Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR meets Machine Learning and Explanation

# Model-Agnostic Explanations by Consensus

**Carlos Mencía**[1] , **Ramon Béjar**[2] , **Raúl Mencía**[1] , **Joao Marques-Silva**[3]
[1]Universidad de Oviedo, Spain
[2]Universitat de Lleida, Spain
[3]ICREA & Universitat de Lleida, Spain
{menciacarlos, menciaraul}@uniovi.es, ramon.bejar@udl.cat, jpms@icrea.cat

**Abstract**

We address the fundamental task of computing rigorous, sample-based abductive explanations for machine learning predictions. In this setting, we propose a new class of explanations derived from a generalization of the consensus operation in propositional logic. We prove that these explanations are precisely those that satisfy a monotonicity property ensuring they remain valid as the sample grows. Furthermore, we show that their computation can be performed efficiently. As a direct application, we also show how these explanations can be used to identify necessary and relevant features. The proposed framework provides a robust and scalable approach to formal model-agnostic XAI.

## 1 Introduction

Explainable Artificial Intelligence (XAI) represents a cornerstone of trustworthy AI. Nevertheless, the best-known methods of XAI lack rigor. As a result, computed explanations can mislead (Marques-Silva and Huang 2024) or be logically incorrect (Ignatiev 2020). Model-aware logic-based XAI provides the strongest guarantees of rigor, but with a few known pitfalls. These include the complexity of reasoning, and the size of explanations. Despite these pitfalls, progress in model-aware logic-based XAI has resulted in practically efficient methods for computing explanations for different families of machine learning (ML) models (Marques-Silva 2022; Darwiche 2023; Marques-Silva 2024). At present, scalability of model-aware logic-based explainability tracks that of deciding adversarial robustness (Wu, Wu, and Barrett 2023; Izza et al. 2024). Still, the fact that the computation of explanations requires reasoning about the logic representation of the ML model poses a continued challenge, especially given the ever-increasing complexity of ML models.

Recent years have witnessed the development of model-agnostic methods for computing explanations (Geng, Schleich, and Suciu 2022; Rudin and Shaposhnik 2023; Cooper and Amgoud 2023; Koriche et al. 2024; Marques-Silva, Lefebre-Lobaina, and Martinez 2025), where explanations are obtained from a sample of the ML model's behavior and not from a logic representation of the model. In contrast to other non-rigorous model-agnostic methods, e.g. (Ribeiro, Singh, and Guestrin 2018), logic-based model-agnostic explanations offer some guarantees of rigor. However, the fact that explanations are obtained from a sample raises a number of practical limitations (Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024).

One important limitation of model-agnostic explanations is the lack of coherence. Indeed, for explanations computed from (incomplete) samples of the ML model's behavior, it can happen that two explanations computed for two points of feature space exhibiting different predictions are consistent with each other. In practice, lack of coherence represents a source of distrust. The solutions proposed in earlier work to tackle incoherence (Amgoud, Cooper, and Debbaoui 2024) exhibit a number of shortcomings, including computation time, explanation size, as well as assumptions made about unknown points of feature space.

Another desirable property is monotonicity, which guarantees that explanations remain valid after adding more instances to the sample, providing a level of reliability that current sample-based approaches lack, except for trivial explanations containing all the features. To the best of our knowledge, no existing solutions fulfill this property for non-trivial, model-agnostic explanations.

This paper proposes a new approach to computing model-agnostic explanations that ensures both properties. Specifically, we introduce a novel class of sample-based abductive explanations, defined as those derivable through a generalization of the consensus operation in propositional logic. We prove that these consensus-based explanations are the only monotonic sample-based explanations, providing a unique theoretical characterization. Furthermore, they are the only guaranteed ground-truth explanations for any classification function consistent with the available sample. We demonstrate that both enumerating the complete set of explanations and finding the subset- or cardinality-minimal ones can be done in polynomial time. Moreover, we show that these explanations can be leveraged to identify necessary and relevant features, a fundamental problem in XAI.

The paper is structured as follows: Section 2 introduces the preliminaries and notation. Section 3 discusses related work. The novel consensus-based explanations are introduced and analyzed in Section 4, and Section 5 focuses on their computation. Section 6 describes how the new explanations can be used to identify necessary and relevant features. Section 7 reports the results of an experimental study. Finally, the paper concludes in Section 8.

## 2 Preliminaries

This section presents the necessary background and notation used throughout the paper.

**Basic definitions.** We consider a classifier $M = (\mathcal{F}, \mathbb{F}, \mathbb{K}, \kappa)$, where $\mathcal{F} = \{1, \ldots, m\}$ is a set of features. Each feature $i \in \mathcal{F}$ takes values from a domain $\mathbb{D}_i$. Domains are assumed to be finite and contain at least two values (i.e., $|\mathbb{D}_i| \geq 2$ for all $i \in \mathcal{F}$). Moreover, real-valued features are assumed to be discretized.[1] Feature space $\mathbb{F} = \mathbb{D}_1 \times \cdots \times \mathbb{D}_m$ is the Cartesian product of the domains. $\mathbb{K}$ is a set of classes. Finally, $\kappa : \mathbb{F} \to \mathbb{K}$ is a non-constant classification function mapping points in feature space to classes. An instance is a tuple $I = (\mathbf{v}, c)$, with $\mathbf{v} = (v_1, \ldots, v_m) \in \mathbb{F}$ and $c = \kappa(\mathbf{v})$.

A *sample* (or dataset) is a sequence of $n$ instances $\mathcal{S} = (I_1, \ldots, I_n)$, where each $I_k = (\mathbf{v}_k, p_k)$ consists of a point $\mathbf{v}_k \in \mathbb{F}$ and its prediction $p_k = \kappa(\mathbf{v}_k) \in \mathbb{K}$. From a computational perspective, $\mathcal{S}$ is represented by an $n \times m$ matrix $\mathbf{V}$ whose $k$-th row is $\mathbf{v}_k$, and an $n \times 1$ vector $\mathbf{p}$ where the $k$-th entry is $p_k$. We assume that every point in the sample is unique, i.e., there are no duplicate instances nor contradictory predictions for the same point. We denote by $\mathbb{S} = \{\mathbf{v}_k \mid k = 1, \ldots, n\}$ the finite set of points present in the sample; clearly, $\mathbb{S} \subseteq \mathbb{F}$ and $|\mathbb{S}| = |\mathcal{S}|$. For a given class $c \in \mathbb{K}$, we define $\mathbb{S}_c = \{\mathbf{v} \in \mathbb{S} \mid \kappa(\mathbf{v}) = c\}$ as the set of points in the sample with prediction $c$.

Given an ML model, a sample can be obtained by iteratively sampling the model, according to a given sampling distribution. In this sense, $\mathcal{S}$ represents a *sample of the classifier's behavior*. By considering the predictions as responses from the model, the sample is consistent by construction, avoiding issues such as contradictory labels or duplicate instances. Moreover, samples may exist even when the ML model is unknown. For example, this would be the case when the sample corresponds to the (training) data from which an ML model is induced, assuming the model is consistent with the sample. In such cases, we assume that any potential inconsistencies and duplicate instances have been resolved beforehand. Throughout, unless otherwise stated, we are not concerned with how the sample was generated; we treat $\mathcal{S}$ as an input.

**Logic-based explanations.** We follow standard definitions of logic-based XAI (Marques-Silva 2022; Darwiche 2023), and focus on *feature selection* explanations (Molnar 2020) within this framework.

A *weak abductive explanation* (WAXp) is a pair $(\mathcal{X}, I)$, where $\mathcal{X} \subseteq \mathcal{F}$ and $I = (\mathbf{v}, c)$, such that fixing the features in $\mathcal{X}$ to the values dictated by $I$ is sufficient for the prediction $c$, with $\mathbf{v} = (v_1, \ldots, v_m) \in \mathbb{F}$ and $c = \kappa(\mathbf{v})$. Formally, given $I$, $(\mathcal{X}, I)$ is a WAXp if the following predicate holds:

$$\mathsf{WAXp}(\mathcal{X}, I) \coloneqq \forall(\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i) \rightarrow (\kappa(\mathbf{x}) = c) \quad (1)$$

[1]The discretization of real-valued features is common practice in XAI (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2018).

Each element $i \in \mathcal{X}$ of a WAXp $(\mathcal{X}, I)$ encodes a literal $(x_i = v_i)$, with $v_i$ taken from the instance $I = (\mathbf{v}, c)$.

For clarity, we will slightly abuse notation. When clear from the context, a WAXp will be represented solely by $\mathcal{X}$, with $I$ left implicit. Furthermore, a WAXp may be also referred to as a consistent set (or conjunction) of literals $\mathcal{T} \in 2^{\mathcal{L}}$, where $\mathcal{L}$ is the set of all possible literals, given the features and their domains. In this case, $\mathcal{T}$ is a WAXp for a class $c \in \mathbb{K}$ if there exists an instance $I = (\mathbf{v}, c)$ such that $\mathcal{T}$ is consistent with $\mathbf{v}$ and $(\mathcal{X}, I)$ satisfies Eq. (1), where $\mathcal{X} \subseteq \mathcal{F}$ is the set of features involved in the literals of $\mathcal{T}$. Similarly, we say that $\mathcal{X}$ is a WAXp for a class $c \in \mathbb{K}$ if there exists an instance $I = (\mathbf{v}, c)$ such that $(\mathcal{X}, I)$ is a WAXp.

The domain of a WAXp $(\mathcal{X}, I)$ is given by all the points in feature space consistent with it:

$$\mathsf{dom}(\mathcal{X}, I) \coloneqq \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{X}} (x_i = v_i)\} \quad (2)$$

By extension, we denote by $\mathsf{dom}(\mathcal{T})$ (or $\mathsf{dom}(\mathcal{X})$ when the literals are clear from the context) the set of points in $\mathbb{F}$ consistent with the literals in the WAXp.

An *abductive explanation* (AXp) is a pair $(\mathcal{X}, I)$ such that $\mathcal{X}$ is subset-minimal with respect to Eq. (1).

Besides WAXps, we also define *weak contrastive explanations* (WCXps). A pair $(\mathcal{Y}, I)$, with $\mathcal{Y} \subseteq \mathcal{F}$ and instance $I = (\mathbf{v}, c)$, is a WCXp if:

$$\mathsf{WCXp}(\mathcal{Y}, I) \coloneqq \exists(\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{i \in \mathcal{F} \setminus \mathcal{Y}} (x_i = v_i) \land (\kappa(\mathbf{x}) \neq c) \quad (3)$$

Clearly, a pair $(\mathcal{Y}, I)$ is a WCXp if there is a way to modify the values of the features in $\mathcal{Y}$ such that the prediction is also modified. A contrastive explanation (CXp) is a pair $(\mathcal{Y}, I)$, such that $\mathcal{Y}$ is a minimal set with respect to Eq. (3).

It is well-established that for every AXp $(\mathcal{X}, I)$, $\mathcal{X}$ is a minimal hitting set of all the sets $\mathcal{Y}$ in all CXps $(\mathcal{Y}, I)$ and vice-versa (Ignatiev et al. 2020; Marques-Silva 2022). This *minimal-hitting set duality* relationship between AXps and CXps builds on the seminal work of R. Reiter in model-based diagnosis (Reiter 1987).

Throughout, we will refer to these explanations as *ground-truth* WAXps or WCXps, or simply as WAXps and WCXps, to distinguish them from the sample-based explanations defined below.

**WAXps and WCXps over samples.** Logic-based explanations have also been applied in model-agnostic settings where the model is characterized solely by a sample. Well-known examples include works focused on computing rule-based explanations (Ribeiro, Singh, and Guestrin 2018; Schleich et al. 2021; Geng, Schleich, and Suciu 2022; Rudin and Shaposhnik 2023; Cooper and Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024; Marques-Silva, Lefebre-Lobaina, and Martinez 2025).

Given a sample $\mathcal{S}$, defined over $\mathbb{S} \subseteq \mathbb{F}$, *sample-based* WAXps and WCXps are defined by replacing $\mathbb{F}$ by $\mathbb{S}$ in Eqs. (1) and (3) respectively, i.e., quantification ranges over the points included in the sample, and not over feature space. That is, given $\mathbb{S}$ and an instance $I = (\mathbf{v}, c) \in \mathcal{S}$, a pair

| Inst. | $x_1$ | $x_2$ | $x_3$ | $\kappa(\mathbf{x})$ |
|---|---|---|---|---|
| $I_1$ | 0 | 0 | 0 | 0 |
| $I_2$ | 0 | 0 | 1 | 0 |
| $I_3$ | 0 | 1 | 1 | 0 |
| $I_4$ | 1 | 0 | 0 | 0 |
| $I_5$ | 1 | 0 | 1 | 1 |
| $I_6$ | 1 | 1 | 0 | 1 |

Table 1: Sample $\mathcal{S}$ for Example 1.

$(\mathcal{X}, I)$, is a sample-based WAXp (sbWAXp), if the following predicate holds:

$$\mathsf{sbWAXp}(\mathcal{X}, I) \coloneqq \forall(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i) \rightarrow (\kappa(\mathbf{x}) = c) \quad (4)$$

Sample-based WCXps are defined by an analogous definition of a predicate $\mathsf{sbWCXp}(\mathcal{Y}, I)$. Recent work (Marques-Silva, Lefebre-Lobaina, and Martinez 2025) showed that the number of sbWCXps for a given instance $I$ is polynomial in the size of the sample. Also, notably, sbWCXps are ground-truth WCXps, i.e., they fulfill Eq. (3), whereas this does not hold in general for sbWAXps. Subset-minimal sbWAXps (and sbWCXps) are referred to as sbAXps (and sbCXps).

**Example 1.** *Consider features $\mathcal{F} = \{1, 2, 3\}$, with domains $\mathbb{D}_i = \{0, 1\}$ for all $i \in \mathcal{F}$, $\mathbb{K} = \{0, 1\}$ and the sample $\mathcal{S} = \{I_1, \ldots, I_6\}$ shown in Table 1. We address the task of explaining the instance $I_1 = ((0, 0, 0), 0)$. The differences between $I_1$ and the instances of class 1 ($I_5$ and $I_6$) yield the sbWCXps $\{1, 3\}$ and $\{1, 2\}$. Both are subset-minimal (i.e., sbCXps). The sbAXps for $I_1$ are the minimal hitting sets of the sbCXps: $\{1\}$ and $\{2, 3\}$. Any superset of an sbAXp (or sbCXp) consistent with the point $(0, 0, 0)$ is an sbWAXp (or sbWCXp) for $I_1$. We may represent the sbWAXp $\{2, 3\}$ using the notation $\{(x_2 = 0), (x_3 = 0)\}$ or as the conjunction of literals $(x_2 = 0) \land (x_3 = 0)$.*

**Explanation functions.** Only a fraction of all possible sbWAXps may be of interest depending on the properties they fulfill. In this regard, earlier work (Amgoud and Ben-Naim 2022; Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024), provided an axiomatic characterization of explanation functions (or *explainers*). Given a sample $\mathcal{S}$ and an instance $I = (\mathbf{v}, c)$, an explainer $\xi$ maps a pair $(\mathcal{S}, I)$ to a set of explanations $\xi(\mathcal{S}, I) \subseteq 2^{\mathcal{L}}$.

Based on the explanations $\xi$ may produce, several axioms for explainers have been established, among which we highlight the following: *Feasibility*, requiring that explanations are consistent with $\mathbf{v}$; *Validity*, requiring that the explanations are sbWAXps, i.e., they fulfill Eq. (4); *Success*, meaning that the explainer always returns at least one explanation; *Coherence*, ensuring that explanations for different classes are never mutually consistent; and *Monotonicity*, requiring that any explanation returned for a given sample remains an explanation for any extension of that sample. We focus on explainers producing (subsets of) sbWAXps. Thus, Feasibility, Validity and Success hold by definition. Coherence is formally defined as follows:

**Definition 1** (Coherence)**.** *An explainer $\xi$ is* coherent *if for any sample $\mathcal{S}$ and any two instances $I, J \in \mathcal{S}$, with $I = (\mathbf{v}_I, c)$, $J = (\mathbf{v}_J, c')$ and $c \neq c'$, it holds that any pair of sbWAXps $(\mathcal{X}_I, I)$ and $(\mathcal{X}_J, J)$ computed by $\xi$ satisfies $\mathsf{dom}(\mathcal{X}_I, I) \cap \mathsf{dom}(\mathcal{X}_J, J) = \emptyset$.*

Notably, (Amgoud, Cooper, and Debbaoui 2024) introduced three polynomial-time coherent explainers: (i) *trivial* explainers, returning only explanations involving all features; (ii) *irrefutable* explainers, satisfying a stronger requirement by returning explanations that cannot be contradicted by *any* valid sbWAXp within the sample for a different class (regardless of whether those sbWAXps are returned by $\xi$); and (iii) *surrogate-based explainers*, which rely on training a decision tree and extracting explanations from it.

We focus primarily on Monotonicity, which is formally defined as follows:

**Definition 2** (Monotonicity)**.** *An explainer $\xi$ is* monotonic *if for any sample $\mathcal{S}$, any superset $\mathcal{S}' \supseteq \mathcal{S}$, and any instance $I \in \mathcal{S}$, any sbWAXp $(\mathcal{X}, I)$ w.r.t. $\mathcal{S}$ computed by $\xi$ is also an sbWAXp w.r.t. $\mathcal{S}'$; that is, $\xi(\mathcal{S}, I) \subseteq \xi(\mathcal{S}', I)$.*

Trivial explainers are the only sample-based monotonic explainers among those proposed in earlier work.

For ease of presentation, we attribute the nature and properties of an explainer directly to the explanations it produces. Accordingly, we say that an sbWAXp is *trivial*, *irrefutable*, *coherent* or *monotonic* if it is produced by an explainer satisfying the corresponding property.

## 3 Related Work

Logic-based explanations in ML have been studied since 2018 (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019; Ignatiev 2020; Audemard et al. 2022; Amgoud and Ben-Naim 2022; Marques-Silva 2022; Darwiche and Hirth 2023; Bassan and Katz 2023), and they address the lack of rigor of non-symbolic methods of explainability (Bach et al. 2015; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2018; Guidotti et al. 2019).

Rigorous sample-based model-agnostic explanations have been extensively studied in recent years (Rudin and Shaposhnik 2019; Geng, Schleich, and Suciu 2022; Rudin and Shaposhnik 2023; Amgoud 2023; Cooper and Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024; Koriche et al. 2024; Marques-Silva, Lefebre-Lobaina, and Martinez 2025). Nevertheless, finding explanations from samples (or operations that are equivalent) can be traced to other domains of AI&ML (Chikalov et al. 2013), including patterns in logical analysis of data (Crama, Hammer, and Ibaraki 1988; Boros et al. 2000), decision rules in rough sets (Pawlak 1982; Pawlak 2002), and also testor theory (Lazo-Cortés et al. 2015). Past work on rule-based explanations addressed cardinality-minimal explanations, but also heuristic methods (Ribeiro, Singh, and Guestrin 2018; Geng, Schleich, and Suciu 2022). Among these,

some (Geng, Schleich, and Suciu 2022) employ a heuristically greedy approach, and so they do not guarantee subset or cardinality minimality, and some (Ribeiro, Singh, and Guestrin 2018) may yield logically incorrect explanations (Geng, Schleich, and Suciu 2022). More recent works (Cooper and Amgoud 2023; Rudin and Shaposhnik 2023; Marques-Silva, Lefebre-Lobaina, and Martinez 2025) proposed algorithms for computing subset-minimal as well as cardinality-minimal explanations.

Following the axiomatic characterization of sample-based explainers proposed in (Amgoud and Ben-Naim 2022; Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024), we focus on essential properties that explanations often fail to satisfy, including the requirement of coherence. Our work is closely related to this line of research, and extends it by focusing on monotonicity.

# 4 Explanations by Consensus

We present a novel class of sample-based explanations and analyze their properties.

## 4.1 The Consensus Operation

In propositional logic, the consensus operation derives new implicants—terms (conjunctions of literals) that satisfy a formula $\phi$—from existing implicants (Blake 1937; Quine 1952; Brown 1990). Given $t_1 = x \wedge \Gamma_1$ and $t_2 = \neg x \wedge \Gamma_2$ the consensus operation yields $t = \Gamma_1 \wedge \Gamma_2$. Since $t \models t_1 \vee t_2$, it follows that $t$ is also an implicant of $\phi$.

The consensus operation can be extended to classifiers, where implicants correspond to ground-truth WAXps. Given feature $i \in \mathcal{F}$ with domain $\mathbb{D}_i = \{v_{i1}, v_{i2}, \ldots, v_{ik}\}$, and $k$ WAXps for a class $c$ acting as premises, $\mathcal{T}_j = (x_i = v_{ij}) \wedge \Gamma_j$ for $j = 1, \ldots, k$, the consensus operation w.r.t. feature $i$ yields $\mathcal{R} = \bigwedge_{j=1}^{k} \Gamma_j$.

**Example 2.** *Let $i \in \mathcal{F}$ with $\mathbb{D}_i = \{1, 2, 3\}$ and three WAXps for a class $c$: $\mathcal{T}_1 = (x_1 = 1) \wedge (x_i = 1)$, $\mathcal{T}_2 = (x_2 = 1) \wedge (x_i = 2)$, and $\mathcal{T}_3 = (x_3 = 2) \wedge (x_i = 3)$. The consensus w.r.t. feature $i$ yields $\mathcal{R} = (x_1 = 1) \wedge (x_2 = 1) \wedge (x_3 = 2)$.*

The result $\mathcal{R}$ is also a WAXp, provided it is consistent:

**Proposition 1.** *Let $\mathcal{T}_1, \ldots, \mathcal{T}_k$ be WAXps for a class $c$. If their consensus $\mathcal{R} = \Gamma_1 \wedge \cdots \wedge \Gamma_k$ over feature $i$ is consistent, then $\mathcal{R}$ is a WAXp for $c$.*

*Proof.* Suppose $\mathcal{R}$ is not a WAXp for $c$. Then $\exists(\mathbf{x} \in \mathbb{F}).\Gamma_1 \wedge \cdots \wedge \Gamma_k \wedge (\kappa(\mathbf{x}) \neq c)$. That is, $\mathbf{x}$ is a point that satisfies $R$, but for which $\kappa(\mathbf{x}) \neq c$. Since $\mathcal{T}_j = (x_i = v_{ij}) \wedge \Gamma_j$ is a WAXp for $c$ for $j \in \{1, \ldots, k\}$, it follows that $(x_i = v_{ij}) \wedge \Gamma_j \rightarrow (\kappa(\mathbf{x}) = c)$. Given that $\mathbf{x}$ satisfies $\mathcal{R}$ (and so every $\Gamma_j$) but $\kappa(\mathbf{x}) \neq c$, it must hold that $\mathbf{x}$ does not satisfy $(x_i = v_{ij})$ for any $j \in \{1, \ldots, k\}$. As a consequence, $x_i \neq v_{ij}$ for all $v_{ij} \in \mathbb{D}_i$, which is a contradiction, since $x_i$ must take some value from its domain. □

The proposition above proves that consensus is a sound inference rule for deriving WAXps. Note that it requires $\mathcal{R}$ to be consistent; otherwise, $\mathcal{R}$ would not cover any point in $\mathbb{F}$ and thus would not satisfy the definition of a WAXp.

Given a set of WAXps $\mathbb{W}$ for a class $c$, its *consensus closure* $C(\mathbb{W})$ is the set containing $\mathbb{W}$ and all WAXps derivable through the application of the consensus operation.

## 4.2 Consensus-Based WAXps

Consider the sample-based setting, where only a sample $\mathcal{S}$, defined over $\mathbb{S} \subseteq \mathbb{F}$ is available. Recall that $\mathbb{S}_c = \{\mathbf{v} \in \mathbb{S} \mid \kappa(\mathbf{v}) = c\}$. Each point $\mathbf{v} \in \mathbb{S}_c$ can be associated with a trivial WAXp $\mathcal{T}_{\mathbf{v}} = \bigwedge_{i \in \mathcal{F}} (x_i = v_i)$. Let $\mathbb{W}_c = \{\mathcal{T}_{\mathbf{v}} \mid \mathbf{v} \in \mathbb{S}_c\}$, denoting the set of all trivial WAXps corresponding to the points in $\mathbb{S}_c$.

We define *consensus-based* WAXps as those derivable through the consensus operation:

**Definition 3** (cbWAXp)**.** *$\mathcal{T}$ is a consensus-based WAXp (or cbWAXp) for class $c$ if and only if $\mathcal{T} \in C(\mathbb{W}_c)$. $\mathcal{T}$ is a cbWAXp for an instance $I = (\mathbf{v}, c)$ if and only if, in addition, $\mathcal{T}$ is consistent with $\mathbf{v}$.*

Note that any cbWAXp for a class $c$ is a cbWAXp for at least one instance $I = (\mathbf{v}, c) \in \mathcal{S}$.

**Example 3.** *Consider again the sample $\mathcal{S}$ from Table 1. For readability, we represent cbWAXps as sets of literals. Initially, $\mathbb{W}_0 = \{\mathcal{T}_{I_1}, \ldots, \mathcal{T}_{I_4}\}$, where: $\mathcal{T}_{I_1} = \{(x_1 = 0), (x_2 = 0), (x_3 = 0)\}$, $\mathcal{T}_{I_2} = \{(x_1 = 0), (x_2 = 0), (x_3 = 1)\}$, $\mathcal{T}_{I_3} = \{(x_1 = 0), (x_2 = 1), (x_3 = 1)\}$ and $\mathcal{T}_{I_4} = \{(x_1 = 1), (x_2 = 0), (x_3 = 0)\}$. The consensus of $\mathcal{T}_{I_1}, \mathcal{T}_{I_2}$ w.r.t. feature 3 yields the cbWAXp $\mathcal{R}_1 = \{(x_1 = 0), (x_2 = 0)\}$. Similarly, $\mathcal{T}_{I_1}, \mathcal{T}_{I_4}$ w.r.t. feature 1 yield $\mathcal{R}_2 = \{(x_2 = 0), (x_3 = 0)\}$, and $\mathcal{T}_{I_2}, \mathcal{T}_{I_3}$ w.r.t. feature 2 yield $\mathcal{R}_3 = \{(x_1 = 0), (x_3 = 1)\}$. Notice that the consensus of $\mathcal{R}_2, \mathcal{R}_3$ w.r.t. feature 3 produces the already known cbWAXp $\mathcal{R}_1$. The cbWAXps for class 0 are $C(\mathbb{W}_0) = \mathbb{W}_0 \cup \{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3\}$. The cbWAXps for $I_1$ are those consistent with its coordinates, i.e., $\{\mathcal{T}_{I_1}, \mathcal{R}_1, \mathcal{R}_2\}$.*

Consensus-based explanations are ground-truth WAXps, i.e., they satisfy Eq. (1) over the entire feature space $\mathbb{F}$.

**Proposition 2.** *Every cbWAXp is a ground-truth WAXp.*

*Proof.* By definition, a cbWAXp for a class $c$ is either a trivial WAXp $\mathcal{T}_{\mathbf{v}}$ corresponding to a point $\mathbf{v} \in \mathbb{S}_c$ or it is derived through the consensus operation starting from the set of trivial WAXps $\mathbb{W}_c$. In the latter case, by Proposition 1, the resulting cbWAXp is a ground-truth WAXp. □

It follows that every cbWAXp $(\mathcal{X}, I)$ is also a valid sbWAXp. Since it is a ground-truth WAXp, it satisfies Eq. (1) quantifying universally over $\mathbb{F}$. Consequently, Eq. (4), which characterizes sbWAXps, trivially holds for any $\mathbb{S} \subseteq \mathbb{F}$. Now, we state two key properties:

**Proposition 3.** *cbWAXps are coherent.*

*Proof.* Let $I, J \in \mathcal{S}$ be two instances with $I = (\mathbf{v}_I, c_I)$, $J = (\mathbf{v}_J, c_J)$ and $c_I \neq c_J$. Let $\mathcal{T}_I$ and $\mathcal{T}_J$ be two cbWAXps for $I$ and $J$ respectively. By Proposition 2, they are ground-truth WAXps. Suppose $\mathsf{dom}(\mathcal{T}_I) \cap \mathsf{dom}(\mathcal{T}_J) \neq \emptyset$. Thus, there exists a point $\mathbf{u} \in \mathbb{F}$ such that $\mathbf{u} \in \mathsf{dom}(\mathcal{T}_I) \cap \mathsf{dom}(\mathcal{T}_J)$. Since $\mathcal{T}_I$ is a WAXp for $I$ and $\mathbf{u} \in \mathsf{dom}(\mathcal{T}_I)$, it holds that $\kappa(\mathbf{u}) = c_I$. Analogously, as $\mathcal{T}_J$ is a WAXp for $J$ and $\mathbf{u} \in \mathsf{dom}(\mathcal{T}_J)$, it holds that $\kappa(\mathbf{u}) = c_J$. So, $c_I = c_J$, which is a contradiction. □

**Proposition 4.** *cbWAXps are monotonic.*

*Proof.* Let $\mathcal{T}_I$ be a cbWAXp for an instance $I = (\mathbf{v}, c) \in \mathcal{S}$. By definition, $\mathcal{T}_I \in C(\mathbb{W}_c)$. Let $\mathcal{S}' \supseteq \mathcal{S}$, defined over $\mathbb{S}' \supseteq \mathbb{S}$, and define $\mathbb{W}'_c = \{\mathcal{T}_\mathbf{v} \mid \mathbf{v} \in \mathbb{S}'_c\}$. Clearly, $\mathbb{W}_c \subseteq \mathbb{W}'_c$, which implies $C(\mathbb{W}_c) \subseteq C(\mathbb{W}'_c)$. Consequently, $\mathcal{T}_I \in C(\mathbb{W}'_c)$ and $\mathcal{T}_I$ is a cbWAXp for $I$ w.r.t. sample $\mathcal{S}'$. □

These properties make cbWAXps inherently robust explanations. Now we demonstrate that cbWAXps are, in fact, the *only* sample-based WAXps that satisfy monotonicity. For ease of presentation, we define an *extension* of a sample $\mathcal{S}$ as any function $f_\mathcal{S} : \mathbb{F} \to \mathbb{K}$ consistent with $\mathcal{S}$; that is, $f_\mathcal{S}(\mathbf{x}) = \kappa(\mathbf{x})$ for all points $\mathbf{x} \in \mathbb{S}$. Note that $\kappa$ is itself an extension of $\mathcal{S}$.

**Lemma 1.** *Let $\mathcal{T} \in 2^\mathcal{L}$ be an sbWAXp for class $c$. $\mathcal{T}$ is a cbWAXp if and only if $\mathsf{dom}(\mathcal{T}) \subseteq \mathbb{S}_c$.*

*Proof.* (*Only If*) $\mathcal{T}$ is a cbWAXp for class $c$. By definition, $\mathcal{T}$ is derived from the set of trivial cbWAXps $\mathbb{W}_c$ through the consensus operation. We show $\mathsf{dom}(\mathcal{T}) \subseteq \mathbb{S}_c$ by induction on the derivation length $l$ (i.e., the number of applications of the consensus operation):

- Base case ($l = 0$). $\mathcal{T}$ is a trivial cbWAXp $\mathcal{T}_\mathbf{v}$ corresponding to a point $\mathbf{v} \in \mathbb{S}_c$, so $\mathsf{dom}(\mathcal{T}_\mathbf{v}) = \{\mathbf{v}\}$. Since $\mathbf{v} \in \mathbb{S}_c$, it holds that $\mathsf{dom}(\mathcal{T}) \subseteq \mathbb{S}_c$.
- Inductive case ($l > 0$). $\mathcal{T}$ is the result of the consensus operation on the cbWAXps $\mathcal{T}_1, \ldots, \mathcal{T}_k$ derived previously. As the inductive hypothesis, we assume that $\mathsf{dom}(\mathcal{T}_j) \subseteq \mathbb{S}_c$, for $j \in \{1, \ldots, k\}$. Since $\mathsf{dom}(\mathcal{T}) = \mathsf{dom}(\mathcal{T}_1) \cup \cdots \cup \mathsf{dom}(\mathcal{T}_k)$, it follows that $\mathsf{dom}(\mathcal{T}) \subseteq \mathbb{S}_c$.

(*If*) We prove the contrapositive: if $\mathcal{T}$ is an sbWAXp but not a cbWAXp for class $c$, then $\mathsf{dom}(\mathcal{T}) \not\subseteq \mathbb{S}_c$. The proof proceeds by backward induction on the cardinality of $\mathcal{T}$.

- Base case ($|\mathcal{T}| = m$). Since $\mathcal{T}$ is an sbWAXp for $c$ and all features appear in $\mathcal{T}$, it follows that $\mathsf{dom}(\mathcal{T}) = \{\mathbf{v}\}$ (i.e., a single point) and that $\mathbf{v} \in \mathbb{S}_c$ (otherwise $\mathcal{T}$ could not possibly be consistent with $\mathbf{v}$). Any such $\mathcal{T}$ is a trivial WAXp and, by definition, a cbWAXp. In this case, the antecedent is false, and the implication holds.
- Inductive case ($|\mathcal{T}| < m$). We assume as the inductive hypothesis that for any sbWAXp $\mathcal{T}'$ of size $i + 1$ that is not a cbWAXp, there exists a point $\mathbf{u} \in \mathsf{dom}(\mathcal{T}')$ such that $\mathbf{u} \notin \mathbb{S}_c$. Now, let $\mathcal{T}$ be an sbWAXp for class $c$ of size $i$ that is not a cbWAXp, and let $\mathcal{X}_\mathcal{T} \subseteq \mathcal{F}$ denote the features present in $\mathcal{T}$. Since $\mathcal{T}$ is not a cbWAXp, there must exist at least one feature $d \in (\mathcal{F} \setminus \mathcal{X}_\mathcal{T})$ and a value $v_{dl} \in \mathbb{D}_d$ such that $(x_d = v_{dl}) \wedge \mathcal{T}$ is not a cbWAXp for class $c$; otherwise we could apply the consensus operation over the values of $d$ to derive $\mathcal{T}$ as a cbWAXp. Furthermore, since $\mathcal{T}$ is an sbWAXp for class $c$, $(x_d = v_{dl}) \wedge \mathcal{T}$ is also an sbWAXp for $c$ (recall $d \notin \mathcal{X}_\mathcal{T}$). So, $(x_d = v_{dl}) \wedge \mathcal{T}$ is an sbWAXp for class $c$ of size $i+1$ that is not a cbWAXp. By the inductive hypothesis, there exists $\mathbf{u} \in \mathsf{dom}((x_d = v_{dl}) \wedge \mathcal{T})$ such that $\mathbf{u} \notin \mathbb{S}_c$. Since $\mathsf{dom}((x_d = v_{dl}) \wedge \mathcal{T}) \subseteq \mathsf{dom}(\mathcal{T})$, it follows that $\mathbf{u} \in \mathsf{dom}(\mathcal{T})$.

□

We can now establish that being consensus-based is a necessary condition for monotonicity.

**Proposition 5.** *Let $\mathcal{T} \in 2^\mathcal{L}$ be an sbWAXp for an instance $I = (\mathbf{v}, c)$. $\mathcal{T}$ is monotonic only if $\mathcal{T}$ is a cbWAXp.*

*Proof.* Suppose $\mathcal{T}$ is not a cbWAXp for $I$. By Lemma 1, there exists a point $\mathbf{u} \in \mathsf{dom}(\mathcal{T})$ such that $\mathbf{u} \notin \mathbb{S}_c$. Consider the sample $\mathcal{S}' = \mathcal{S} \cup \{(\mathbf{u}, c')\}$, with $c' \neq c$. Clearly, $\mathcal{S} \subsetneq \mathcal{S}'$. However, $\mathcal{T}$ is not an sbWAXp for $I$ w.r.t. $\mathcal{S}'$. Hence, $\mathcal{T}$ is not monotonic. □

From Propositions 4 and 5 we obtain:

**Theorem 1.** *Let $\mathcal{T} \in 2^\mathcal{L}$ be an sbWAXp for an instance $I = (\mathbf{v}, c)$. $\mathcal{T}$ is monotonic if and only if $\mathcal{T}$ is a cbWAXp.*

Notably, cbWAXps are not only ground-truth WAXps (as stated in Proposition 2), but they are also the *only* sbWAXps guaranteed to be ground-truth WAXps for *every* classification function consistent with the sample.

**Proposition 6.** *Let $\mathcal{T} \in 2^\mathcal{L}$ be an sbWAXp for the instance $I = (\mathbf{v}, c)$. $\mathcal{T}$ is guaranteed to be a ground-truth WAXp only if $\mathcal{T}$ is a cbWAXp.*

*Proof.* We prove the contrapositive. Suppose $\mathcal{T}$ is not a cbWAXp. By Lemma 1, there exists a point $\mathbf{u} \in \mathsf{dom}(\mathcal{T})$ such that $\mathbf{u} \notin \mathbb{S}_c$. Let $f_\mathcal{S}$ be an extension of the sample such that $f_\mathcal{S}(\mathbf{u}) = c'$, with $c' \neq c$. Clearly, $\mathcal{T}$ is not a ground-truth WAXp for $f_\mathcal{S}$. Therefore, $\mathcal{T}$ cannot be guaranteed to be a ground-truth WAXp. □

Together, these results imply:

**Theorem 2.** *Let $\mathcal{T} \in 2^\mathcal{L}$ be an sbWAXp for an instance $I = (\mathbf{v}, c)$. $\mathcal{T}$ is guaranteed to be a ground-truth WAXp if and only if $\mathcal{T}$ is a cbWAXp.*

The previous results precisely characterize cbWAXps as the monotonic sbWAXps. By definition, these include the trivial explanations. We now prove that cbWAXps are a subset of the irrefutable explanations.

**Proposition 7.** *Let $\mathcal{T}$ be an sbWAXp for an instance $I = (\mathbf{v}, c)$. If $\mathcal{T}$ is a cbWAXp, then $\mathcal{T}$ is irrefutable.*

*Proof.* $\mathcal{T}$ is a cbWAXp. Then, by Lemma 1, $\mathsf{dom}(\mathcal{T}) \subseteq \mathbb{S}_c$. Suppose $\mathcal{T}$ is not irrefutable. Then, there exists an sbWAXp $\mathcal{T}'$ for an instance $I' = (\mathbf{v}', c')$, with $\mathbf{v}' \neq \mathbf{v}$ and $c' \neq c$, such that $\mathsf{dom}(\mathcal{T}) \cap \mathsf{dom}(\mathcal{T}') \neq \emptyset$. So, there exists a point $\mathbf{u} \in \mathsf{dom}(\mathcal{T}')$ such that $\mathbf{u} \in \mathbb{S}_c$. Since $c \neq c'$, the point $\mathbf{u}$ acts as a counterexample for $\mathcal{T}'$ in $\mathbb{S}$. As a consequence, $\mathcal{T}'$ is not an sbWAXp for $I'$. A contradiction. □

Note that while Coherence (see Proposition 3) ensures that cbWAXps do not contradict each other, the fact that they are irrefutable explanations establishes a stronger property: they cannot be contradicted by any sbWAXp.

# 5 Computing Consensus-Based Explanations

This section focuses on the computation of consensus-based WAXps, given a sample $\mathcal{S}$ and an instance $I = (\mathbf{v}, c) \in \mathcal{S}$. Specifically, we will study the problems of: (i) deciding whether a given set of features corresponds to a cbWAXp; (ii) computing a subset-minimal cbWAXp; (iii) enumerating all cbWAXps; and (iv) computing the smallest possible of such explanations. We will show that all these problems are solvable in polynomial time in the size of the input.[2]

For ease of presentation, throughout we will use the predicate $\mathsf{cbWAXp}(\mathcal{X}, I)$, where $\mathcal{X} \subseteq \mathcal{F}$. This predicate holds if and only if the conjunction of literals corresponding to $\mathcal{X}$ for the instance $I$ is a cbWAXp. We will also refer to the set of features $\mathcal{X}$ itself as the cbWAXp it represents.

## 5.1 Deciding the cbWAXp Predicate

First, we address the problem of deciding whether a given set $\mathcal{X} \subseteq \mathcal{F}$ is a cbWAXp. Noticeably, this problem is solvable in polynomial time. We simply need to count the instances in the sample that match the values of the instance $I$ for all features in $\mathcal{X}$ and have prediction $c$.

**Proposition 8.** *Given instance $I = (\mathbf{v}, c)$ and $\mathcal{X} \subseteq \mathcal{F}$, $\mathsf{cbWAXp}(\mathcal{X}, I)$ holds if and only if $|\{(\mathbf{x}, c) \in \mathcal{S} : \bigwedge_{i\in\mathcal{X}}(x_i = v_i)\}| = \prod_{u\in\mathcal{F}\setminus\mathcal{X}} |\mathbb{D}_u|$.*

*Proof.* By Lemma 1, $\mathsf{cbWAXp}(\mathcal{X}, I)$ holds if and only if for all points $\mathbf{x} \in \mathsf{dom}(\mathcal{X})$, $(\mathbf{x}, c) \in \mathcal{S}$. On the other hand, $\mathsf{dom}(\mathcal{X}) = \{\mathbf{x} \in \mathbb{F} \mid \bigwedge_{i\in\mathcal{X}}(x_i = v_i)\}$, and so $|\{(\mathbf{x}, c) \in \mathcal{S} : \bigwedge_{i\in\mathcal{X}}(x_i = v_i)\}| = |\mathsf{dom}(\mathcal{X})| = \prod_{u\in\mathcal{F}\setminus\mathcal{X}} |\mathbb{D}_u|$. □

Consequently, the problem is reduced to a straightforward counting task over the sample. Performing such a count can be done in $O(nm)$ time, with $n = |\mathbb{S}|$ and $m = |\mathcal{F}|$. Thus:

**Proposition 9.** *The* cbWAXp *predicate can be decided in $O(nm)$ time.*

## 5.2 Subset-Minimal cbWAXps

In practice, one often seeks explanations that do not contain redundant information. We address the problem of computing a *subset-minimal* cbWAXp, i.e., a cbWAXP $\mathcal{M} \subseteq \mathcal{F}$ such that no proper subset $\mathcal{M}' \subsetneq \mathcal{M}$ is a cbWAXp. These explanations are referred to as cbAXps.

Clearly, cbAXps are minimal sets that satisfy the predicate cbWAXp. To devise an efficient algorithm, we first prove that it is monotonically increasing w.r.t. set inclusion:

**Proposition 10.** *If $\mathsf{cbWAXp}(\mathcal{X}, I)$ holds, then $\mathsf{cbWAXp}(\mathcal{X}', I)$ holds for any $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{F}$.*

*Proof.* Suppose $\mathsf{cbWAXp}(\mathcal{X}, I)$ holds. By Lemma 1, $\mathsf{dom}(\mathcal{X}) \subseteq \mathbb{S}_c$. Given $\mathcal{X}'$ such that $\mathcal{X} \subseteq \mathcal{X}'$, it follows that $\mathsf{dom}(\mathcal{X}') \subseteq \mathsf{dom}(\mathcal{X})$. Hence, $\mathsf{dom}(\mathcal{X}') \subseteq \mathbb{S}_c$, and by Lemma 1, $\mathsf{cbWAXp}(\mathcal{X}', I)$ holds. □

This allows us to compute a cbAXp with a linear number of predicate tests by using a *deletion* approach:[3] Initially, $\mathcal{M} = \mathcal{F}$. For each feature $i$, if $\mathsf{cbWAXp}(\mathcal{M} \setminus \{i\}, I)$ holds, remove $i$ from $\mathcal{M}$; otherwise, keep $i$ in $\mathcal{M}$. Upon termination, $\mathcal{M}$ is guaranteed to be a cbAXp. This algorithm performs $m$ predicate tests, each taking $O(nm)$ time, resulting in a total complexity of $O(nm^2)$. Thus:

**Proposition 11.** *Computing a single cbAXp for an instance $I$ can be done in $O(nm^2)$ time.*

## 5.3 Enumeration and Smallest cbWAXps

In some scenarios, a single explanation may not suffice, and an enumeration of several (or even all) may be required. The sheer number of such explanations often poses a significant challenge. For instance, the number of sbAXps can be exponential w.r.t. the size of the sample. In contrast, we show that the number of cbWAXps is bounded by the number of instances in the sample.

To prove this, we introduce the concept of *maximal witness* for a cbWAXp. Given a cbWAXp $(\mathcal{X}, I)$, with $I = (\mathbf{v}, c)$, a *maximal witness* is an instance $J = (\mathbf{u}, c) \in \mathcal{S}$, such that $\mathbf{u}$ agrees with $\mathbf{v}$ exactly on the features in $\mathcal{X}$. That is, $u_i = v_i$ if and only if $i \in \mathcal{X}$. By definition, each instance in $\mathcal{S}$ can serve as a maximal witness for at most one cbWAXp, for a given instance $I$.

**Example 4.** *Consider the cbWAXps $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{T}_{I_1}$ for $I_1 = ((0, 0, 0), 0)$ from Example 3. For $\mathcal{R}_1 = \{(x_1 = 0), (x_2 = 0)\}$, its maximal witness is $I_2 = ((0, 0, 1), 0)$, since $I_2$ agrees with $I_1$ exactly on features $\{1, 2\}$. Similarly, for $\mathcal{R}_2 = \{(x_2 = 0), (x_3 = 0)\}$, the maximal witness is $I_4 = ((1, 0, 0), 0)$, as they agree exactly on $\{2, 3\}$. $I_1$ itself is the maximal witness for the trivial cbWAXp $\mathcal{T}_{I_1} = \{(x_1 = 0), (x_2 = 0), (x_3 = 0)\}$. However, $I_3 = ((0, 1, 1), 0)$ is not a maximal witness for any cbWAXp of $I_1$, as their common coordinates $\{(x_1 = 0)\}$ do not form a cbWAXp.*

Such witnesses are guaranteed to exist:

**Lemma 2.** *Every cbWAXp $\mathcal{X}$ for an instance $I = (\mathbf{v}, c)$ has at least one maximal witness in the sample $\mathcal{S}$.*

*Proof.* Suppose $\mathcal{X}$ does not have a maximal witness in $\mathcal{S}$. Then, for every point $\mathbf{u} \in \mathbb{F}$ that agrees with $\mathbf{v}$ exactly on the features in $\mathcal{X}$ (i.e., $u_i = v_i$ if and only if $i \in \mathcal{X}$), it holds that $(\mathbf{u}, c) \notin \mathcal{S}$. If $\mathcal{X} = \mathcal{F}$, the only such point is $\mathbf{v}$ itself. Since $(\mathbf{v}, c) \in \mathcal{S}$, as it is the instance to be explained, it follows that it is a maximal witness for $\mathcal{F}$, contradicting the initial assumption. Otherwise, if $\mathcal{X} \subsetneq \mathcal{F}$, at least one such point $\mathbf{u}$ is guaranteed to exist in $\mathbb{F}$ (since $|\mathbb{D}_i| \geq 2$ for all $i \in \mathcal{F}$, we can always pick $u_j \neq v_j$ for $j \in \mathcal{F} \setminus \mathcal{X}$). By construction, any such $\mathbf{u}$ satisfies that $\mathbf{u} \in \mathsf{dom}(\mathcal{X})$ but $\mathbf{u} \notin \mathbb{S}_c$ (by our initial assumption). By Lemma 1, $\mathcal{X}$ cannot be a cbWAXp for class $c$ and, consequently, not for $I$ either. A contradiction. □

[2] The input is a sample $\mathcal{S}$, consisting of $n$ instances with $m$ features and their predictions, and an instance to be explained, so the complexity is defined in terms of this size. This is common in the literature on sample-based explanations, e.g., (Cooper and Amgoud 2023; Amgoud, Cooper, and Debbaoui 2024; Marques-Silva, Lefebre-Lobaina, and Martinez 2025). Notice that $n$ may be exponentially larger than $m$.

[3] Other general-purpose algorithms for computing a minimal set over a monotone predicate (Marques-Silva, Janota, and Mencía 2017), such as QuickXplain or Progression, could be used as well.

Now, we are ready to establish the following bound:

**Theorem 3.** *The number of distinct cbWAXps for $I$ is at most $n = |\mathcal{S}|$.*

*Proof.* By Lemma 2, every cbWAXp has at least one maximal witness in $\mathcal{S}$. By definition, each instance $J \in \mathcal{S}$ can serve as the maximal witness for at most one cbWAXp for $I$. Therefore, the number of distinct cbWAXps cannot exceed the number of instances in the sample. □

The result above provides the basis for a polynomial-time algorithm to enumerate all the cbWAXps for a given instance $I = (\mathbf{v}, c)$. Specifically, each instance $J = (\mathbf{u}, c) \in \mathcal{S}$ yields a candidate set $\mathcal{X}_u = \{i \in \mathcal{F} \mid v_i = u_i\}$, representing the features where $\mathbf{v}$ and $\mathbf{u}$ agree. If $\mathsf{cbWAXp}(\mathcal{X}_u, I)$ holds, then $J$ is a maximal witness for $\mathcal{X}_u$. Since every cbWAXp must have a maximal witness in the sample, it suffices to traverse the instances in $\mathcal{S}$ with prediction $c$ and test whether each resulting candidate set is indeed a cbWAXp for $I$. As this test takes $O(nm)$ and there are at most $n$ candidates, the entire process takes $O(n^2m)$ time.

If some feature domains are not binary, two different instances might yield the same cbWAXp. Such duplicates can be filtered out in $O(n^2m)$ time, maintaining the same overall complexity. Furthermore, finding all subset-minimal cbWAXps (i.e., the cbAXps) can also be achieved in $O(n^2m)$ by a pairwise comparison of the cbWAXps. Notably, finding the cardinality-minimal cbWAXps is also in $O(n^2m)$, as it only requires selecting the smallest ones in the list. This is in stark contrast to finding the smallest sbAXps, which is known to be NP-hard (Rudin and Shaposhnik 2023; Marques-Silva, Lefebre-Lobaina, and Martinez 2025).

In summary, we state the following result:

**Theorem 4.** *All distinct, subset-minimal, and cardinality-minimal cbWAXps for an instance $I$ can be enumerated in $O(n^2m)$ time.*

## 6 Deciding Feature Necessity and Relevance

Our framework can be used to identify *necessary* and *relevant* features, providing insights into their importance for a prediction. A feature is necessary if it belongs to *all* (subset-minimal) AXps, and it is relevant if it belongs to *at least one* of them. Necessary features are thus also relevant.

We consider these properties in terms of *ground-truth* AXps, and propose efficient sufficient conditions to identify them from a sample. Specifically, we first consider the case of a fixed sample, and then address the problem of deciding feature necessity by querying an ML model as an oracle.

### 6.1 Determination Given a Fixed Sample

By leveraging cbWAXps we can identify, in certain cases, whether a feature is necessary or relevant. The first result relies only on the existence of a cbWAXp:

**Proposition 12.** *Let $\mathcal{X} \subseteq \mathcal{F}$ be a cbWAXp for an instance $I$. Then, for all $i \in \mathcal{F} \setminus \mathcal{X}$, $i$ is not necessary for $I$.*

*Proof.* Since $\mathcal{X}$ is a cbWAXp, by Proposition 2 it is a ground-truth WAXp for $I$. Let $i \in \mathcal{F} \setminus \mathcal{X}$. Since $i \notin \mathcal{X}$, it follows that $i$ cannot belong to any AXp $\mathcal{X}' \subseteq \mathcal{X}$. Notice there exists at least one such $\mathcal{X}'$. Hence, $i$ does not belong to all AXps and, by definition, $i$ is not necessary for $I$. □

Thus, for any feature $i \in \mathcal{F}$, we can test $\mathsf{cbWAXp}(\mathcal{X} \setminus \{i\}, I)$ in polynomial time. If it holds, $i$ is safely declared as not necessary, due to the monotonicity of the predicate. Testing all features takes $O(nm^2)$.

So far, we have only considered the instances in the sample that have the same prediction $c$ as the instance $I$ to be explained. However, the instances with prediction $c' \neq c$ may also provide useful information.

Each such instance yields an sbWCXp $\mathcal{Y} \subseteq \mathcal{F}$, consisting of the features that disagree with the point $\mathbf{v}$ in $I$. By definition, any sbWCXp is also a ground-truth WCXp (Marques-Silva, Lefebre-Lobaina, and Martinez 2025). We can leverage this property to detect necessary features when the sample is sufficiently informative:

**Proposition 13.** *Let $\mathcal{Y} = \{i\}$ be an sbWCXp for an instance $I$. Then, the feature $i$ is necessary for $I$.*

*Proof.* By the minimal hitting set duality relationship, any AXp $\mathcal{X} \subseteq \mathcal{F}$ hits all CXps. Since $\mathcal{Y} = \{i\}$ is a ground-truth WCXp, it must contain at least one minimal WCXp (i.e., a CXp). Given that $|\mathcal{Y}| = 1$, the only possible CXp is $\mathcal{Y}$ itself (note that the empty set cannot be a CXp, since keeping all original values from $I$ cannot change the prediction). So, every AXp must contain $i$; otherwise it would have an empty intersection with $\mathcal{Y}$. Thus, the feature $i$ is necessary. □

This result provides a sufficient condition for efficiently identifying necessary features from the sample. It suffices to traverse the instances with a prediction other than $c$ and identify those that yield an sbWCXp of size 1. This process can be performed in $O(nm)$ time.

We can go further by noting that both cbWAXps and sbWCXps are ground-truth WAXps and WCXps, respectively.

**Proposition 14.** *Let $\mathcal{X}$ be a cbWAXp and $\mathcal{Y}$ be an sbWCXp for an instance $I$. If $\mathcal{X} \cap \mathcal{Y} = \{i\}$, then $i$ is a relevant feature.*

*Proof.* Since $\mathcal{X}$ is a cbWAXp, it is a ground-truth WAXp. Also, since $\mathcal{Y}$ is an sbWCXp, it is a ground-truth WCXp. We have that $\mathcal{X} \cap \mathcal{Y} = \{i\}$, so $i$ must belong to any possible AXp $\mathcal{X}' \subseteq \mathcal{X}$; otherwise, the WCXp $\mathcal{Y}$ would not be hit by $\mathcal{X}'$, nor would any CXp $\mathcal{Y}' \subseteq \mathcal{Y}$ (violating the hitting set duality). Thus, there exists at least one AXp containing $i$, which implies that $i$ is a relevant feature. □

This result enables an efficient procedure for detecting feature relevance. It suffices to compute all cbWAXps and sbWCXps and perform pairwise intersections, identifying as relevant any feature that forms a resulting singleton. These operations maintain the $O(n^2m)$ complexity of computing all cbWAXps (all sbWCXps can be computed in $O(nm)$), as the number of cbWAXps and sbWCXps is bounded by $n$, and the intersection of any two such sets can be performed in $O(m)$ by representing them as bitvectors of size $m$.

The following example illustrates the application of the three sufficient conditions:

| Inst. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\kappa(\mathbf{x})$ |
|---|---|---|---|---|---|
| $I_1$ | 0 | 0 | 0 | 0 | 0 |
| $I_2$ | 0 | 0 | 1 | 0 | 0 |
| $I_3$ | 0 | 1 | 1 | 0 | 1 |
| $I_4$ | 1 | 0 | 0 | 0 | 1 |

Table 2: Sample $\mathcal{S}$ for Example 5.

**Example 5.** *Consider the sample $\mathcal{S}$ shown in Table 2 and the instance $I_1 = ((0,0,0,0),0)$ to be explained. The cbWAXps for $I_1$ are $\mathcal{X}_1 = \{1,2,3,4\}$ and $\mathcal{X}_2 = \{1,2,4\}$. The sbWCXps for $I_1$ are $\mathcal{Y}_1 = \{2,3\}$ and $\mathcal{Y}_2 = \{1\}$, obtained from $I_3$ and $I_4$ respectively. Since $3 \notin \mathcal{X}_2$, by Proposition 12 it follows that the feature 3 is not necessary. Conversely, as $|\mathcal{Y}_2| = 1$, Proposition 13 implies that 1 is a necessary feature. Finally, the intersection $\mathcal{X}_2 \cap \mathcal{Y}_1 = \{2\}$ is a singleton, which by Proposition 14 guarantees that the feature 2 is relevant. The sufficient conditions do not allow us to determine the necessity of 2, the relevance of 3, nor the necessity and relevance of 4 for $I_1$ given the sample $\mathcal{S}$.*

### 6.2 Deciding Feature Necessity by Sampling

The sufficient conditions introduced so far rely on the evidence available in a fixed sample $\mathcal{S}$. As a consequence, certain properties may remain undecided, as seen in Example 5 regarding the necessity of the features 2 and 4.

However, the conditions for feature necessity can be leveraged to overcome this limitation efficiently if the ML model $\kappa$ is available as an oracle. We denote by $\mathbf{v}[i \leftarrow v']$ the point in feature space obtained by replacing the $i$-th component of $\mathbf{v}$ with the value $v' \in \mathbb{D}_i$. Then:

**Theorem 5.** *Let $I = (\mathbf{v}, c)$ be an instance such that $\kappa(\mathbf{v}) = c$. A feature $i \in \mathcal{F}$ is necessary for $I$ if and only if there exists a value $v' \in \mathbb{D}_i$ such that $\kappa(\mathbf{v}[i \leftarrow v']) \neq c$.*

*Proof.* If there exists $v' \in \mathbb{D}_i$ such that $\kappa(\mathbf{v}[i \leftarrow v']) \neq c$, the instance $(\mathbf{v}[i \leftarrow v'], \kappa(\mathbf{v}[i \leftarrow v']))$ yields the sbWCXp $\{i\}$. So, by Proposition 13, $i$ is necessary for $I$. Conversely, if $\kappa(\mathbf{v}[i \leftarrow v']) = c$ for all $v' \in \mathbb{D}_i$, by the consensus operation it follows that $\mathcal{F} \setminus \{i\}$ is a cbWAXp. So, by Proposition 12, $i$ is not necessary for $I$. □

From a computational perspective, this result ensures that deciding feature necessity requires at most $|\mathbb{D}_i| - 1$ oracle queries, as the prediction for the original value $v_i$ is already known. In particular, for binary domains, a single query suffices. Since model inference is typically polynomial in $m$ (e.g., in neural networks or random forests), our approach remains highly scalable. This stands in sharp contrast to the NP-hardness of checking feature necessity in model-based settings for many families of classifiers.

Furthermore, whenever a feature $i$ is found to be not necessary for $I$, the procedure effectively constructs a cbWAXp $\mathcal{F} \setminus \{i\}$, yielding a ground-truth WAXp that is strictly more informative than the trivial explanation $\mathcal{F}$.

| Dataset | zoo | kvk | car |
|---|---|---|---|
| $m$ | 16 | 36 | 6 |
| $n$ | 59 | 3196 | 1728 |
| Avg. # | 3.03 | 11.86 | 24.12 |
| Avg. Smallest | 14.98 | 33.68 | 2.10 |
| Min. Size | 14 | 31 | 1 |
| % I. nec | 6.78 | 32.04 | 75.98 |
| % I. not nec | 72.88 | 93.55 | 99.65 |
| % I. rel | 3.39 | 27.60 | 24.02 |
| Avg. nec | 0.07 | 0.42 | 1.66 |
| Avg. not nec | 1.49 | 3.39 | 4.34 |
| Avg. rel | 0.03 | 0.83 | 1.10 |

Table 3: Summary of results for the selected datasets.

## 7 Experimental Results

This section reports empirical results to assess consensus-based explanations.[4] We implemented a prototype to compute all distinct, subset-minimal, and cardinality-minimal cbWAXps for any given instance in a sample, as well as to verify the sufficient conditions for feature necessity and relevance from Section 6. The prototype is implemented in Python 3, and all the experiments were run on a Linux machine (AMD EPYC 9654 2.40 GHz, 512 GB RAM). The experimental study considers a number of well-known datasets as samples that represent a model in the sample-based setting. In all cases, we removed duplicate rows.

The study is divided into two parts: a fixed-sample analysis to assess the number, size, and usefulness of the existing cbWAXps, and an experiment to observe their evolution as the available evidence grows.

### 7.1 Fixed-Sample Analysis

We first consider three datasets: `zoo`, `kr-vs-kp` (referred to as `kvk`), and `car`. The `zoo` dataset contains $m = 16$ features (15 binary and one categorical with 6 values) and few instances ($n = 59$); it classifies animals into 7 categories. `kvk` is a larger dataset, with $m = 36$ binary features and $n = 3196$; it represents a chess game where the goal is to predict if a player wins or not. Finally, `car` is an exhaustive dataset with $m = 6$ features (three with domain size 4 and three with domain size 3), representing a model that classifies cars into four categories. Since all $n = 1728$ possible combinations of feature values are present, the sample represents the entire feature space.

Table 3 summarizes the results for each dataset. We report the average number of cbWAXps per instance (Avg. #), the average size of the smallest cbWAXp found for each instance (Avg. Smallest), and the size of the smallest explanation found overall in the whole dataset (Min. Size). Furthermore, the table shows the percentage of instances in the sample for which the proposed sufficient conditions identified at least one necessary (% I. nec), not necessary (% I. not nec),

[4] Detailed results and experimental data are available at https://github.com/raulmencia/kr26-cbwaxps.git.

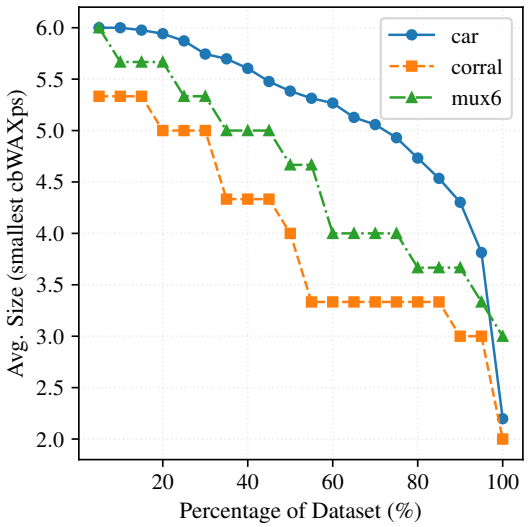


Figure 1: Avg. size (smallest cbWAXps).

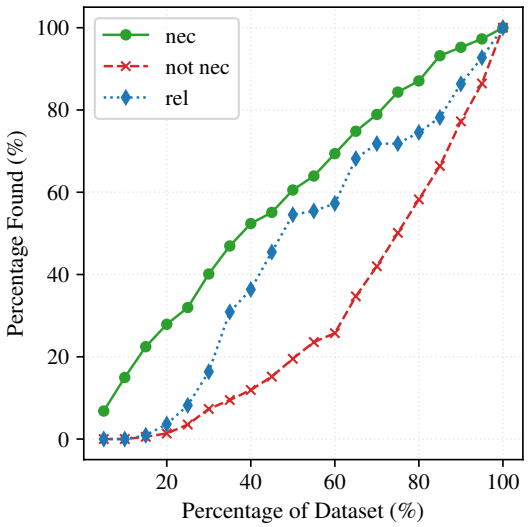


Figure 2: Feature necessity and relevance (car).

or relevant feature (% I. rel), along with the average number of features identified in each category (Avg. nec, Avg. not nec, and Avg. rel). Regarding relevant features, we do not include those identified as necessary, even though they are indeed relevant. This is to better highlight the effectiveness of each of the sufficient conditions presented in Section 6.

The results for the zoo dataset are the most conservative, as could be expected. Clearly, the cbWAXps are few and large, with the smallest one containing 14 out of the 16 features. However, they still provide useful, sound information where other methods may offer shorter but potentially incorrect explanations. For 43 out of 59 instances, at least one non-trivial cbWAXp exists, which leads to identifying at least one non-necessary feature in more than 70% of cases. Also, relevant and necessary features are identified for a few instances. In kvk, we observe a clear increase in the number of explanations per instance. In this case, the smallest cbWAXps are more effective at discarding features, with the overall smallest explanation containing 31 out of the 36 features. This shows that as the sample size grows, the framework provides more succinct and informative explanations. Furthermore, the sufficient conditions for necessity and relevance are much more effective, identifying these properties for a significant portion of the dataset, and in higher numbers on average. Finally, for car, since the sample covers the entire feature space, cbWAXps coincide with the complete set of ground-truth WAXps. Notably, our sufficient conditions correctly identify all necessary, not necessary, and relevant features. This was confirmed by comparing the results against the full set of cbAXps (which in this case represent ground-truth AXps).

Regarding computation times, cbWAXps are computed efficiently. Specifically, on average, enumerating all cbWAXps and identifying necessary and relevant features for a single instance took 0.0005 s, 2.13 s and 0.22 s for zoo, kvk, and car, respectively.

## 7.2 Evolution Under Sample Expansion

We conducted an additional experiment to evaluate how cbWAXps evolve as the sample grows. For this purpose, we employed the car, mux6, and corral datasets, as all these are exhaustive—i.e., they contain all possible points in feature space. Both mux6 and corral consist of $n = 64$ instances, $m = 6$ features, and two classes.

In each case, we initially selected a random subset comprising 5% of the instances in the dataset. We then computed all cbWAXps for these selected instances using samples of increasing size: $5\%, 10\%, 15\%, \ldots, 100\%$. At each step, the sample was expanded by adding instances chosen at random, among the ones not yet included. Notice that at 100%, all ground-truth WAXps are found.

Figure 1 shows the evolution of the average size of the smallest cbWAXps for the selected instances as the sample size increases. As can be observed, the size of the discovered explanations consistently decreases as the sample grows. This reduction is more gradual for mux6 and corral, where cbWAXps are refined more regularly as new instances are added. In contrast, the trend for car is more pronounced towards the end. This behavior may be due to the larger size and dimensionality of car. Furthermore, as these results are averaged across all selected instances and they are based on random additions, a more efficient reduction in size could be expected if a sampling strategy focused on the instance being explained were used.

Despite the trend observed for the size of explanations in car, the computed cbWAXps are useful for identifying the necessary and relevant features throughout the expansion process. Figure 2 shows the percentage of all necessary, not necessary, and relevant (but not necessary) features discovered as the sample grows, averaged for the selected instances. Notably, the progression here is much more gradual than the one observed for explanation size.

## 8 Conclusions

This paper introduces a novel class of sample-based abductive explanations, namely those derivable by consensus. Arguably, these explanations attain the highest possible degree of rigor: they are provably correct and provide formal guarantees that no other sample-based approach can offer. Specifically, they are the only sample-based abductive explanations that fulfill the monotonicity property, ensuring they remain valid as the sample grows. Furthermore, they are the only guaranteed ground-truth explanations for any classification function consistent with the sample. In critical scenarios where misinterpretation is not an option, this guarantee of soundness becomes indispensable.

From a computational perspective, cbWAXps offer a significant advantage over many existing logic-based approaches. We have shown that they can be computed efficiently; beyond finding a single explanation, the framework allows for the enumeration of all such explanations and the identification of the smallest ones in polynomial time in the size of the input. Furthermore, even in cases where only a fixed sample is available, cbWAXps can be effectively used to discover necessary and relevant features. This capability is further enhanced when the ML model is accessible as an oracle: in such cases, feature necessity can be decided with a query complexity linear in the size of the corresponding domain. Such efficiency encourages further research into *logic-based* sampling strategies that could exploit this framework in order to strategically query the model in a targeted and efficient manner.

These explanations can be large in practice, as was experimentally observed. However, their size is the necessary price to be paid for the guarantees they bring, and a reflection of the available evidence in the sample. Consequently, cbWAXps are not intended to replace existing XAI methods, but rather to serve as a complement to them within a robust explainability pipeline. Their unique formal guarantees of correctness and low computational cost allow them to serve as a verification layer for other sample-based explanations. For instance, our framework could be used to *enforce* that other (smaller) sample-based explanations contain the necessary features identified, making them more reliable.

Finally, we envision a number of promising directions for future work: (i) conducting a comprehensive experimental study to further analyze the advantages and limitations of cbWAXps compared to other sample-based explanations in the literature; (ii) extending our current theoretical analysis on the identification of necessary and relevant features; (iii) developing active sampling strategies to reduce explanation size while minimizing the total number of queries required when the model is used as an oracle; (iv) exploring heuristics and parallelization techniques to further enhance the scalability of our algorithms when dealing with very large samples; and (v) investigating probabilistic relaxations of the monotonicity property, which would potentially allow for a broad range of explanations balancing conciseness and soundness guarantees in a principled way, thereby providing a more flexible framework.

## Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments and constructive suggestions, which helped improve the quality of this paper. This work is partially supported by the Spanish Government under grants PID2023-152814OB-I00, PID2022-141746OB-I00, and PID2022-139835NB-C22, by the Principality of Asturias under grant SEK-25-GRU-GIC-24-018, and by the Catalan Government under grant 2021-SGR-1615.

## AI Declaration

Generative AI was used only for linguistic editing to improve readability. All ideas, results, and conclusions were developed by the authors, who take full responsibility for the content.

## References

Amgoud, L., and Ben-Naim, J. 2022. Axiomatic foundations of explainability. In *IJCAI*, 636–642.

Amgoud, L.; Cooper, M. C.; and Debbaoui, S. 2024. Axiomatic characterisations of sample-based explainers. In *ECAI*, 770–777.

Amgoud, L. 2023. Explaining black-box classifiers: Properties and functions. *Int. J. Approx. Reason.* 155:40–65.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.; and Marquis, P. 2022. On the explanatory power of boolean decision trees. *Data Knowl. Eng.* 142:102088.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

Bassan, S., and Katz, G. 2023. Towards formal XAI: formally approximate minimal explanations of neural networks. In *TACAS*, 187–207.

Blake, A. 1937. *Canonical expressions in Boolean algebra*. Ph.D. Dissertation, The University of Chicago.

Boros, E.; Hammer, P. L.; Ibaraki, T.; Kogan, A.; Mayoraz, E.; and Muchnik, I. B. 2000. An implementation of logical analysis of data. *IEEE Trans. Knowl. Data Eng.* 12(2):292–306.

Brown, F. M. 1990. *Boolean reasoning - the logic of boolean equations*. Kluwer.

Chikalov, I.; Lozin, V. V.; Lozina, I.; Moshkov, M.; Nguyen, H. S.; Skowron, A.; and Zielosko, B. 2013. *Three Approaches to Data Analysis - Test Theory, Rough Sets and Logical Analysis of Data*. Springer.

Cooper, M. C., and Amgoud, L. 2023. Abductive explanations of classifiers under constraints: Complexity and properties. In *ECAI*, 469–476.

Crama, Y.; Hammer, P. L.; and Ibaraki, T. 1988. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research* 16:299–325.

Darwiche, A., and Hirth, A. 2023. On the (complete) reasons behind decisions. *J. Log. Lang. Inf.* 32(1):63–88.

Darwiche, A. 2023. Logic for explainable AI. In *LICS*, 1–11.

Geng, Z.; Schleich, M.; and Suciu, D. 2022. Computing rule-based explanations by leveraging counterfactuals. *Proc. VLDB Endow.* 16(3):420–432.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51(5):93:1–93:42.

Ignatiev, A.; Narodytska, N.; Asher, N.; and Marques-Silva, J. 2020. From contrastive to abductive explanations and back again. In *AIxIA*, 335–355.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. On relating explanations and adversarial examples. In *NeurIPS*, 15857–15867.

Ignatiev, A. 2020. Towards trustable explainable AI. In *IJCAI*, 5154–5158.

Izza, Y.; Huang, X.; Morgado, A.; Planes, J.; Ignatiev, A.; and Marques-Silva, J. 2024. Distance-restricted explanations: Theoretical underpinnings & efficient implementation. In *KR*.

Koriche, F.; Lagniez, J.; Mengel, S.; and Tran, C. 2024. Learning model agnostic explanations via constraint programming. In *ECML*, 437–453.

Lazo-Cortés, M. S.; Trinidad, J. F. M.; Carrasco-Ochoa, J. A.; and Sánchez-Díaz, G. 2015. On the relation between rough set reducts and typical testors. *Inf. Sci.* 294:152–163.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, 4765–4774.

Marques-Silva, J., and Huang, X. 2024. Explainability is *Not* a game. *Commun. ACM* 67(7):66–75.

Marques-Silva, J.; Janota, M.; and Mencía, C. 2017. Minimal sets on propositional formulae. Problems and reductions. *Artif. Intell.* 252:22–50.

Marques-Silva, J.; Lefebre-Lobaina, J. A.; and Martinez, M. V. 2025. Efficient and rigorous model-agnostic explanations. In *IJCAI*, 2637–2646.

Marques-Silva, J. 2022. Logic-based explainability in machine learning. In *Reasoning Web*, 24–104.

Marques-Silva, J. 2024. Logic-based explainability: Past, present and future. In *ISoLA*, 181–204.

Molnar, C. 2020. *Interpretable machine learning*. Lulu.com.

Pawlak, Z. 1982. Rough sets. *Int. J. Parallel Program.* 11(5):341–356.

Pawlak, Z. 2002. Rough sets and intelligent data analysis. *Inf. Sci.* 147(1-4):1–12.

Quine, W. V. 1952. The problem of simplifying truth functions. *American mathematical monthly* 521–531.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artif. Intell.* 32(1):57–95.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*, 1527–1535.

Rudin, C., and Shaposhnik, Y. 2019. Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *SSRN* (3395422).

Rudin, C., and Shaposhnik, Y. 2023. Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *J. Mach. Learn. Res.* 24:16:1–16:44.

Schleich, M.; Geng, Z.; Zhang, Y.; and Suciu, D. 2021. GeCo: Quality counterfactual explanations in real time. *Proc. VLDB Endow.* 14(9):1681–1693.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, 5103–5111.

Wu, M.; Wu, H.; and Barrett, C. W. 2023. VeriX: Towards verified explainability of deep neural networks. In *NeurIPS*.