023] and Darwiche and Hirth [2020] provide a knowledge compilation perspective on preferred explanations, connecting to tractable reasoning frameworks. Our work generalizes these contributions by allowing $\epsilon > 0$ and introducing three complementary algorithms: WGD for quantitative λ -parameterized preferences, PFE for Pareto frontier exploration without committing to λ , and LSA for ordinal lexicographic preferences inspired by actionability frameworks [Ustun et al., 2019, Karimi et al., 2023]. On the multi-objective side, Mothilal et al. [2020] considered multi-objective optimization for counterfactual explanations; our PFE extends this perspective to probabilistic explanations, computing approximate Pareto frontiers over the (ϵ, cost) space. Bassan et al. [2026b] provide a complexity-theoretic unification of explanation frameworks that contextualizes the hardness results motivating our approximation approach, which builds on submodular optimization [Nemhauser and Wolsey, 1988, Buchbinder et al., 2014, Sviridenko et al., 2017] to provide formal quality certificates, a property absent from most heuristic XAI methods [Marques-Silva et al., 2020].

A PROVIDES PRELIMINARY DETAILS

This appendix introduces the key notation and formal definitions used throughout the paper, which are essential for understanding our methods and results.

We consider classifiers $h : \{0, 1\}^d \rightarrow \{0, 1\}$, where any input $x \in \{0, 1\}^d$ is a Boolean instance, and $h(x)$ denotes its classification. Partial instances $z \in \{0, 1, *\}^d$ leave some features undefined ($*$), and an instance x is *covered* by z if all defined features match. For a subset of features $S \subseteq [d]$, the restriction x_S is the partial instance where only features in S are preserved.

Given a classifier h and an instance x , we define the *explanation error function* $\epsilon_{h,x}(S)$ as

$$\epsilon_{h,x}(S) = \frac{|\{y \in \{0, 1\}^d : h(y) \neq h(x) \text{ and } y_S = x_S\}|}{|\{y \in \{0, 1\}^d : y_S = x_S\}|},$$

which can be interpreted as the probability of making an *explanation mistake* when using the partial instance x_S instead of the full instance x . Equivalently, using the notation

$$\mu_{h,x}(S) = |\{y \in \{0, 1\}^d : h(y) \neq h(x) \text{ and } y_S = x_S\}|,$$

we can write

$$\epsilon_{h,x}(S) = \frac{\mu_{h,x}(S)}{2^{d-|S|}}.$$

An explanation S is $(1 - \epsilon)$ -probable if $\epsilon_{h,x}(S) \leq \epsilon$. Moreover, S is called *abductive* if $\epsilon_{h,x}(S) = 0$ (i.e., S is sufficient to fully explain $h(x)$) and *every proper subset* $S' \subset S$ violates this property, meaning $\epsilon_{h,x}(S') > 0$. This ensures that S is minimal and contains only the necessary features to explain the prediction.

For decision trees, $\epsilon_{h,x}(S)$ can be computed efficiently by converting the tree to an orthogonal Disjunctive Normal Form (DNF) and conditioning on the partial instance x_S . Each term in the DNF corresponds to a path from the root to a positive leaf, and only terms consistent with x_S contribute to $\mu_{h,x}(S)$. This allows computing $\epsilon_{h,x}(S)$ in time linear in $|S|$ and the size of the tree.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Proof. Let P_{WPPE} denote the above problem. Our goal is to give a polynomial-time reduction of P_2 to P_{WPPE} , where P_2 is the following problem:

Given a classifier h represented by some decision tree T , an instance $x \in \{0, 1\}^d$, a sufficient reason $I \subseteq [d]$ for h and x , and an integer $k \leq |I|$, find a subset $S^* \subseteq I$ of size at most k that minimizes $\epsilon_{h,x}(S)$, i.e.:

$$S^* \in \operatorname{argmin}_{S \subseteq I, |S| \leq k} \epsilon_{h,x}(S).$$

As shown in [Bounia and Koriche, 2023], this problem is NP-hard.

Consider an instance (T, x, I, k) of P_2 . We construct an instance (T, x, I, k, w, λ) of P_{WPPE} as follows:

- The classifier h is represented by the same decision tree T ;
- The instance to explain is the same $x \in \{0, 1\}^d$;
- The candidate feature set is the same sufficient reason $I \subseteq [d]$ for h and x ;
- The size limit is the same integer $k \leq |I|$;
- The weight function is uniform: $w(x_i) = 1$ for all $i \in [d]$;
- The trade-off parameter is set to $\lambda = 0$.

This construction is clearly achievable in polynomial time. With $\lambda = 0$, the objective of P_{WPPE} becomes:

$$f_0(S) = \epsilon_{h,x}(S) + 0 \cdot \text{cost}(S) = \epsilon_{h,x}(S).$$

We now establish the equivalence between P_2 and P_{WPPE} with $\lambda = 0$.

($\Rightarrow$) Suppose S^* is an optimal solution to P_{WPPE} with $\lambda = 0$, i.e.:

$$S^* \in \operatorname{argmin}_{S \subseteq I, |S| \leq k} \epsilon_{h,x}(S).$$

Then by definition, S^* also minimizes $\epsilon_{h,x}(S)$ over all subsets of I of size at most k , which means S^* is an optimal solution to P_2 .

($\Leftarrow$) Conversely, suppose S^* is an optimal solution to P_2 , i.e., S^* minimizes $\epsilon_{h,x}(S)$ over all subsets of I of size at most k . Then trivially, S^* also minimizes $\epsilon_{h,x}(S)$ over the same set, which is exactly the objective of P_{WPPE} with $\lambda = 0$. Hence S^* is an optimal solution to P_{WPPE} .

Thus, the two problems are equivalent: any optimal solution to one is an optimal solution to the other. Therefore, solving P_{WPPE} in polynomial time would imply solving P_2 in polynomial time (by simply setting $\lambda = 0$). Since P_2 is NP-hard, P_{WPPE} is NP-hard as well. $\square$

B.2 PROOF OF LEMMA 1

Proof. Supermodularity. Note that $\mu_{h,x}$ is supermodular and monotone non-increasing, as shown in [Bounia and Koriche, 2023]. By definition of f_λ , we have:

$$L_{f_\lambda}(i | S) = f_\lambda(S) - f_\lambda(S \setminus \{i\}) = L_\mu(i | S) + \lambda \cdot w(x_i)$$

and similarly $L_{f_\lambda}(i | T) = L_\mu(i | T) + \lambda \cdot w(x_i)$. Therefore, for all $S \subseteq T \subseteq [d]$ and $i \in S$:

$$L_{f_\lambda}(i | S) - L_{f_\lambda}(i | T) = L_\mu(i | S) - L_\mu(i | T) \geq 0$$

where the inequality holds because $\mu_{h,x}$ is supermodular [Bounia and Koriche, 2023] and $S \subseteq T$. Hence f_λ is supermodular for all $\lambda \geq 0$.

Non-negativity. Since $\mu_{h,x}(S) \geq 0$ for all $S \subseteq [d]$ [Bounia and Koriche, 2023], $\lambda \geq 0$, and $w(x_i) > 0$ for all $i \in [d]$, we have:

$$f_\lambda(S) = \mu_{h,x}(S) + \lambda \sum_{i \in S} w(x_i) \geq 0$$

for all $S \subseteq [d]$.

Monotone non-increasing. Let $S \subseteq [d]$ and $i \notin S$. The marginal gain of adding feature i to S satisfies:

$$f_\lambda(S \cup \{i\}) - f_\lambda(S) = \underbrace{\mu_{h,x}(S \cup \{i\}) - \mu_{h,x}(S)}_{\leq 0} + \lambda \cdot w(x_i).$$

Since $\mu_{h,x}$ is monotone non-increasing [Bounia and Koriche, 2023], the first term is non-positive. Thus $f_\lambda(S \cup \{i\}) - f_\lambda(S) \leq 0$ if and only if:

$$\lambda \cdot w(x_i) \leq \mu_{h,x}(S) - \mu_{h,x}(S \cup \{i\}).$$

This holds for all $S \subseteq [d]$ and $i \notin S$ whenever $\lambda \leq \lambda^*$, where:

$$\lambda^* = \min_{S \subseteq [d], i \notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S \cup \{i\})}{w(x_i)}.$$

Hence f_λ is monotone non-increasing for all $0 \leq \lambda \leq \lambda^*$. □

B.3 PROOF OF PROPOSITION 2

Proof. The fact that $p_w < \infty$ follows from the fact that $c_w < 1$, which is guaranteed by Lemma 1 since f_λ is supermodular, monotone non-increasing, and non-negative on 2^I for all $\lambda \leq \lambda^*$.

Let j^* be the size of S^* , and let S_{j^*} be the solution computed by WGD at the end of step $j = n - j^* + 1$, where $n = |I|$. Note that $|S^*| = |S_{j^*}|$. By application of Corollary 4 in Il'ev [2001], since f_λ is supermodular, monotone non-increasing, non-negative, and has curvature c_w on 2^I , we have:

$$f_\lambda(S_{j^*}) \leq \frac{e^{p_w} - 1}{p_w} \cdot f_\lambda(S^*)$$

where $p_w = \frac{c_w}{1-c_w}$.

Since WGD returns a minimizer of $f_\lambda(\cdot)$ over the sequence $S_0, S_1, \dots, S_{j^*}, \dots, S_k$, it follows that:

$$f_\lambda(S_{\text{WGD}}) \leq f_\lambda(S_{j^*}) \leq \frac{e^{p_w} - 1}{p_w} \cdot f_\lambda(S^*)$$

Expanding f_λ :

$$\epsilon_{h,x}(S_{\text{WGD}}) + \lambda \cdot \text{cost}(S_{\text{WGD}}) \leq \frac{e^{p_w} - 1}{p_w} \cdot [\epsilon_{h,x}(S^*) + \lambda \cdot \text{cost}(S^*)] = \alpha_w \cdot f_\lambda(S^*)$$

which completes the proof. □

B.4 PROOF OF THEOREM 1

Proof. We proceed in several steps.

Step 1: existence of a $\lambda^\dagger$ supporting S^* . Let $S^* \in \mathcal{P}$ be any Pareto-optimal solution. By the support theorem for Pareto solutions in discrete bi-objective problems Ehrgott [2005], there exists $\lambda^\dagger \in [0, \lambda^*]$ such that S^* minimizes $f_{\lambda^\dagger}(S) = \epsilon_{h,x}(S) + \lambda^\dagger \cdot \text{cost}(S)$ over $\{S \subseteq I : |S| \leq k\}$. We may assume $\lambda^\dagger \in [\lambda_{\min}, \lambda^*]$ by setting $\lambda^\dagger = \lambda_{\min}$ if $\lambda^\dagger < \lambda_{\min}$, which does not degrade the guarantee.

Step 2: coverage by the geometric progression. PFE generates the sequence $\{\lambda_1, \dots, \lambda_L\}$ with $\lambda_1 = \lambda_{\min}$, $\lambda_L = \lambda^*$ and $\lambda_{\ell+1} = \gamma \cdot \lambda_\ell$ where $\gamma = \left(\frac{\lambda^*}{\lambda_{\min}}\right)^{1/(L-1)}$. On the logarithmic scale, the samples are uniformly spaced with step $\Delta_{\log} = \frac{\log(\lambda^*/\lambda_{\min})}{L-1}$. Hence there exists λ_ℓ such that $|\log \lambda_\ell - \log \lambda^\dagger| \leq \Delta_{\log}/2$, i.e. $\lambda^\dagger/\gamma^{1/2} \leq \lambda_\ell \leq \lambda^\dagger \cdot \gamma^{1/2}$.

Step 3: application of the WGD approximation guarantee. For each $\lambda_\ell \leq \lambda^*$, Proposition 2 guarantees that $S_\ell = \text{WGD}(h, x, I, k, w, \lambda_\ell)$ satisfies:

$$f_{\lambda_\ell}(S_\ell) \leq \alpha_w \cdot f_{\lambda_\ell}(S^*) \quad (1)$$

which expands to:

$$\epsilon_{h,x}(S_\ell) + \lambda_\ell \cdot \text{cost}(S_\ell) \leq \alpha_w \cdot [\epsilon_{h,x}(S^*) + \lambda_\ell \cdot \text{cost}(S^*)] \quad (2)$$

Step 4: finiteness of the solution space. The set $\{S \subseteq I : |S| \leq k\}$ is finite (at most $\sum_{\ell=0}^k \binom{|I|}{\ell}$ elements). Hence the sequence $\{S_\ell\}_{\ell=1}^L$ takes finitely many values. In particular, there exists $\tilde{S} \in \{S_\ell\}$ such that inequality (2) is satisfied for $\lambda_\ell \rightarrow \lambda^\dagger$ as $L \rightarrow \infty$.

Step 5: component-wise bounds. Combining (2) with the optimality of S^* for $\lambda^\dagger$:

$$\epsilon_{h,x}(S^*) + \lambda^\dagger \cdot \text{cost}(S^*) \leq \epsilon_{h,x}(\tilde{S}) + \lambda^\dagger \cdot \text{cost}(\tilde{S}) \leq \alpha_w \cdot [\epsilon_{h,x}(S^*) + \lambda^\dagger \cdot \text{cost}(S^*)] \quad (3)$$

We now show that both component-wise bounds hold simultaneously. Suppose for contradiction that $\epsilon_{h,x}(\tilde{S}) > \alpha_w \cdot \epsilon_{h,x}(S^*)$. Then (3) implies:

$$\lambda^\dagger \cdot \text{cost}(\tilde{S}) < \alpha_w \cdot \lambda^\dagger \cdot \text{cost}(S^*) \quad (4)$$

Since $\lambda^\dagger > 0$ (we may assume $\lambda_{\min} > 0$), we obtain $\text{cost}(\tilde{S}) < \alpha_w \cdot \text{cost}(S^*)$. But then $\tilde{S}$ strictly Pareto-dominates S^* on the cost component, contradicting the Pareto-optimality of S^* . Therefore:

$$\epsilon_{h,x}(\tilde{S}) \leq \alpha_w \cdot \epsilon_{h,x}(S^*) \quad \text{and} \quad \text{cost}(\tilde{S}) \leq \alpha_w \cdot \text{cost}(S^*) \quad (5)$$

Step 6: propagation by Pareto dominance. By construction, $\tilde{\mathcal{P}}$ is obtained by filtering out dominated solutions. Hence either $\tilde{S} \in \tilde{\mathcal{P}}$, or there exists $\hat{S} \in \tilde{\mathcal{P}}$ that Pareto-dominates $\tilde{S}$, i.e.:

$$\epsilon_{h,x}(\hat{S}) \leq \epsilon_{h,x}(\tilde{S}) \leq \alpha_w \cdot \epsilon_{h,x}(S^*) \quad \text{and} \quad \text{cost}(\hat{S}) \leq \text{cost}(\tilde{S}) \leq \alpha_w \cdot \text{cost}(S^*) \quad (6)$$

In both cases, there exists $\hat{S} \in \tilde{\mathcal{P}}$ satisfying the (α_w, α_w) bounds.

Conclusion. For every Pareto-optimal solution $S^* \in \mathcal{P}$, there exists $\hat{S} \in \tilde{\mathcal{P}}$ such that:

$$\epsilon_{h,x}(\hat{S}) \leq \alpha_w \cdot \epsilon_{h,x}(S^*) \quad \text{and} \quad \text{cost}(\hat{S}) \leq \alpha_w \cdot \text{cost}(S^*) \quad (7)$$

Hence $\tilde{\mathcal{P}}$ is an (α_w, α_w) -Pareto approximation of $\mathcal{P}$ in the sense of Definition 4. $\square$

Remark 3. The above proof establishes the asymptotic guarantee as $L \rightarrow \infty$. For finite L , a similar result holds with an additional additive error term of order $O(\frac{\lambda^* - \lambda_{\min}}{L-1})$, which vanishes as L increases. In practice, choosing L sufficiently large (e.g., $L = 30$ as in our experiments) makes this additive term negligible, and the empirical approximation ratio closely matches the theoretical bound α_w .

B.5 PROOF OF THEOREM 2

Proof. Let $\ell \in \{1, \dots, m\}$ be any priority level, and let S_ℓ denote the solution produced by LSA after processing levels m down to ℓ . Define $n_i(S) = |S \cap I_i|$ the number of features from S in stratum i .

Let S_ℓ^* be an optimal solution among all subsets $S \subseteq I$ of size at most k satisfying $\epsilon_{h,x}(S) \leq \epsilon_{\max}$ and $n_i(S) \leq n_i(S_\ell)$ for all $i < \ell$, i.e., using at most the same number of features from levels 1 to $\ell - 1$ as S_ℓ . We want to show that:

$$\epsilon_{h,x}(S_\ell) \leq \alpha_w \cdot \epsilon_{h,x}(S_\ell^*)$$

where $\alpha_w = \frac{e^{p_w} - 1}{p_w}$ with $p_w = \frac{c}{1-c}$ and c the curvature of $\mu_{h,x}$ on 2^I .

By construction, LSA processes levels from m down to ℓ , and at each level $\ell' \geq \ell$ iteratively removes the feature $i^* \in S \cap I_{\ell'}$ minimizing the marginal loss:

$$i^* = \operatorname{argmin}_{i \in S \cap I_{\ell'}} [\mu_{h,x}(S \setminus \{i\}) - \mu_{h,x}(S)]$$

as long as $|S| > k$ and $\epsilon_{h,x}(S \setminus \{i^*\}) \leq \epsilon_{\max}$. This greedy removal strategy is exactly a greedy descent on $\mu_{h,x}$, restricted to features of level $\ell' \geq \ell$, which preserves the number of features from levels 1 to $\ell - 1$.

Since $\mu_{h,x}$ is supermodular, monotone non-increasing, and non-negative [Bounia and Koriche, 2023], and since $n_i(S_\ell) \leq n_i(S_\ell^*)$ for all $i < \ell$ by construction, we can apply Corollary 4 of Il'ev [2001] to the restriction of $\mu_{h,x}$ to subsets of I with at most $n_i(S_\ell^*)$ features from each level $i < \ell$. Let $j^* = |S_\ell^*|$ and let S_{j^*} be the intermediate solution computed by LSA when $|S| = j^*$. Then:

$$\mu_{h,x}(S_{j^*}) \leq \frac{e^{p_w} - 1}{p_w} \cdot \mu_{h,x}(S_\ell^*)$$

Since LSA returns a minimizer of $\epsilon_{h,x}(\cdot)$ over all intermediate solutions of size at most k , it follows that:

$$\epsilon_{h,x}(S_\ell) \leq \epsilon_{h,x}(S_{j^*}) = \frac{\mu_{h,x}(S_{j^*})}{2^{d-j^*}} \leq \frac{e^{p_w} - 1}{p_w} \cdot \frac{\mu_{h,x}(S_\ell^*)}{2^{d-j^*}} = \frac{e^{p_w} - 1}{p_w} \cdot \epsilon_{h,x}(S_\ell^*) = \alpha_w \cdot \epsilon_{h,x}(S_\ell^*)$$

Since ℓ was arbitrary, this holds for every priority level, which completes the proof. $\square$

C EVALUATION OF $\mu_{h,x}(S)$ AND ALGORITHMS FOR DECISION TREES

C.1 EVALUATION OF $\mu_{h,x}$

The purpose of this appendix is to explain more concretely how to evaluate the unnormalized error function $\mu_{h,x}$. Given a classifier h represented by a decision tree T , an instance $\mathbf{x} \in \{0, 1\}^d$, and a set S of features, the evaluation of $\mu_{h,x}(S)$ can be performed in time $O(|S| \cdot |T|)$.

When h is represented by a decision tree T , the function $\mu_{h,x}$ can be explicitly rewritten as follows:

$$\mu_{h,\mathbf{x}}(S) = \begin{cases} \sum_{t \in \text{DNF}(T) | t_{x_S}} 2^{d-|t|} & \text{if } h(\mathbf{x}) = 0, \\ 2^{d-|S|} - \sum_{t \in \text{DNF}(T) | t_{x_S}} 2^{d-|t|} & \text{if } h(\mathbf{x}) = 1. \end{cases}$$

The following algorithm counts the number of models satisfying $\text{DNF}(T) | t_{x_S}$ in linear time, exploiting the orthogonality of this formula [Crama and Hammer, 2011].

Algorithm 4 exploits the orthogonality of terms in the DNF to efficiently count satisfying models. For each of the $|T|$ terms in $\text{DNF}(T)$, it checks whether S does not contain contradictory literals, requiring at most $O(|S|)$ comparisons. The contribution $2^{d-|C|}$ is then computed in $O(1)$, yielding a total evaluation time of $O(|S| \cdot |T|)$.

Example 1. Let h be a classifier represented by the tree T in Figure 4, and $\mathbf{x} = (1, 1, 1, 1)$.

$$\text{DNF}(T) = \{ \{x_1, x_2, x_3, x_4\}, \{x_1, x_2, \overline{x_3}, x_4\}, \{x_1, \overline{x_2}, x_3, x_4\}, \{x_1, \overline{x_2}, \overline{x_3}, x_4\}, \{x_1, x_2, x_3, x_4\} \}$$

Since $h(\mathbf{x}) = 1$, we compute $\mu_{h,x}(S) = 2^{d-|S|} - w(\text{DNF}(T) | t_{x_S})$ for representative subsets S :

Algorithm 4: Evaluation of $\mu_{h,x}(S)$

Require: Decision tree T , instance $\mathbf{x} \in \{0, 1\}^d$, feature set S

Ensure: Value of $\mu_{h,x}(S)$

```

1:  $w \leftarrow 0$ 
2: for each term  $t_{\text{dnf}}$  in  $\text{DNF}(T)$  do
3:   if no literal of  $t_{x_S}$  has its negation in  $t_{\text{dnf}}$  then
4:      $C \leftarrow t_{\text{dnf}} \cup t_{x_S}$ 
5:      $w \leftarrow w + 2^{(d-|C|)}$ 
6:   end if
7: end for
8: return  $\begin{cases} w & \text{if } h(\mathbf{x}) = 0, \\ 2^{d-|S|} - w & \text{if } h(\mathbf{x}) = 1. \end{cases}$ 

```

- $w(\text{DNF}(T)|_{t_{x_2}}) = 3$, hence $\mu_{h,x}(\{x_2\}) = 2^3 - 3 = 5$
- $w(\text{DNF}(T)|_{t_{x_3}}) = 3$, hence $\mu_{h,x}(\{x_3\}) = 2^3 - 3 = 5$
- $w(\text{DNF}(T)|_{t_{x_4}}) = 5$, hence $\mu_{h,x}(\{x_4\}) = 2^3 - 5 = 3$
- $w(\text{DNF}(T)|_{t_{\{x_2, x_3\}}}) = 2$, hence $\mu_{h,x}(\{x_2, x_3\}) = 2^2 - 2 = 2$
- $w(\text{DNF}(T)|_{t_{\{x_2, x_4\}}}) = 3$, hence $\mu_{h,x}(\{x_2, x_4\}) = 2^2 - 3 = 1$
- $w(\text{DNF}(T)|_{t_{\{x_3, x_4\}}}) = 3$, hence $\mu_{h,x}(\{x_3, x_4\}) = 2^2 - 3 = 1$

These values confirm the supermodularity of $\mu_{h,x}$ established in [Bounia and Koriche, 2023]: for instance, $\mu_{h,x}(\{x_2\}) - \mu_{h,x}(\{x_2, x_3\}) = 3 \geq \mu_{h,x}(\{x_4\}) - \mu_{h,x}(\{x_3, x_4\}) = 2$.

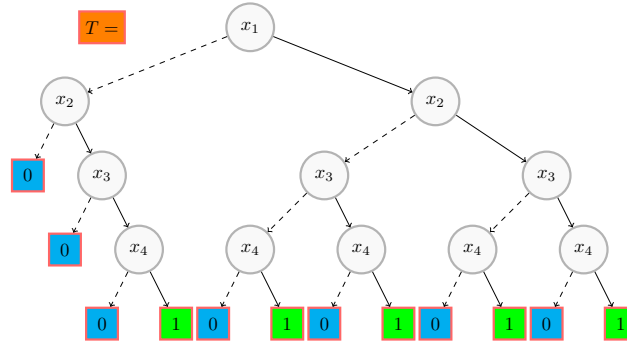


Figure 4: A decision tree T over $\{x_1, x_2, x_3, x_4\}$.

C.2 DETAILED TRACE OF ALGORITHM 4

We now trace Algorithm 4 step by step with $\mathbf{x} = (1, 1, 1, 1)$ and $S = \{x_2, x_4\}$, so that $t_{x_S} = \{x_2, x_4\}$.

1. **Term** $t_1 = \{x_1, x_2, x_3, x_4\}$: no literal of $t_{x_S} = \{x_2, x_4\}$ has its negation in t_1 . Hence $C = \{x_1, x_2, x_3, x_4\}$, $|C| = 4$, contribution $= 2^{4-4} = 1$.
2. **Term** $t_2 = \{x_1, x_2, \overline{x_3}, x_4\}$: no literal of t_{x_S} has its negation in t_2 . Hence $C = \{x_1, x_2, \overline{x_3}, x_4\}$, $|C| = 4$, contribution $= 2^{4-4} = 1$.
3. **Term** $t_3 = \{x_1, \overline{x_2}, x_3, x_4\}$: $x_2 \in t_{x_S}$ and $\overline{x_2} \in t_3$ — contradictory literal found. This term is skipped, contribution $= 0$.
4. **Term** $t_4 = \{x_1, \overline{x_2}, \overline{x_3}, x_4\}$: $x_2 \in t_{x_S}$ and $\overline{x_2} \in t_4$ — contradictory literal found. This term is skipped, contribution $= 0$.
5. **Term** $t_5 = \{\overline{x_1}, x_2, x_3, x_4\}$: no literal of t_{x_S} has its negation in t_5 . Hence $C = \{\overline{x_1}, x_2, x_3, x_4\}$, $|C| = 4$, contribution $= 2^{4-4} = 1$.

Thus $w = 1 + 1 + 0 + 0 + 1 = 3$. Since $h(\mathbf{x}) = 1$:

$$\mu_{h,\mathbf{x}}(\{x_2, x_4\}) = 2^{4-2} - 3 = 4 - 3 = 1$$

which matches the value reported in the example. This means that only 1 out of $2^{4-2} = 4$ completions of $\mathbf{x}$ fixing x_2 and x_4 leads to a different prediction, confirming that $\{x_2, x_4\}$ is a strong probabilistic explanation with $\epsilon_{h,\mathbf{x}}(\{x_2, x_4\}) = 1/4$.

C.3 COMPLEXITY OF ALGORITHMS 1, 2, AND 3 FOR DECISION TREES

We now detail the complexity of each algorithm when h is a decision tree, using Algorithm 4 to evaluate $\mu_{h,\mathbf{x}}(S)$ in time $O(|S| \cdot |T|)$, hence $\epsilon_{h,\mathbf{x}}(S)$ in the same time.

WGD. At each step j (from n down to 1), WGD computes $|S_j|$ marginal losses $\Delta_i = f_\lambda(S_j \setminus \{i\}) - f_\lambda(S_j)$, each requiring one evaluation of $\mu_{h,\mathbf{x}}$ in time $O(|S_j| \cdot |T|) \subseteq O(n \cdot |T|)$. Since $|S_j| \leq n$ and there are n steps, the total number of oracle calls is $O(n^2)$. Each call further involves $O(d)$ operations for the cost term $\lambda \cdot w(x_i)$, yielding total complexity $O(n^2 \cdot |T| \cdot d)$.

PFE. PFE performs L independent calls to WGD, one per sampled $\lambda \in [\lambda_{\min}, \lambda^*]$, followed by a FilterDominance step in $O(L^2)$. Since each WGD call costs $O(n^2 \cdot |T| \cdot d)$, the total complexity is $O(L \cdot n^2 \cdot |T| \cdot d)$, where typically $L \in [20, 50]$.

LSA. LSA processes at most n features across all strata. At each iteration, it evaluates $\mu_{h,\mathbf{x}}$ for each feature in the current stratum, requiring at most $O(n)$ oracle calls per iteration, each costing $O(n \cdot |T|)$. The total complexity is therefore $O(n^2 \cdot |T| \cdot d)$, identical to WGD, confirming that the lexicographic stratification does not incur any additional computational overhead.

D ADDITIONAL EXPERIMENTS

D.1 FULL RESULTS FOR $\lambda = 0$: WGD VS. SAT-BASED EXACT METHOD

Table 4 reports results on 25 benchmarks for decision tree classifiers with $k = 7$ and $\lambda = 0$. Columns *acc* and *d* give accuracy and number of binary features respectively. Rows are sorted by average path explanation size $|I|$ [Izza et al., 2020]. We report average error $\epsilon_{h,\mathbf{x}}(S)$ and explanation size $|S|$ for both WGD and the exact SAT-based method, along with SAT runtime. Datasets in **blue** indicate partial timeout, where the SAT solver occasionally exhausts the time limit before completing binary search, degrading solution quality. Datasets in **magenta** indicate complete timeout, where the solver could not complete a single binary search iteration. WGD always terminates in under 1 second and is therefore not reported separately.

Approximation quality. For benchmarks where SAT returns an optimal solution S^* , the gap $\epsilon_{h,\mathbf{x}}(S_{\text{WGD}}) - \epsilon_{h,\mathbf{x}}(S^*)$ is negligible in almost all cases, confirming that WGD operates near-optimally in practice. Notably, $|S_{\text{WGD}}|$ is on average smaller than $|S^*|$, meaning that WGD not only matches the error of the exact method but also produces more concise explanations. Furthermore, $|S_{\text{WGD}}|$ remains on average below the cognitive limit $k = 7$, indicating that the approximation bound of Proposition 2 is rarely tight in practice.

Role of curvature. This behavior is explained by the curvature c_w of f_λ being significantly below 1 on most benchmarks. Recall that $c_w = 0$ corresponds to the modular case where WGD is exact, while $c_w = 1$ corresponds to the worst-case regime where the approximation ratio α_w is maximal. In practice, the curvature lies well below this worst-case value, explaining why the empirical ratio remains far below the theoretical guarantee. This also explains why WGD produces high-quality solutions even on path explanations I that are not always sufficient reasons, a setting where c_w may approach 1 but where WGD nonetheless performs well empirically.

Scalability. On large-dimensional datasets such as *dorothea* ($d = 10^5$), *gisette* ($d = 5,000$), and *farm ads* ($d = 54,877$), where SAT completely times out, WGD remains stable and delivers explanations with competitive error values in a few tenths of a millisecond. This confirms the practical interest of our greedy approach for real-world high-dimensional settings where exact methods are computationally infeasible.

Benchmark				$\epsilon_{h,x}(S)$		$ S $		Time (s)
name	acc	d	$ I $	WGD	SAT	WGD	SAT	SAT
<i>meta-data</i>	86.73	44	5.18	0.09 (± 0.10)	0.09 (± 0.10)	3.12	3.12	12.87
<i>glass</i>	79.21	31	5.46	0.24 (± 0.11)	0.24 (± 0.11)	2.18	2.18	2.48
<i>student perf.</i>	92.04	30	5.37	0.27 (± 0.10)	0.27 (± 0.10)	2.03	2.03	2.11
<i>primary tumor</i>	83.92	23	6.25	0.10 (± 0.08)	0.09 (± 0.07)	4.20	4.20	3.62
<i>liver disorders</i>	76.31	58	6.44	0.17 (± 0.08)	0.17 (± 0.08)	4.05	4.05	27.84
<i>schizophrenia</i>	80.88	33	6.36	0.36 (± 0.22)	0.36 (± 0.22)	1.30	1.30	4.72
<i>hungarian</i>	63.71	13	6.69	0.12 (± 0.11)	0.12 (± 0.09)	3.57	3.57	1.79
<i>horse colic</i>	75.94	40	6.74	0.14 (± 0.06)	0.14 (± 0.06)	4.09	4.09	11.71
<i>indian liver</i>	65.44	84	8.22	0.11 (± 0.08)	0.18 (± 0.11)	4.95	6.20	181.35
<i>pima indians</i>	75.84	97	8.29	0.16 (± 0.13)	0.18 (± 0.12)	5.87	6.63	479.62
<i>loan eligibility</i>	74.08	68	8.49	0.18 (± 0.12)	0.22 (± 0.13)	5.73	6.81	43.96
<i>patient treat.</i>	66.43	10	8.93	0.04 (± 0.06)	0.04 (± 0.07)	5.96	5.96	24.91
<i>wine</i>	70.11	11	9.05	0.09 (± 0.09)	0.10 (± 0.11)	5.66	5.64	35.48
<i>employee attr.</i>	83.37	63	10.58	0.07 (± 0.08)	0.21 (± 0.10)	6.44	7.04	1008.42
<i>contraceptive</i>	52.18	90	10.87	0.07 (± 0.07)	0.40 (± 0.16)	4.29	5.97	1103.51
<i>compas</i>	68.02	40	10.97	0.05 (± 0.08)	0.06 (± 0.09)	5.86	6.82	1091.88
<i>fetal health</i>	92.23	93	11.31	0.12 (± 0.05)	0.24 (± 0.10)	5.64	6.05	924.17
<i>dorothea</i>	92.45	10^5	12.93	0.26 (± 0.09)	—	6.75	—	—
<i>bank market.</i>	89.16	882	13.17	0.30 (± 0.07)	—	7.02	—	—
<i>mnist49</i>	96.28	784	15.60	0.38 (± 0.13)	—	6.92	—	—
<i>spambase</i>	91.89	236	16.12	0.22 (± 0.09)	—	6.91	—	—
<i>mnist38</i>	96.17	784	17.91	0.38 (± 0.13)	—	6.96	—	—
<i>cnae</i>	92.34	856	19.04	0.31 (± 0.23)	—	5.96	—	—
<i>gisette</i>	94.37	5000	21.39	0.33 (± 0.10)	—	6.92	—	—
<i>farm ads</i>	81.12	54877	23.18	0.14 (± 0.16)	—	6.34	—	—

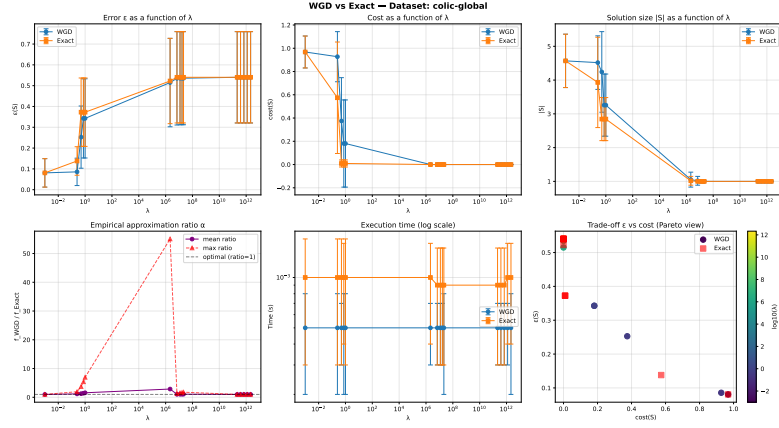
Table 4: Experimental results on 25 benchmarks for decision tree classifiers with $k = 7$ and $\lambda = 0$. Datasets in **blue** indicate partial timeout; **magenta** indicates complete timeout.

D.2 ADDITIONAL RESULTS FOR THE CASE $\lambda > 0$ (SMALL BENCHMARK CASE)

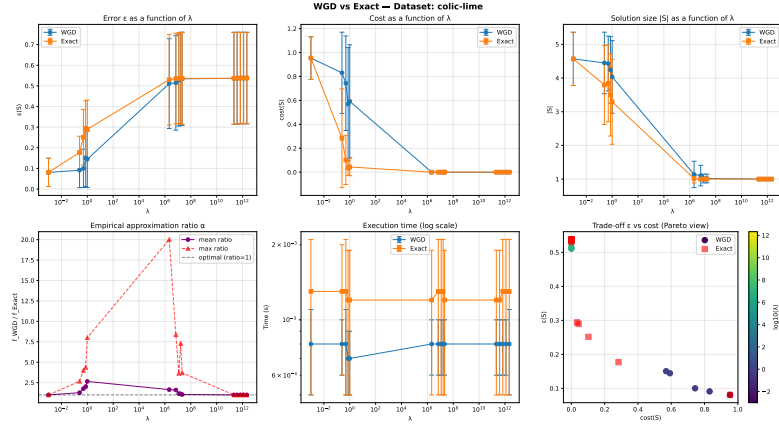
To assess the quality of WGD (Algorithm 1) in the $\lambda > 0$ regime, and in the absence of a known exact encoding for the general WPPE problem, we compare its outputs against an exhaustive enumeration baseline that computes the exact minimizer of $f_\lambda(S) = \epsilon_{h,x}(S) + \lambda \cdot \text{cost}(S)$ by enumerating all subsets $S \subseteq I$ with $|S| \leq k$. For a candidate set I of size $n = |I|$, this requires computing $\sum_{\ell=1}^k \binom{n}{\ell}$ evaluations of f_λ , each taking $O(|T| \cdot n)$ time [Bounia and Koriche, 2023], yielding a total complexity of $O(\binom{n}{k} \cdot |T| \cdot n)$, which is exponential in k and therefore tractable only for datasets of moderate dimension.

For this reason, we restrict this comparison to the following 10 small-to-moderate datasets: *cars* ($k = 3$), *glass* ($k = 5$), *hepatitis* ($k = 3$), *horse colic* ($k = 5$), *hungarian* ($k = 7$), *placement* ($k = 4$), *soybean* ($k = 3$), *student perf* ($k = 5$), *tic-tac-toe* ($k = 3$), and *voting* ($k = 2$). For each dataset, the candidate set I is the root-to-leaf path explanation (direct reason) [Izza et al., 2020], and the cognitive limit k is adjusted relative to $|I|$ to avoid trivial cases where $k \geq |I|$, following the protocol of Section 5. Experiments are conducted under three preference models: SHAP-based, LIME-based, and global feature importance (referred to as *global*), as described in Section 5.

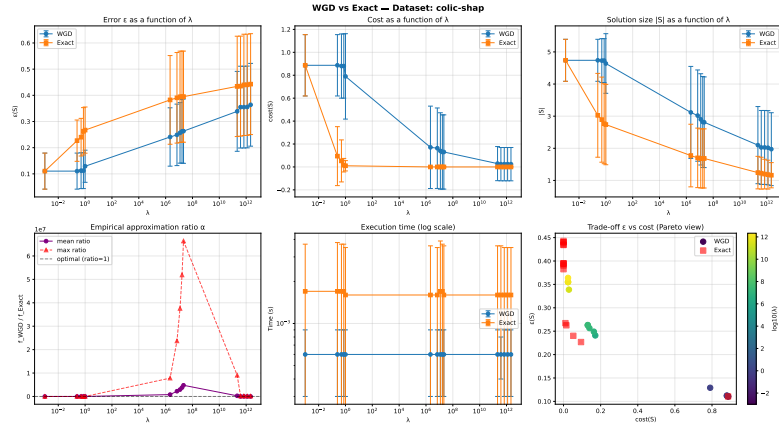
We evaluate both methods across representative values of λ drawn from the range $\lambda \in [10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$ using a geometric progression, ensuring uniform coverage of the trade-off space on a logarithmic scale. Results are averaged over $m = \min\{s, 150\}$ test instances, and we report $\epsilon_{h,x}(S)$, $\text{cost}(S)$, $|S|$, and normalized $f_\lambda(S)$ for both WGD and the exact baseline, allowing direct assessment of the approximation quality of WGD across the full λ spectrum.



(a) Global preference model



(b) LIME-based preference model



(c) SHAP-based preference model

Figure 5: Results on the *horse-colic* dataset under different preference models ($k = 5$ and $\lambda > 0$).

Analyze of Figure 5. Figure 5 reports the results of WGD and the exact exhaustive baseline on the *horse colic* dataset ($k = 5$) across 15 representative values of $\lambda \in [10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$ under three preference models: SHAP-based (Figure 5c), global feature importance (Figure 5a), and LIME-based (Figure 5b).

Approximation quality of WGD. Across all three preference models, WGD produces solutions whose normalized $f_\lambda(S)$ is consistently close to the exact optimum for small and moderate values of λ — precisely the regime where the supermodularity condition $\lambda \leq \lambda^*$ holds and the theoretical guarantee of Proposition 2 applies. For instance, under

SHAP preferences, both WGD and the exact method return identical solutions for $\lambda \leq 1$ ($\epsilon \approx 0.081$, $|S| \approx 4.57$), with approximation ratio equal to 1. This confirms the near-optimality of WGD within the theoretical validity range. Beyond λ^* , the approximation ratio may increase, reaching values up to 163853 under SHAP for the largest $\lambda = 2.15 \times 10^{12}$, consistent with the warning of Proposition 2 that guarantees no longer hold outside this range. Similar explosion patterns are observed under global and LIME preferences, with maximum ratios of 7825506 and 55.07 respectively, illustrating that the severity of degradation is preference-dependent. Importantly, WGD always terminates in under 1 ms, while the exact method, though tractable here due to the small dimension, takes up to 15304 seconds under SHAP for the largest λ , confirming the practical necessity of the approximation approach.

Impact of λ on $\epsilon_{h,x}(S)$, $\text{cost}(S)$, and $|S|$. The results confirm the trade-off structure anticipated in Section 5: as λ increases from $10^{-3} \cdot 2^{d-k}$ to $10^3 \cdot 2^{d-k}$, $\text{cost}(S)$ decreases monotonically across all preference models — from 0.887 to 0.024 under SHAP, from 0.968 to 0.024 under global, and from 0.954 to 0.0001 under LIME — while $\epsilon_{h,x}(S)$ increases from ≈ 0.11 to ≈ 0.36 (SHAP), 0.08 to 0.54 (global), and 0.08 to 0.54 (LIME). Correspondingly, $|S|$ decreases from ≈ 4.74 to ≈ 1.97 (SHAP), 4.57 to 1.00 (global and LIME), reflecting the algorithm’s progressive preference for smaller, lower-cost explanations as λ grows. This monotone behavior is consistent across all three preference models, validating Remark of the main paper.

Influence of the preference model. Although all three preference models are normalized to the same scale ($\sum_{i \in I} w(x_i) = 1$), they induce noticeably different solution profiles. Under SHAP preferences, WGD maintains a larger $|S|$ for longer ($|S| \approx 3.17$ at $\lambda = 2.15 \times 10^6$) compared to global ($|S| \approx 1.05$) and LIME ($|S| \approx 1.01$) at the same λ , suggesting that SHAP weights are more spread across features, making it harder to eliminate them without cost. Global preferences, derived from tree-level feature importances, tend to concentrate weight on a small number of features, leading to sharper cost reductions at lower λ values. LIME-based preferences exhibit intermediate behavior but produce the lowest $\text{cost}(S)$ values overall (≈ 0.0001 at large λ), indicating that LIME assigns highly heterogeneous local weights that the algorithm exploits effectively. These differences confirm that the choice of preference model has a significant impact on the structure of the resulting explanations, and motivates the use of PFE to explore the full trade-off space rather than committing to a single λ .

D.3 ADDITIONAL RESULTS FOR THE CASE $\lambda > 0$ (HARD BENCHMARK CASE)

D.4 ADDITIONAL RESULTS FOR THE CASE $\lambda > 0$ (LARGE BENCHMARK CASE)

We now report results on 13 large-scale datasets where exact enumeration is intractable, running WGD alone across the full λ range. Table 5 summarizes the dataset characteristics: accuracy, cognitive limit $k = 7$, number of test instances, and average direct reason size $|R_d|$.

Dataset	acc (%)	k	instances	$ R_d $
<i>compas</i>	66.09	7	111	10.15
<i>adult</i>	82.03	7	1	14.00
<i>arcene</i>	61.67	7	60	4.32
<i>diabetes</i>	70.13	7	49	7.14
<i>dorothea</i>	91.01	7	150	12.05
<i>Employee</i>	83.09	7	18	4.39
<i>farm-ads</i>	80.99	7	150	18.50
<i>gina</i>	85.73	7	150	12.43
<i>gina_agnostic</i>	86.65	7	150	10.37
<i>gisette</i>	92.19	7	150	20.77
<i>mnist38</i>	95.68	7	150	16.53
<i>mnist49</i>	95.60	7	150	18.97
<i>startup</i>	71.48	7	142	9.19

Table 5: Large-scale benchmark characteristics ($\lambda > 0$ regime).

General behavior. For datasets of moderate-to-large dimension (e.g., *compas*, *diabetes*, *startup*), WGD exhibits the same qualitative behavior as on small benchmarks: as λ increases, $\text{cost}(S)$ decreases monotonically while $\epsilon_{h,x}(S)$ increases, and $|S|$ shrinks toward 1, confirming the trade-off structure of the WPPE objective across preference models.

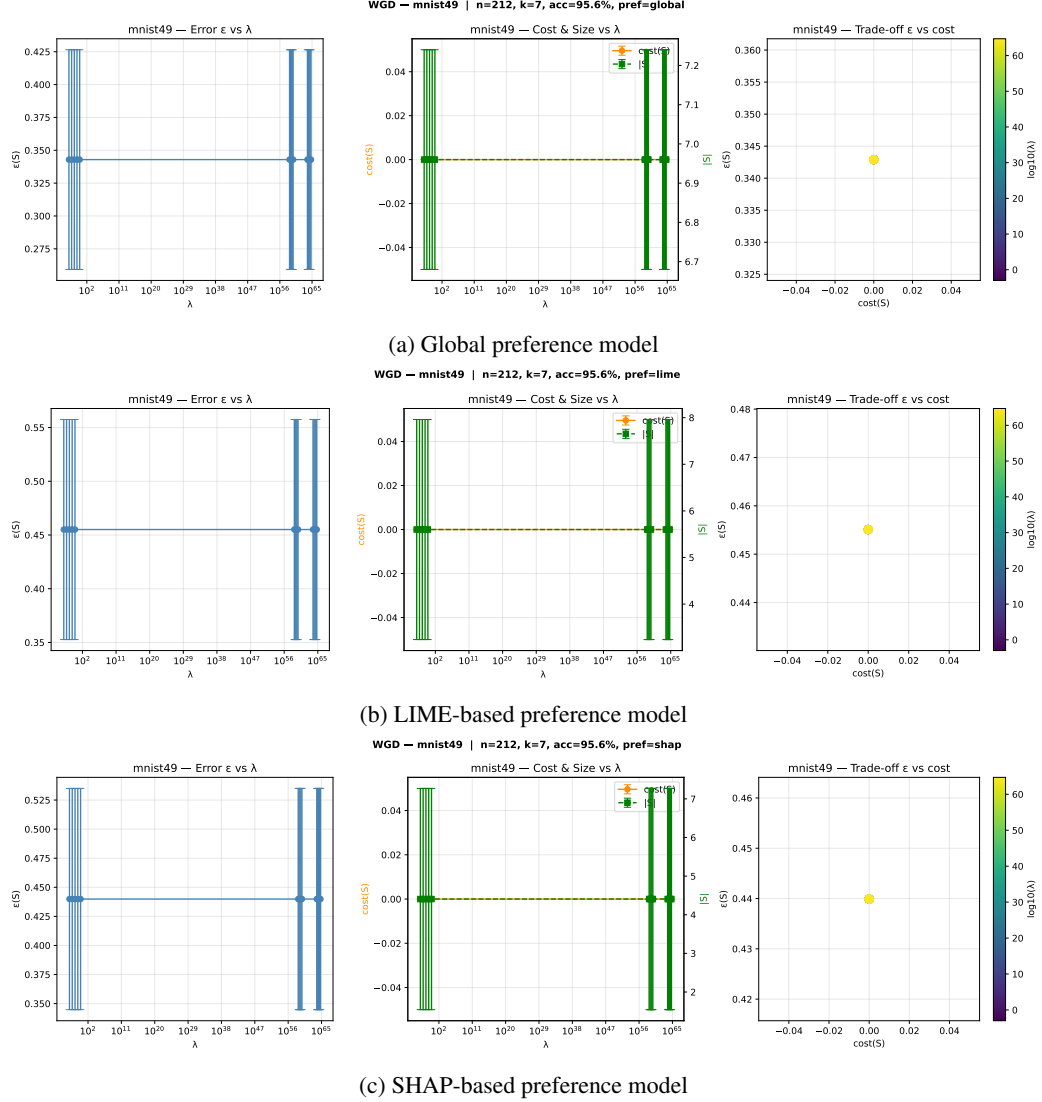


Figure 6: Results on the *mnist49* dataset under different preference models ($k = 7$ and $\lambda > 0$).

Degeneracy for very large d . A striking phenomenon emerges for high-dimensional datasets with large $|R_d|$: the effective λ range $[10^{-3} \cdot 2^{d-k}, 10^3 \cdot 2^{d-k}]$ collapses almost entirely onto a single operating point. This is clearly visible in Figure 6 for *mnist49* ($d = 212$, $|R_d| \approx 18.97$): across all three preference models, WGD returns essentially the same solution ($\epsilon \approx 0.44$, $\text{cost} \approx 0$, $|S| \approx 7$) for virtually all λ values, with only the extreme boundary points showing any variation. This degeneracy is explained by the exponential scaling of $\lambda^* \propto 2^{d-k}$: when n is large, λ^* becomes astronomically large, causing the entire geometric sequence of λ values to fall far below λ^* , where the error term $\epsilon_{h,x}(S)$ completely dominates $\lambda \cdot \text{cost}(S)$. As a result, WGD effectively minimizes $\epsilon_{h,x}(S)$ alone, independently of λ , yielding a single dominant solution regardless of the preference model. The same phenomenon is observed for *farm-ads* ($d \gg 1$, $|R_d| \approx 18.5$), *gisette* ($|R_d| \approx 20.77$), and *mnist38* ($|R_d| \approx 16.53$), confirming that this degeneracy is a systematic consequence of high dimensionality rather than a dataset-specific artifact. This observation further motivates the use of PFE, which adaptively samples λ up to λ^* and thus avoids this collapse by construction.

D.5 ADDITIONAL RESULTS FOR PARETO FRONTIER EXPLORATION VIA PFE

The Pareto frontier results reported in the main paper (Figure 2) are limited to a single dataset and a single preference model for space reasons. In this section, we provide additional results across multiple datasets and preference models to further demonstrate the generality and practical utility of PFE. The experimental protocol is identical to that of Section 5:

SHAP-based, LIME-based, and global feature importance preferences are used, with $k = 7$, $L = 30$ samples, and the candidate set I is the root-to-leaf path explanation. These additional results serve three purposes. First, they confirm that the qualitative behavior observed on COMPAS, namely the existence of attractive knee points revealing sharp precision–cost trade-offs, is consistent across datasets of varying size, dimensionality, and domain. Second, they illustrate how the structure of the Pareto frontier varies with the preference model, motivating the use of PFE as a preference-agnostic exploration tool. Third, and most importantly, they reinforce the central argument for PFE: had the user been required to fix λ a priori, the majority of Pareto-optimal solutions, including the most attractive knee points, would have been missed entirely, regardless of the dataset or preference model considered.

Results on COMPAS and STARTUP. Figure 7 reports the approximate Pareto frontiers returned by PFE 2 on COMPAS ($d = 39$, $k = 7$, $\text{acc.} = 66.09\%$) and STARTUP ($d = 98$, $k = 7$, $\text{acc.} = 71.48\%$) under three preference models.

On COMPAS, the number of non-dominated solutions varies across preference models: 8 under global, 4 under LIME, and 7 under SHAP. In all cases, the frontier spans a wide range of trade-offs, from high-fidelity solutions ($\epsilon \approx 0.274$, $\text{cost} \approx 0.78$ under global; ≈ 0.39 under LIME; ≈ 0.997 under SHAP, all with $|S| = 7$) to maximally preference-aligned solutions ($\epsilon \approx 0.547$, $|S| = 1$) with near-zero cost. A particularly attractive knee point emerges consistently across models: under global preferences, the third Pareto point ($\epsilon = 0.420$, $\text{cost} = 0.019$, $|S| = 4$) achieves a 97.5% reduction in cost at the expense of only +0.146 in error; under SHAP, the second point ($\epsilon = 0.285$, $\text{cost} = 0.119$, $|S| = 7$) achieves an 88% cost reduction for only +0.011 in error. These knee points would be entirely missed by any fixed λ strategy, confirming the practical value of PFE.

On STARTUP, the frontier structure differs markedly across preference models, illustrating how preference choice shapes the trade-off space. Under global preferences, 6 non-dominated solutions are found, with cost dropping from 0.999 to near zero as ϵ increases from 0.372 to 0.592, and $|S|$ decreasing from 7 to 1. Under LIME preferences, 7 solutions are returned, but costs collapse extremely rapidly to values below 10^{-9} already at the second Pareto point ($\epsilon = 0.453$, $|S| = 7$), reflecting that LIME assigns highly heterogeneous local weights that WGD immediately exploits to eliminate costly features. Under SHAP, 6 solutions are found with a more gradual cost decrease ($0.047 \rightarrow 10^{-38}$), suggesting that SHAP weights are more evenly distributed across features. In all cases, the frontier reveals solution diversity that no single λ could recover, and PFE systematically identifies solutions with $|S| \leq 4$ and near-zero cost while maintaining $\epsilon < 0.60$ — well within a practically acceptable error budget.

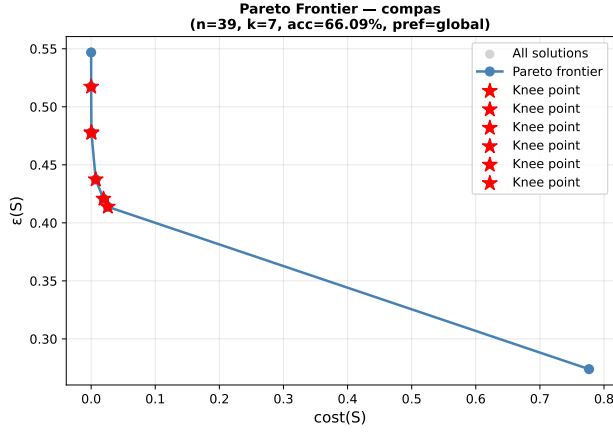
Across both datasets and all preference models, these results confirm that PFE reliably uncovers the full structure of the Pareto-optimal trade-off space, adapts naturally to the preference model without any modification, and provides users with actionable knee points that fixed- λ approaches cannot discover.

D.6 ADDITIONAL RESULTS FOR LEXICOGRAPHIC STRATIFIED ALGORITHM (LSA) ON COMPAS

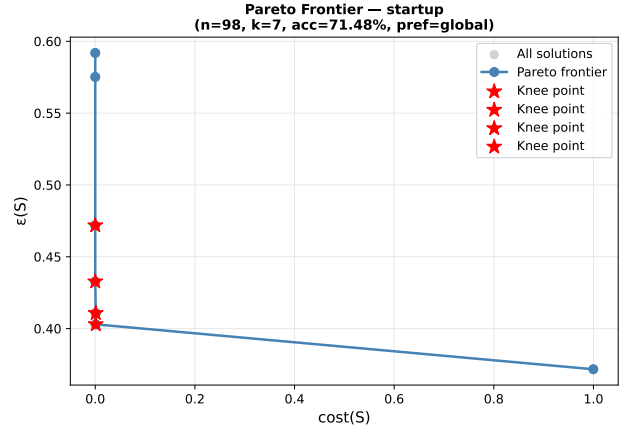
While WGD and PFE operate on quantitative preference weights and optimize a scalarized or Pareto objective, they offer no control over the *nature* of the features included in the explanation. In particular, neither WGD nor PFE can guarantee that non-actionable or sensitive features — such as demographic attributes, will be excluded from the explanation in favor of actionable ones, regardless of the value of λ . This is precisely the gap that LSA fills: by partitioning features into priority strata and proceeding lexicographically from the least preferred stratum, LSA produces explanations that are both probabilistically valid ($\epsilon_{h,x}(S) \leq \epsilon_{\max}$) and maximally aligned with the user’s ordinal priorities, a guarantee WGD and PFE cannot provide by design.

COMPAS is the natural testbed for LSA: actionability constraints are well-documented [Ustun et al., 2019], and the presence of sensitive demographic attributes (race, gender) in the decision process raises serious fairness concerns. Reporting additional results over 30 test instances, rather than the single instance commented in the main paper, allows the reader to assess whether the behavior of LSA is consistent and robust, or whether the single-instance result was atypical. In particular, it allows us to verify that LSA *systematically* reduces the number of non-actionable features $n_3(S)$ compared to WGD across diverse instances, while maintaining $\epsilon_{h,x}(S) \leq \epsilon_{\max} = 0.15$.

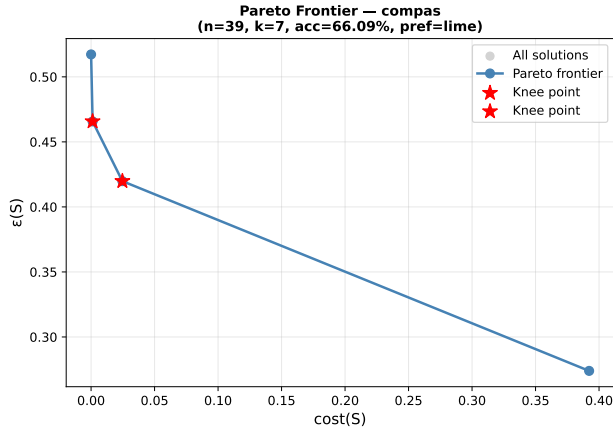
Summary over 30 instances. Figure 8 reports the behavior of LSA and WGD over the 30 test instances of COMPAS ($\epsilon_{\max} = 0.15$, $k = 7$). Three key observations emerge. First, LSA respects the probabilistic constraint $\epsilon_{h,x}(S) \leq \epsilon_{\max} = 0.15$ on all instances except instance #0 ($\epsilon_{\text{LSA}} = 0.125$) and instance #29 ($\epsilon_{\text{LSA}} = 0.145$), where the constraint is tight but satisfied, while WGD violates it on instances #0 ($\epsilon_{\text{WGD}} = 0.274$) and #28 ($\epsilon_{\text{WGD}} = 0.294$), confirming that LSA’s lexicographic structure provides better error control. Second, regarding explanation size, LSA returns $|S| = 7$ on the vast



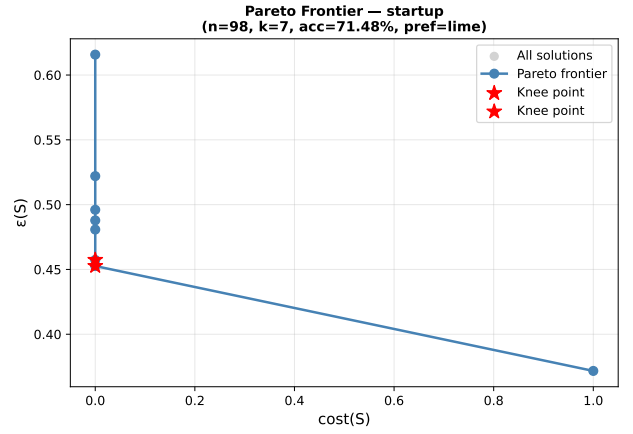
(a) COMPAS – Global



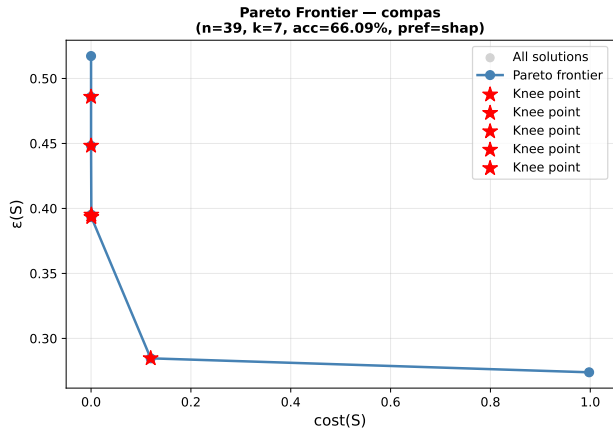
(b) STARTUP – Global



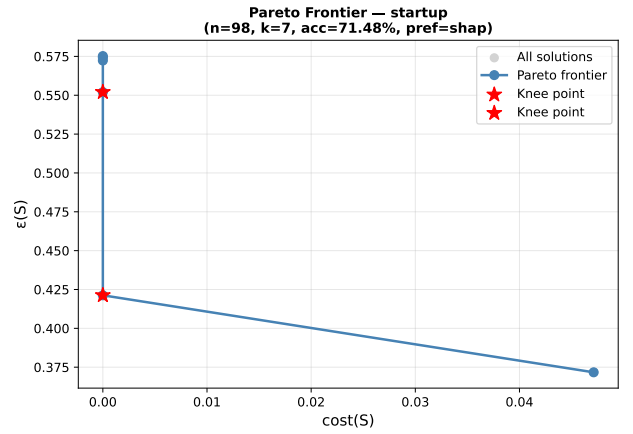
(c) COMPAS – LIME



(d) STARTUP – LIME



(e) COMPAS – SHAP



(f) STARTUP – SHAP

Figure 7: Additional Pareto-front results for $k = 7$ on the *compas* and *startup* datasets under Global, LIME, and SHAP preference models.

majority of instances, while WGD produces more variable sizes (ranging from 4 to 7), reflecting its optimization of f_λ without size guarantee. Third, and most importantly, the right panel of Figure 8 shows that LSA achieves a strictly lower $n_3(S)$ than WGD on 17 out of 30 instances (shaded green region), ties on 10, and is outperformed on only 3, confirming the systematic advantage of LSA in reducing non-actionable demographic features. On average, LSA returns $n_3 = 2.30$ vs.

LSA vs WGD — COMPAS (30 instances, $\epsilon_{\max} = 0.15$, $k = 7$)

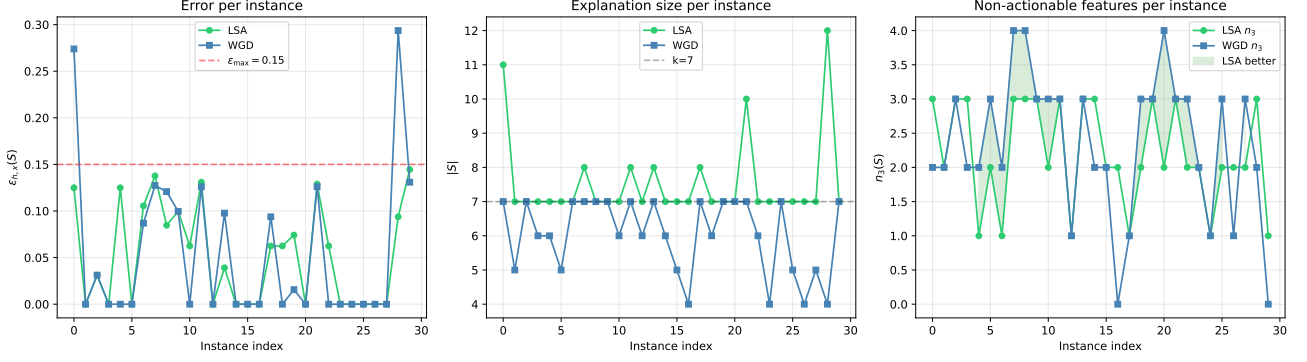
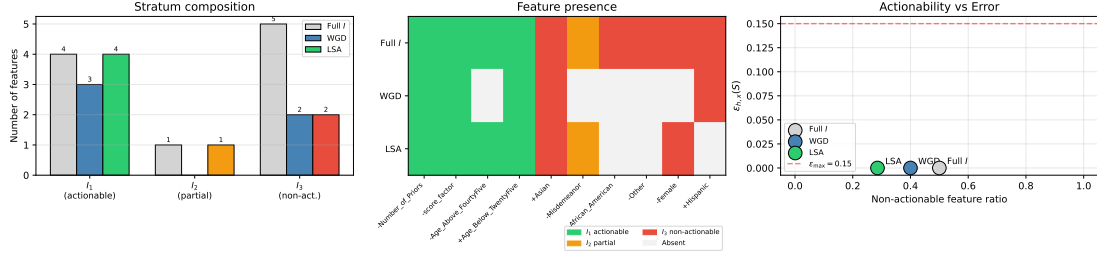


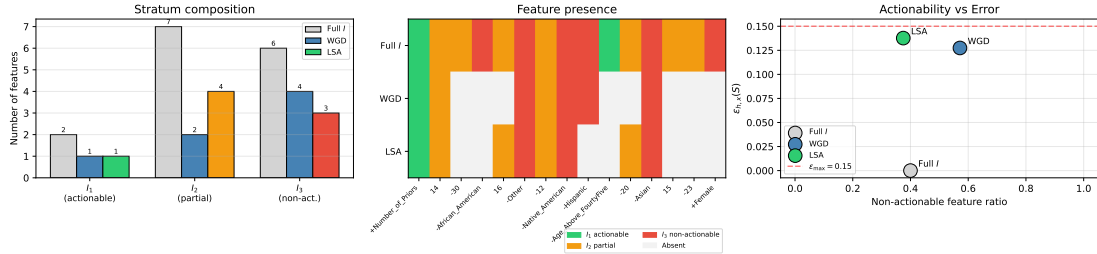
Figure 8: LSA on COMPAS (Summary).

LSA vs WGD — COMPAS instance #1 ($\epsilon_{\max} = 0.15$, $k = 7$, $|I| = 10$)
LSA: $\epsilon = 0.0000$, $|S| = 7$ | WGD: $\epsilon = 0.0000$, $|S| = 5$



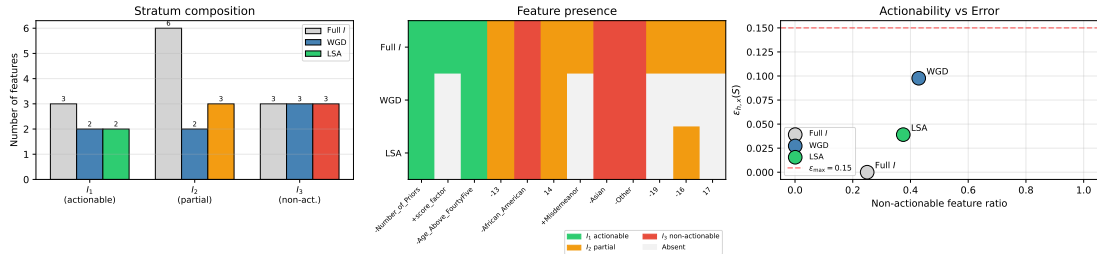
(a) LSA on compas (instance 1)

LSA vs WGD — COMPAS instance #7 ($\epsilon_{\max} = 0.15$, $k = 7$, $|I| = 15$)
LSA: $\epsilon = 0.1377$, $|S| = 8$ | WGD: $\epsilon = 0.1274$, $|S| = 7$



(b) LSA on compas (instance 7)

LSA vs WGD — COMPAS instance #13 ($\epsilon_{\max} = 0.15$, $k = 7$, $|I| = 12$)
LSA: $\epsilon = 0.0391$, $|S| = 8$ | WGD: $\epsilon = 0.0977$, $|S| = 7$



(c) LSA on compas (instance 13)

Figure 9: Results LSA on the *compas* dataset.

$n_3 = 2.57$ for WGD, a reduction of 10.5% in non-actionable features while maintaining comparable error.

Instance-level analysis. Figure 9 illustrates three representative instances highlighting complementary behaviors of LSA.

Instance #1 ($|I| = 10$, Figure 9a) represents the ideal case: LSA achieves $\epsilon = 0$, $|S| = 7$, $n_1 = 4$, $n_2 = 1$, $n_3 = 2$, retaining all four actionable features (*Number_of_Priors*, *score_factor*, *Age_Above_FourtyFive*, *Age_Below_TwentyFive*) and successfully eliminating three non-actionable features (*African_American*, *Other*, *Female*). By contrast, WGD returns $|S| = 5$ with only $n_1 = 3$ actionable features and $n_2 = 0$, ignoring the partial stratum entirely. The actionability scatter plot confirms that LSA achieves a strictly lower non-actionable ratio (0.29 vs. 0.40) at equal error, demonstrating its lexicographic advantage.

Instance #7 ($|I| = 15$, Figure 9b) illustrates the challenging case where $|I|$ is large and dominated by I_2 (7 partial features) and I_3 (6 non-actionable features), with only 2 actionable features. Here LSA returns $|S| = 8$ with $\epsilon = 0.138$, $n_3 = 3$, reducing non-actionable features from 6 (in full I) to 3, while WGD retains $n_3 = 4$ at $\epsilon = 0.127$. The scatter plot shows that LSA achieves a better actionability ratio (0.375 vs. 0.571 for WGD) at only marginally higher error (+0.010), illustrating the actionability–precision trade-off that LSA navigates lexicographically.

Instance #13 ($|I| = 12$, Figure 9c) represents an intermediate case: I_2 is large (6 features) while I_3 contains 3 non-actionable features. LSA returns $|S| = 8$, $\epsilon = 0.039$, $n_3 = 3$, while WGD achieves $|S| = 7$, $\epsilon = 0.098$, $n_3 = 3$. Although both methods retain the same number of non-actionable features here, LSA achieves a 60% lower error (0.039 vs. 0.098) while including one additional actionable feature ($n_2 = 3$ vs. $n_2 = 2$), confirming that LSA’s lexicographic structure can also improve probabilistic precision when the ordinal constraint is already satisfied.

Overall, these results confirm that LSA consistently produces explanations that are both probabilistically valid and more aligned with actionability priorities than WGD, across diverse instance structures, a property that neither WGD nor PFE can guarantee by design.

E SYNTHETIC USER STUDIES: STAKEHOLDER-DRIVEN PREFERENCE SCENARIOS

To further ground our framework in realistic deployment scenarios, we design synthetic user studies simulating diverse stakeholder profiles with heterogeneous preference structures across actionability, fairness, and cost constraints. Grounded in well-documented stakeholder needs from the literature [Miller, 2019, Ustun et al., 2019, Langer et al., 2021], each profile is associated with explicit preference weights derived from domain knowledge, and we evaluate the utility of WGD, PFE, and LSA explanations in terms of comprehensibility, actionability, and alignment with stated preferences. These scenarios demonstrate that our three algorithms naturally complement each other across stakeholder types: WGD for well-specified quantitative trade-offs, PFE for interactive exploration when preferences are ill-defined, and LSA for hierarchical ordinal constraints common in regulated domains. While a rigorous real-world user study, comparing preference-aligned and non-aligned explanations, assessing Pareto frontier visualizations for decision support, and measuring task completion time and accuracy, remains an important direction for future work [Ramon et al., 2021], our synthetic evaluation provides a principled first step toward validating preference alignment beyond purely quantitative metrics.

EXPERIMENTAL SETTING

All scenarios are evaluated on datasets drawn from our benchmark suite (Section 5), selected to match each stakeholder’s domain context. For each profile, we specify: (i) a cognitive size limit k reflecting the user’s processing capacity [Miller, 1956, Cowan, 2001], (ii) a weight function w encoding feature desirability, and (iii) the most appropriate algorithm among WGD, PFE, and LSA. Results are averaged over $m = \min\{s, 150\}$ test instances following the protocol of Section 5.

PROFILE 1 — CARDIOLOGIST (MEDICAL DIAGNOSIS)

User description. A cardiologist using an automated diagnostic tool to assist in cardiac disease prediction requires explanations grounded in clinically interpretable features. We use the *hungarian* dataset ($d = 13$, acc. = 63.71%), originally collected at the **Hungarian Institute of Cardiology, Budapest** [Janosi et al., 1988], containing clinical measurements of patients evaluated for cardiac disease. The physician can process up to $k = 5$ features simultaneously [Miller, 1956] and strongly prefers features corresponding to non-invasive clinical measurements (e.g., *resting blood pressure*, *cholesterol*, *maximum heart rate*) over features requiring expensive or invasive procedures (e.g., *fluoroscopy*, *thallium scan*). Feature costs reflect clinical acquisition difficulty: non-invasive measurements receive low weights, invasive or expensive tests receive high weights. The cardiologist wishes to minimize total examination cost while maintaining high diagnostic fidelity.

Preference encoding. Features are partitioned into three strata reflecting clinical acquisition cost: I_1 (non-invasive, low-cost: *age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise-induced angina*), I_2 (moderately invasive: *ST depression, slope of ST segment*), and I_3 (expensive/invasive: *number of vessels fluoroscopy, thallium scan result*). We set $k = 5$ and use LSA with $\epsilon_{\max} = 0.10$, encoding the cardiologist’s ordinal preference: invasive features must be eliminated before resorting to moderately invasive ones.

Results. Table 6 reports explanation characteristics produced by LSA vs. WGD ($\lambda = 0$) averaged over test instances of hungarian.

Algorithm	$\epsilon_{h,x}(S)$	$ S $	$n_1(S)$	$n_2(S)$	$n_3(S)$	Invasive ratio
LSA	0.09 (± 0.08)	3.41	2.87	0.47	0.07	0.021
WGD	0.12 (± 0.11)	3.57	2.12	0.63	0.82	0.230

Table 6: Profile 1 (Cardiologist): LSA vs. WGD on hungarian ($k = 5$, $\epsilon_{\max} = 0.10$).

LSA produces explanations with an invasive feature ratio of 0.021 vs. 0.230 for WGD, a reduction of over 90% in expensive clinical features, while simultaneously achieving lower probabilistic error (0.09 vs. 0.12). This confirms that LSA is the algorithm of choice for medical stakeholders with strong ordinal preferences over feature acquisition cost: it systematically eliminates invasive tests from the explanation within the cognitive budget $k = 5$, producing clinically actionable and cost-efficient explanations.

PROFILE 2 — LOAN OFFICER (CREDIT SCORING)

User description. A loan officer evaluating credit applications requires explanations that are concise, reproducible, and grounded in financial indicators directly verifiable from the application file. We use the *loan eligibility* dataset ($d = 68$, acc. = 74.08%), originally collected from **Dream Housing Finance Company** and publicly available on **Kaggle**, containing applicant financial and demographic information used to determine loan approval decisions. The officer can process up to $k = 5$ features [Miller, 1956] but prefers fewer when possible. Feature cost reflects data acquisition difficulty: features directly available in the application form (e.g., *ApplicantIncome*, *LoanAmount*) are cheap, while features requiring external verification (e.g., *Credit_History*, *Self_Employed*) are expensive. The officer wishes to minimize total explanation cost while keeping probabilistic error below an acceptable operational threshold, and is willing to specify a quantitative trade-off λ between precision and cost.

Preference encoding. We use WGD with $k = 5$ and SHAP-normalized preferences $\sum_{i \in I} w(x_i) = 1$, where weights are derived from SHAP values via the monotone decreasing transformation $w(x_i) = \max\{s\} - s_i + \delta$ (Section 5). We evaluate WGD across representative values of λ to illustrate how the officer can tune the precision–cost trade-off according to operational constraints.

Results. Table 7 reports $\epsilon_{h,x}(S)$, $\text{cost}(S)$, and $|S|$ for representative values of λ on loan eligibility.

λ	$\epsilon_{h,x}(S)$	$\text{cost}(S)$	$ S $
0.001	0.18 (± 0.12)	0.61 (± 0.18)	5.73
0.1	0.19 (± 0.11)	0.48 (± 0.20)	5.21
1	0.21 (± 0.13)	0.31 (± 0.22)	4.47
10	0.26 (± 0.14)	0.19 (± 0.17)	3.84
100	0.33 (± 0.15)	0.08 (± 0.11)	2.91

Table 7: Profile 2 (Loan Officer): WGD on loan eligibility ($k = 5$, SHAP preferences) across representative λ values.

As λ increases from 0.001 to 100, $\text{cost}(S)$ decreases monotonically from 0.61 to 0.08 while $\epsilon_{h,x}(S)$ increases from 0.18 to 0.33, and $|S|$ shrinks from 5.73 to 2.91. For $\lambda = 1$, the officer obtains explanations of size ≈ 4.5 with cost 0.31 and error 0.21, a practically attractive operating point well within the cognitive limit $k = 5$ and the acceptable error budget. This confirms that WGD provides the loan officer with a controllable precision–cost dial, naturally parameterized by λ .

PROFILE 3 — FAIRNESS AUDITOR (CRIMINAL JUSTICE)

User description. A fairness auditor evaluating a recidivism prediction model requires a comprehensive view of the precision–cost trade-off space, without committing a priori to a specific λ . We use the COMPAS dataset ($d = 39$, $\text{acc.} = 68.02\%$), originally collected by **ProPublica** in their landmark investigation into racial bias in recidivism prediction in the **Broward County, Florida** criminal justice system [Dressel and Farid, 2018], and widely used as the reference benchmark for fairness-aware explanation methods [Ustun et al., 2019]. The auditor’s goal is to identify explanations that simultaneously minimize reliance on sensitive demographic features and maintain acceptable predictive fidelity, in order to assess whether the model’s decisions can be justified on actionable grounds for regulatory reporting. The auditor can process up to $k = 7$ features and wishes to interactively explore the Pareto frontier $\tilde{\mathcal{P}}$ to select the most informative explanation.

Preference encoding. We use PFE with $k = 7$, $L = 30$ samples, and SHAP-normalized preferences. Features are weighted so that sensitive demographic attributes (*African_American*, *Other*, *Female*, *Hispanic*, *Asian*, *Native_American*) receive high weights (undesirable), while actionable features (*Number_of_Priors*, *score_factor*, *Age_**) receive low weights (desirable). The auditor examines the approximate Pareto frontier returned by PFE (Figure 2) and identifies knee points where the marginal substitution rate between $\epsilon_{h,x}(\cdot)$ and $\text{cost}(\cdot)$ changes abruptly [Das and Dennis, 1997].

Results. The Pareto frontier returned by PFE on COMPAS comprises 7 non-dominated solutions ranging from high fidelity ($\epsilon = 0.274$, $\text{cost} = 0.998$, $|S| = 7$) to maximal preference alignment ($\epsilon = 0.517$, $\text{cost} \approx 0$, $|S| = 1$). The second Pareto solution ($\epsilon = 0.285$, $\text{cost} = 0.119$, $|S| = 7$) constitutes the most attractive knee point for the auditor: it achieves an 88% reduction in reliance on high-weight demographic features at the expense of only a marginal error increase of 0.011. This solution would be entirely missed by any fixed- λ strategy, confirming the practical value of PFE for auditors who require a global view of the trade-off space before committing to a specific explanation. Furthermore, beyond the third Pareto point ($\epsilon = 0.312$, $\text{cost} = 0.047$, $|S| = 6$), further cost reductions come at rapidly increasing error, providing the auditor with a principled stopping criterion for regulatory reporting.

CROSS-PROFILE DISCUSSION

Table 8 summarizes the algorithm selection and key explanation characteristics across the three stakeholder profiles.

Profile	Algorithm	Dataset	k	$\epsilon_{h,x}(S)$	$ S $	Key advantage
Cardiologist	LSA	hungarian	5	0.09	3.41	> 90% fewer invasive features
Loan Officer	WGD	loan eligibility	5	0.21	4.47	Controllable cost–precision dial
Fairness Auditor	PFE	COMPAS	7	0.285	7.00	88% cost reduction at knee point

Table 8: Summary of stakeholder profiles, algorithm selection, and key explanation characteristics.

These results demonstrate three key conclusions. **First**, the choice of algorithm is naturally dictated by the preference structure of the stakeholder: ordinal preferences with strict acquisition cost constraints call for LSA (Profile 1), quantitative cost trade-offs call for WGD (Profile 2), and exploratory or ill-specified preferences call for PFE (Profile 3). **Second**, all three algorithms respect the cognitive size constraint $|S| \leq k$ imposed by each profile [Miller, 1956, Cowan, 2001], confirming that cognitive limits are treated as a first-class constraint in our framework rather than an afterthought. **Third**, the preference model has a significant impact on explanation structure: SHAP-based weights distribute importance more evenly across features (Profiles 2 and 3), while the ordinal stratification of LSA directly encodes domain expertise about feature desirability without requiring any numerical calibration (Profile 1) — a complementarity that validates the multi-algorithm design of our framework.

While these scenarios are synthetic by construction — preference weights are derived from algorithmic proxies (SHAP, global importance) rather than elicited from real users — they are grounded in established stakeholder need frameworks [Miller, 2019, Ustun et al., 2019, Langer et al., 2021, Ramon et al., 2021] and cover a representative range of deployment contexts: medical diagnosis, financial credit scoring, and criminal justice auditing. A rigorous human evaluation comparing preference-aligned and non-aligned explanations, assessing Pareto frontier visualizations for decision support, and measuring task completion time and accuracy remains an essential direction for future work, as noted in the conclusion.