Scientific Theory of a Black-Box: A Life Cycle-Scale XAI Framework Based on Constructive Empiricism

Anonymous authors
Paper under double-blind review

Abstract

Explainable AI (XAI) offers a growing number of algorithms that aim to answer specific questions about black-box models. What is missing is a principled way to consolidate explanatory information about a fixed black-box model into a persistent, auditable artefact, that accompanies the black-box throughout its life cycle. We address this gap by introducing the notion of a *scientific theory of a black box* (SToBB). Grounded in Constructive Empiricism, a SToBB fulfils three obligations: (i) *empirical adequacy* with respect to all available observations of black-box behaviour, (ii) *adaptability* via explicit update commitments that restore adequacy when new observations arrive, and (iii) *auditability* through transparent documentation of assumptions, construction choices, and update behaviour. We operationalise these obligations as a general framework that specifies an extensible observation base, a traceable hypothesis class, algorithmic components for construction and revision, and documentation sufficient for third-party assessment. Explanations for concrete stakeholder needs are then obtained by querying the maintained record through interfaces, rather than by producing isolated method outputs. As a proof of concept, we instantiate a complete SToBB for a neural-network classifier on a tabular task and introduce the Constructive Box Theoriser (CoBoT) algorithm, an online procedure that constructs and maintains an empirically adequate rule-based surrogate as observations accumulate. Together, these contributions position SToBBs as a life cycle-scale, inspectable point of reference that supports consistent, reusable analyses and systematic external scrutiny.

1 Introduction

Explainable AI (XAI) provides methods for analysing black-box decisions. Each method builds on its own assumptions and evaluation criteria and is intended to answer a specific question for specific users (Speith, 2022; Tomsett et al., 2018; Nauta et al., 2023; Mersha et al., 2024). While XAI offers methods for the individual needs, these methods are typically applied in an ad hoc manner, producing isolated outputs that may need to be recomputed for similar questions and that can yield incompatible characterisations (Krishna et al., 2024) of the same model. To address this problem, we suggest a perspective that treats explanatory information as an evolving artefact maintained across the black-box's life cycle, rather than as a sequence of disconnected method outputs. Accordingly, such a foundation should aim to ensure adequacy, adaptability, and auditability over time, providing a stable basis for consistent analyses while remaining open to new forms of inquiry. We argue that one way forward is to approach explainability for machine learning systems in the same role that scientific theories play in science: designed not to be answers to any particular question outright, but to be persistent and auditable consolidations of knowledge that provide a shared point of reference and support diverse forms of inquiry. To this end, we propose and develop the idea of a *scientific theory of a black-box* (SToBB).

To consider a more concrete scenario, one can imagine a fixed classifier trained on the Abalone dataset (Nash et al., 1994) that predicts whether an individual abalone has reached a target age, relevant to harvesting decisions and pricing. During development, model developers probe the classifier globally to understand which input features (e.g., physical measurements of shell size or weight) influence predictions and where the model fails, using global summaries and debugging-oriented explanations. After deployment, operators

and employees in an aquaculture setting may request local or contrastive explanations for specific batches, e.g., why a given lot is classified as not ready for harvest and what changes in measurements would flip that assessment. At the same time, sustainability officers or external auditors may periodically inspect the model's behaviour over time, asking whether different abalone subpopulations are treated consistently and whether model decisions reflect sustainability obligations. Wholesalers seeking to buy abalone may request justification for the asked price, which could be supported by explanations that demonstrate the relevant qualities of the batch, such as predicted age, size, and compliance with sustainability constraints. All of these inquiries concern the same underlying classifier and lead to similar probings of it, but arise in different contexts, i.e., different stages of the system's life cycle and from different stakeholders who vary in intent and background knowledge. To address a diverse set of inquiries, a SToBB organises shared explanatory information as a maintained record that can accumulate over time and be re-used in different contexts; concrete, context-appropriate explanations are then obtained by querying this record.

For the conceptual grounding of the SToBB we turn to philosophy of science, which offers established accounts of scientific theories and their function. In particular, *Constructive Empiricism* (CE) (van Fraassen, 1980) aligns with two central requirements in XAI: fidelity to observed model behaviour and pragmatic usefulness to different users. In CE, scientific theories are judged by their *empirical adequacy*, that is, their agreement with all observed phenomena, and by their *usefulness* to the scientist. Acceptance of a theory rests on two commitments: conviction in the theory's empirical adequacy and a willingness to use and adapt it as new observations arise. A theory is therefore part of a continuous process in which observations either reaffirm its adequacy or demand refinement. Notably, CE does not assume that theories describe an underlying truth that lies beyond what is observable. This stance avoids overclaiming access to the supposed hidden "true reasoning" of a black-box while still providing a structured and auditable basis for analysis.

To operationalise the concept of a SToBB for XAI, *surrogate models* offer a natural starting point. Surrogates are inherently interpretable models designed to approximate the black-box's input-output mapping (Bastani et al., 2017; Molnar, 2020). They are typically applied post hoc and independently of the black-box's architecture (Speith, 2022) and may approximate behaviour locally (Ribeiro et al., 2016; Lundberg & Lee, 2017; Ribeiro et al., 2018) or globally (Bastani et al., 2017). In contrast to theories in CE, they are not generally required to achieve perfect agreement with the black-box on all observed data, nor are they systematically adapted when inconsistencies appear. This limits their epistemic value over time. Constructive Empiricism provides a principled basis to recast surrogates as scientific theories: to achieve empirical adequacy, the surrogate must agree with the black-box on all available observations, and to justify acceptance demands adaptability to preserve adequacy as new observations accumulate.

Scientific theories are also documented, scrutinised, and evaluated within a community. The same principle motivates a SToBB, which must make explicit its assumptions, construction choices, and update behaviour to enable third parties (e.g., auditors or operators) to independently assess its merit, scope, and limitations. Detailed transparency requirements and their connection to existing documentation practices are discussed in Section 2.

It is important to emphasise that we consider the SToBB a complementary perspective in XAI and not a universal solution. It addresses a specific gap: the lack of a persistent, expanding, auditable representation of explanatory information that accompanies a black-box as observations accumulate. As a result, SToBBs serve as a life cycle-scale, inspectable point of reference for explanation: empirical adequacy provides a checkable baseline, while decoupling explanatory information from its context-dependent presentation enables consistent, reusable analyses across stakeholders. Throughout this paper, we treat the black-box model as fixed, meaning that its decision behaviour does not change while explanatory information is collected. We return to this assumption and its implications in the discussion.

In this paper we make the following contributions:

- We introduce the notion of a SToBB, a documented and evolving artefact that represents a fixed black-box model via an interpretable surrogate, an observation base, and explicit update and documentation commitments.

- As a conceptual prerequisite to this definition, we operationalise CE for XAI by translating empirical adequacy, acceptance as commitment, and pragmatic virtues into concrete design and documentation obligations for constructing and maintaining SToBBs.
- We present a proof-of-concept that instantiates a complete SToBB for a neural-network classifier on a tabular task; to this end, we introduce the *Constructive Box Theoriser* (CoBoT) algorithm as a generic procedure for tabular domains, which may be of independent interest.

Outline: Section 2 motivates the structural gap in the current XAI landscape that a SToBB aims to address. Subsequently, Section 3 introduces the central notions of Constructive Empiricism and transfers the concepts to XAI. This yields an overview of the properties a SToBB must fulfill. Based on this, Section 4 defines the constituents of the SToBB and discusses its properties. Section 5 goes on to describe and exemplify the process to create a SToBB, also introducing the CoBoT algorithm. Section 6 provides the discussion and conclusion.

2 Context

In this paper we argue for a perspective that treats explanatory information regarding a black-box as something that is collected and integrated over time. In this section we examine practical motivations for this perspective, focusing on two aspects: (i) heterogeneous stakeholder needs, and (ii) transparency needs that arise throughout the AI system life cycle.

(i) Stakeholders and explanatory needs The XAI literature recognizes a diverse set of stakeholders who require insight into the behaviour of AI systems. These groups include, but are not limited to, developers, regulatory entities, operators, domain experts, and affected users (Tomsett et al., 2018; Barredo Arrieta et al., 2020; Langer et al., 2021; Mersha et al., 2024). Langer et al. (2021) emphasize that individuals may belong to several stakeholder groups and that individuals within a group can differ in expertise, background knowledge, and expectations. Research on user needs further shows that these stakeholders ask different kinds of explanatory questions and rely on different levels of abstraction (Liao et al., 2020; 2021).

(ii) Transparency needs across the AI life cycle Stakeholder groups largely identical to those identified in the XAI literature also appear in several Trustworthy ML frameworks and standards.[1] All frameworks embed AI systems within a broader *life cycle* that spans design, development, testing, verification, deployment, and monitoring. Although the frameworks differ in focus – from risk management to organisational governance – they identify similar qualities for trustworthiness, such as accountability, human agency, technical robustness, compliance, and transparency. Importantly, transparency is regarded not only as a property in itself but also as an enabler for these other qualities, and transparency needs arise *continuously* at all stages of the life cycle (NIST, 2023; EU HLEG AI, 2019; OECD, 2024). Explainability plays an important role in addressing some of these transparency needs. As the Abalone example in the introduction motivates: Developers may rely on debugging-oriented analyses early in the life cycle, end-users may require local or contrastive decision explanations to verify fair treatment or obtain actionable information, and operators or auditors may need system-level evidence over time to assess compliance and orderly operation.

The structural gap in the XAI landscape While XAI acknowledges diverse stakeholders, the life cycle perspective is largely absent. Surveys of XAI methods (Barredo Arrieta et al., 2020; Speith, 2022; Mersha et al., 2024) show that research predominantly develops techniques for isolated explanatory tasks. Although these methods see successful practical use (Liao et al., 2021; Arya et al., 2019), they remain compartmentalised even when probing the *same* black-box model. In particular, explanations are produced independently of each other, which may lead to incompatible descriptions of model behaviour, a phenomenon known as the *disagreement problem* (Krishna et al., 2024), or to redundant computations over time.

[1] IEEE P7001 (IEEE, 2022), ISO/IEC 42001 (ISO/IEC, 2023), the OECD AI Recommendation (OECD, 2024), the NIST AI Risk Management Framework (NIST, 2023), the Fraunhofer Guidelines for Trustworthy AI (Poretschkin et al., 2023), and the EU Ethics Guidelines for Trustworthy AI (EU HLEG AI, 2019) that now underpin the EU AI Act (European Parliament and European Council, 2024)

Recall that stakeholders may have different explanatory goals and differ in their background knowledge. Despite these differences, all inquiries concern the same underlying model. If explanatory information is collected only for the immediate question at hand, it remains siloed and must be recomputed when similar questions arise later in the life cycle. A structured approach that records and consolidates explanatory information therefore enhances efficiency, consistency, and reuse across stakeholders. We propose to treat explanatory information not as a set of disconnected method outputs, but as a persistent *record* that aggregates observations and remains consistent as it evolves. Such a record is not itself an explanation, but a representational foundation from which diverse explanations can be derived. To the best of our knowledge, there is no artefact in the current XAI landscape intended to simultaneously (i) record, (ii) represent, and (iii) continuously update explanatory information for a fixed black-box as observations accumulate, while remaining auditable and reusable across the system life cycle. The SToBB framework fills this gap by treating explanatory information as a maintained record rather than a sequence of disconnected method outputs.

What would such an artefact be? From the perspective of the philosophy of science, something that represents the behaviour of a system and serves as a common foundation from which diverse explanations can be derived is a *scientific theory*. Following this perspective we turn, in particular, toward the operationalisation of CE (van Fraassen, 1980) as a framework for structuring such a record. The semantic conception of theories in CE and the emphasis on empirical adequacy and updateability align closely with the requirements identified above. The next section introduces the core notions of CE and discusses their translation to XAI.

Documentation as a second strand of transparency While some transparency needs are addressed by XAI methods, others are met by structured documentation aimed at stakeholders not directly involved in system development, e.g., to help understand a system's overall design, limitations, and intended use. Documentation frameworks exist that aim at a life cycle-scale solution (Chmielinski et al., 2024) or that support transparency for specific components of the ML pipeline: datasheets document datasets (Gebru et al., 2021), model cards document trained models (Mitchell et al., 2019), and care labels summarise resource usage (Fischer et al., 2023). On the one hand, a SToBB itself serves as a piece of evolving documentation, falling in line with the works just mentioned but with a different focus. On the other hand, as the SToBB aims to remain useful throughout the black-box's life cycle, it itself will require accompanying documentation.

3 Adopting Constructive Empiricism to Explainable AI

In this section, we translate the central notions of CE into requirements for explanatory artefacts in XAI. This translation of requirements will serve as the conceptual basis for the *Scientific Theory of a black-box* (SToBB), introduced in Section 4. CE introduced by van Fraassen, is a philosophy of science that provides a definition of a scientific theory and a description of its relation to the world and to the scientist using it. A key feature of CE is that it does not require a theory to be literally true about the world, because the theory may include highly abstract elements or unobservable properties and processes whose correspondence with reality we cannot ascertain. Rather, a theory must be *empirically adequate* and pragmatically useful for scientific practice.

We will discuss *observational* and *relational substructures* as the two necessary constituents of a theory, and the notions of *empirical adequacy*, *theory acceptance*, and *pragmatic virtues*. For the reader's convenience, Table 1 aligns the central concepts of CE with their counterparts in our setting and already anticipates the design obligations that are defined formally in Section 4 and instantiated as a process in Section 5. It can already serve as a high-level guide for first-time readers, indicating the direction of subsequent sections, and it may also be referenced as a compact summary once the details have been introduced.

3.1 Observational and Relational Substructures

In CE's semantic view, a theory is presented as a family of structures. It distinguishes *observational* from *relational* substructures. Observational substructures represent the measurable states and outcomes accessible through instruments. Measurements may be erroneous and such observations may be rejected. *Relational substructures* in turn capture the rules or relations connecting those observations (think of variables and

CE concept	CE definition	Constructing a Scientific Theory of a Black-Box
Observational substructures	The measurable states and outcomes accessible to science.	Define the observation space, including inputs, targets, possible auxiliary measurements derived from the black box and respective extraction methods. Make explicit what counts as an observation and how its quality is ensured.
Relational substructures	The rules or relations acting on observations.	Specify the hypothesis class of the surrogate and the structural constraints that make its input–output mapping interpretable.
Empirical adequacy	All observable phenomena must be representable and consistent with the theory.	Ensure that the surrogate reproduces the black box on all available observations. Adequacy is binary and must be restored through updates whenever inconsistencies appear.
Theory acceptance	Belief in empirical adequacy and commitment to its continued use.	Ground belief in demonstrable agreement with all observations and define a clear update policy to preserve adequacy when new data arise.
Pragmatic virtues	Criteria to choose between empirically adequate theories on grounds of usefulness.	Identify and justify pragmatic criteria shaping the surrogate, including user-centered design choices (e.g., compactness, coherence, comprehensibility), auxiliary observables added for pragmatic reasons, and documentation with diagnostic measures that record adequacy, adaptation, and design evolution.

Table 1: Conceptual alignment between CE and the design obligations of a SToBB. It summarizes how each CE concept informs both structure and process.

their composition to a formula). Together, they provide both the descriptive surface of a theory and the framework that supports explanation.

Observational substructures: For a SToBB, this requires making explicit what counts as an observation: the input variables, the target outputs, and the procedures by which they are measured. As we will later discuss, an observation must include, but need not be limited to, the inputs used by the black-box (Section 3.4).

Relational substructures: The semantic view of CE aligns with the notion of *surrogate models*: a hypothesis class forms a family of admissible structures, and each instantiated surrogate corresponds to a concrete structure within that family. For a SToBB, this requires specifying the surrogate's hypothesis class and the constraints imposed on it so that the mapping from inputs to outputs can be traced along an interpretable structure. We use *traceability* to refer to the ability to follow how observations are represented and processed by the surrogate in an interpretable way and to link specific surrogate components back to the observations that constrain them.

3.2 Empirical adequacy

According to CE, the central aim of a scientific theory is to be *empirically adequate*, that is, to be consistent with *all available* observations. A theory is empirically adequate if all observable phenomena can be represented within one of its observational substructures. This requires the theoretical ability to represent any possible measurement outcome as well as achieving internal consistency with all actually observed data. Importantly, CE does not claim that theories describe unobservable truths about the world; it is sufficient

that they consistently account for all observed behaviour. No theory that conflicts with accepted observations can be advocated as correct. But there is no claim that the relational substructures represent the true reasoning that is being executed by the black-box or the hidden rules that really govern its behaviour.

Transferring this to surrogate models in XAI, the surrogate must reproduce the black-box on all currently available observations, but it needs not give guarantees for unobserved black-box behaviour. This is stricter than typical surrogate modelling in XAI, where approximation error on observed data is tolerated, and looser than standard ML practice, where performance on a held-out test set is used as a proxy for generalisation to an unknown data distribution. In the context of a SToBB, the black-box represents a fixed target function, and the task is not to generalise beyond it but to reproduce its behaviour on all available observations. As the next subsection makes more explicit, this adequacy is an important building block to provide justification for trust in the surrogate.

3.3 Theory acceptance

For a scientist to *accept* a theory means (a) to believe that it is empirically adequate, and (b) to commit to using it. Committing entails a willingness to work with the theory in the expectation that it will be applicable to future observations, and to adapt it if inconsistencies arise. In van Fraassen's view, theory building is an iterative process informed by observations and testing.

While empirical adequacy is concerned with the relation between the theory and the world, acceptance pertains to the relation between the theory and the scientist. We see the acceptance of a theory in CE as analogous to accepting a model or an explanation. Accordingly, a SToBB needs to give grounds to justify its acceptance, i.e., the belief in its empirical adequacy and the decision to commit to its use. Belief in adequacy is grounded in the surrogate *demonstrably agreeing* with all available observations; the traceability requirement of the hypothesis class ensures that this agreement can be checked by linking surrogate behaviour back to specific observations. Commitment extends this stance into the future by requiring to maintain the surrogate's adequacy: if new observations reveal inconsistencies, the surrogate must be adapted to restore adequacy and the adaptation policy must be outlined comprehensibly. Thus, under CE, grounds for trusting the surrogate shift from one-off predictive performance on unseen data to transparent and continuous process-based guarantees.

3.4 Pragmatic virtues

Since multiple empirically adequate theories can coexist, CE recognises additional criteria for preference. If faced with two empirically adequate theories, it is rational to prefer one over the other solely on the basis of *usefulness*, whatever this may entail for the scientist. Properties that are determined to be relevant to this end are subsumed under the notion of *pragmatic virtues*. This may for example include choice of language, simplicity constraints, or other features that aid scientific inquiry. Pragmatic virtues do not concern the truth of the theory but only its practical value to the scientist.

In the context of a SToBB, two broader categories of virtues are particularly relevant:

(i) *User-centred criteria.* With the goal of being useful for explanation, stakeholder-centred criteria enter already at the design stage, where several choices must be made. For example, a general consensus exists regarding which broader hypothesis classes are considered inherently interpretable (Molnar, 2020; Barredo Arrieta et al., 2020). Designers may take these as inspiration when specifying the surrogate of a SToBB. Each hypothesis class is typically associated with established explanation formats for which further characterisations exist that describe desirable properties (Nauta et al., 2023). With the goal of later supporting explanation, the designer might therefore consider how their chosen hypothesis class could best enable such desirable properties.

Another important aspect is the inclusion of *auxiliary measures* in the observation space that quantify black-box behaviours beyond the raw prediction. The motivation for this arises from the Rashomon effect: multiple surrogates can achieve perfect fidelity to the black-box while relying on different internal logics (Breiman, 2001; Marx et al., 2020). Extending the observation space with auxiliary

measures, such as feature-attribution information or other observations derivable from the black-box, constrains adequacy on a richer set of observables, reducing this multiplicity. Such enrichment does not reveal whether the surrogate's reasoning coincides with the "real reasoning" that underlies the behaviour of the black-box, but it adds observable intermediates that make the surrogate's behaviour empirically testable at finer resolution. From a CE perspective, this is a pragmatic virtue: it increases empirical discipline and user confidence without claiming access to unobservable reasoning.

(ii) *Documentation.* A SToBB is intended to serve not only its creators but also external stakeholders such as operators, auditors, domain experts, and affected users. To enable independent judgment and informed use, a SToBB should therefore be accompanied by documentation, similar to model cards and datasheets mentioned in Section 2. It has to provide a detailed account of information related to its design, as to make assumptions, intended use, and scope explicit. It should detail how adequacy of the surrogate is determined, how its update process operates, and how other pragmatic design choices affect its behaviour. Analogous to how standard ML practices report learning curves, model size, or stability metrics to characterise training dynamics, the documentation can include diagnostic measures of its surrogate that record the evolution over time, e.g., how often new data triggers updates or how its structural components take shape. These diagnostics are not part of empirical adequacy but complement the descriptive documentation with quantitative evidence.

Utilising this alignment of CE to XAI concepts, we introduce SToBB, a novel artefact, in the next section.

4 Constituents and Properties of a SToBB

The previous section derived a set of concrete obligations from Constructive Empiricism (CE) and explained how they translate into operational requirements for XAI components. We now bring these elements together to define a SToBB and describe the properties that follow from its structure. Section 5 will discuss and exemplify the construction of such an artefact.

Constituents. A *scientific theory of a black-box* (SToBB) is a structured artefact that represents the behaviour of a fixed black-box through an interpretable surrogate and the processes that maintain it. A SToBB consists of the following five components:

1. ***Observation base*:** An extensible record of black-box input-output behaviour and auxiliary measures. These observations form the empirical basis that the surrogate must match to demonstrate empirical adequacy.
2. ***Hypothesis class*:** The specification of a traceable model family that operates on the observation space.
3. ***Algorithmic components*:** Procedures for *constructing* an initial surrogate and for *updating* it when new observations are added, ensuring that empirical adequacy is established and maintained over time.
4. ***Adequate surrogate*:** A concrete model instantiated from the hypothesis class that agrees with all records in the observation base.
5. ***Documentation*:** A description of all theoretical, functional, and operational properties and requirements, sufficient to enable a third party to audit, deploy, and fully maintain the SToBB.

Each component of a SToBB is required to satisfy one or more obligations derived in Section 3. In brief: *Empirical adequacy* requires an explicit observation base, an instantiated surrogate, and a clearly defined adequacy criterion. *Acceptance as commitment* requires an update procedure that restores adequacy when new observations arrive, and a hypothesis class that constrains the form of revisions. *Pragmatic virtues* require interpretable structure, diagnostics, and documentation that support transparency, auditability, and structured use.

4.1 Interfaces: Extracting explanations

We deliberately define the role of a SToBB to be the collection and representation of explanatory information in an empirically adequate and accessible manner; it is *not* its task to answer any particular question outright. When a concrete explanatory need arises for a given stakeholder and context, the information in the SToBB is accessed through *interfaces*. An interface is a procedure that takes the current SToBB and a query and returns an answer tailored to that concrete application demand. It is analogous to a way of accessing a scientific theory or applying it for some specific purpose. Thus, interfaces are not core components of a SToBB, and the SToBB does not need to anticipate all future interfaces. They are important to the use of a SToBB but not for the conceptualisation in this paper. The minimal requirement for the SToBB is that the documentation is detailed enough to enable third parties to implement interfaces for their own use-cases.

As to not undermine the intent of the SToBB, a few general constraints for interfaces seem justified: To provide context-appropriate answers, interfaces may transform and structure information contained in the SToBB and add background information that the explainee requires. Since interfaces may answer questions regarding previously seen as well as unseen samples, newly encountered samples can be inconsistent with the current SToBB. In such cases, interfaces must not directly modify the SToBB; instead, an update must proceed via the documented update procedure. To preempt the disagreement problem, an interface should, if possible, preserve in its output the traceability to both the underlying observations and the surrogate structure.

In line with pragmatic virtues, interfaces defined over time may be included with the SToBB. In that case, they should themselves be described in the documentation (purpose, assumptions, and dependencies on the SToBB), so that their purpose and behaviour are transparent for third parties. Creators of a SToBB will likely design interfaces for their own use, which may serve as useful examples for later maintainers.

4.2 Properties of a SToBB

The following properties stem directly from the definition of a SToBB, making it particularly valuable for XAI practice by providing a maintained and auditable point of reference across the black-box's life cycle.

Baseline quality guarantees. Because the surrogate is required to be empirically adequate, any analysis derived from it is grounded in a model that reproduces all known behaviour of the black-box. Interfaces may impose additional quality checks, but these build on a demonstrably consistent foundation.

Extensibility. The usefulness of the SToBB is not limited to only the use-cases its designers anticipated. Explanations are obtained through interfaces defined over the surrogate and its documentation. New types of analyses can therefore be introduced without rebuilding the SToBB, as long as they rely on the existing structures.

Accumulation of knowledge. As observations accumulate, the SToBB is strengthened because adequacy must hold for an increasingly rich empirical base. Existing results remain available, and new analyses can reuse previously established structure, which is especially valuable in time-sensitive settings.

Traceability. Every structural element of the surrogate can be linked back to the observations and update steps that shaped it. Explanations produced through transparent interfaces can likewise be traced to specific components of the SToBB, forming a continuous chain of reference to underlying data.

Common point of reference. Because all interfaces draw from the same surrogate and documentation, their results are comparable and interpretable in a shared context. If discrepancies arise between interface outputs, the transparency of the SToBB provides a principled basis for diagnosing and resolving them.

These properties collectively illustrate how a SToBB offers a coherent foundation for explanation with an emphasis on usefulness throughout the life cycle of the black-box. They integrate the obligations of CE: empirical adequacy as the minimal requirement, acceptance through transparency and a commitment to continued adaptation, and pragmatic virtues as criteria that guide design. The next section describes and exemplifies the practical process by which a SToBB is constructed and maintained.

5 Process and Proof of Concept: Constructing a SToBB via Constructive Box Theoriser

Section 4 defined a SToBB through five components. We now describe a process by which a researcher may construct this artefact in practice and how it is maintained over the life cycle of a black-box. To make the process concrete, we exemplify each component immediately in a proof-of-concept for a small neural network trained on a tabular task. We introduce the CoBoT algorithm and use it as the algorithmic component to build a SToBB. Note that the formal description of CoBoT is in Appendix B, this section provides a high-level description as to not distract from the overall proof-of-concept.

The subsections are organised around the SToBB components *observation base*, *hypothesis class*, *algorithmic components*, and the resulting *adequate surrogate*. In each subsection, we first state what must be specified or created, and then illustrate how this is realised in the running example. Documentation aspects are discussed within the respective component subsections. In addition, we discuss operational requirements, as well as the optional diagnostic information and interfaces separately at the end of the section. We collect a list of potential questions the documentation should answer in Appendix A.

Our proof-of-concept builds around a black-box neural network trained on the three class version of the Abalone dataset (Nash et al., 1994). The network has a feed-forward architecture with four hidden layers, 32 neurons each, and ReLU activation functions. The final black-box validation accuracy is 0.64. The SToBB-artefact for our proof-of-concept comprises the information given after "**Example.**" in each subsection, together with supplementary material in Appendix B, the documented code base[2] and the therein contained answers to questions in Appendix A

5.1 Observation Base

The first step is to set up the *observation base*, i.e., the structures that will store observations as they are collected over the black-box's life cycle. To implement this storage, one must specify the observation *space*, i.e., the relevant black-box input- and output variables and, where applicable, auxiliary measures derived from the black-box.

The documentation should explain why each variable or measure is included, how it is obtained, and what quality criteria apply. This enables third parties to assess whether the SToBB rests on a relevant and reliable empirical basis. As the black-box is used, new observations are recorded in the observation base and must be incorporated into the SToBB, expanding its empirical base.

Example. The observation space consists of the input of the black-box (a 7 dimensional real-valued feature vector) and the corresponding class label predicted by the black-box. Further, each sample is associated with attribution scores computed with LIME (Ribeiro et al., 2016). The attribution scores act as auxiliary observables used to constrain the surrogate to subspaces marked as important to the black-box decision. The scores are obtained using the official LIME implementation[3], using default parameters of an exponential kernel with a bandwidth of 1.93 and sampling 5000 perturbations. An observation is rejected if all attribution scores are smaller than zero.

The observations are stored inside data fields of a "CoBoT class" object (see Section 5.5).

5.2 Hypothesis Class

The researcher needs to specify the hypothesis class of surrogate models, which determines the structures that operate on the observations. The documentation should describe how a surrogate processes an observation step by step, making the trace logic explicit, and specify conditions under which an observation is *not* covered. It should also record how user-centred criteria or other pragmatic virtues informed the design of the class and how these choices constrain its expressivity.

[2]Code available at anonymous.4open.science/r/cobot_tmlr-F583
[3]Version 0.2.0.1 from pypi.org/project/lime

Example. As the black-box operates on tabular data, we choose a rule based hypothesis class, which is a popular format for explanation (Lakkaraju et al., 2016). Rules are represented as axis-aligned bounding boxes defined over subspaces of the input domain.

Formally, let $d \in \mathbb{N}$ be the input dimensionality and let $[1..d] := \{1, \ldots, d\}$. We define the hypothesis class as a *set of boxsystems*

$$\mathcal{B} \;=\; \big\{\, (I, B) \;\big|\; I \subseteq [1..d],\; B = \{b_1, \ldots, b_m\} \,\big\},$$

where each pair (I, B) is called a *boxsystem*. Here, I denotes a fixed subspace of input dimensions and B is a finite set of *non-overlapping, axis-aligned, fully bounded boxes* b defined on that subspace. Each $b \in B$ is associated with a class label $c \in \mathrm{C}$.

The subspaces I are derived from a local explainer Φ via an indicator function $\mathcal{I}$. Given an input sample x, Φ yields an attribution vector $a \in \mathbb{R}^d$, and

$$I \;=\; \mathcal{I}\big(a\big) \subseteq [1..d]$$

denotes the set of dimensions deemed relevant for the corresponding black-box prediction. For each distinct feature set I encountered in the observation base, exactly one corresponding boxsystem (I, B) is maintained.

All observations whose local explanations induce the same subspace I are associated with the corresponding boxsystem and constrain its boxes. By construction, each observation is associated with exactly one subspace I and, within the corresponding boxsystem, with at most one box b.

Tracing logic of $\mathcal{B}$. Given a sample x, let

$$I_x \;=\; \mathcal{I}\big(\Phi(x)\big) \quad \text{and} \quad c_x \;=\; f(x).$$

Tracing succeeds if there exists a boxsystem $(I_x, B) \in \mathcal{B}$ and a box $b \in B$ such that

$$x \in b \quad \text{and} \quad \textsc{Label}(b) = c_x.$$

If no such box exists, tracing fails and an update is triggered (see *update mechanism* Section 5.3).

User-centric considerations. The size of I directly dictates the length or complexity of each rule, a factor that plays an important role in the user-friendliness of rules (Nauta et al., 2023; Ribeiro et al., 2018). To control this, we introduce a parameter $k \in \mathbb{N}$ that limits the number of selected dimensions.

For a given k, we define a constrained indicator function $\mathcal{I}^{k\uparrow}$ such that, for an attribution vector produced by Φ,

$$\mathcal{I}^{k\uparrow}(a) \;\subseteq\; \{\, i \in [1..d] \mid a_i > 0 \,\}, \qquad |\mathcal{I}^{k\uparrow}(a)| \leq k,$$

where $\mathcal{I}^{k\uparrow}(a)$ contains the indices of the *highest-scoring positive* components of a.

While this does limit the expressive power of the hypothesis class, we describe in the following subsection a principled way by which the algorithm chooses the smallest value for k that allows empirical adequacy, automating the trade-off between adequacy and user-friendliness.

5.3 Algorithmic Components

The algorithmic components define how a surrogate is *constructed* from a set of observations and how it is *updated* when adequacy fails. The description must make the procedure reproducible and should specify whether the algorithm can represent the full hypothesis class. The update procedure has to be described in equal detail, stating under what conditions updates are triggered, what operations may apply during updates, and how updates restore empirical adequacy.

Example. For the sake of clarity and simplicity, we focus on describing the general properties and guarantees of the algorithm. For a full technical description of CoBoT the reader is referred to Appendix B.1.

Construction: CoBoT incrementally builds the box system, refining it when new data points are added. The algorithm ensures that adequacy is always preserved while optimising for complexity. We initialise CoBoT

with $k = 3$ and an empty set $\mathcal{B} = \emptyset$. Processing the observations one at a time, CoBoT incrementally adds boxsystems to $\mathcal{B}$. For each boxsystem the algorithm guarantees that 1) none of the boxes overlap, 2) each processed observation is unambiguously associated with one box, and 3) all samples associated with one box have the same label. The $k = 3$ dimensions used are the three dimensions with the highest attribution scores greater than zero.

Update Mechanism: Revision is conducted in case inconsistencies arise. These can occur in three ways:

1. When processing the local explanation: The local explanation indicates a subspace that has not been encountered before. For this subspace, a boxsystem is created, within which a single box, containing the new observation, is placed.
2. When processing the feature values: A boxysstem for the subspace exists, but within the subspace the sample lies outside of any box. The algorithm attempts to expand the boundaries of existing boxes (of the correct label) to include the point. If an expansion is found that leads to no inconsistencies with other observations, the algorithm stops. If no possible expansion is found, the sample is placed inside a new box containing only itself.
3. Misclassification: The observation is mapped to a box with a different class label than predicted by the black-box. The erroneous box is then dissolved, and its contained samples are merged into existing boxes whenever possible; otherwise, new boxes are created.

To illustrate the update mechanism, Figure 2 (Appendix B.2) shows several updates of the box-system for $I = \{2, 6\}$ until its current state.

Adapting maximal rule-complexity parameter k: If an incoming observation shares the same important subspace and identical features values with a previous observations, yet is associated with a different label, CoBoT returns an error. In this case, the two observations are indistinguishable in the selected subspace. However, since the black-box assigns them different values, this indicates that either (i) the dimensionality of the subspace is too small (i.e., the pragmatic requirement was too strict), or (ii) the feature attribution method is not working as intended. To address case (i), CoBoT increases k by one and rebuilds the surrogate from scratch using all stored observations, while recording its previous state and the conflicting sample for later inspection. Case (ii) occurs if CoBoT fails to build an empirically adequate $\mathcal{B}$ despite using all important dimensions in an observation. In this case, the observation space is inherently ambiguous and has to be modified externally, e.g., by the maintainers of the SToBB that then need to rebuilt the observation base.

5.4 Adequate Surrogate

Once the observation space, hypothesis class, and algorithmic components are defined, the researcher instantiates the first surrogate. The surrogate must be empirically adequate with respect to all observations available at this stage. If adequacy fails, adjustments to the algorithm or the hypothesis class may are required before continuing. After deployment, the surrogate must be maintained: when new observations arrive that break adequacy, an update must proceed via the documented update procedure.

Example. In our example, CoBoT instantiates the first surrogate by incrementally constructing $\mathcal{B}$ from the available observations, starting from $\mathcal{B} = \emptyset$. Adequacy is maintained incrementally by the revision operations described before.

5.5 Operation, distribution and maintenance

To be useful to parties other than its original creators, a SToBB must also be documented in terms of how it is operated and distributed. This includes specifying how the artefact is stored and versioned, and how external stakeholders can access it, e.g., via APIs or dedicated tooling. The documentation should outline the technical preconditions for using and updating the SToBB (such as required software, model access, or computational resources) and the processes that govern who may perform updates and under which conditions. Together, these provisions support structured use and maintenance of the SToBB by third parties over the life cycle of the black-box.

Example. As mentioned previously, interaction with the SToBB centres around a "CoBoT class" object. The object not only stores the observation base (Section 5.1), but also implements the algorithmic components (Section 5.3), contains the current surrogate (Section 5.4) and the record of its evolution as well as further diagnostic information (Section 5.6 below). The documentation of the code base contains the necessary information to load and interact with the provided object. The code base can be found at anonymous.4open.science/r/cobot_tmlr-F583.

5.6 Optional: Diagnostic Information

After the adequate surrogate has been successfully instantiated, additional diagnostic information may be collected and documented, such as coverage statistics, structural complexity measures, or the number of updates applied. These measures do not test adequacy but complement the descriptive documentation with quantitative information.

Example. Each time CoBoT processes an update, we track internal statistics about outcome and state of the surrogate. We track the *total number of features sets*, the *total number of boxes* across all features sets, and the *total number of singletons* contained within those boxes. Figure 1 summarises the diagnostics over time for the current surrogate model. The current surrogate is based on 4 177 observations, of which 187 triggered updates (4.5%). Maximal subspace size remains as initialised, $k = 3$. The boxsystems of the surrogate occupy 16 distinct subspaces and contain a cumulative sum of 20 individual boxes. Two of those boxes are singletons, containing only a single sample. This yields a compression gain of $1-(20/4177) \approx 0.995$.

5.7 Optional: Interfaces

As discussed in Section 4.1, interfaces are not part of the SToBB itself but may be provided and documented to better support downstream use.

Example. We include example interfaces answering three common questions in XAI that can be attributed to the leading questions *Why*, *How to be that*, and *How*, respectively, as presented in Liao et al. (2020):

- **Local interface:** *Q: Why is this input given that prediction?* What features were important? Given an input, the interface returns a rule by extracting the box constraints relevant for that instance. This query can trigger an update on the surrogate. A formal listing is given in Algorithm 2 (Appendix B.3).
- **Contrastive interface:** *Q: Why is this input given this prediction and to what values would the important features need to change to lead to a different outcome?* For a given input, the interface extracts the applicable box and the closest box with a different label from the same boxsystem. This query can trigger an update on the surrogate. A formal listing can be found in Algorithm 3 (Appendix B.3).
- **Global interface:** *Q: What subspaces are considered important and how do they vary with class label and regions of the input space?* To answer this question, the interface first computes a 2D-UMAP projection of all input-coordinates in the observation basis. Each sample is then assigned to its corresponding box and coloured according to the respective subspace (important dimensions), while class labels are encoded by marker shape. This provides a visual account of how the importance of different dimensions changes over the data manifold. This interface cannot trigger updates. A listing of the procedure and its output are provided in Algorithm 4 and Figure 3, respectively (Appendix B.3).

6 Discussion and Conclusion

We introduced the notion of a *scientific theory of a black box* (SToBB), a structured and auditable artefact that consolidates explanatory information about a fixed machine-learning model. Building on CE, we identified three obligations that guide the design of such an artefact: empirical adequacy with respect to observed

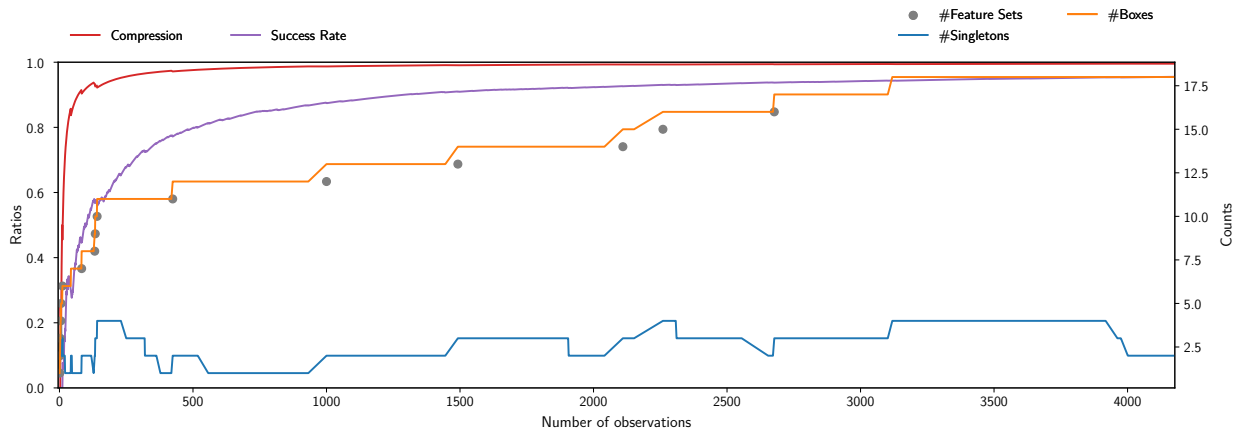

Figure 1: Diagnostic information on the current surrogate model. Left axis ("Ratios"): Compression is defined as the gain $1 - \frac{\#boxes}{\#samples}$; success rate is the fraction of samples that did *not* trigger an update. Right axis ("Counts"): #Feature Sets denotes the number of unique feature sets/distinct subspaces; #singletons and #boxes count the number of singleton and the total boxes, respectively. After processing all available 4 177 samples, the surrogate comprises less than 20 boxes. Across all observations, 187 updates were performed.

behaviour, adaptability to new observations, and transparency and auditability through documentation. We translated these obligations into a general process for constructing and maintaining a SToBB and exemplified through CoBoT how this process can be instantiated in practice.

Following existing XAI taxonomies (Speith, 2022), a SToBB can be described as a global, post-hoc artefact that centres a surrogate model. This characterisation, however, neither captures the temporal scope of the SToBB perspective nor the particular CE-imposed obligations the surrogate has to fulfil. More importantly, it misses what the SToBB perspective adds beyond a surrogate description: it makes explicit a maintained representational basis that supports a clear distinction between *explanatory information* and *explanation*. The explanatory information, represented by the observation base and the surrogate, can be designed and evaluated primarily with respect to empirical adequacy and traceability, independently of the interfaces that must take user-centred qualities into account. The resulting baseline guarantees give interface designers an explicit point of reference for making principled trade-offs between information richness and the users' need for comprehensible answers, a well-known pain point in XAI research (Jacovi & Goldberg, 2020; Doshi-Velez & Kim, 2017). As illustrated by the CoBoT example, a SToBB may still be constructed with user-centred criteria in mind, but whenever there is tension, empirical adequacy takes precedence.

The aim of this paper was the conceptual development of a SToBB. The accompanying proof-of-concept is necessarily limited in its scope: The CoBoT example targets one data-domain, incorporates a single auxiliary measure, and provides an online learner that covers construction and update in one procedure. Because the SToBB framework formulates only abstract, qualitative constraints, it leaves considerable freedom in how auxiliary measures, hypothesis classes, and algorithmic procedures are designed. These components provide several directions for future work. Two avenues seem particularly natural: (i) developing other hypothesis classes and algorithmic components directly derived from established interpretable model families, such as linear or logistic regression and prototype-based models (Molnar, 2020); and (ii) exploring existing explanation methods to assess whether and how they can be adapted into components of a SToBB. In our proof-of-concept, we used a local explanation method to extend the observation space with auxiliary signals of black-box behaviour; other methods could be adopted in a similar fashion. Conversely, for methods such as Anchors (Ribeiro et al., 2018) one could imagine SToBB variants whose hypothesis class consists of sets of local rules, with algorithmic components that generalise from individual anchors and update them as new observations arrive. Working out such adaptations requires rethinking these methods as parts of a cumulative and updateable process rather than as single-shot explainers. The suitability of a particular observation-space design, hypothesis class, or algorithmic component will depend on the black-box at hand

and the domain in which the AI system operates. Over time, SToBBs of different shapes can be explored and best practices established.

The hypothesis class of a SToBB naturally induces a language bias that favours some types of explanations over others. In the case of CoBoT, we used the black-box's application domain to deliberately target rule-based explanations, which motivated choosing axis-aligned boxes as the surrogate structure. As we also saw, this choice doesn't strictly limit interfaces to only produce this one type of explanation. Interfaces are free to transform the information contained in the SToBB as the context requires. The work of Naik & Turán may provide a useful starting point for interface design that complements the SToBB framework particularly well. Where our paper transfers van Fraassen's Constructive Empiricism to XAI, Naik & Turán (2020) transfer van Fraassen's *theory of explanation* to XAI that guides the formalisation of questions and answers.

Specifically designed to be empirically adequate and extensible, SToBBs address use-cases that differ from those of many existing XAI methods. Their focus on accumulating information during the life cycle of a black-box positions them primarily as tools for AI governance. At a minimum, a SToBB presupposes stable access to a fixed model instance whose input-output behaviour can be logged, the ability to store and process observations in compliance with applicable legal and organisational rules, and an interpretable hypothesis class and update procedure that relevant stakeholders can inspect. Using SToBBs in practice will require decisions about how to regulate access, updates, and versioning, e.g., who may read or modify a SToBB, how documentation integrity is ensured over time, and how privacy and regulatory constraints affect what observations can be stored and which parties may query which parts of the artefact. A SToBB provides a structured target for such access within the bounds set by the surrounding governance regime. It cannot compensate for missing access to model behaviour, cannot override constraints on data retention and sharing, and does not in itself guarantee that all parties obtain the level of access they might ideally want. The concrete answers to these governance questions will depend on the individual AI system and its application setting and thus lie beyond our discussion here, but they are of practical relevance to the adoption of SToBBs and will require exploration in the future.

In this work we treated the black-box as fixed: once trained, its input-output behaviour does not change, i.e., all collected observations remain valid for the lifetime of the SToBB. In practice, an AI system may make use of different black-box versions over its life cycle, e.g., through fine-tuning or re-training. In such cases, a SToBB would naturally accompany a single black-box instance; if the model is replaced, a new SToBB can be instantiated, potentially reusing parts of the existing observation base. For example, in our proof-of-concept, changes to the underlying neural network would affect predictions and local explanations, but the input samples themselves remain pointers to regions of interest in the input space; this information could be used to build a new observation base. We expect the conceptual structure of SToBBs to be flexible enough to support such scenarios, e.g., by integrating mechanisms in the algorithmic components, such as checks on when the observation base has become stale, or designing algorithms that allow efficient reconstruction under model change. Working out concrete strategies for non-stationary settings is left for future work.

To conclude, by separating the persistent representational foundation of explanation from the methods that present answers in context, the SToBB framework offers a new direction for XAI research. It complements existing XAI techniques and focuses on a coherent basis for life cycle-spanning transparency, reuse of explanatory information, and systematic external scrutiny. Future work needs to follow several directions: explore richer relational structures, extended auxiliary measures, and principled development of interfaces, as well as investigating the practical aspects of deployment and maintenance.

References

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012*, 2019.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and

challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*, 2017.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Kasia Chmielinski, Sarah Newman, Chris N Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, et al. The CLeAR Documentation Framework for AI Transparency. *Harvard Kennedy School Shorenstein Center Discussion Paper*, 2024.

Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.

European Commission, High-Level Expert Group on Artificial Intelligence EU HLEG AI. Ethics Guidelines for Trustworthy AI, 2019. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Parliament and European Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Technical report, 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

Raphael Fischer, Matthias Jakobs, Sascha Mücke, and Katharina Morik. A Unified Framework for Assessing Energy Efficiency of Machine Learning. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022*, pp. 39–54. Springer Nature Switzerland, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, November 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

IEEE. IEEE Standard for Transparency of Autonomous Systems. *IEEE Std 7001-2021*, pp. 1–54, 2022. doi: 10.1109/IEEESTD.2022.9726144.

International Organization for Standardization and International Electrotechnical Commission ISO/IEC. Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023). https://www.iso.org/standard/42001, 2023.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=jESY2WTZCe.

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939874. URL https://doi.org/10.1145/2939672.2939874.

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. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103473. URL https://www.sciencedirect.com/science/article/pii/S0004370221000242.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376590. URL https://doi.org/10.1145/3313831.3376590.

Q. Vera Liao, Milena Pribic, Jaesik Han, Sarah Miller, and Daby Sow. Question-driven design process for explainable ai user experiences. *CoRR*, abs/2104.03483, 2021. URL https://arxiv.org/abs/2104.03483.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

Melkamu Mersha, Khang Lam, Joseph Wood, Ali K. AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, September 2024. ISSN 09252312. doi: 10.1016/j.neucom.2024.128111. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231224008828.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

Harish Naik and György Turán. Explanation from specification. *arxiv:2012.07179*, 2020.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone [Dataset]. UCI ML Repository, 1994. URL 10.24432/C55C7W.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s), July 2023. ISSN 0360-0300. doi: 10.1145/3583558. URL https://doi.org/10.1145/3583558.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. doi: 10.6028/NIST.AI.100-1. URL https://doi.org/10.6028/NIST.AI.100-1.

OECD. Recommendation of the council on artificial intelligence, 2024. URL https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, et al. Guideline for Trustworthy Artificial Intelligence–AI Assessment Catalog. *arXiv preprint arXiv:2307.03681*, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Timo Speith. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2239–2250, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534639. URL https://doi.org/10.1145/3531146.3534639.

Eike Stadtländer, Tamás Horváth, and Stefan Wrobel. Learning weakly convex sets in metric spaces. *arXiv:2105.06251*, 2024.

Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems. *arXiv preprint arXiv:1806.07552*, 2018.

Bas. C. van Fraassen. *The Scientific Image*. Oxford University Press, 1980. ISBN 9780198244271. URL 10.1093/0198244274.001.0001.

A Example Documentation Questions

The goal of the documentation is to maximize transparency and informed use of third parties. The questions listed below may serve as a starting point:

Scientific Theory of a Black Box (SToBB) – Documentation Question Sheet

Observations

- What variables, auxiliary measures and target outputs constitute the observation space?
- How are auxiliary measures of the black-box behaviour obtained?
- What quality criteria apply to these measurements?
- How is the observation base stored?

Hypothesis Class

- What interpretable model family defines the surrogate's relational structure?
- How does the surrogate process an observation step by step?
- Under which conditions is an observation not covered by the surrogate?
- Which user-centred criteria or pragmatic virtues shaped the design?

Algorithmic Components

- How is the surrogate initially constructed from the observation base?
- How does the surrogate demonstrate empirical adequacy?
- How is the surrogate stored?
- What is the update policy when new observations appear?
- Does the algorithm represent the full hypothesis class?
- How are trade-offs or approximations documented?

Operational Requirements

- What resources or runtime requirements must be met to deploy and maintain the SToBB?
- Does access to parts of the SToBB have to be restricted for specific users or interfaces?
- How are updates to the SToBB performed?
- How are updates to the SToBB recorded?

Diagnostics (if any)

- Which diagnostic metrics are monitored during construction and updates?
- How frequently are diagnostics evaluated and recorded?
- How are diagnostic records stored and made accessible?

Interfaces (if available)

- What interfaces are implemented and what questions do they answer?
- Who is the target audience of the interface?
- What is the purpose, method, and assumption of each interface?
- How can external parties audit or replicate the analyses?

The CoBoT repository anonymous.4open.science/r/cobot_tmlr-F583 contains answers to these questions for our proof-of-concept.

B Extended documentation for the example SToBB

This section of the appendix contains a detailed description of the CoBoT algorithm, more detailed diagnostic information from the concrete example and a formal description of the interface functions.

B.1 Algorithmic Component CoBoT — Step-by-Step Description

Following the brief description of CoBoT in Section 5.3, we now give a detailed description of the algorithm, going through Algorithm 1 step by step.

Algorithm 1 Construction and Update procedure of Constructive Box Theoriser (CoBoT)

Input: Black-box model $f : \mathbb{R}^d \to \mathrm{C}$ for some positive integer d and finite set C, local explainer Φ for f, a binarization function $\mathcal{I}^{k\uparrow}$ for local explanation and k, input sample x, set of boxsystems $\mathcal{B}$, set of encountered observations $\mathcal{O}$

Output: Set of boxsystems $\mathcal{B}$

1: $c \leftarrow f(x)$
2: $a \leftarrow \Phi(f, x, c)$
3: $I_x \leftarrow \mathcal{I}(a)$ ▷ obtain set of important dimensions
4: $\mathcal{O} \leftarrow \mathcal{O} \cup \{(x, c, I_x)\}$
5: $B_{I_x} \leftarrow B$ **if** $\exists (I, B) \in \mathcal{B}$ with $I = I_x$, **else** None
6: **if** $B_{I_x} =$ None **then** ▷ None on first encounter of itemset
7: $\quad b_x \leftarrow \textsc{CreateSingleton}(x, c, I_x)$
8: $\quad \mathcal{B} \leftarrow \mathcal{B} \cup \{(I_x, \{b_x\})\}$
9: **else**
10: $\quad$ **if** $\exists b \in B_{I_x}$ with $x \in b$ **then**
11: $\quad\quad$ **if** $\textsc{Label}(b) \neq c$ **then**
12: $\quad\quad\quad B_{\mathcal{O}_b} \leftarrow \{\textsc{CreateSingleton}(o, c_o, I_o) \mid (o, c_o, I_o) \in \mathcal{O} \wedge I_o = I_x \wedge o \in b\}$
13: $\quad\quad\quad B'_{I_x} \leftarrow \textsc{Merge}((B_{I_x} \setminus \{b\}) \cup B_{\mathcal{O}_b})$
14: $\quad\quad\quad \mathcal{B} \leftarrow (\mathcal{B} \setminus \{(I_x, B_{I_x})\}) \cup \{(I_x, B'_{I_x})\}$
15: $\quad$ **else** ▷ no existing box contains sample
16: $\quad\quad b_x \leftarrow \textsc{CreateSingleton}(x, c, I_x)$
17: $\quad\quad B'_{I_x} \leftarrow \textsc{Merge}(B_{I_x} \cup \{b_x\})$
18: $\quad\quad \mathcal{B} \leftarrow (\mathcal{B} \setminus \{(I_x, B_{I_x})\}) \cup \{(I_x, B'_{I_x})\}$
19: **return** $\mathcal{B}$, $\mathcal{O}$

1) Initialization As input, CoBoT requires a black-box model f, an input sample x, a local explainer function Φ, an indicator function $\mathcal{I}^{k\uparrow}$, a list of observations $\mathcal{O}$, and a (possibly empty) set of boxsystems $\mathcal{B}$. The indicator function $\mathcal{I}^{k\uparrow}$ transforms a local explanation into a set of indices I by extracting the top k most important feature dimensions. In lines 1 to 3, CoBoT computes the class label c for the input sample x and the local explanation a, from which it derives the feature set I_x of most important dimensions. The observation triple (x, c, I_x) is then added to the set of observations $\mathcal{O}$ in line 4.

2) Creating a new boxsystem Line 6 begins with the consistency check and the updating logic. It checks whether the feature set I_x is present in the set of boxsystems. If not, line 7 creates a singleton-box b_x around x in the subspace determined by I_x, associates it with the corresponding class label, and adds the tuple $(I_x, \{b_x\})$ to $\mathcal{B}$. CoBoT then returns the updated set of boxsystems $\mathcal{B}$ and observations $\mathcal{O}$.

3) Resolving inconsistency In case I_x had been encountered before, CoBoT proceeds in line 10 and checks if there exists a box b in the corresponding boxsystem B_{I_x} that covers x. If so, line 11 checks if the box's label matches the black-box output c. If the labels match, the algorithm returns. Otherwise, the boxsystem is updated: Line 12 selects all samples from the current list of observations that lie within the affected box and places each into its own singleton box. In the next step (line 13), the inconsistent box b is removed from B_{I_x} and all singleton-boxes in $B_{\mathcal{O}_b}$ are attempted to be *merged with* the remaining boxes in B_{I_x}. To this end, we use an adaptation of the algorithm presented in Stadtländer et al. (2024). Given a set of boxes,

the algorithm greedily selects the closest two boxes having the same class, and attempts to join them. If a join leads to an overlap with any other box, the join operation is reverted. This process continues until no further consistent mergers are possible. While the original algorithm by Stadtländer et al. was developed for binary classifications, we adapt it mutatis mutandis to multi-class problems. The algorithm has several desirable properties for our scenario:

i It naturally guarantees that *all and only indicated* dimensions in I_x are used. This guarantees full control over rule complexity.

ii The algorithm guarantees to return a boxsystem that is *fully* consistent with the training data, thereby fulfilling empirical adequacy.

iii Boxes are unambiguous: None of the computed boxes overlap, meaning that 1) each sample induces the construction of at most one box and 2) no sample can fall into two boxes at the same time.

iv All boxes are bounded on all sides and do not extent beyond values observed in the input data. Thus, each box is strictly tied to observations.

If boxes from two different boxsystems, say B_{I_1} and B_{I_2}, overlap, this is of no concern, because Φ indicated that f used different sets of features for its decision. Continuing on line 15, the set of boxsystems is updated with the adapted boxsystem and CoBoT returns.

4) Expanding coverage The final *else* case is entered if a boxsystem does exist for I_x, but none of the boxes in B_{I_x} cover the sample. In that case line 16 creates a new singleton box around x and line 17 performs the merge operation described above in an attempt to extend any existing consistent box by the singleton-box. The set of boxsystems is updated and CoBoT returns.

B.2 Further Diagnostic Information

This further complements the diagnostic information shown in the curves in Figure 1, Section 5.6.

Increments of maximal subspace size CoBoT was initialized with $k = 3$, this has not been increased.

Evolution of a box-system Despite the current subspace size being $k = 3$, the surrogate uses subspaces with less than 3 dimensions. Figure 2 shows the subspace of the boxsystem for feature set $I = \{2, 6\}$ at different points in time. Two classes are present, indicated by the two different box colors. Samples of individual observations are visualized as dots. The scatterplot indicates that the features are linearly correlated within the subspace. Of the 187 total updates to the surrogate CoBoT performed, 27 were performed on this boxsystem.

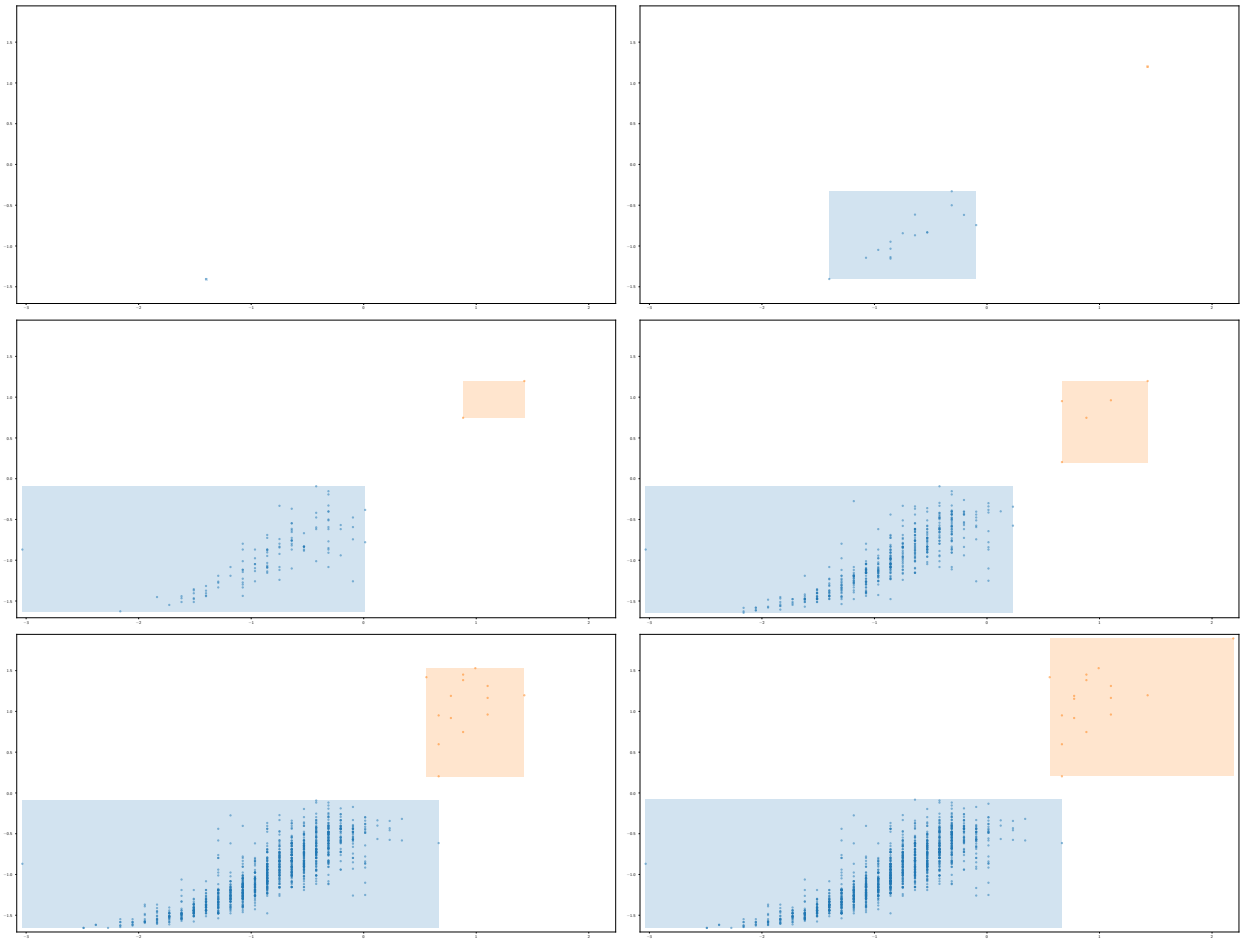

Figure 2: Evolution of the axis-aligned bounding boxes contained in the boxsystem for subapace $I = \{2, 6\}$. Each individual image shows the same boxystem after different updates. Updates number [1, 7, 16, 19, 25, 27 (final)]. Colors encode classes.

B.3 Interfaces

We provide interfaces addressing categories of questions listed in the XAI question bank (Liao et al., 2020). By design, the surrogate's hypothesis class naturally supports extracting rule based explanations. Algorithm 2 and Algorithm 3 extract a local and a contrastive explanation, respectively. Algorithm 4 gives a global view.

Algorithm 2 answers the question "Given x, why did f predict c?" by returning the applicable box. By instead returning the supporting samples of the relevant box, a different interface could give an "explanation by example", answering the question "What kind of instance gets the same prediction" . Algorithm 3 answers the question "Given x, why did f predict c and what other class is the sample most similar to?", supplying in its output also a neighboring box of a different class. It could be adapted to give a targeted answer "Why c and not c' ?" by restricting the search in line 9 to b associated with c'. In case the request from any of the interfaces leads to a failure condition (*itemset unknown, not covered, wrong label*), CoBoT automatically updates the surrogate.

Algorithm 2 LOCAL Explanation

Input: $\mathcal{B}$, I_x, c, x

1: $B_{I_x} \leftarrow B : ((I, B) \in \mathcal{B} \wedge I = I_x)$ **else** None
2: **if** $B_{I_x} \neq$ None **then**
3: $\quad b_x \leftarrow b : b \in B_{I_x} : x \in b$ **else** None
4: $\quad$ **if** $b_x \neq$ None $\wedge$ LABEL$(b_x) = c$ **then**:
5: $\quad\quad$ **return** b_x
6: UPDATESURROGATE(x) ▷ Call CoBoT's update mechanism to integrate x in $\mathcal{B}$
7: LOCAL$(I_x, c, x, \mathcal{B})$

Algorithm 3 CONTRASTIVE Explanation

Input: $\mathcal{B}$, I_x, c, x, a distance function $d(.,.)$ between two boxes

1: $B_{I_x} \leftarrow B : ((I, B) \in \mathcal{B} \wedge I = I_x)$ **else** None
2: **if** $B_{I_x} \neq$ None **then**
3: $\quad b_x \leftarrow b : b \in B_{I_x} : x \in b$ **else** None
4: $\quad$ **if** $b_x \neq$ None $\wedge$ LABEL$(b_x) = c$ **then**:
5: $\quad\quad b_{\neg c} \leftarrow \underset{b' \in B_{I_x} \wedge \text{LABEL}(b) \neq c}{\textbf{argmin}} d(b_x, b')$
6: $\quad\quad$ **return** $(b_x, b_{\neg c})$
7: UPDATESURROGATE(x)
8: CONTRASTIVE$(\mathcal{B}, I_x, c, x, d)$

To obtain a global view on $\mathcal{B}$, we define Algorithm 4. It computes a 2d UMAP embedding of all observations, color-coded by the feature set they were mapped to and the marker shape indicating the class label predicted by f. Figure 3 shows the interface outputs for the current set of the SToBB. The color imbalance highlights that the feature sets have vastly different support. Feature set $\{2, 6\}$ (light pink) occupies most of the right. Despite the comparatively large support of the feature set, the number of updates it triggered was below 15% of all updates (27/187). The left side of the plot is occupied by more feature sets, brown, blue and green. Purple, lavender, cyan and orange have very little support and are confined to very small areas. Colours form perceptibly connected regions. Conspicuously, many feature sets seem to use only a single type of marker, indicating that classes are separable not only by value but by feature alone. More detailed statistics could be delivered by another interface.

Algorithm 4 Global Visualization

Input: Set of observations $\mathcal{O} = \{(x, c_x, I_x)\}$, a 2D projection operator UMAP$(\cdot)$
Output: 2D visualization of projected samples colored by I_x and shaped by c_x

$X \leftarrow \{x \mid (x, c_x, I_x) \in \mathcal{O}\}$
$(u_x, v_x)_{x \in X} \leftarrow$ UMAP(X)
Visualization $\leftarrow$ EMPTYPLOT
for each $(x, c_x, I_x) \in \mathcal{O}$ **do**
$\quad$ color$_x$ $\leftarrow$SUBSPACETOCOLOR(I_x) ▷ color encodes important dimensions
$\quad$ marker$_x$ $\leftarrow$CLASSTOMARKER(c_x) ▷ marker encodes class label
$\quad$ Visualization.ADDSCATTER$((u_x, v_x)$, color$_x$, marker$_x$)
return Visualization

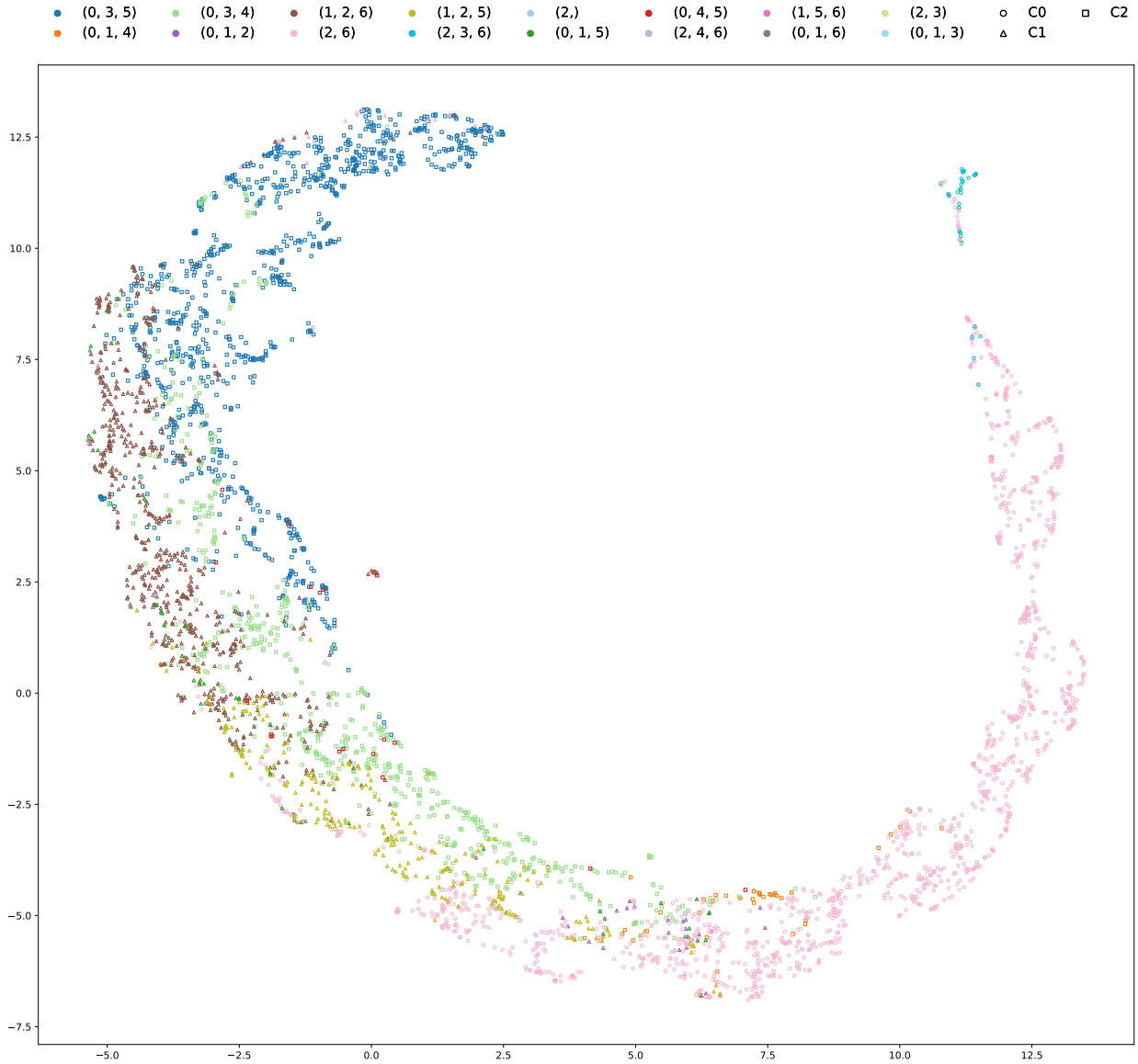

Figure 3: 2d UMAP embedding of the 4 177 observed samples in the CoBoT-SToBB. Colors indicate feature sets from $\mathcal{I}^{3\uparrow}$, marker style indicates class label.