IMPERIAL

Causal Discovery for Trustworthy Artificial Intelligence

Fabrizio Russo

December 2024

Department of Computing
Imperial College London

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Safe and Trusted Artificial Intelligence
of Imperial College London

Abstract

Artificial Intelligence (AI) is increasingly deployed in high-stakes domains such as healthcare, finance, and public policy, where decisions have far-reaching societal consequences. Ensuring trust in these systems requires moving beyond opaque, pattern-driven predictions to methods that are both effective and interpretable, while aligning with expert knowledge. Causal discovery, which uncovers cause-and-effect relationships from data and represents them as graphs, plays a central role in achieving these goals as a foundation for trustworthy AI.

This thesis introduces three novel methodologies to advance causal discovery and enhance the robustness, transparency, and accountability of AI systems. First, we propose **Contestable Neural Networks** (NNs), a framework that empowers domain experts to contest NNs, by validating and refining their mechanisms, using causal graphs. By exposing the causal relationships learned by the NN, this approach unlocks transparency and aligns model outputs with expert knowledge, enhancing robustness and accountability.

However, as systems grow complex, placing the entire burden on experts becomes impractical. To address this, we develop the **Shapley-PC algorithm**, which enhances constraint-based causal discovery by improving the identification of v-structures. By integrating Shapley values, a game-theoretic measure of variable contribution, the algorithm mitigates errors caused by noisy or limited data, ensuring robust causal graph recovery while maintaining theoretical guarantees.

Building on this foundation, we introduce **Causal ABA**, an Assumption-Based Argumentation framework that resolves conflicts between statistical evidence and expert knowledge. While Shapley-PC improves robustness in finite-sample settings, Causal ABA ensures stronger consistency guarantees between causal discovery outputs and input data. By combining structured reasoning with transparent conflict resolution, it unifies diverse information sources to produce trustworthy causal graphs.

Together, these contributions firmly position **causal discovery** as a cornerstone for **trustworthy AI**. By aligning data-driven insights with expert input, improving reliability, and ensuring transparency, this work lays a solid foundation for AI decision-making in high-stakes domains.

Statement of Originality

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Fabrizio Russo (2024)

Copyright Declaration

Acknowledgements

I found an old note in a notebook from 2013, written when I had just arrived in the UK. It said I would pursue a PhD after completing the MSc I was working toward at the time. Back then, I didn't fully understand what a PhD entailed, but, somehow, it felt like the right thing to do. Years later, in 2019, I met my wonderful supervisor, Francesca Toni, who was giving a keynote at a conference. After her talk, I approached her, and in 2020, I began the PhD journey that ends with this thesis.

I am profoundly grateful to Francesca for the opportunity. I thank her for granting me the freedom to explore and the trust that I would find my way, while offering invaluable support and guidance. Pursuing a PhD has been in my plans for some time, but getting here was made possible by many incredible people who helped me build the skills and confidence I needed to embark on this journey.

Firstly, I want to thank Ant for giving me my first "real" job, my first manager Garry for trusting me to learn SAS and automate the repetitive tasks I did not want to do, and Sunit for supporting me in that endeavour. At 4most, I am grateful to Damien, Craig, and Chris for their mentoring, support, and for the opportunity to keep learning. I owe special thanks to Mark for his inspiring leadership, his generosity in funding my MSc, and for introducing me to Pearl's work. To Torgunn, thank you for your friendship and mentoring, and for making me a researcher before I even realised it was what I wanted to be. To all my colleagues at 4most, thank you for making those years so meaningful.

I am also deeply thankful to the STAI CDT for funding this work and for the supportive community it provided, especially during the challenging first year of lockdowns and the pandemic. My friends and colleagues in the CDT and at Imperial have been a constant source of encouragement. I am particularly grateful to CLArg for our weekly meetings, which gave me a sense of belonging, and especially to Francesco, for his ability to create groups, and margaritas. To my first co-authors Antonio, Emanuele, and Pietro—thank you for showing me how research can be fun teamwork. And to Anna, for helping bring my ideas to life, improve them through genuine challenge and rigour, and for the joy we found in the process. Finally, to my examiners, Ben Glocker and Peter Flach, for their insightful comments and for making me feel like my work was worth examining.

Of course, none of this would have been possible without my amazing family: Babbo, Ma', and Peppe. You have always supported me unconditionally, making me feel like I could achieve anything, knowing you would always be there for me. To my Bifolchi and Asco brothers and sisters, and my London family—JD, Ruben, and Michelle—thank you for being my steadfast anchors through thick and thin. Finally, to Ireri: thank you for your love, your patience, and your unwavering support through countless late nights and weekends. Te quiero mucho.

To everyone who has been part of this journey, thank you from the bottom of my heart. I could not have done this without you.

To Family, Friends, Fun, and Filosofia.

Our brains store an incredible amount of causal knowledge which, supplemented by data, we could harness to answer some of the most pressing questions of our time.

– Judea Pearl, 2018

Contents

List of Tables

List of Figures

Notational Conventions

Abbreviations

ABA Assumption-Based Argumentation

ABAF ABA Framework

ADS Automated Decision System

AF Argumentation Framework

AI Artificial Intelligence

ANM Additive Noise Model

ASP Answer Set Programming

AUC Area Under the Curve

BIC Bayesian Information Criterion

BN Bayesian Network

CASTLE Causal Structure Learning

CIT Conditional Independence Test

CPDAG Completed Partially DAG

DAG Directed Acyclic Graph

DGP Data Generating Process

ER Erdös-Rényi

FCI Fast Causal Inference

FCM Functional Causal Model

FFNN Feed-Forward Neural Network

FGS Fast Greedy Search

GES Greedy Equivalence Search

HELOC Home Equity Line of Credit

HITL Human-in-the-loop

LP Logic Programming

MEC Markov Equivalence Class

ML Machine Learning

MSE Mean Squared Error

NN Neural Network

PAG Partial Ancestral Graph

PDAG Partially Oriented DAG

SEM Structural Equation Model

SetAF AF with Collective Attacks

SetAFs AFs with Collective Attacks

SF Scale-Free

SGD Stochastic Gradient Descent

SHD Structural Hamming Distance

SID Structural Interventional Distance

SIV Shapley Independence Value

SME Subject-Matter Expert

std Standard Deviation

UT Unshielded Triple

XAI Explainable AI

List of Symbols

$\perp\!\!\!\perp_{\mathcal{G}}$ d-Separated

$\perp\!\!\!\perp$ Independent

Θ Set of Neural Network Weights

$\mathbb{P}$ Probability Distribution

$\mathbb{R}$ Set of Real Numbers

$\mathbb{Z}$ Set of Integer Numbers

$\mathbf{E}$ Set of edges in a DAG

$\mathbf{S}$ Set of conditioning variables

$\mathbf{V}$ Set of nodes in a DAG

$\mathbf{W}$ Weighted Adjacency matrix

$\mathcal{A}$ Set of Assumptions

$\mathcal{C}$ Skeleton of a graph

$\mathcal{D}$ Dataset

$\mathcal{G}$ Directed Acyclic Graph

$\mathcal{H}_0$ Null Hypothesis

$\mathcal{H}_1$ Alternative Hypothesis

$\mathcal{L}$ Language

$\mathcal{R}$ Set of Rules

adj Adjacency Set

pa Parent Set

$\not\perp\!\!\!\perp_{\mathcal{G}}$ d-Connected

$\not\perp\!\!\!\perp$ Dependent

Chapter 1

Introduction

The transformative impact of Artificial Intelligence (AI) on modern society is undeniable. From reshaping everyday life to revolutionising industries, AI has become integral to critical sectors such as healthcare, finance, and public policy (Bhatnagar and Gajjar, 2024; UK Government, 2024; West and Allen, 2018). Its potential to improve decision-making and optimise processes has become evident (Bryant, 2023). However, as AI systems increasingly influence high-stakes decisions, pressing questions arise regarding trust, accountability, and the broader consequences of automated decisions (Bernstein and Raman, 2015; Brynjolfsson and McAfee, 2017; OECD, 2021).

Central to AI's success is its ability to strengthen decision-making processes (Bhatnagar and Gajjar, 2024; Meissner and Narita, 2023). By processing vast amounts of data quickly and accurately, AI systems generate insights and recommendations that enhance efficiency (Cavalcanti et al., 2024), reduce human error (Gursel et al., 2023; Shademan et al., 2016; Yalçın et al., 2023), and uncover hidden patterns (Davies et al., 2021; Jumper et al., 2021; Kovalevskiy et al., 2024). These capabilities have driven their adoption in critical domains, such as medical diagnostics, consumer finance, and legal judgments, where decisions often have far-reaching consequences.

To ensure that these benefits are realised without compromising ethical and societal values, **Trustworthy AI** has emerged as a guiding framework. Attributes such as transparency, robustness, accountability, and fairness are essential to ensure AI's responsible and reliable deployment (High-Level Expert Group on AI, 2019; National Institute of Standards and Technology, 2023). Transparency ensures that AI systems are understandable and their decisions traceable, while robustness guarantees reliable operation under diverse conditions. Accountability enables the individuals or organisations deploying these systems to be audited and held responsible for their actions and decisions. Finally,

fairness ensures that AI systems operate equitably and respect user rights. These principles are vital for building public trust and managing risks as AI becomes increasingly embedded in critical decision-making processes (European Commission, 2021b).

International efforts reflect a growing consensus on the importance of these principles. The European Union (EU) has spearheaded initiatives such as the *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on AI, 2019) and the *AI Act* (European Parliament, 2024), emphasising key areas such as **Human Agency and Oversight**, **Accountability**, **Transparency**, **Technical Robustness and Safety**, and **Diversity, Non-Discrimination, and Fairness** to align AI development with societal values (European Commission, 2021a,b).

Similarly, in the United States (US), the *AI Risk Management Framework* by NIST (National Institute of Standards and Technology, 2023), outlines principles including **Accountability and Transparency**, **Explainability and Interpretability**, **Validity and Reliability**, and **Fairness with Managed Bias**. The EU and US frameworks share significant overlap, underscoring a global commitment to ensuring AI systems are **reliable**, **transparent**, **accountable**, and aligned with societal values.

Across these initiatives, a shared understanding emerges of robustness, transparency, and accountability as foundational pillars of trustworthy AI. Collaborative efforts by organisations such as the OECD (OECD, 2024a,b) highlight the global resolve to align AI systems with ethical and societal expectations. By adhering to these principles, trustworthy AI empowers stakeholders—including consumers, governments, and domain experts—to evaluate and engage with these systems effectively, enabling their ethical integration into society.

This thesis builds upon the foundational principles of trustworthy AI by addressing critical challenges outlined in the EU and US frameworks. It focuses on three key attributes: **robustness**, **transparency**, and **accountability**. These attributes collectively ensure that AI systems are reliable, open to scrutiny, and aligned with societal values, fostering public trust and enabling their deployment in high-stakes domains such as healthcare, finance, and policymaking. Transparency, in particular, serves as a foundation for other attributes, as it allows stakeholders to evaluate and understand AI systems, thereby enabling accountability and reinforcing robustness.

A central concept in this work is **contestability**, which empowers stakeholders to challenge and refine AI decisions. Rooted in global AI ethics frameworks, such as those from the OECD (OECD, 2024c) and ACM (Baeza-Yates and Matthews, 2022), and supported by regulations like GDPR (European Union, 2016), contestability is especially crucial in high-stakes contexts such as credit scoring

and parole decisions, where errors or biases can have profound societal impacts. By creating mechanisms for oversight and redress, contestability allows users—whether decision-makers, regulators, or domain experts—to interact with and influence AI systems. This alignment with human expectations, ethical norms, and domain requirements directly promotes accountability and strengthens public confidence (Hicks, 2022).

Contestability fundamentally depends on transparency, as stakeholders can only meaningfully engage with decisions they fully understand (Hicks, 2023; Yurrita et al., 2023b). The synergy between these attributes also advances **fairness** (Yurrita et al., 2023a). Transparency reveals the inner workings of AI systems, while contestability empowers stakeholders to address biases and inequities. Together, these mechanisms empower stakeholders to ensure ethical outcomes, reinforcing trustworthiness in high-stakes applications.

This thesis identifies **causal discovery** as a critical enabler of transparency, accountability, and robustness. By uncovering relationships between variables and presenting them as interpretable causal structures, causal discovery makes complex AI decisions understandable, particularly for domain experts. Rooted in causal reasoning, it enhances robustness by ensuring reliability under noisy or changing conditions. Importantly, causal discovery provides a clear and actionable foundation for contestability, allowing stakeholders to validate, challenge, and refine model behaviour.

The next section builds on these foundational principles to explore how AI systems enhance decision-making processes. It examines their dual role in analysing data and designing interventions, highlighting their transformative potential across diverse domains.

1.1 AI for Decision-Making

AI has transformed decision-making processes across diverse domains, significantly enhancing the ability to analyse and act on complex data (Bryant, 2023; Meissner and Narita, 2023). By processing vast datasets efficiently, AI can unearth patterns and insights that were previously inaccessible. However, these systems fulfil two distinct functions in decision-making: **predictions** and **causal inference**.

Prediction involves estimating the likelihood of future events based on historical data, the primary domain of Machine Learning (ML), which excels at identifying patterns and trends to forecast outcomes. In contrast, causal inference seeks to uncover the underlying causes of observed phenomena, guiding decisions about effective interventions. While prediction anticipates trends, inference provides the insights necessary to design strategies that actively shape outcomes (Pearl and Mackenzie, 2018).

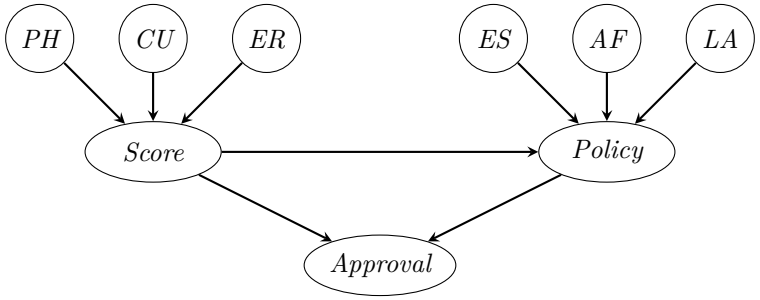
This distinction is captured by Pearl and Mackenzie's *Ladder of Causation*, which categorises reasoning into three levels. At the first rung are associations, addressing questions like "what if I see...?" The second rung focuses on interventions, asking "what if I do...?" The third and highest level involves counterfactuals, exploring questions of the likes of "what if I had done...?" Prediction aligns closely with association, while causal inference operates at the levels of interventions (*doing*) and counterfactuals (*imagining*). This progression—from recognising patterns to understanding mechanisms—highlights the complementary roles of prediction and causal reasoning.

A practical example of prediction in high-stakes decision-making is credit scoring. Financial institutions use ML models to evaluate the creditworthiness of loan applicants, analysing factors like credit history and financial stability to estimate the likelihood of default. These models automate decision-making, enabling faster and more accurate credit assessments and expanding access to credit. However, while ML excels at predicting likelihood of default, it is limited to identifying correlations and patterns. For instance, an ML model can rank applicants by risk but cannot determine which loans to approve to minimise overall default rates. Questions like "what if I do...give this customer a loan?" require causal inference, which reveals the mechanisms driving observed outcomes and informs actionable strategies (Pearl and Mackenzie, 2018). To illustrate, consider the following example, which highlights the need for causal reasoning in lending decisions.

Example 1.1.1 (Berkson's Paradox in Lending Decisions)**.** *A bank seeks to redevelop its score model to better predict the likelihood of loan approval and target creditworthy customers more effectively. Loan approval is determined by two main factors: the applicant's credit score and the bank's lending policy.*

The score aggregates financial metrics such as payment history (PH), time on electoral roll (ER), and credit utilisation (CU), which aim to assess creditworthiness. Meanwhile, the policy evaluates contextual factors like employment status (ES), affordability (AF), and loan amount (LA), ensuring alignment with the bank's risk tolerance and strategic goals.

These relationships are depicted in the causal graph below:



In this context, the score is calculated using an ML model trained on data from approved loans, as this is the only subset where complete information is available. However, this training process inherently conditions on the approval outcome, which is determined by the policy. This introduces a significant challenge: the approval process acts as a collider, creating a spurious correlation between the score and the variables influencing the policy. This phenomenon, known as selection bias or Berkson's paradox (Berkson, 1946; Peters et al., 2017), can distort the relationships learned by the model.

For example, within the approved population, the bank might observe a negative correlation between the score and employment status (ES). This misleading relationship could suggest that applicants with unstable ES are more likely to be approved, even though this bias arises solely from the conditioning effect. Relying on such flawed correlations could lead the bank to adjust its lending criteria in ways that perpetuate biases and unfairly disadvantage certain groups of applicants.

To address this selection bias, financial institutions often train an auxiliary model, with fewer data, where the approval outcome is the target variable. This model is trained on applicant data that includes both approved and rejected applications (Baesens et al., 2016), allowing it to capture the influence of the policy. The auxiliary model then predicts the likelihood of approval for the entire applicant population. These predictions are used to adjust the score, ensuring it reflects the applicant's financial stability independently of the selection process.

Integrating causal inference provides a more principled approach to understanding and mitigating such biases. By leveraging causal discovery, the bank can identify the true relationships between variables and disentangle spurious correlations introduced by the selection process. This causal insight enables the bank to refine both its score model and lending policies, ensuring they better represent the general applicant population. Such adjustments not only improve the fairness of lending decisions but also enhance their effectiveness.

The loan approval example highlights a broader challenge in data-driven decision-making: biases inherent in the data itself. Selection bias, as illustrated in the example, is one of many forms of bias that can distort the relationships learned by ML models, leading to spurious correlations and flawed predictions. In this case, conditioning on a collider—loan approval—introduced a false correlation between policy factors and credit scores. More generally, historical data often embeds systemic biases or reflects the policies and processes that shaped its generation. These limitations underscore the need for approaches that move beyond surface-level pattern recognition to uncover the causal mechanisms driving outcomes.

This thesis takes a dual approach to address the challenges of bias and reliability in decision-making systems. It focuses on enabling transparency and accountability in predictive models while facilitating causal reasoning through the discovery of causal graphs. These graphs reveal the underlying mechanisms that drive observed outcomes, feeding into causal models and guiding the refinement of decision-making processes. By tackling these challenges, this work provides a foundation for developing AI systems that align more closely with ethical and societal values, particularly in high-stakes applications such as credit scoring.

The interplay between ML and causal reasoning, as demonstrated in the earlier discussion and the loan approval example, underscores the complementary strengths and limitations of these paradigms. To fully understand how these approaches intersect and where gaps remain, the following section examines their respective capabilities and challenges in greater depth. This analysis sets the stage for the thesis's focus on advancing causal discovery as a unifying framework for trustworthy AI systems.

1.1.1 Machine Learning and Causal Models

ML has transformed industries such as healthcare and finance by leveraging data to identify patterns and make accurate predictions. For example, diagnostic tools powered by ML achieve expert-level performance in detecting diseases from imaging data (Esteva et al., 2017), while fraud detection systems monitor millions of transactions in real time to flag anomalies (Le et al., 2024). The scalability and efficiency of ML make it invaluable for tasks that require rapid and precise pattern recognition in data-rich environments. However, ML models primarily focus on associations, limiting their ability to address questions on the second and third rung of the Ladder of Causality.

Causal models, in contrast, aim to uncover the mechanisms driving observed phenomena, enabling actionable insights. These models are indispensable in contexts where understanding root causes is essential for designing effective interventions. For instance, causal reasoning can optimise public health campaigns by identifying disease drivers (Lewis and McCormick, 2012) or inform environmental policies by evaluating human impacts on climate (Stott et al., 2004). By focusing on causation rather than correlation, causal models enhance interpretability and robustness, particularly in high-stakes settings where actionable insights are essential (Peters et al., 2017). However, causal models often struggle with scalability and big data, where ML excels.

We compare the strengths and limitations of ML and causal models in Table 1.1, highlighting their distinct attributes and implications for trustworthy AI systems.

Causal models stand out for their high transparency, achieved through explicit representations

Table 1.1: Comparison of Machine Learning and Causal Models

Attribute	Machine Learning	Causal Models
Interpretability	Limited	High
Predictive Accuracy	High	Moderate
Generalisation	Moderate	High
Actionable Insights	Limited	High
Data Requirements	High	Moderate
Scalability	High	Low
Expert Knowledge	Moderate	High

of cause-and-effect relationships. This clarity supports interpretability by enabling stakeholders to scrutinise decision-making processes, fostering accountability in critical domains. In contrast, ML systems are often opaque, which complicates efforts to ensure transparency and restricts stakeholder engagement in validating or contesting decisions. These limitations highlight the importance of balancing ML's efficiency with the interpretability and actionability offered by causal reasoning.

Predictive accuracy is a key strength of ML, driven by its ability to process large datasets and identify fine-grained patterns (Goodfellow et al., 2015). However, its focus on correlations often limits its ability to provide actionable insights essential for accountability. Causal models, while less focused on prediction, excel in explaining the mechanisms underlying observed outcomes, enabling targeted interventions that align with desired goals. This explanatory capability bolsters robustness, as causal models generalise effectively under distributional shifts, which drive changes in correlations, by relying on stable relationships–the causal mechanisms. In contrast, ML systems frequently encounter challenges such as overfitting and degraded performance in dynamic environments (Magliacane et al., 2018).

Scalability is another area where ML demonstrates significant advantages, allowing it to handle large-scale, automated applications with ease. However, this scalability comes with high data requirements, which can limit its relevance in resource-constrained settings. Causal models are more efficient with data and integrate domain expertise to explicitly define relationships, further supporting transparency and accountability. At the same time, causal models require expert knowledge, which often limits their scalability. While ML incorporates expertise indirectly through feature engineering, this approach often lacks the directness and clarity provided by causal reasoning.

The complementary strengths of ML and causal models suggest that integrating these paradigms can address many of the challenges associated with trustworthy AI. For example, in healthcare, ML

can efficiently process large patient datasets to identify risk factors, while causal models provide explanations for these risks and guide interventions (Morris et al., 2019). This synergy is particularly critical in high-stakes domains where both predictive power and interpretability are essential. By combining ML's scalability and accuracy with the transparency and robustness of causal models, it becomes possible to design AI systems capable of more reliable and ethically aligned decision-making.

However, leveraging this synergy requires addressing persistent challenges that stem from the limitations of each paradigm. Issues such as opacity in ML, difficulties with noisy data, and the integration of domain expertise remain barriers to creating systems that meet the demands of high-stakes applications. These challenges reflect trade-offs between scalability, interpretability, and the ability to incorporate domain-specific knowledge, which must be carefully balanced.

1.1.2 Challenges and Opportunities

While ML and causal models have complementary strengths, their inherent limitations can hinder broader applicability in decision-making contexts. Integrating the strengths of both paradigms while mitigating their weaknesses is critical for developing systems that are both accurate and interpretable, as well as adaptable to complex, real-world conditions.

Scalability ML models are scalable because they only require more data to improve their performance. However, this scalability comes at the cost of interpretability, as the models become increasingly complex and opaque (Assis et al., 2025). Causal models, by contrast, are less scalable due to their reliance on expert knowledge, not always available or easy to elicit and codify at scale. This limitation can hinder their applicability when the problem is complex enough that expert knowledge is insufficient to capture all relevant causal relationships (Sgaier et al., 2020).

Opacity and Lack of Interpretability The opacity of ML models, particularly Neural Networks (NNs), undermines stakeholder trust by limiting transparency and accountability (Rudin, 2019). While efforts can improve interpretability, they often fall short of the explicit clarity provided by causal models. Even causal models, despite their structured nature, can face difficulties when attempting to represent complex systems (Rago et al., 2023). However, their structure remains accessible.

Handling Noisy and Conflicting Information Real-world data is frequently noisy and inconsistent, posing challenges for both paradigms. ML models risk overfitting to noisy patterns, leading to reduced generalisation (Szegedy et al., 2014). Causal models are instead more robust. However, the discovery

of causal assumptions can be challenging, particularly when data is noisy and statistical evidence is inconsistent (Tsagris, 2019). Reconciling these conflicts remains essential for robust decision-making.

Incorporating Domain Expertise Driving accountability, the integration of expert knowledge is a critical requirement for trustworthy AI. While causal models explicitly leverage domain expertise, they can encounter conflicts between data-driven discoveries and stakeholder insights (Alrajeh et al., 2020). ML, conversely, often lacks mechanisms to directly expertise, relying instead on feature engineering or other inductive biases (von Rüden et al., 2023). These gaps hinder contestability and limit the alignment of AI systems with ethical and societal values.

Addressing these limitations is crucial for advancing trustworthy AI. By enhancing interpretability, managing data uncertainty, and integrating domain expertise more effectively, it is possible to create systems that are both scalable and reliable. The following sections outline the strategies and contributions proposed in this thesis to tackle these challenges.

1.2 Thesis Objectives

Both ML and causal models face challenges in high-stakes decision-making contexts, such as a lack of interpretability, difficulties with noisy and inconsistent data, and limited mechanisms for integrating domain expertise. These issues hinder their robustness, transparency, and accountability, limiting their ability to support ethical and reliable decision-making. This thesis addresses these challenges by developing methodologies that improve interpretability, strengthen reliability, and promote accountability, enabling stakeholders to critically evaluate and refine AI decisions.
The primary objective of this thesis is:

> *To develop methodologies that strengthen the robustness and transparency of both ML and Causal Models, while empowering human agency and fostering accountability.*

This overarching goal is supported by six thematic objectives, each addressing pivotal gaps in stakeholder engagement, interpretability, and reliability, contributing to the advancement of trustworthy AI systems.

Human Agency and Accountability Ensuring that AI systems empower stakeholders is crucial for aligning decision-making processes with societal and ethical standards.

- **Objective 1:** Develop contestable systems that allow stakeholders to scrutinise and refine decisions.

- **Objective 2:** Provide formal guarantees for the alignment of expert knowledge and statistical methods.

Transparency and Interpretability Improving interpretability ensures that AI systems remain transparent and open to scrutiny, fostering trust and enabling effective oversight.

- **Objective 3:** Enable interpretable representations of NN decision-making processes.
- **Objective 4:** Enable interpretability of Causal Discovery methods.

Robustness and Reliability Ensuring robust and reliable performance in noisy or data-scarce conditions is critical for the safe deployment of AI systems.

- **Objective 5:** Improve the robustness of NN models in noisy or data-scarce settings.
- **Objective 6:** Ensure causal discovery remains reliable under noisy or insconsistent data.

1.3 Thesis Contributions

The methodological choices in this thesis were driven by our objective to address key limitations in AI systems. NNs were chosen for their demonstrated capability in predictive tasks (Goodfellow et al., 2016), underpinned by their capability to approximate any function (Hornik et al., 1989). This universal approximation property allows NNs to model complex relationships across diverse domains, making them suitable for a wide range of applications. Additionally, causal regularisation was shown to improve NNs performance and generalisation capabilities (Kyono et al., 2020), demonstrating the value of causal discovery in enhancing NN robustness.

For causal discovery, the PC-Stable algorithm (Colombo and Maathuis, 2014) was selected as our base algorithm due to its theoretical guarantees (Harris and Drton, 2013; Kalisch and Bühlmann, 2007; Spirtes et al., 2000), which align with the principles of trustworthy AI. The algorithm's modular structure and ability to handle high-dimensional data make it particularly suitable for real-world scenarios. To enhance causal discovery, Shapley values and Assumption-Based Argumentation (ABA) were employed as complementary frameworks.

Shapley values were chosen for their interpretability and ability to decompose the contributions of individual components within a system. Originally developed in cooperative game theory (Shapley,

1953), they have become prominent in ML for attributing importance to features in complex models (Lundberg and Lee, 2017). Their framework of players, coalitions and marginal contributions was pivotal in extracting more granular insights from statistical tests in causal discovery.

Finally, ABA was employed for its ability to resolve conflicts through non-monotonic reasoning (Cyras et al., 2017). This framework can integrate statistical evidence and expert knowledge, allowing for the formal representation and resolution of disagreements. By reconciling diverse sources of information, ABA has the potential to increase systems' transparency and consistency. This capability is critical in noisy datasets, ensuring that the resulting outputs are robust, interpretable, and aligned with stakeholder expectations.

These methodological choices underpin the contributions of this thesis, addressing key gaps in predictive modelling and causal discovery.

Theoretical Contributions

This thesis introduces principled methodologies to enhance the robustness, transparency and accountability of NNs and causal discovery frameworks.

- **Causal Knowledge Injection (Chapter 4):** This contribution formalises the integration of causal knowledge into Feed-Forward Neural Networks (FFNNs). This approach provides a principled mechanism for leveraging expert feedback, aligning predictive systems with domain insights, hence fostering accountability.
  Addresses Objectives: 1, 2.

- **Robust Causal Discovery Rules (Chapter 5):** A novel Shapley-based decision rule to orient v-structures given the skeleton of a causal graph is introduced. This interpretable rule reduces the reliance on single independence tests, enhancing robustness in causal discovery under noisy conditions, while retaining theoretical guarantees.
  Addresses Objectives: 4, 6.

- **Framework for Causal Argumentation (Chapter 6):** A novel ABA formalisation integrates statistical evidence with expert knowledge, enabling transparent argumentation about causal relationships. This framework identifies and resolves conflicts within data and between data and external knowledge, improving the interpretability, robustness, and contestability of discovered causal graphs.
  Addresses Objectives: 1, 2, 4, 6.

Novel Algorithms and Methodologies

Building on these theoretical foundations, this thesis introduces practical algorithms and methodologies that operationalise these concepts, demonstrating their application in real-world scenarios.

- **Contestable Neural Networks (Chapter 4):** A human-in-the-loop framework is proposed for extracting and injecting causal graphs into FFNNs, enabling stakeholders to contest and refine predictions. This two-way interaction strengthens the robustness, interpretability, and domain alignment of NNs in decision-making contexts.
  Addresses Objectives: 1, 2, 3, 5.

- **Shapley-PC Algorithm (Chapter 5):** Building upon the PC-Stable algorithm (Colombo and Maathuis, 2014), the Shapley-PC algorithm incorporates our novel Shapley-based decision rule to improve robustness in noisy and quasi-unfaithful settings. It provides superior causal graph extraction capabilities while retaining formal guarantees.
  Addresses Objective: 6.

- **ABA-PC Framework (Chapter 6):** By combining argumentation and causal discovery, the ABA-PC framework handles noisy and conflicting datasets. This integration of formal methods and statistical metrics improves the consistency and transparency of causal graph reconstruction, providing a principled approach to integrate external knowledge and facilitate contestability.
  Addresses Objectives: 2, 6.

Impact and Results

The proposed methodologies translate theoretical advancements into practical solutions, addressing key limitations in predictive modelling and causal discovery. Each approach has been rigorously validated, demonstrating measurable improvements in robustness and transparency:

- The Contestable NNs framework achieves up to a 2.4x improvement in predictive performance and an 85% reduction in model complexity on real and synthetic datasets.
- The Shapley-PC algorithm improves causal graph reliability by addressing errors in independence tests, improving v-structure identification up to a 6x compared to PC-based predecessors, while using the same statistical tests results from data.
- The ABA-PC algorithm reduces structural error by up to 100% compared to state-of-the-art methods, while providing formal guarantees and greater transparency.

These contributions demonstrate the potential for deploying transparent, robust, and accountable AI systems in high-stakes domains. By addressing core limitations of both MLand Causal Models, the methodologies empower stakeholders—including medical professionals, regulators, and policymakers—to better understand the causal mechanisms underlying predictions, ensure fairness in automated decisions, and derive actionable insights more aligned with ethical and societal values.

1.4 Thesis Structure and Bibliographic Notes

This thesis is structured to address the challenges of robustness, transparency, and accountability in AI systems, progressing from foundational concepts to methodological innovations and their validation.

Chapters 2 and 3 lay the groundwork for the research. Chapter 2 introduces essential concepts, including causal discovery, NN regularisation, Shapley values, and argumentation frameworks. Chapter 3 critically evaluates the literature on contestable AI and causal discovery, identifying gaps that further motivate the proposed methodologies.

Building on this foundation, Chapter 4 presents **Contestable NNs**, enabling stakeholders to inspect and refine predictions through causal graphs. Chapter 5 addresses robustness in causal discovery with the **Shapley-PC** algorithm, designed to handle noisy data effectively. Chapter 6 introduces **Causal ABA**, an argumentation-based framework that integrates expert knowledge with statistical evidence, producing consistent and transparent causal graphs.

Chapter 7 concludes the thesis by summarising key contributions, reflecting on their implications for trustworthy AI, and outlining future research directions.

This thesis builds upon several peer-reviewed publications that underpin the research presented herein. The following papers contributed directly to the development of the methodologies and findings:

- Chapter 4 builds on:
    1. Russo, F., & Toni, F. (2023). Causal Discovery and Knowledge Injection for Contestable Neural Networks. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, FAIA, vol. 372, pp. 2025–2032. IOS Press. (Russo and Toni, 2023a)
    2. Leofante, F., Ayoobi, H., Dejl, A., Freedman, G., Gorur, D., Jiang, J., Paulino-Passos, G., Rago, A., Rapberger, A., Russo, F., Yin, X., Zhang, D., & Toni, F. (2024). Contestable AI needs computational argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR 2024)*. pp. 888–896. (Leofante et al., 2024).

- Chapter 5 builds on:

    3. Russo, F., & Toni, F. (2024). Shapley-PC: Constraint-based Causal Structure Learning with Shapley Values. *arXiv preprint arXiv:2312.11582*. Submitted to the 4th Conference on Causal Learning and Reasoning (CLeaR 2025). (Russo and Toni, 2023b)

- Chapter 6 builds on:

    4. Russo, F., Rapberger, A., & Toni, F. (2024). Argumentative Causal Discovery. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR 2024)*. pp. 938–949. (Russo et al., 2024)
    5. Russo, F. (2023). Argumentation for Interactive Causal Discovery. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pp. 7091–7092. (Russo, 2023)

In addition, the following publications provided high-level motivation, discussed in §3.3.4, and informed future research directions discussed in Chapter 7:

6. Rago, A., Russo, F., Albini, E., Baroni, P., & Toni, F. (2021). Forging Argumentative Explanations from Causal Models. In *Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021)*, CEUR Workshop Proceedings, vol. 3086. (Rago et al., 2021)
7. Rago, A., Russo, F., Albini, E., Toni, F., & Baroni, P. (2023). Explaining Classifiers' Outputs with Causal Models and Argumentation. *FLAP*, 10(3), 421–509. (Rago et al., 2023)

Chapter 2

Background

In this chapter, we introduce the key concepts and foundational theories that underpin the methods and contributions presented in Chapters 4, 5, and 6. We begin by outlining the fundamentals of **causal graphs** (§2.1)—a central theme of this thesis—and their connection to statistical methods (§2.2). These foundations naturally lead to the **PC algorithm** (§2.3), a pioneering method for causal discovery that underlies the enhancements proposed in Chapter 5. We then explore the role of **causal discovery as an auxiliary task** (§2.4) in improving the robustness of FFNNs, which supports the Contestable NNs introduced in Chapter 4. Following this, we introduce **Shapley values** (§2.5), a game-theoretic measure of variable contribution that we integrate into the PC algorithm to address finite-sample challenges and improve its robustness (Chapter 5). Finally, we present **Assumption-Based Argumentation (ABA)** (§2.6), a knowledge representation formalism that supports the argumentation-based causal discovery framework developed in Chapter 6.

2.1 Causal Graphs

Causal graphs are central to this thesis, providing a unifying framework for representing and reasoning about cause-and-effect relationships. At their core, causal graphs capture the structure of causal relationships within a system. Nodes represent factors of interest, while directed edges denote the causal influence of one factor on another. This structure facilitates the construction of *graphical causal models* (Peters et al., 2017), such as Structural Equation Models (SEMs) and Functional Causal Models (FCMs) (Pearl, 2009; Spirtes et al., 2000), which are capable of simulating interventions and analysing counterfactual behaviours. These capabilities make causal graphs a foundational tool for understanding complex systems and designing interventions across domains.

The history of causal graphs is deeply rooted in disciplines such as epidemiology, economics, and social sciences, where they have been used for decades to formalise causal relationships (Spirtes et al., 2000; Wright, 1934). More recently, causal models have been used to improve ML models' performance on tasks such as optical caractecter recognition or gene perturbation (Peters et al., 2017). This interdisciplinary relevance underscores their utility as a framework for reasoning about causality in both theoretical and applied contexts.

The structured representation provided by causal graphs not only enhances interpretability and transparency, but also enables contestability by providing stakeholders with a clear view of the underlying relationships forming causal assumptions. These properties align closely with the principles of trustworthy AI, making causal graphs a critical component of the proposed methodologies.

To illustrate the practical implications of causal graphs, the following example demonstrates their application in modelling socioeconomic factors.

Example 2.1.1. *An illustrative example of a causal graph is shown in the figure below, where nodes such as* Education*,* Race*,* Occupation*, and* Income *are connected by arrows that represent causal relationships between these variables. In this hypothetical scenario, an individual's* Education *and* Race *both influence their* Occupation*, which in turn directly impacts their* Income*. Additionally,* Education *has a direct effect on* Income*.*

In this scenario, we could estimate the causal effect of Race *on* Income *by observing its indirect impact through* Occupation*. For instance, systemic biases may restrict access to certain occupations for individuals of different racial backgrounds, even when their levels of* Education *are comparable. Similarly, the direct influence of* Education *on* Income *highlights its role in providing opportunities for career advancement and access to higher-paying roles.*

A policymaker analysing this graph might identify targeted interventions to address income disparities. For example, efforts to reduce racial bias in occupational opportunities or to improve access to quality education could have measurable impacts on income equality between different demographic groups.

This example highlights how causal graphs enable reasoning about complex interdependencies and the effects of potential interventions. In the remainder of this section we introduce basic terminology and notions for graphs in general, and causal graphs in particular.

In this thesis, causal graphs are used as analytical tools for integrating expert knowledge into predictive models and as desired outputs of causal discovery methods. Causal discovery algorithms aim to infer the structure of causal relationships from observational data, providing data-driven approaches to constructing causal graphs. These algorithms leverage statistical methods to identify patterns in the data that reflect causal relationships, enabling the construction of causal graphs that capture the underlying mechanisms of the system.

For instance, in Example 2.1.1, causal discovery seeks to determine the relationships among *Education*, *Race*, *Occupation*, and *Income* from observational data, reconstructing the causal structure represented in the graph. This process is fundamental for formulating the structural assumptions required to build causal models (Pearl, 2009; Peters et al., 2017), but also to improve predictions, as we demonstrate in Chapter 4. In this section we cover the graphical and statistical foundations of causal discovery, while a review of the literature is presented in §3.3.

Graph Types and Terminology A graph is a pair $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \{1, \dots, d\}$ is a finite set of *nodes* and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is a set of *edges*. Our graphs are *simple*, meaning they have no self-loops or multiple edges between nodes (Lauritzen, 1996, p.4). Edges can be either directed or undirected: an edge is *undirected* if both $(i, j) \in \mathbf{E}$ and $(j, i) \in \mathbf{E}$, denoted $i - j$, while an edge is *directed* if $(i, j) \in \mathbf{E}$ but $(j, i) \notin \mathbf{E}$, denoted $i \to j$.

Graphs can be classified as *directed* (containing only directed edges ($\leftrightarrows$)), *undirected* (containing only undirected edges ($-$)), or *partially directed* (containing both types). The *skeleton*, denoted $\mathcal{C}$, of a (partially) directed graph is obtained by replacing all directed edges with undirected ones. A graph is *complete* if every pair of nodes is connected by an edge (Pearl, 2009, p.13).

A *path* between two nodes i and j is a sequence of distinct nodes $i = k_1, k_2, \dots, k_n = j$ such that $(k_m, k_{m+1}) \in \mathbf{E}$ for all $m = 1, \dots, n-1$. A *cycle* is a path where $i = j$. A path is *directed* if it contains only directed edges and *undirected* if all edges are undirected (Peters et al., 2017, p.82). A graph is *acyclic* if there is no directed path that begins and ends with the same node. Such a graph is called a DAG, or a Partially Oriented DAG (PDAG) if it is partially directed.

If an edge exists between two nodes, then these nodes are *adjacent* or *neighbours*. The *degree* of a node is the number of its neighbours. The set of nodes adjacent to i in the graph $\mathcal{G}$ is denoted by $adj(\mathcal{G}, i)$. If one of these adjacent nodes, j, points to i with a directed edge, j is called a *parent* of i. The set of parents of node i is denoted by $pa(\mathcal{G}, i)$. A node i is a *descendant* of node j if there is a directed path from j to i, and an *ancestor* if the path is in the opposite direction. *Non-descendants* of a node are all nodes that

are not its descendants, including the node itself (Lauritzen, 1996, p.6). In Example 2.1.1, the adjacency set of *Race* is $adj(\mathcal{G}, Race) = \{Occupation\}$, while $adj(\mathcal{G}, Occupation) = \{Education, Race, Income\}$. The parent set of *Occupation* is $pa(\mathcal{G}, Occupation) = \{Education, Race\}$, while *Race* is an ancestor of *Income*. Finally, *Income* is both a descendant (via *Occupation*) and a child of *Education*.

Causal Graphs A causal graph is a graph where nodes represent factors or entities in a system, and edges represent relationships between these factors (Pearl, 2009; Spirtes et al., 2000). For instance, Example 2.1.1 illustrates socioeconomic factors using a causal graph, where nodes correspond to variables like *Race*, *Education*, and *Occupation*, and edges capture causal relationships between them.

A triple (i, j, k) is called an *Unshielded Triple (UT)* if two nodes i and k are not adjacent, but both are adjacent to the third node j, denoted $i - j - k$. A *v-structure*, also known as an *immorality*, occurs when $i \to j \leftarrow k$, making j a *collider*. When graphs are causal, then a collider is also referred to as a *common effect* (Pearl, 2009, p.17).

Definition 2.1.2 (Collider)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and $\not{p} = i_1, \ldots, i_m$ a directed path. If $(i_{k-1}, i_k) \in \mathbf{E}$ and $(i_{k+1}, i_k) \in \mathbf{E}$, then i_k is a* collider *wrt $\not{p}$, and the UT $i_{k-1} - i_k - i_{k+1}$ is a* v-structure*: $i_{k-1} \to i_k \leftarrow i_{k+1}$.*

If two nodes i, k have a common ancestor j, i.e., there is a directed path from j to both i and k that does not pass through either i or k (Peters et al., 2017), then j is called a *confounder* or *common cause*. This means that i and k are dependent but not causally related, as j is 'confounding' the relationship observed between i and k.

Example 2.1.3. *In Example 2.1.1, the triple* (Education, Occupation, Race) *forms a UT that becomes a v-structure when considering the orientations of the arrows Education $\to$ Occupation $\leftarrow$ Race. Intuitively, given the causal graph in the example, if we know the occupation of an individual, as well as their race, we can infer their level of education.* Education*, instead, is a confounder for the relationship between* Occupation *and* Income*, affecting both these factors.*

Infer (In)Dependencies from Causal Graphs The relationships between nodes in a causal graph depend on the paths connecting them and whether these paths are *active* or *blocked*. To understand how causal relationships between variables are represented in graphs, and how information flows between them, it is essential to define an *active* path (Pearl, 1989, p.118). Intuitively, an active path allows information to flow between two nodes, depending on the presence of colliders along the path and whether these colliders or their descendants are included in the *conditioning set*, which represents the current state of knowledge. Formally:

Definition 2.1.4 (**S**-active Path)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and let $\not{p}$ be a directed path between distinct nodes $i, j \in \mathbf{V}$. We say that $\not{p}$ is* **S**-active *for a conditioning set $\mathbf{S} \subseteq \mathbf{V} \setminus \{i, j\}$ if and only if the following conditions hold for each node $k \in \not{p}$:*

1. *If k is a* collider *(i.e., $i \to k \leftarrow j$), then either:*
   - *$k \in \mathbf{S}$, or*
   - *a descendant of k is in $\mathbf{S}$.*
2. *If k is not a collider, then $k \notin \mathbf{S}$.*

An active path is one in which no node acts as a collider, unless it is conditioned upon, or one of its descendants is in the conditioning set. Colliders *block* paths, but conditioning on them *opens* the path, making them active.

Using these notions, we can define *d-separation* (Pearl, 1989, p.117): two nodes i and j are *d-separated* given a set $\mathbf{S}$ if no $\mathbf{S}$-active path exists between them. Formally:

Definition 2.1.5 (d-Separation)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG. Two distinct nodes $i, j \in \mathbf{V}$ are said to be* d-separated *given a conditioning set $\mathbf{S} \subseteq \mathbf{V} \setminus \{i, j\}$ if there is no $\mathbf{S}$-active path between i and j in $\mathcal{G}$. We denote this as $i \perp\!\!\!\perp_{\mathcal{G}} j \mid \mathbf{S}$. Conversely, if an $\mathbf{S}$-active path exists between i and j, they are said to be* d-connected*, denoted as $i \not\perp\!\!\!\perp_{\mathcal{G}} j \mid \mathbf{S}$.*

We illustrate these key concepts with an example:

Example 2.1.6. *Consider the causal graph in Example 2.1.1. From the graph structure, using Definition 2.1.5, we can deduce the following d-separation relationships:*

- ***d-Separation without Conditioning:*** *In the path Education $\to$ Occupation $\leftarrow$ Race, both* Race *and* Education *influence* Occupation*, which acts as a collider. Without conditioning,* Race *and* Education *are* ***d-separated****, meaning* Race *provides no information about* Education*.*
- ***Collider:*** *In Education $\to$ Occupation $\leftarrow$ Race,* Occupation *is a collider. If unconditioned,* Education *and* Race *are independent (****d-separated****). Conditioning on* Occupation *opens the path, making* Education *and* Race ***d-connected****, introducing an artificial dependency.*
- ***d-Separation with Conditioning:*** *Conditioning on* Occupation *blocks the path Race $\to$ Occupation $\to$ Income. Once* Occupation *is known,* Race *provides no additional information about* Income*, rendering them* ***d-separated****.*

- ***Multiple Paths:*** *The direct path Education → Income ensures that* Education *and* Income *remain* ***d-connected****, even if* Occupation *is conditioned upon. This indicates* Education *still influences* Income *directly, regardless of* Occupation.

This example demonstrates how d-separation enables us to deduce conditional independencies from a graph structure. With this understanding, we can now explore the statistical tools needed to infer graphical relationships from data.

2.2 Causal Graphs and Statistics

As we mentioned, causal graphs encode the causal relationships between factors in a system. These factors, when measured and recorded, assume an aleatoric nature that is represented by random variables (Casella and Berger, 2002; Pearl, 2009).

Conditional Independence We denote a set of random variables as $\mathbf{X} = \{X_1, \ldots, X_d\}$. Each variable X_k, for $k \in \{1, \ldots, d\}$, takes values in $\mathcal{X}_k \subseteq \mathbb{R}$ if they are continuous and in $\mathcal{X}_k \subseteq \mathbb{Z}$ if they are discrete, with $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. Intuitively, two variables X_i, X_j are *conditionally independent*, given a conditioning set $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$, if fixing the values of the variables in $\mathbf{S}$, X_i or X_j does not provide any additional information about X_j or X_i, respectively. In this case, we write $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}$ (Dawid, 1979; Lauritzen, 1996). Otherwise, the variables are dependent and are denoted by $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S}$. If the conditioning set is empty, we write $X_i \perp\!\!\!\perp X_j$ and $X_i \not\perp\!\!\!\perp X_j$ to denote *marginal* independence and dependence, respectively. We can formalise the notion of independence in probabilistic terms as follows:

Definition 2.2.1 (Conditional Independence)**.** *Let* $\mathbb{P}(\mathbf{X})$ *be a joint probability distribution,* $X_i, X_j \in \mathbf{X}$ *and* $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$*, then*

$$X_i \perp\!\!\!\perp X_j \mid \mathbf{S} \iff \mathbb{P}(X_i \mid X_j, \mathbf{S}) = \mathbb{P}(X_i \mid \mathbf{S}) \tag{2.1a}$$

$$X_i \perp\!\!\!\perp X_j \mid \mathbf{S} \iff \mathbb{P}(X_i, X_j \mid \mathbf{S}) = \mathbb{P}(X_i \mid \mathbf{S})\mathbb{P}(X_j \mid \mathbf{S}) \tag{2.1b}$$

The formulation in Equation 2.1a reflects the intuition we mentioned earlier: the conditional probability of X_i given X_j and $\mathbf{S}$ is the same as the conditional probability of X_i given $\mathbf{S}$ alone, i.e. knowing X_j does not provide any additional information about X_i beyond what is already known in $\mathbf{S}$. The formulation in Equation 2.1b implies that two variables X_i, X_j are independent if the joint probability distribution of X_i, X_j can be factorised into the product of the marginal probabilities of

each variable. Lauritzen (1996, p.29) provides alternative formulations and more details. We use this latter definition to establish the connection between a DAG and the joint probability of the variables represented in the DAG. Indeed, the connection lies in how we can factorise the joint probability of all variables, given the structure of the DAG, as we discuss next.

Markov Property A distribution is said to be *Markovian* with respect to a graph when this graph encodes the conditional independencies that are implied by the graph's structure (Peters et al., 2017, p.100). In essence, the conditional independencies found in the distribution must align with the d-separation relationships entailed by the graph (Peters et al., 2017, Def. 6.21).

Definition 2.2.2 (Markov Property)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $\mathbb{P}(\mathbf{V})$ be a joint probability distribution. With respect to $\mathcal{G}$, $\mathbb{P}(\mathbf{V})$ is said to satisfy the following:*

*(i) The **global Markov property** if, for all disjoint subsets $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z} \subseteq \mathbf{V}$, we have:*

$$\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$$

*(ii) The **local Markov property** if each variable is independent of its non-descendants, given its parents.*

*(iii) The **Markov factorisation property** if:*

$$\mathbb{P}(\mathbf{V}) = \prod_{i=1}^{d} \mathbb{P}(V_i \mid pa(\mathcal{G}, V_i))$$

Here is what these properties mean in simpler terms:

(i) The **global Markov property** states that if two sets of variables are d-separated in a graph, then they are independent in the distribution, generated by that graph. This is a powerful property that allows us to deduce independencies from the graph structure.

(ii) The **local Markov property** states that each variable is independent of its non-descendants, given its parents. This simplifies the consideration of dependencies to direct relationships, reducing the search space for causal discovery algorithms (see §2.3).

(iii) The **Markov factorisation property** shows that the joint probability of all variables can be broken down into simpler terms—each variable is conditionally dependent only on its direct parents. This drastically simplifies the calculation of joint probabilities, reducing it to the product of these conditional probabilities.

The Markov property connects d-separation relationships to conditional independencies. However, different DAGs can encode the same set of conditional independencies. When this happens, the DAGs are said to belong to the same Markov Equivalence Class (MEC) (Peters et al., 2017, p.102).

A compact characterisation of MECs is provided in (Verma and Pearl, 1990, Theorem 1), providing the necessary and sufficient conditions for DAG equivalence:

Theorem 2.2.3 (DAG Equivalence)**.** *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures.*

This concept is crucial for understanding the limitations of causal discovery using observational data alone. Observational data cannot distinguish between DAGs in the same MEC, as they represent the same set of conditional independence relationships. This ambiguity, known as *observational equivalence* (Pearl, 2009, Theorem 1.2.8), arises because the conditional independencies encoded by a DAG depend only on its skeleton and v-structures, not on the directionality of all edges. Consequently, the boundaries of causal discovery from observational data are defined by the inability to determine edge directions beyond these constraints.

To resolve this ambiguity, additional information—such as expert knowledge or experimental data from randomised control trials—can be used to orient the edges in the graph. Notably, there are also cases where relationships deduced from observational data cannot be established through experiments alone (Spirtes et al., 2000, Ch. 9). This happens when observational dependencies arise due to a specific Data Generating Process (DGP) that the experiment could not fully replicate.

To represent MECs we use *CPDAGs* (Chickering, 2002b). A CPDAG is a PDAG with the following characteristics (Peters et al., 2017, p.102):

Definition 2.2.4 (CPDAG)**.** *Let $\mathcal{G}$ be a DAG and MEC($\mathcal{G}$) be the set of all DAGs in the MEC of $\mathcal{G}$. CPDAG($\mathcal{G}$) $= (\mathbf{V}, \mathbf{E})$ is the CPDAG of $\mathcal{G}$, where $(i, j) \in \mathbf{E}$ if and only if $i \to j$ in one of the DAGs in MEC($\mathcal{G}$).*

This definition implies that:

- The skeleton of the CPDAG is the same as the skeleton of the DAG.
- Directed edges correspond to edge directions that are consistent across all DAGs in the MEC.
- Undirected edges represent arrows that are oriented both ways in the DAGs of the MEC.

CPDAGs provide a compact way to represent all the DAGs in a given MEC. They capture the essential features of the equivalence class while acknowledging the limitations of observational data.

Example 2.2.5. *In Example 2.1.1, the only v-structure of the graph is Education → Occupation ← Race. The remaining edges (Education − Income, Occupation − Income) could be oriented in either direction without violating the conditional independencies (d-separations) implied by the graph and cannot be distinguished from observational data alone. Hence, the CPDAG would be the following:*

The MEC of this DAG would include all the possible orientations of the edges between Occupation *and* Income*, totalling four DAGs.*

Infer DAGs from (In)Dependencies Another important concept in relating graphs and data is *Faithfulness*. As the converse of the Markov property, this assumption ensures that all and only the independence relationships in the data are reflected in the graph, enabling us to infer the graph structure directly from the (in)dependencies. Formally (Peters et al., 2017, Def. 6.33):

Definition 2.2.6 (Faithfulness)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $\mathbb{P}(\mathbf{V})$ be a joint probability distribution.* $\mathbb{P}(\mathbf{V})$ *is* faithful *to $\mathcal{G}$ if, for all disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$:*

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}.$$

A distribution that is Markovian but not faithful to a DAG $\mathcal{G}$ arises when two paths cancel each other out. This can occur when the combined effects through different path sum to zero, leading to an independence relationship in the data that is not captured in $\mathcal{G}$. In these scenarios, the graph fails to represent the observed independence relationships in the data. Peters et al. (2017, §6.5.3) provide more details and examples. However, Spirtes et al. (2000, Theorem 3.2) demonstrate that such scenarios where the causal effects work against faithfulness occur with probability zero in linear models with coefficients drawn randomly from positive densities.

Ramsey et al. (2006) decompose the faithfulness assumption into two components: *adjacency faithfulness* and *orientation faithfulness*. This breakdown is useful for understanding how graph structures can be inferred from data. Additionally, this decomposition serves as a foundation for testing and validating our causal discovery algorithm in Chapter 5.

Definition 2.2.7 (Adjacency Faithfulness)**.** *Let $\mathbb{P}(\mathbf{V})$ be a joint probability distribution Markovian with respect to a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and let $X_i, X_j \in \mathbf{V}$. Then:*

$$X_i \in adj(\mathcal{G}, X_j) \Longrightarrow X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S} \quad \forall \mathbf{S} \subseteq \mathbf{V} \setminus \{X_i, X_j\}.$$

This property ensures that if two variables are adjacent in the graph, they remain dependent given any conditioning set. It guarantees that the graph captures all and only the undirected paths necessary to create dependencies among variables. Furthermore, it justifies the *adjacency search* phase of constraint-based causal discovery algorithms, which we will discuss in §2.3. A similar formalisation can also be found in (Spirtes et al., 2000, Lemma 5.1.2).

Definition 2.2.8 (Orientation Faithfulness)**.** *Let $\mathbb{P}(\mathbf{V})$ be a joint probability distribution Markovian with respect to a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and let $X_i - X_j - X_k \in \mathcal{G}$ be a UT. Then:*

(i) If $X_i \to X_j \leftarrow X_k \in \mathcal{G}$, then $X_i \not\perp\!\!\!\perp X_k \mid \mathbf{S}$ for all $\mathbf{S} \subseteq \mathbf{V} \setminus \{X_i, X_k\} \cup X_j$.

(ii) Otherwise, $X_i \perp\!\!\!\perp X_k \mid \mathbf{S}$ for all $\mathbf{S} \subseteq \mathbf{V} \setminus \{X_i, X_k, X_j\}$.

Orientation faithfulness ensures that the role of colliders in opening or blocking paths is consistently reflected in the observed dependencies. When a collider is conditioned upon, it opens the path between its parent nodes, making them dependent, regardless of the other variables in the conditioning set. This property underpins the v-structure orientation step of constraint-based causal discovery and is the focus of Chapter 5. Spirtes et al. (2000, Lemma 5.1.3) provide a similar formalisation.

Faithfulness, combined with the Markov property and notion of d-separation, enables us to orient edges in a DAG based on conditional independencies observed in data. For instance, by analysing dependencies, we can infer whether a UT can be oriented as a v-structure: conditioning on a collider X_j renders the path $X_i \to X_j \leftarrow X_k$ active. Hence, if we observe $X_i \perp\!\!\!\perp X_k$ and $X_i \not\perp\!\!\!\perp X_k \mid X_j$ we can infer that X_j is a collider for X_i and X_k, and can orient the UT.

Example 2.2.9. *Let us illustrate how the graph structure in Example 2.1.1 can be inferred from observed conditional independencies:*

- ***Adjacency Faithfulness:** If we observe Race $\perp\!\!\!\perp$ Education, we can infer that there is no edge between these variables in the graph. Similarly, if Race $\perp\!\!\!\perp$ Income $\mid \{$Occupation, Education$\}$, we can deduce that no direct edge exists between Race and Income. If these were connected, they would be dependent given any conditioning set, since the path driving the dependency would be open no matter the conditioning set.*

- ***Orientation Faithfulness:*** *If we then observe* $\text{Race} \not\perp\!\!\!\perp \text{Education} \mid \{\text{Occupation}\}$*, we infer that the UT* $\text{Race} - \text{Occupation} - \text{Education}$ *should be oriented as* $\text{Race} \to \text{Occupation} \leftarrow \text{Education}$*. Under orientation faithfulness, we would also expect* $\text{Race} \not\perp\!\!\!\perp \text{Education} \mid \{\text{Occupation}, \text{Income}\}$*, as conditioning on the collider Occupation opens the path between Race and Education, making them dependent.*

This example demonstrate how observed (in)dependencies, coupled with the faithfulness assumption, allow us to infer the underlying causal graph structure from data.

Causal Sufficiency Another important assumption when working with causal graphs is *causal sufficiency* (Spirtes et al., 2000). Causal sufficiency holds if all relevant variables are included in the analysis, particularly those that might act as latent confounders if omitted.

Example 2.2.10. *Consider again our DAG from Example 2.1.1, focusing on the relationship between* Occupation *and* Income*. Observational data might show that individuals in higher-status occupations generally earn higher incomes. However, omitting the variable* Education*, which influences both* Occupation *and* Income*, could lead to an incorrect estimate of the direct effect of* Occupation *on* Income*. Specifically,* Education *confounds this relationship because it impacts both the type of jobs an individual is eligible for, and their potential income. Including* Education *in the analysis enables us to correctly infer the direct relationship between* Occupation *and* Income*.*

Failing to ensure causal sufficiency can lead to paradoxical results, such as those exemplified by Simpson's paradox, where aggregate data trends differ from subgroup trends (Peters et al., 2017, Example 6.37). Let us formalise the causal sufficiency assumption (Spirtes et al., 2000, p.45):

Definition 2.2.11 (Causal Sufficiency)**.** *Let* $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ *be a set of variables, with* $X \notin \mathbf{V}$ *as a common cause of* $V_i, V_j \in \mathbf{V}$*. Let* X *take values in domain* $\mathcal{X}$*, and let* $\mathbb{P}(\mathbf{V})$ *be the joint probability distribution of* $\mathbf{V}$*. The set* $\mathbf{V}$ *is* causally sufficient *if and only if:*

$$\mathbb{P}(\mathbf{V} \mid X = x) = \mathbb{P}(\mathbf{V}) \quad \forall x \in \mathcal{X}.$$

In essence, the joint distribution of the variables in $\mathbf{V}$ remains invariant across the values of any common cause X that is not included in $\mathbf{V}$. This invariance is critical for correctly modelling the causal relationships without introducing bias from unobserved confounders.

While the causal sufficiency assumption is used throughout this thesis, the proposed algorithms can be extended to relax this requirement. Additional rules such as the ones from Fast Causal Inference

(FCI) (Spirtes et al., 2000) (see §3.3), or incorporating additional prior knowledge, could address cases where the assumption does not hold.

A key implication of the faithfulness and causal sufficiency assumptions is the one-to-one correspondence between variables in the data and nodes in the true causal graph. Consequently, we use the terms *variable* and *node* interchangeably and, where context permits, use a variable's index to refer to the variable or its corresponding node.

Conditional Independence Tests (CITs) Identifying conditional independencies from data is central to constraint-based causal discovery. As we showed earlier, given faithfulness, a MEC can be reliably inferred from a set of independencies. However, these relations must be determined empirically from the observed sample if no prior knowledge is available. This is where CITs play a pivotal role.

Definition 2.2.12 (CIT)**.** *Let $X_i, X_j \in \mathbf{X}$ be two random variables, $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$ be a conditioning set. A CIT tests the hypothesis:*

$$\mathcal{H}_0 : X_i \perp\!\!\!\perp X_j \mid \mathbf{S} \quad \textit{versus} \quad \mathcal{H}_1 : X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S}.$$

We denote a CIT by $I(X_i, X_j \mid \mathbf{S})$, and assume it outputs a valid p-value $p \in [0, 1]$ *(Casella and Berger, 2002, Def. 8.3.26 and 8.3.27). For a chosen significance level α:*

$$I(X_i, X_j \mid \mathbf{S}) = p > \alpha \implies X_i \perp\!\!\!\perp X_j \mid \mathbf{S}, \quad (\textit{fail to reject } \mathcal{H}_0),$$
$$I(X_i, X_j \mid \mathbf{S}) = p \leq \alpha \implies X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S}, \quad (\textit{reject } \mathcal{H}_0).$$

In statistical hypothesis testing, a *Type I error* occurs if one rejects $\mathcal{H}_0$ when it is actually true, and a *Type II error* occurs if one fails to reject $\mathcal{H}_0$ when $\mathcal{H}_1$ holds (Casella and Berger, 2002, Def. 8.1.3). Under $\mathcal{H}_0$, the p-value distribution is uniform in idealised conditions, thus allowing to control the Type I error rate through the significance level α. Under $\mathcal{H}_1$, p-values are typically skewed towards zero, making small p-values more likely as the departure from independence increases (Hung et al., 1997). Consequently, p-values are monotonically informative: smaller values provide stronger evidence against $\mathcal{H}_0$ (Casella and Berger, 2002, p.397).

The probability of correctly rejecting $\mathcal{H}_0$ when $\mathcal{H}_1$ is true is given by the *power function* (Casella and Berger, 2002, Def. 8.3.1), which characterises the CIT's overall efficacy. An ideal or *perfect CIT*—discussed in Chapter 5—would have zero Type I error and perfect power, though this ideal is unattainable in practice. Typically, one fixes α to limit Type I errors, and relies on larger samples or

more sensitive tests to enhance power and mitigate Type II errors (Wasserman, 2004).

Numerous CITs have been proposed. Fisher's Z-test (Fisher, 1970) is widely used due to its simplicity and efficiency in the Gaussian setting (Colombo and Maathuis, 2014; Ramsey et al., 2006; Spirtes et al., 2000). Alternatives include the χ^2 test (Pearson, 1900), rank-based methods (van der Waerden, 1940), kernel-based tests (Gretton et al., 2007), and various non-parametric approaches (Chen et al., 2024; Shah and Peters, 2018; Zhang et al., 2023, 2011). For the empirical evaluations in Chapters 5 and 6, we use Fisher's Z-test (Fisher, 1970) as our CIT, however, the methods proposed in this thesis are flexible and can accommodate alternative CITs as needed.

In summary, CITs provide the statistical foundations for discovering and validating the conditional independencies that underlie causal structures. By identifying where conditional independencies hold or fail, these tests supply the building blocks necessary to constrain the space of candidate models. In the next section, we build on this foundation by introducing the PC[1] algorithm (Spirtes et al., 2000), which leverages such constraints to infer causal graphs from observational data.

2.3 The PC Algorithm

The PC algorithm is a foundational constraint-based causal discovery method (Spirtes et al., 2000). Constraint-based algorithms leverage CITs (the constraints) and graphical rules based on d-separation to recover the ground-truth causal graph, providing provable guarantees under specific assumptions.

The PC algorithm operates under three core assumptions: acyclicity, faithfulness, and sufficiency. Under these assumptions, the algorithm has been shown to be sound and complete (Spirtes et al., 2000). Furthermore, consistency in the large-sample limit has been proven under additional assumptions about the DGP, specifically for Gaussian (Kalisch and Bühlmann, 2007) and Gaussian Copula (Harris and Drton, 2013) distributions.

The algorithm proceeds in three main steps: skeleton construction, v-structure orientation, and pattern completion. Pseudocode for the PC algorithm is provided in Algorithm 1.

1. **Skeleton construction via adjacency search:** Starting from a complete graph, the algorithm performs CITs for each pair of adjacent variables. If an independence is detected, the edge between the variables is removed, and the separating set responsible for the independence is recorded. This step produces a skeleton, denoted as $\mathcal{C}$. Leveraging Definition 2.2.7—which states that adjacent variables are dependent given any conditioning set—the algorithm prioritises

[1] From its creators' names: **P**eter Spirtes and **C**lark Glymour.

Algorithm 1: The Peter-Clark (PC) Algorithm

Input: $I(X_i, X_j \mid \mathbf{S}) \; \forall \; X_i, X_j \in \mathbf{V}, \mathbf{S} \subseteq \mathbf{V} \setminus (X_i, X_j)$; α

Step 1: Adjacency Search

$\mathcal{C} \coloneqq \langle \mathbf{V}, \mathbf{E} \rangle, \mathbf{E} = \mathbf{V} \times \mathbf{V}$ ▷ Complete Graph over $\mathbf{V}$

$\mathbf{C} = \emptyset$ ▷ Separating Sets

for $X_i \in \mathcal{C}$ **do**

  for $X_j \in \text{adj}(\mathcal{C}, X_i)$ **do**

    for $\mathbf{S} \in \mathbf{N}$ **do**

      if $I(X_i, X_j \mid \mathbf{S}) \geq \alpha$ **then**

        Remove edge from $\mathcal{C}$

        $\mathbf{C} = \mathbf{C} \cup \mathbf{S}$

return $\mathcal{C}$ ▷ Skeleton

Step 2: Orient v-structures

for $X_i - X_j - X_k \in \mathcal{C}$ **do**

  if $X_j \notin \mathbf{C}$ **then**

    orient: $X_i \rightarrow X_j \leftarrow X_k$

return $\mathcal{C}$ ▷ PDAG

Step 3: Pattern Completion

Apply Meek's rules (Meek, 1995) to $\mathcal{C}$ until no more edges can be oriented

return $\mathcal{C}$ ▷ CPDAG

marginal independence tests, as they are both computationally cheaper and statistically more reliable (Casella and Berger, 2002). It incrementally increases the size of the conditioning sets only after testing all pairs of adjacent variables, yielding significant efficiency improvements compared to the earlier SGS[2] algorithm (Spirtes et al., 2000).[3]

2. **Orienting v-structures:** Using the principle from Definition 2.2.8—that a collider is not included in any separating set of the other two variables in a UT—the algorithm examines each UT $X_i - X_j - X_k$ in the skeleton $\mathcal{C}$ output from step 1. If X_j is not in any of the separating sets saved in step 1 for X_i and X_k, the UT is oriented as a v-structure: $X_i \rightarrow X_j \leftarrow X_k$.

3. **Completing patterns:** Following the v-structure orientation, the algorithm analyses all triangles (sets of three adjacent variables) and kites (sets of four adjacent variables) in the graph. It applies the orientation rules described in (Meek, 1995) to deduce further edge directions. These rules ensure that the resulting graph is a sound and complete CPDAG, accurately representing the true MEC under the faithfulness assumption (Meek, 1995).

The PC algorithm represents a foundational constraint-based causal discovery method, offering strong theoretical guarantees in the sample limit. However, in finite-sample settings, PC has been

[2]Named after its inventors: Peter **S**pirtes, Clark **G**lymour, and Richard **S**cheines

[3]The part of the procedure relative to the size of the conditioning set is omitted from Algorithm 1 for conciseness. It can be found in (Colombo and Maathuis, 2014, Algorithm 3.2).

extended in various ways, such as guaranteeing order-independence (PC-Stable (Colombo and Maathuis, 2014)) or making sure that step 2 did not violate faithfulness (CPC (Ramsey et al., 2006)). See §3.3 for more details on the various extensions to PC.

In this thesis, we build upon these advances and propose a novel enhancement to constraint-based causal discovery, the Shapley-PC algorithm (Chapter 5). By integrating Shapley values into the v-structure orientation phase, Shapley-PC mitigates finite-sample errors while maintaining theoretical guarantees. This approach strengthens the reliability of causal graph recovery, addressing key limitations of existing methods.

2.4 Causal Discovery as an Auxiliary Task

Having examined contraint-based causal discovery, we now explore the integration of causal discovery with predictive models, focusing on FFNNs. In Chapter 4, we will introduce Contestable NNs, which leverage causal graphs to enhance interpretability, contestability, and robustness via expert knowledge injection. We begin by introducing our prediction setting.

Consider a prediction task with input variables $\mathbf{X} = \{X_1, \ldots, X_d\}$ and a target variable Y. Each input variable X_k takes values $x_k \in \mathcal{X}_k \subseteq \mathbb{R}$, while the target variable Y takes values $y \in \mathcal{Y} \subseteq \mathbb{R}$ for regression tasks or $\mathcal{Y} \subseteq \mathbb{Z}$ for classification tasks. The input space is defined as $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. This setup is a specialisation of the general one in Section 2.2, where the target variable Y is treated separately for prediction purposes, and the input space is assumed to be continuous, while the target can be either continuous or discrete.[4]

A predictive model learns a function $f_Y : \mathcal{X} \mapsto \mathcal{Y}$ that maps input values $[x_1, \ldots, x_d]$ to a predicted target value y. In practice, f_Y is learned from a training dataset:

$$\mathcal{D} = \{(x_{j1}, \ldots, x_{jd}, y_j) \mid j = 1, \ldots, N; x_j \in \mathcal{X}, y_j \in \mathcal{Y}\},$$

drawn from a joint probability distribution $\mathbb{P}(\mathbf{X}, Y)$ over the input and target variables.

In this setting, NNs can be leveraged to model the relationship between $\mathbf{X}$ and Y. Among various NN architectures, FFNNs—consisting of multiple layers of interconnected computational units (neurons)—form a foundational class (Goodfellow et al., 2016). Each neuron computes a weighted sum of its inputs plus a bias term, followed by a non-linear activation function. By stacking several layers, FFNNs can approximate a broad class of functions, capturing highly non-linear DGPs. Under suitable

[4]We keep d *input* variables instead of $d - 1$ for ease of exposition.

conditions, FFNNs are universal approximators (Hornik et al., 1989). In this thesis, when the context is clear, we use NN and FFNN interchangeably.

Training a NN involves optimising its parameters (weights and biases) to minimise a loss function that quantifies the task's objective. Gradient-based algorithms such as back-propagation and Stochastic Gradient Descent (SGD) iteratively adjust these parameters to achieve better performance on the training data. Once training converges, the resulting model can predict target values for previously unseen inputs. Beyond this initial training, a well-trained network may also be *fine-tuned* for related tasks or to adapt to new domains, enabling more efficient and flexible reuse of learned representations (Yosinski et al., 2014). We explore both these aspects in the context of *contesting* NNs in Chapter 4.

Despite their strengths, NNs often require large datasets (Picha et al., 2023), typically operate as black boxes that limit interpretability and make it difficult to incorporate domain knowledge (von Rüden et al., 2023), and may struggle with generalisation (Szegedy et al., 2014). A common methodology to improve generalisation and stability is *regularisation* (Hastie et al., 2009, Ch. 7). Given the complementary strengths of causal models when compared to ML (cf. Table 1.1), a natural way forward is to consider integrating causal insights into the training process. Indeed, (Kyono et al., 2020) demonstrated the value of embedding causal discovery into FFNNs, as we see next.

Causal Structure Learning (CASTLE) Kyono et al. (2020) introduce CASTLE, a neural architecture that integrates causal discovery as an auxiliary task alongside the primary prediction objective. This dual-task approach enables the model to predict a target variable while simultaneously uncovering the causal structure among input variables by leveraging causal discovery as a form of *regularisation*. Specifically, CASTLE penalises weights that would create cycles in the graph representing variable relationships, thereby inducing a DAG structure. This encourages the learned relationships to adhere to causal principles while maintaining low prediction error on the target variable.

While traditional causal discovery aims to identify genuine causal dependencies, CASTLE demonstrates that enforcing acyclicity alone can significantly enhance predictive performance. Empirical evaluations on both synthetic and real-world datasets show that encouraging the NN to learn an acyclic representation of inter-variable relationships leads to improvements in both accuracy and generalisation (Kyono et al., 2020).

The CASTLE architecture, illustrated in Figure 2.1, simultaneously predicts a target variable Y and reconstructs the input variables $\mathbf{X}$. It comprises $d + 1$ subnetworks, each with d input neurons corresponding to all but one of the variables. Each subnetwork excludes a specific variable (among

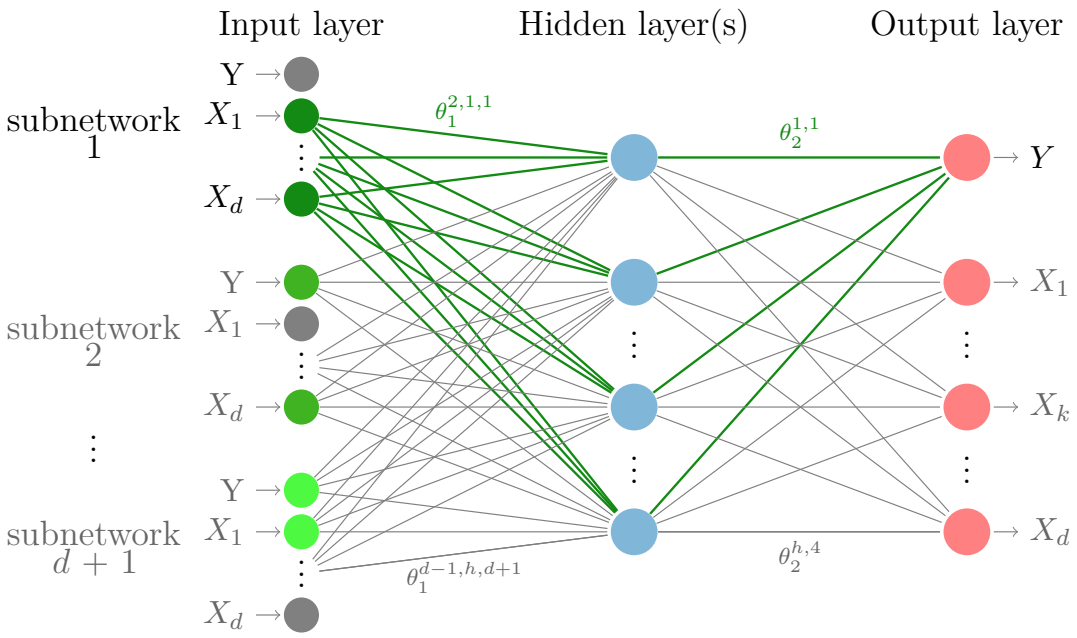


Figure 2.1: Joint NN Structure. Green arrows refer to the highlighted subnetwork 1, predicting Y (Output layer), while having Y masked (grayed out in Input layer). Biases are omitted for clarity of presentation.

$X_1, \ldots, X_d, Y$) from its inputs (indicated by a grayed-out neuron in Figure 2.1) and must reconstruct that excluded variable. Thus, each subnetwork learns to infer one variable solely from the others, without direct access to the variable it aims to recover.

All subnetworks have $M+1$ layers, and differ only in their input and output layers. The intermediate layers (layers $2, \ldots, M$) are common across all subnetworks, enabling the model to efficiently capture inter-variable relationships. We refer to this design as the Joint NN.

During training, the weights of all layers are optimised to achieve the best performance across all subnetworks simultaneously, thereby encoding the structure of interactions among all variables. This design draws inspiration from multi-task learning, where shared representations enable related tasks to inform and improve each other (Maurer et al., 2016). In the context of CASTLE, this shared learning mechanism facilitates capturing common patterns across variable relationships, which enhances both predictive accuracy and allows the estimation of the causal structures.

As in standard NN training, optimisation employs back-propagation and SGD. However, CASTLE introduces additional components into the loss function to achieve its dual objectives of prediction and causal discovery. The total loss combines two key elements: the prediction loss and the *DAG loss*. The prediction loss evaluates the accuracy of the predicted target, using Mean Squared Error (MSE) for regression tasks and cross-entropy for classification tasks.

The DAG loss, adapted from (Zheng et al., 2018), can be subdivided into three components:

- The **reconstruction loss**, calculated as the MSE for the outputs of each subnetwork, assuming they are continuous.
- The **acyclicity loss**, a regularisation term that penalises cycles in the graph structure. This is a function of the weight matrix and induces the resulting graph to be a DAG.
- An L_1 **sparsity loss** that encourages sparsity in the weight matrix, promoting simpler and more interpretable models.

This combination of prediction and reconstruction tasks allows CASTLE to balance predictive performance with structural regularisation (Kyono et al., 2020). However, the causal graph inferred by CASTLE is solely data-driven and derived directly from the weight matrix, resulting in limited control over the relations discovered and used by the NN for prediction. In this thesis, we address this limitation by focusing on the explicit construction of the weight matrix and its subsequent translation into an actual causal graph that can be both inspected and validated. The weight matrix is a crucial component of our Contestable NNs framework, introduced in Chapter 4, where it serves as the foundation for building interpretable and contestable NN models. Let us formally define the weight matrix and its associated *Weighted Adjacency Matrix* ($\mathbf{W}$), as used in CASTLE.

The sets of weights for each layer $l \in \{1, \ldots, M\}$ are denoted by Θ_l, where $\theta_l^{i,j}$ represents the individual weight connecting neuron i in layer l to neuron j in layer $l+1$. Additionally, $\theta_1^{i,j,k}$ refers to the weight from input neuron i in the first layer to neuron j in the first hidden layer of the k-th subnetwork. Then, the adjacency matrix $\mathbf{W}$ is defined as follows.

Definition 2.4.1 (Weighted Adjacency Matrix)**.** *Given a NN with weights Θ, where $h = |\Theta_2|$, $j = 1, \ldots, h$, and $i, k \in \{1, \ldots, d+1\}$, the (i,k)-th entry of $\mathbf{W}$ is defined as:*

$$w_{ik} = \sqrt{\sum_{j=1}^{h} \left(\theta_1^{i,j,k}\right)^2}.$$

Hence, $\mathbf{W}$ is a square, hollow matrix of order $d+1$, holding a non-negative weight for each variable, including the target, in relationship to all others. Each entry, w_{ik}, for $i, k \in \{1, \ldots, d+1\}$, is calculated as the square root of the sum of the squared input layer weights across all hidden neurons.

For standardised input data, each entry $w_{ik} \in \mathbf{W}$ represents the magnitude of the effect of variable i on variable k within the NN. The causal interpretation of $\mathbf{W}$ relies on the *acyclicity* term in the

loss function, which induces $\mathbf{W}$ to approximate a DAG. However, within the CASTLE framework, acyclicity is not strictly enforced but merely encouraged through the loss function. This means that if the acyclicity loss is not exactly 0, the resulting structure may still contain cycles. This limitation contrasts with the Contestable NNs framework proposed in this thesis, where acyclicity is strictly enforced to ensure a valid DAG.

To illustrate these limitations, we revisit Example 2.1.1. The weighted adjacency matrix, constructed according to Definition 2.4.1, does not directly produce a DAG that faithfully represents the causal relationships in the data. However, the resulting graph remains a close approximation to the underlying causal structure.

Example 2.4.2. *Consider the following weighted adjacency matrix as output by CASTLE when fitted to synthetic data generated from the DAG in Example 2.1.1.*

	Income	Education	Race	Occupation
Income		0.05	0.02	0.07
Education	0.11		0.00	0.15
Race	0.06	0.00		0.12
Occupation	0.21	0.08	0.07	

The matrix illustrates the magnitude of the relationships learned from the data, where each cell w_{ik} quantifies the influence between variables. Higher values indicate stronger connections. For instance:

- *The strongest dependency is observed between* Occupation *and* Income *(0.21), with* Occupation *exerting a significant influence on the prediction of* Income.
- *The effect of* Education *on* Income *is both direct (0.11) and indirect, through* Occupation *(0.15). We know that* Education *is confounding the relationship between* Occupation *and* Income *in the true causal graph.*
- *The indirect effect of* Race *on* Income *is captured as if it were direct (0.06), although the effect on* Income *is mediated through* Occupation *(0.12) in the true causal graph.*

While the adjacency matrix provides some insights into the relationships between variables, interpreting it as a (causal) DAG introduces challenges. A thresholding process is often applied to convert a weighted matrix into a binary adjacency matrix, identifying the presence or absence of edges (Zheng et al., 2018). This process is not part of the CASTLE framework, which focuses on weight regularisation for predictive performance. As a result, the inferred graph structure may contain cycles, rendering it unsuitable for causal interpretation.

Threshold selection can significantly impact the resulting graph structure, introducing a hyperparameter that requires careful tuning. For example:

- *A threshold of 0.1 recovers the true causal structure.*
- *A threshold of 0.2 only retains the edge between* Occupation *and* Income*, missing 3 edges.*
- *A threshold of 0.05 introduces 4 spurious edges, and as many cycles.*

Thus, while CASTLE can produce weighted adjacency matrices that reflect variable relationships, its inability to directly enforce acyclicity or ensure valid causal graphs limits its utility for causal reasoning.

These limitations highlight the need for systematic approaches to translate weighted adjacency matrices into valid, inspectable DAGs that accurately represent causal relationships. In Chapter 4, we address these challenges through Contestable NNs. Our proposed framework enables the extraction of meaningful and verifiable causal graphs from NN weights, fostering contestability and interpretability through expert engagement.

Having explored the integration of causal discovery with NNs, underpinning the contestation algorithm in Chapter 4, we now shift our focus to the methods used to develop transparent and robust causal discovery frameworks in Chapters 5 and 6: Shapley values and ABA.

2.5 Shapley Values

Understanding how each variable influences an outcome is fundamental to developing interpretable and trustworthy models. The Shapley value, a concept rooted in cooperative game theory (Shapley, 1953) and later adapted for ML explainability (Lundberg and Lee, 2017), is a well-established method for quantifying and attributing contributions to individual variables.

The Shapley value provides a fair method for distributing the total value generated by a team among its individual players, based on their unique contributions. It captures the marginal contribution of a player $i \in \mathbf{N}$ when joining any potential coalition $\mathbf{S}$, averaged across all possible combinations to build $\mathbf{S}$. Contributions are weighted by the likelihood of each coalition forming (Shapley, 1953).

Consider a team $\mathbf{N}$ of players working together to achieve a collective value $v(\mathbf{N})$. The function v assigns a real number $v(\mathbf{S})$ to any coalition of players. Formally:

Definition 2.5.1 (Shapley Value)**.** *Let* $\mathbf{N} = \{1, \ldots, n\}$ *be a set of players,* $\mathbf{S} \subseteq \mathbf{N}$*, and* $v : \mathbf{S} \mapsto \mathbb{R}$ *a value function. The Shapley value* $\phi_v(i)$ *of player* $i \in \mathbf{N}$ *is defined as:*

$$\phi_v(i) = \sum_{\mathbf{S} \subseteq \mathbf{N} \setminus \{i\}} \frac{|\mathbf{S}|!(n - |\mathbf{S}| - 1)!}{n!} \left[v(\mathbf{S} \cup \{i\}) - v(\mathbf{S}) \right].$$

The weighting factor $\frac{|\mathbf{S}|!(n-|\mathbf{S}|-1)!}{n!}$ represents the probability of a coalition of size $|\mathbf{S}|$ forming, accounting for all possible orderings of players. This ensures that every coalition is treated equitably, and the contribution of a player is evaluated independently of the coalition's composition.
The Shapley value satisfies four key properties (Shapley, 1953):

- **Symmetry:** $\phi_v(i) = \phi_v(j)$ if $v(\mathbf{S} \cup \{i\}) = v(\mathbf{S} \cup \{j\})\ \forall\ \mathbf{S} \subseteq \mathbf{N} \setminus \{i, j\}$.
- **Efficiency:** $\sum_{i \in \mathbf{N}} \phi_v(i) = v(\mathbf{N}) - v(\{\})$.
- **Additivity:** $\phi_{u+v}(i) = \phi_u(i) + \phi_v(i)$ for any value functions u, v.
- **Nullity:** $\phi_v(i) = 0$ if $v(\mathbf{S} \cup \{i\}) = v(\mathbf{S})\ \forall\ \mathbf{S} \subseteq \mathbf{N} \setminus \{i\}$.

These axioms ensure fair and consistent credit attribution. **Symmetry** guarantees that players contributing equally to all coalitions receive identical Shapley values, while **Efficiency** ensures that the total value created is fully distributed among all players. **Additivity** allows for consistent attribution when combining value functions, and **Nullity** assigns a Shapley value of zero to players who contribute nothing to any coalition. Due to these properties, Shapley values are widely regarded as a fair and robust solution for credit attribution (Young, 1985).

To illustrate their application, we revisit the causal graph from Example 2.1.1, demonstrating how Shapley values can quantify contributions in an ML context, as proposed by Lundberg and Lee (2017).

Example 2.5.2 (Shapley Value Calculation for Income Prediction)**.** *Consider again the causal graph from Example 2.1.1, and the task of predicting the likelihood of changing* Income *based on the input variables* Education *(E),* Race *(R), and* Occupation *(O). Shapley values quantify the marginal contribution of each variable to this prediction by evaluating the model's output across different subsets of inputs.*

Let the value function v *for each subset* $\mathbf{S} \subseteq \{E, R, O\}$ *represent the predicted probability of achieving a higher* Income*, with values in the range* $[0, 1]$*. These values are:*

$$v(\emptyset) = 0.5, \quad v(\{E\}) = 0.7, \quad v(\{R\}) = 0.65, \quad v(\{O\}) = 0.6,$$

$$v(\{E,R\}) = 0.75, \quad v(\{E,O\}) = 0.72, \quad v(\{R,O\}) = 0.68, \quad v(\{E,R,O\}) = 0.8.$$

Using these values, the Shapley value for E is computed as follows:

$$\phi_v(E) = \sum_{\mathbf{S}\subseteq\{R,O\}} \frac{|\mathbf{S}|!(3-|\mathbf{S}|-1)!}{3!} \left[v(\mathbf{S}\cup\{E\}) - v(\mathbf{S})\right].$$

Calculating the marginal contributions of variable E for each subset $\mathbf{S}$*:*

- *For* $\mathbf{S} = \emptyset$*, contribution:* $v(\{E\}) - v(\emptyset) = 0.7 - 0.5 = 0.2$*, weighted by* $\frac{1}{3}$*.*
- *For* $\mathbf{S} = \{R\}$*, contribution:* $v(\{E,R\}) - v(\{R\}) = 0.75 - 0.65 = 0.1$*, weighted by* $\frac{1}{6}$*.*
- *For* $\mathbf{S} = \{O\}$*, contribution:* $v(\{E,O\}) - v(\{O\}) = 0.72 - 0.6 = 0.12$*, weighted by* $\frac{1}{6}$*.*
- *For* $\mathbf{S} = \{R,O\}$*, contribution:* $v(\{E,R,O\}) - v(\{R,O\}) = 0.8 - 0.68 = 0.12$*, weighted by* $\frac{1}{3}$*.*

Summing the weighted contributions gives:

$$\phi_v(E) = \frac{1}{3}(0.2) + \frac{1}{6}(0.1) + \frac{1}{6}(0.12) + \frac{1}{3}(0.12) = 0.143.$$

Hence, the Shapley value for E is approximately 0.143*. Similarly, the Shapley values for the other variables are:*

$$\phi_v(R) = 0.1, \quad \phi_v(O) = 0.157.$$

The Shapley value calculation in this example demonstrates that *Occupation* contributes the most to predicting *Income*, with a Shapley value of 0.157, followed by *Education* and *Race*. These results reflect the relative importance of each input variable, highlighting that *Occupation* is the strongest predictor. By fairly attributing each variable's contribution, Shapley values provide interpretable insights into how input factors influence the prediction, enabling informed analysis of model behaviour.

2.6 Assumption-Based Argumentation (ABA)

Chapter 6 explores how knowledge representation techniques offer an alternative to purely statistical methods for causal discovery. These techniques provide structured frameworks for representing complex relationships, integrating domain expertise, and handling inconsistencies in data (Brachman and Levesque, 2004). Among these, *Computational Argumentation* has emerged as a formal model

for reconciling differences in non-monotonic reasoning and constructing interpretable models and explanations (Dung, 1995; Toni, 2014).

Argumentation Frameworks (AFs) are particularly valuable for causal discovery due to their ability to systematically evaluate and resolve conflicts, a frequent challenge when dealing with diverse and noisy data sources. Furthermore, their transparency supports the construction of coherent and interpretable models, even in the presence of uncertainty (Amgoud and Prade, 2009; Krause et al., 1995).

The two most prominent frameworks within computational argumentation are Dung's Abstract Argumentation (Dung, 1995) and ABA (Bondarenko et al., 1997). Abstract Argumentation considers arguments as abstract entities, focusing on their relationships without regard to internal structure. ABA, on the other hand, extends this approach by incorporating assumptions and rules, allowing arguments to be explicitly defined and enabling more detailed reasoning processes.

This thesis leverages ABA as a formal mechanism for managing inconsistencies and combining insights from both data and expert knowledge in the construction of causal graphs. Its structured approach facilitates the representation and resolution of conflicts inherent in causal reasoning (Alrajeh et al., 2020). By integrating ABA into causal discovery, this work seeks to derive interpretable causal graphs that align with both empirical evidence and domain expertise, enabling reliable and transparent decision-making.

Let us introduce the basics of ABA necessary to understand our proposed Causal ABA (Chapter 6). A more detailed account can be found in (Cyras et al., 2017).

We assume a deductive system $(\mathcal{L}, \mathcal{R})$, where $\mathcal{L}$ is a formal language, i.e., a set of sentences, and $\mathcal{R}$ is a set of rules defined over $\mathcal{L}$. A rule $r \in \mathcal{R}$ is expressed in the form $a_0 \leftarrow a_1, \ldots, a_n$ where each $a_i \in \mathcal{L}$. The head of the rule is $head(r) = a_0$, and the body is $body(r) = \{a_1, \ldots, a_n\}$. Formally:

Definition 2.6.1 (ABAF)**.** *An ABA Framework (ABAF) is a tuple* $(\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$*, where* $(\mathcal{L}, \mathcal{R})$ *is a* deductive system*,* $\mathcal{A} \subseteq \mathcal{L}$ *is a set of* assumptions*, and* $^- : \mathcal{A} \mapsto \mathcal{L}$ *is a function mapping each assumption* $a \in \mathcal{A}$ *to a sentence in* $\mathcal{L}$ *(contrary function).*

A sentence $q \in \mathcal{L}$ is *tree-derivable* from $S \subseteq \mathcal{A}$ and rules $R \subseteq \mathcal{R}$, denoted as $S \vdash_R q$, if there exists a finite rooted labelled tree T satisfying the following conditions:

- The root of T is labelled with q;
- The set of labels for the leaves of T is equal to S or $S \cup \{\top\}$;
- For every inner node v of T, there exists a rule $r \in R$ such that:

- v is labelled with $head(r)$,
- The number of successors of v is $|body(r)|$,
- Each successor of v is labelled with a distinct $a \in body(r)$ or $\top$ if $body(r) = \emptyset$.

When the rule set R is clear from the context, we simplify notation and write $S \vdash q$ instead of $S' \subseteq S \vdash_R q$.

The contrary function captures the negation of an assumption, allowing for the representation of conflicting information, formalised by the notion of *attack*.

Definition 2.6.2 (Attack)**.** *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ be an ABAF. A set $S \subseteq \mathcal{A}$* attacks *$T \subseteq \mathcal{A}$ if there exists $S' \subseteq S$ and $a \in T$ such that $S' \vdash \overline{a}$. A set S* defends *T if it attacks every attacker of T.*

With a slight notational abuse, we say that a set $S \subseteq \mathcal{A}$ attacks an assumption $a \in \mathcal{A}$ if S attacks the singleton $\{a\}$, and we define $\overline{S} = \{\overline{a} \mid a \in S\}$.

A fundamental property in ABA is the *closure* of assumption sets. A set $S \subseteq \mathcal{A}$ in an ABAF D is *closed* if $S \vdash a$ implies $a \in S$, meaning all assumptions derivable from S are contained in S. This property ensures consistency and completeness within the reasoning framework.

Building on closure, *flatness* characterises the nature of assumption sets in an ABAF. Specifically, an ABAF is called *flat* if every set $S \subseteq \mathcal{A}$ is closed; otherwise, it is called *non-flat*. Flat ABAFs are easier to analyse due to their straightforward reasoning paths, whereas non-flat ABAFs may contain circular dependencies, where assumptions indirectly support each other, creating loops that complicate the analysis of their relationships. This distinction influenced the implementation of the causal discovery algorithm in Chapter 6, where we employed Answer Set Programming (ASP) and the `clingo` framework (Gebser et al., 2019) to handle such complexities.

ASP provides a declarative programming paradigm suitable for encoding and solving complex combinatorial problems inherent in non-flat ABAFs. As we will detail in Chapter 6, there is a mapping between stable models (solutions of ASP) and stable extensions (solutions of ABAFs). We leverage this mapping to implement our Causal ABA in `clingo`, a powerful solver for ASP programs, which enables efficient grounding of rules and the search for stable models, thereby facilitating the resolution of circular dependencies.

To reason effectively within ABAFs, it is crucial to identify sets of assumptions that are *admissible*. Admissibility builds on the notions of attacks and defence to define which sets of assumptions can coexist without internal conflicts and are capable of defending themselves against external attacks.

Definition 2.6.3 (Admissible Set)**.** *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ be an ABAF. A set $S \subseteq \mathcal{A}$ is* conflict-free *if it does not attack itself. S is* admissible *if it is conflict-free and defends itself, denoted as $S \in ad(D)$.*

Intuitively, *admissible* sets consist of assumptions that neither attack one another nor permit undefended attacks from external assumptions. These sets form the foundation for defining different ABA semantics, which provide criteria for determining which assumptions to accept or reject within an ABAF.

In ABA, *semantics* offer various perspectives for analysing ABAFs. *Extensions*—sets of assumptions acceptable under a given semantics—capture these perspectives. Below, we describe four key ABA semantics: grounded, complete, preferred, and stable (denoted as *gr*, *co*, *pr*, and *stb*, respectively).

Definition 2.6.4 (ABA Semantics)**.** *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ be an ABAF and $S \in ad(D)$. Then:*

- *$S \in co(D)$ if S contains every assumption it defends.*
- *$S \in gr(D)$ if S is $\subseteq$-minimal in $co(D)$.*
- *$S \in pr(D)$ if S is $\subseteq$-maximal in $co(D)$.*
- *$S \in stb(D)$ if S attacks each $\{x\} \subseteq \mathcal{A} \setminus S$.*

Each semantics reflects a different reasoning standpoint (Baroni et al., 2011):

- **Complete extensions** strengthen the admissible definition by including precisely the arguments that the set collectively defends.
- **Grounded extensions** are the $\subseteq$-minimal complete extension, the most conservative stand.
- **Preferred extensions** are the least conservative, aiming at maximising acceptance ($\subseteq$-maximal complete extension).
- **Stable extensions** adopt a more defensive approach, requiring that accepted assumptions attack all assumptions not included in the set.

These semantics are central to understanding and analysing ABAFs. By examining different semantics and their corresponding extensions, we can explore how varying degrees of conservatism and aggressiveness in accepting assumptions influence the resulting extensions/conclusions. Relationships between these semantics have been extensively studied, with findings such as $cf(D) \supseteq ad(D) \supseteq co(D) \supseteq pr(D) \supseteq stb(D)$, as detailed in the comprehensive overview provided by Cyras et al. (2017), which highlights their implications for reasoning within ABAFs.

In Chapter 6, we leverage these semantics to develop a method for reasoning with causal structures and statistical relationships in the context of causal discovery.

Graphical ABA Representation To visualise the attack structure between assumptions in the ABAFs introduced in Chapter 6, we employ *AFs with Collective Attacks (SetAFs)* (Nielsen and Parsons, 2006). SetAFs generalise Dung's abstract AFs by allowing sets of arguments to collectively attack other arguments, making them particularly well-suited for representing the attack relations in ABAFs, as demonstrated by König et al. (2022).

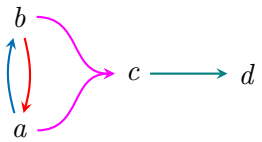
Formally, an AF with Collective Attacks (SetAF) is a pair (A, R), where A is a set of arguments and $R \subseteq 2^A \times A$ is an attack relation. The attack relation specifies that a set of arguments $S \subseteq A$ attacks an argument $a \in A$ if $(S, a) \in R$. An ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ can be instantiated as a SetAF by setting $A = \mathcal{A}$ and defining the attack relation R as follows: $S \subseteq A$ attacks $a \in A$ if $S \vdash \overline{a}$. This representation captures the underlying interactions and dependencies between assumptions in a graphical manner, enabling intuitive reasoning about conflicts and defences (König et al., 2022).

Let us consider an example ABAF to explore the extensions derived from its attack structure and illustrate the utility of SetAFs.

Example 2.6.5 (Collective Attacks and Extensions)**.** *Consider an ABAF* $(\mathcal{A}, \mathcal{R}, ^-)$ *with assumptions* $\mathcal{A} = \{a, b, c, d\}$*, contraries* $\overline{a} = p$*,* $\overline{b} = q$*,* $\overline{c} = r$*,* $\overline{d} = s$*, and rules:*

$$p \leftarrow b; \quad q \leftarrow a; \quad r \leftarrow a, b; \quad s \leftarrow c.$$

This ABAF can be represented as a SetAF, capturing the attack relations between assumptions:



- $\{a, b\}$ *collectively attacks* c *as* $\{a, b\} \vdash r$*;*
- a *attacks* b *as* $\{a\} \vdash q$*;*
- b *attacks* a *as* $\{b\} \vdash p$*;*
- c *attacks* d *as* $\{c\} \vdash s$*.*

Next, we examine the extensions derived from this ABAF under different semantics:

- ***Conflict-Free Sets:*** *These include all sets of assumptions that do not attack each other:* $\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a, d\}, \{b, d\}, \{a, c\}, \{b, c\}$*. Note that sets such as* $\{a, b\}$ *are excluded because* a *and* b *attack each other.*

- ***Admissible Extensions:*** *Admissible sets are conflict-free and defend themselves against attacks:* $\emptyset, \{a\}, \{b\}, \{a, c\}, \{b, c\}$*. For instance,* $\{d\}$ *is not admissible because it is attacked by* c *and does*

not defend itself. $\{a, d\}$ *and* $\{b, d\}$ *are not admissible because they do not defend* d *against the attack from* c*.*

- ***Complete Extensions:*** *Complete extensions are conflict-free sets that include exactly the arguments they defend:* $\emptyset, \{a, c\}, \{b, c\}$*. The empty set* $\emptyset$ *is a complete extension because all arguments are attacked, leaving no opportunity to defend all arguments in the set while staying conflict-free. Sets like* $\{a\}$ *and* $\{b\}$ *are not complete because they defend* d *but do not include it.*

- ***Grounded Extension:*** *The grounded extension is the* $\subseteq$*-minimal complete extension:* $\emptyset$

- ***Preferred Extensions:*** *Preferred extensions are the* $\subseteq$*-maximal complete extensions:*$\{a, c\}, \{b, c\}$

- ***Stable Extensions:*** *Stable extensions are admissible, and attack all assumptions outside the extension:* $\{a, c\}, \{b, c\}$

This example demonstrates how different semantics yield varying sets of accepted assumptions. Conflict-free sets avoid internal attacks, admissible extensions defend against external attacks, complete extensions fully incorporate defended assumptions, and grounded extensions represent the most cautious approach. Preferred extensions, on the other hand, strive for maximal acceptance, while stable extensions require that every assumption outside the set is attacked, making it the most aggressive form of defence.

As detailed in Chapter 6, we use ABA to reason about observed (in)dependence relations and causal relations extracted from data or provided by experts. This mix of statistical and graphical relations serve as assumptions, which we combine with d-separation rules to infer causal relationships.

Stable semantics play a key role in our work as they identify the strongest possible positions in an ABAF, aligning with the need for robust decision-making in causal discovery. It is worth noting that stable extensions may be absent in some cases due to their strict requirements. However, in some cases also multiple stable extensions may exist, reflecting the non-uniqueness of stability.

Another reason for focusing on stable semantics is their correspondence with stable semantics in Logic Programming (LP) (Caminada et al., 2015; Schulz and Toni, 2015), recently extended to non-flat instances (Rapberger et al., 2024). In standard ABA-LP translations, assumptions are linked to default negated contraries, e.g., an assumption $a \in \mathcal{A}$ with contrary $a_c \in \mathcal{L}$ corresponds to the default negated literal *not* a_c. Our framework extends this approach by leveraging ABA's expressive power while utilising efficient LP tools for reasoning with non-flat ABAFs. This enables the implementation and

evaluation of our causal discovery algorithm, using existent LP solvers to calculate semantics for our Causal ABAFs.

2.7 Summary

In this chapter, we provided an overview of the foundational concepts necessary to understand the methods presented in this thesis. These concepts included causal graphs, causal discovery, regularised NNs, Shapley values, and ABA. Building on this foundation, the following chapters introduce novel methods that integrate these concepts: injecting and contesting NNs using causal graphs (Chapter 4), leveraging Shapley values for causal discovery (Chapter 5), and employing ABA for causal discovery (Chapter 6). Before delving into these contributions, we first review the relevant literature in Chapter 3 to contextualise our work within the broader research landscape.

Chapter 3

Related Work

This chapter reviews three key areas foundational to this thesis: **Contestable AI**, **Knowledge Injection**, and **Causal Discovery**. We position this work within the existing literature, highlighting critical gaps that motivate our research.

We begin with **Contestable AI** (§3.1), examining frameworks and challenges in enabling stakeholders to scrutinise and refine AI systems. Special attention is given to operationalisation and redress. Next, we explore **Knowledge Injection** (§3.2) particularly as a mechanism to support contestability and stakeholder-driven redress. Finally, we turn to **Causal Discovery** (§3.3), providing a high-level overview of the field before delving into its three key methodological families: continuous score-, constraint-, and logic-based approaches. Within logic-based methods, we investigate the role of Computational Argumentation as a framework for enhancing explainability while advancing causal discovery. Together, these intersecting areas frame the challenges and opportunities addressed by this thesis in advancing trustworthy AI systems.

3.1 Contestable AI

Contestable AI emphasises the importance of enabling stakeholders to interact with, challenge, and revise Automated Decision Systems (ADSs) for a safe and equitable deployment AI. Contestability is unlocked by transparency and fosters accountability—core principles of trustworthy AI. This section examines contestability through the lenses of the framework established by Leofante et al. (2024), which provides the foundation for the methodologies proposed in Chapter 4 to operationalise contestability for NNs. The choice of this framework reflects its comprehensive formalisation of contestable AI, addressing key dimensions such as contested and contesting entities, as well as methods of engagement.

Contested Entities Contestability in ADSs spans various aspects, including:

- **System design**, focusing on ADSs' structure (Alfrink et al., 2023b; Almada, 2019; Yurrita et al., 2023b).
- **Training data**, addressing the provenance and fairness of data (Kluttz et al., 2020).
- **Inputs**, scrutinising the data provided for specific predictions (Alfrink et al., 2023b).
- **Outputs**, examining the system's decisions (Henin and Métayer, 2022; Hicks, 2022; Hirsch et al., 2017; Lyons et al., 2021; Ploug and Holm, 2020; Tubella et al., 2020).
- **Entire ADSs**, engaging with the holistic operation of the system (Alfrink et al., 2023a; Lyons et al., 2021; Vaccaro et al., 2021).

Existing research on contestability predominantly focuses on static entities, such as outputs or inputs, rather than entire ADSs. While the literature acknowledges the need for contesting ADSs as a whole, operationalising such contestability remains largely unexplored, with the exception of (Ioannou and Michael, 2024), which applies argumentation to dialogues for arbitration purposes. This thesis addresses these gaps by operationalising contestability for NNs as an example of a complete ADS.

Our approach in Chapter 4 enables a dynamic, iterative two-way learning process facilitated by causal graphs. This allows experts to contest the NNs while simultaneously addressing gaps in their own knowledge through learned relationships. Although the contestation of experts by the ADS is not supported—a limitation noted in (Leofante et al., 2024)—the collaborative interaction fostered by our methodology promotes robust, interpretable, and adaptable AI systems.

Contesting Entities The literature identifies several types of contesting entities, including:

- **Decision subjects**, individuals affected by ADS decisions, who benefit from simplified explanations and accessible interactions (Almada, 2019; Henin and Métayer, 2022; Hicks, 2022; Lyons et al., 2021; Ploug and Holm, 2020).
- **Third-party representatives or collective actions**, enabling group-based advocacy and oversight (Alfrink et al., 2023b; Lyons et al., 2021).
- **Subject-Matter Experts (SMEs) and regulators**, whose technical expertise is critical for contesting decisions in high-stakes domains (Alfrink et al., 2023a; Kluttz et al., 2020; Tubella et al., 2020).

In Chapter 4, we prioritise SMEs as contesting entities due to their ability to provide technical scrutiny and domain-specific insights. Our methodology equips SMEs with causal graphs learned by the NN, enabling them to challenge and refine model relationships during training. By integrating expert input, our approach introduces a novel form of **Human-in-the-loop (HITL)** debugging, addressing calls for more meaningful and impactful contestation methods (Henin and Métayer, 2022; Leofante et al., 2024). Our methodology is further motivated by the value of human input in shaping robust and trustworthy AI systems (Kluttz et al., 2020).

Although Alfrink et al. (2023b); Lyons et al. (2021) highlight the importance of enabling collective actions, our focus is on individual SMEs or small groups that can reach consensus during the iterative process that refines the model relations. Arbitration methods, such as those proposed by Ioannou and Michael (2024), offer a promising direction for extending our methodology to support group-based contestation in the future.

Contestation Methods The literature highlights various methods supporting contestation, including participatory design (Alfrink et al., 2023a; Almada, 2019; Vaccaro et al., 2021), operation logs (Hicks, 2022), and model explanations (Almada, 2019; Kluttz et al., 2020; Ploug and Holm, 2020). Explanations are particularly crucial for contestability as they enable stakeholders to challenge decisions beyond examining outputs alone (Leofante et al., 2024; Lyons et al., 2021).

Among explanation methods, *process-centric explanations* (Yurrita et al., 2023b) stand out by providing context-specific justifications tied to the reasoning behind decisions. These explanations explicitly connect inputs, conditions, and mechanisms driving outcomes, enabling stakeholders to engage meaningfully with the decision process. In high-stakes domains, such as healthcare or finance, process-centric explanations are particularly valuable for fostering critical debate and iterative feedback. By contrast, post-hoc methods like counterfactual explanations (Venkatasubramanian and Alfano, 2020) often lack the depth to reveal underlying reasoning or support dynamic engagement.

In Chapter 4, we propose causal graphs as a process-centric explanation method. Causal graphs offer a structured and context-aware representation of a NN's learned relationships, mapping how the model interprets data relative to real-world causal mechanisms. By clearly highlighting the processes influencing the NN's behaviour, causal graphs allow experts to scrutinise, challenge, and refine model assumptions. This approach produces actionable and interpretable insights, directly addressing the need for process-driven contestability (Yurrita et al., 2023b). Additionally, we argue that computational argumentation's intrinsic characteristics naturally supports contestability (Leofante et al., 2024).

Post-hoc and Ex-ante Contestability Post-hoc contestability involves retrospectively reviewing ADS decisions, often in regulatory contexts (Lyons et al., 2021; Tubella et al., 2020). By contrast, ex-ante contestability is a design principle that enables stakeholders to interactively contest decisions during system operation (Alfrink et al., 2023b; Hirsch et al., 2017). Both paradigms address different stages of contestability but are rarely integrated into a cohesive framework.

Our proposed contestation framework in Chapter 4 bridges these two paradigms. It enables contestation after the NN training, conceptualised as a form of post-hoc contestability that allows stakeholders to challenge learned model relations. Simultaneously, it embodies ex-ante contestability by guiding the development of improved ADSs through contestation during the learning process. This dual focus facilitates a more comprehensive contestation framework that supports both retrospective analysis and proactive system refinement.

Interactivity and Iteration Interactive contestation is critical for fostering engagement with ADS decisions at both development and deployment stages. Kluttz et al. (2020) highlight how interactive approaches, particularly in high-stakes domains such as healthcare, enable professionals to better understand and influence algorithmic decisions. Interactivity empowers stakeholders to scrutinise decisions following their intuition and reasoning. When coupled with access to faithful explanations and contextual information, this trait closely aligns with the principles human agency and accountability. Furthermore, iterative interaction facilitates the integration of human knowledge and preferences into the training process, a concept widely supported in prior research (Hirsch et al., 2017; Leofante et al., 2024; Lyons et al., 2021).

Our contestation methodology in Chapter 4 operationalises this principle by enabling iterative refinement of model relations until alignment is achieved between the NN and expert knowledge. This iterative and interactive framework supports a collaborative process for building robust and trustworthy systems, while encouraging experts' accountability.

Redress A significant gap in the literature is the lack of mechanisms for redress in contestable AI (Fanni et al., 2023; Ogunleye, 2022). While some works propose normative reasoning (Tubella et al., 2020) or process modelling (Kluttz et al., 2020) to support contestation, few address corrective actions for ADS decisions that fail to align with stakeholder expectations. The absence of formal methodologies for addressing such misalignments limits the practical utility of contestable AI frameworks.

To address this gap, we propose a novel algorithm for *Causal Graph Injection* into NNs, introduced in Chapter 4. This algorithm provides a principled approach to redressing learned model relations that

deviate from expert knowledge. By leveraging causal links within a causal graph, our method ensures that model corrections are transparent and grounded in domain expertise. This approach advances the principles of transparency and accountability, offering a practical mechanism for aligning model behaviour with stakeholder expectations.

With this foundation, we now turn to the literature on Knowledge Injection to contextualise and further motivate our approach to redress NN relations through causal graphs.

3.2 Knowledge Injection

Knowledge injection involves integrating domain expertise into ML models to improve their interpretability, robustness, and performance. von Rüden et al. (2023) provide a comprehensive taxonomy for *Informed ML*, categorising knowledge injection along three dimensions: **knowledge source**, **representation**, and **integration**. These dimensions capture the *what* and *how* of knowledge injection, while the authors also address the *why*, emphasising motivations such as reducing data requirements, improving interpretability, and enhancing performance. Our work aligns with this taxonomy under the following categories:

- **Knowledge source**: Expert knowledge, specifically causal relationships provided by SMEs
- **Representation**: Human feedback, specifically on causal relationships, represented as graphs
- **Integration**: Learning algorithms that incorporate causal graphs into NN training

We extend both the *why* and *how* of knowledge injection by introducing a novel methodology: **knowledge injection empowered by causal discovery**. This approach not only enhances causal understanding and predictive performance, but also enables contestability by allowing stakeholders to scrutinise and refine model relations.

Knowledge Injection for Contestability Incorporating feedback into models is essential for contestability, as it provides a mechanism for redress. While contestability often focuses on data subjects (e.g., end-users contesting decisions) (Almada, 2019), extending it to include technical experts and professionals has been strongly advocated (Henin and Métayer, 2022). Existing methods range from structured interactions and model explanations (Kluttz et al., 2020) to normative reasoning and process modelling (Tubella et al., 2020). However, the nature of interactions must be tailored to the expertise of the user (Henin and Métayer, 2022). For instance, data subjects require simplified

explanations, while technical experts benefit from detailed insights, such as causal graphs, feature attributions or counterfactual explanations, to scrutinise and refine model behaviour.

Our approach enables technical experts to contest a NN model at any point during its development cycle. By leveraging causal graphs, experts can validate or disable spurious relationships that the model learns. This form of HITL debugging goes beyond traditional HITL tasks, such as annotation (Wu et al., 2022), by directly influencing the learning process itself. For example, Lertvittayakumjorn et al. (2020) propose *FIND*, a method where practitioners disable spurious filters in convolutional NNs using word clouds. In contrast, our methodology refines connections between features through causal graphs, enabling experts to guide model behaviour with a deeper understanding of its internal relations, as compared to what experts expect in the real world. This represents a distinctive and advanced form of HITL interactive model development, improving both transparency and robustness.

3.2.1 Causal Knowledge Injection

Integrating expert knowledge into causal discovery is a longstanding challenge (Meek, 1995). We show that this integration is critical for enabling HITL debugging, fostering contestability, and aligning NN models with domain-specific insights. In the broader causality literature, the integration of expert knowledge has been a central theme, with the modelling process often starting from a set of causal assumptions or constraints (Pearl and Mackenzie, 2018). However, the integration of expert knowledge and data-driven causal discovery remains an open research question.

Early work by Meek (1995) introduced background knowledge into causal discovery tasks, a concept advocated since the inception of structural causal models (Pearl, 1989). Constantinou et al. (2023), evaluate the impact of incorporating prior knowledge into causal discovery across multiple algorithms, while Chowdhury et al. (2023) specifically analyse this effect for a specific causal discovery algorithm. These studies highlight the importance of integrating expert knowledge into causal discovery, particularly in high-stakes domains where domain-specific insights are critical (Rudin, 2019).

Our Contestable NNs build on these insights by enabling SMEs to iteratively refine causal structures, representing the NNs' "view of the data" during training. This differs from CASTLE (Kyono et al., 2020), which focuses on regularising the NN's weights to induce acyclicity, without visibility or constraints over the discovered causal structures. Our approach extends this by enabling direct contestation of the learned relationships, placing SMEs in control of the model's causal structure.

The deep learning literature offers various strategies for integrating domain knowledge into NNs, modifying loss functions and adjusting hyperparameters (Borghesi et al., 2020), or, introducing inductive

biases to approximate structural assumptions (Goyal and Bengio, 2022). CASTLE (Kyono et al., 2020) directly influence the learning process, and could be seen as a form of inductive bias. However, we provide acyclicity guarantees as well as transparency and a structured approach to contesting the learned relationships, which is not addressed by CASTLE.

Other works address injecting causal knowledge as alternatives to CASTLE. For example, Geiger et al. (2022) propose *Interchange Intervention Training (IIT)*, which aligns NNs with counterfactual behaviours derived from causal models, while Zhang et al. (2020) introduce the *deep CAusal Manipulation Augmented model (CAMA)* to make NNs robust to input space manipulations.

Unlike these methods, which assume access to a complete causal model and rely on data augmentation, our approach:

- Focuses solely on the causal structure without making further assumptions about the DGP;
- Modifies model weights directly based on extracted causal graphs and guided by expert feedback;
- Actively involves humans in discovering causal relationships from the data.

This thesis introduces a methodology that integrates causal discovery with knowledge injection to facilitate contestability and improve the robustness of NNs. By enabling SMEs to contest the learned model relations, we provide a novel approach to HITL debugging and principled model refinement. This aligns with the broader objectives of trustworthy AI, fostering transparency and improving the robustness of NNs through contestability. Additionally, it facilitates stakeholder engagement and accountability. Let us now explore the broader landscape of causal discovery to contextualise our methodology within this field.

3.3 Causal Discovery

Causal discovery seeks to uncover the presence and direction of causal relationships between variables from observational data, revealing the underlying causal structure through statistical methods and graphical criteria (Glymour et al., 2019). Such structures are fundamental for constructing causal models, which integrate probabilities or functions to quantify the causal effects depicted in the graph (Pearl, 2009).

The field has developed around three main categories of approaches (Glymour et al., 2019; Guo et al., 2021; Spirtes and Zhang, 2016; Zanga et al., 2022): **constraint-based**, **score-based**, and **Functional Causal Model (FCM)-based** methods. Constraint-based methods rely on CITs and

graphical rules to construct causal graphs (Colombo and Maathuis, 2014; Spirtes et al., 2000). By contrast, score-based methods optimise a predefined score, such as Bayesian Information Criterion (BIC), to identify the best-fitting graph (Chickering, 2002b; Ramsey et al., 2017). Finally, FCM-based methods assume specific functional relationships between variables to infer causal graphs (Bühlmann et al., 2014; Peters et al., 2014).

Within the constraint-based paradigm, logic-based methods provide an alternative perspective, focusing on formal consistency when addressing conflicts arising from multiple and noisy data sources (Eberhardt, 2017). Similarly, continuous optimisation methods extend score-based approaches by incorporating gradient-based techniques from ML, enabling scalable and efficient discovery of causal graphs while satisfying acyclicity constraints (Vowels et al., 2022).

This section reviews these approaches and their extensions, focusing on three families of methods central to this thesis: **continuous optimisation methods**, leveraged in Chapter 4 to integrate causal discovery into predictive modelling tasks; **constraint-based algorithms**, forming the basis for the novel approach developed in Chapter 5; and **logic-based methods**, which underpin the argumentation-based framework proposed in Chapter 6. Comprehensive surveys such as (Eberhardt, 2017), (Glymour et al., 2019), (Vowels et al., 2022), (Guo et al., 2021), and (Zanga et al., 2022) provide further context for foundational methods, algorithmic scalability, and applications.

3.3.1 Score-based Methods

Score-based methods treat causal discovery as an optimisation problem, aiming to maximise a predefined score to identify the best-fitting graph. Greedy Equivalence Search (GES), introduced by Chickering (2002b), is a foundational method in this category, optimising the BIC when evaluating the addition or removal of edges, or changes in causal direction. Under the assumptions of faithfulness (independencies in the data reflect the true graph), acyclicity (causal graphs are non-recursive), and sufficiency (no latent confounders), GES has been shown to be sound and complete (Chickering, 2002b).

Scalable adaptations, such as Fast Greedy Search (FGS) (Ramsey et al., 2017), extend GES's applicability to high-dimensional data, while generalised scoring functions (Huang et al., 2018) broaden its scope to more complex functional forms. Extensions by Ng et al. (2021) and Rolland et al. (2022) relax the faithfulness assumption, enabling exact search under linear Gaussian and non-linear effects, respectively. These methods, while often classified as score-based, also assume specific functional forms of variable relationships, situating them near FCM-based approaches, discussed in §3.3.5. In Chapter 6, we use FGS as a baseline due to its availability and scalability.

Continuous optimisation methods extend traditional score-based approaches by reformulating the acyclicity constraint into a differentiable form, enabling the use of gradient-based optimisation techniques such as SGD. NOTEARS (Zheng et al., 2018), which initiated this line of research, introduced a differentiable acyclicity constraint, enabling efficient causal discovery in high-dimensional settings. Subsequent works have extended NOTEARS to non-linear settings (Lachapelle et al., 2020; Zheng et al., 2020) and addressed uncertainties in mapping continuous matrix outputs to binary structures (Ng et al., 2022). NOTEARS-MLP (Zheng et al., 2020), a baseline in Chapter 6, exemplifies the scalability and flexibility of these methods. Despite their advantages—including the relaxation of faithfulness assumptions—continuous optimisation methods lack guarantees of recovering the true causal structure, even in the sample limit, due to their stochastic nature (Vowels et al., 2022).

Integration of causal discovery with predictive modelling further extends the utility of these methods. For instance, CASTLE (Kyono et al., 2020) incorporates the acyclicity constraint as a regularisation term in its loss function, treating causal discovery as an auxiliary task to predictive modelling (§2.4). By employing a masking scheme to prevent self-loops, CASTLE ensures variables cannot predict themselves, though it does not eliminate other cycles. While CASTLE improves the predictive accuracy of NNs compared to methods like dropout (Srivastava et al., 2014), L1/L2 regularisation (Hastie et al., 2009), and batch normalisation (Ioffe and Szegedy, 2015), it focuses primarily on predictive performance rather than on causal discovery. Furthermore, it lacks transparency and a mechanism for incorporating expert feedback to modify learned relationships.

Other works, such as (Lachapelle et al., 2020) and (Ng et al., 2022), have applied masking schemes for structure recovery—targeting entire paths (Lachapelle et al., 2020) or leveraging matrix-wide probabilistic distributions (Ng et al., 2022). Building on these approaches, in Chapter 4, we propose a novel masking scheme for NN weights that integrates the benefits of CASTLE's predictive modelling with the structure-recovery focus of these methods. Our approach introduces an interactive mechanism for SMEs to refine both relationships and predictions, aligning discovery with domain expertise and advancing continuous score-based methods by fostering contestability and integrating expert input.

3.3.2 Constraint-based Methods

Constraint-based methods construct causal graphs by leveraging CITs and graphical rules derived from d-separation. A notable method in this category is the PC algorithm (Spirtes et al., 2000), which efficiently discovers causal relationships in large-scale settings (see §2.3 for details). Earlier algorithms, such as SGS (Spirtes et al., 2000) and Inductive Causation (IC) (Pearl, 1989), provided

strong theoretical guarantees but are computationally prohibitive. In contrast, the PC algorithm achieves computational efficiency, particularly for sparse graphs (Kalisch and Bühlmann, 2007), and is sound and complete under the assumptions of faithfulness and sufficiency (Spirtes et al., 2000).

However, the results of PC can be sensitive to variable ordering in finite samples, due to the iterative update of the graph throughout the search. This limitation is addressed by extensions such as PC-Stable (Colombo and Maathuis, 2014), which ensures order independence in the adjacency search (Step 1), by removing edges only after all variables have been tested for a given conditioning set size. This modification ensures that the computed skeleton is stable and invariant to the order in which variables are tested, addressing a key limitation of the original algorithm. Other works improve the efficiency or robustness of Step 1 of the PC algorithm. For instance, Tsagris (2019) propose to use heuristics to prioritise tests based on probabilistic measures, while Abellán et al. (2006) introduce Bayesian-inspired methods for edge selection and removal. For Step 1, our proposed Shapley-PC algorithm (Chapter 5) adopts the strategy of PC-Stable, which serves as a baseline for our empirical study.

In the v-structure orientation phase (Step 2), various decision rules have been developed to address the challenges of finite sample variability. Conservative-PC (CPC) (Ramsey et al., 2006) ensures rigorous orientation by requiring that all separating sets confirm adjacency faithfulness (see Definition 2.2.7), while Majority-PC (MPC) (Colombo and Maathuis, 2014) adopts a less stringent approach by orienting v-structures when the collider appears in fewer than half of the separating sets. PC-Max (Ramsey, 2016) selects the conditioning set with the maximum p-value to guide v-structure orientation, and ML4C (Dai et al., 2023) frames the problem as a supervised learning task, using synthetic data to predict orientations. CPC offers a conservative approach, MPC introduces a middle-ground rule, PC-Max is guided by statistical significance, and ML4C leverages ML to enhance decision-making.

Below, we give a detailed overview of the variants of PC closest to our proposed Shapley-PC (Chapter 5). These methods retain Step 1 (using either the original or stable version) and Step 3 of the PC algorithm and only differ from ours in the v-structure orientation rule. Notably, they carry out the same CITs in all phases, and hence serve as baselines for isolating the impact of our novel decision rule.

- **Conservative-PC (CPC)** (Ramsey et al., 2006): builds on the distinction between adjacency and orientation faithfulness (Ramsey et al., 2006), assuming adjacency faithfulness (i.e., edges are correctly identified, Definitions 2.2.7) and explicitly checking for orientation faithfulness (Definitions 2.2.8). Specifically, for a skeleton $\mathcal{C}$ and a UT $X_i - X_j - X_k$, the potential collider X_j is

deemed valid only if it is found in **none of the separating sets S**, with $\mathbf{S} \subseteq adj(\mathcal{C}, X_i) \setminus \{X_j\} \vee \mathbf{S} \subseteq adj(\mathcal{C}, X_k) \setminus \{X_j\}$. This additional check makes CPC more robust in scenarios where orientation faithfulness might not hold, adding a relatively small computational overhead (Ramsey et al., 2006), as confirmed by our the complexity analysis in Chapter 5.

- **Majority-PC (MPC)** (Colombo and Maathuis, 2014) relaxes the stringent orientation faithfulness check of CPC. For a skeleton $\mathcal{C}$ and a UT $X_i - X_j - X_k$, it orients X_j as a collider if it appears in **fewer than half of the separating sets S**, with $\mathbf{S} \subseteq adj(\mathcal{C}, X_i) \setminus \{X_j\} \vee \mathbf{S} \subseteq adj(\mathcal{C}, X_k) \setminus \{X_j\}$. This approach balances strict adherence to orientation faithfulness with the flexibility to handle data with limited consistency.
- **PC-Max** (Ramsey, 2016): focuses on the CIT with the **maximum p-value** for a given UT and the same conditioning sets analysed by CPC and MPC. PC-Max decision rule orients the UT as a v-structure only if the conditioning set for the selected CIT does not include the variable under consideration. This modification follows the insight that higher p-values indicate lower chances of dependency (Hung et al., 1997), aiming at minimising the likelihood of Type I error. Additionally, PC-Max includes a check before orienting a v-structure to avoid bi-directed edges.

PC-Max has been shown to outperform the other variants of the PC algorithm in terms of edge orientation accuracy, with no additional computational overhead compared to CPC and MPC (Ramsey, 2016). However, as with CPC, its reliance on a single independence test introduces brittleness into its decision rule, particularly in the presence of noisy or conflicting data. This highlights a broader limitation of existing constraint-based methods: their sequential structure can amplify the impact of erroneous independence tests, propagating inconsistencies throughout the estimated causal graph.

To address this limitation, we propose a novel Shapley-based decision rule to orient v-structures. By aggregating information across independence tests, our decision rule mitigates the reliance on single tests, providing a more robust foundation for causal discovery. The resulting Shapley-PC algorithm, introduced in Chapter 5, adopts the PC-Stable modification for Step 1 and incorporates the bi-directed edge check from PC-Max. This combination improves both reliability and interpretability, while still allowing for efficient computation. Additionally, the CITs performed during Shapley-PC are leveraged as external knowledge to instantiate the Causal ABA framework introduced in Chapter 6.

Despite these advancements, Shapley-PC shares a common limitation with other constraint-based methods: it does not guarantee consistency between the output DAG and the observed independence test results (Spirtes et al., 2000; Tsagris, 2019). This lack of consistency can lead to situations where the

estimated causal graph contradicts the statistical evidence, undermining robustness and interpretability. These limitations highlight the need for a structured approach to conflict resolution, particularly in settings where noisy or conflicting data are unavoidable.

Addressing these challenges requires moving beyond traditional constraint-based algorithms to methods that can formally reason about inconsistencies and ensure outputs that are both statistically and causally consistent. This motivates the introduction of *Causal ABA* in Chapter 6, which represents causal discovery as a structured set of rules and assumptions within an argumentative framework. By refining the outputs of Shapley-PC, Causal ABA resolves conflicts in independence test results while incorporating expert knowledge to enhance accountability and robustness.

Constraint-based methods have also been extended to handle latent confounder and cycles. The FCI algorithm (Colombo et al., 2012; Spirtes et al., 2000) relaxes the sufficiency assumption, producing Partial Ancestral Graphs (PAGs) that represent latent confounders with bi-directed edges. Variants such as FCI-Max (Raghu et al., 2018) adapt PC's decision rules to settings with latent confounders, while adaptations of FCI (Mooij and Claassen, 2020) and the CCD algorithm (Richardson and Spirtes, 1999) extend constraint-based methods to handle cyclic structures. While these extensions demonstrate the adaptability of constraint-based methods, addressing latent variables and cycles lies beyond the scope of this thesis. However, these challenges highlight promising future directions for enhancing the applicability of constraint- and logic-based algorithms.

By focusing on reliable decision rules and modularity, constraint-based methods provide a robust foundation for causal discovery. These strengths underpin the development of Shapley-PC and motivate its further refinement through logic-based methods, as explored in the next section.

3.3.3 Logic-based Methods

Causal discovery methods must address inconsistencies in statistical metrics caused by noisy or conflicting data. For example, the original PC algorithm assumes no inconsistencies, and once an independence is found between two variables, it will not test the pair again. However, erroneous results can lead to inconsistencies in the output graph, violating independence constraints (see Example 6.1.1 for an illustration).

Formal reasoning frameworks, such as Boolean Satisfiability (SAT), Answer Set Programming (ASP), and constraint or dynamic programming, tackle these challenges by deriving causal graphs that respect a set of constraints. These frameworks align with the argumentative principles proposed in Chapter 6, which prioritise conflict resolution and consistency in outputs, providing stronger guarantees

than constraint-based methods.

SAT-based methods like cSAT+ (Triantafilou et al., 2010) and COmbINE (Triantafillou and Tsamardinos, 2015) encode (in)dependencies as path constraints, leveraging SAT solvers to ensure sound and complete outputs in the sample limit. Extensions such as (Zhalama et al., 2017) implement these approaches in ASP, achieving significant efficiency gains. However, these methods primarily ensure consistency across datasets and do not address the acceptability of statistical and causal relations—central goals of our work in Chapter 6.

Hyttinen et al. (2013) extended SAT-based methods to support directed cycles, latent variables, and background knowledge, demonstrating their flexibility. ASPCR[1] (Hyttinen et al., 2014) builds on this foundation, encoding *conditioning* and *intervening* declaratively in ASP and optimising test results as weak constraints. While ASPCR excels at constraint optimisation, it does not prioritise transparency or stakeholder engagement. Scalability remains an issue, with support limited to six variables. Although Hyttinen et al. (2017) proposed a Max-SAT solver to improve efficiency, the absence of a public implementation limits its practical utility.

Logical inference methods further extend the capabilities of logic-based causal discovery. Logical Causal Inference (LoCI) (Claassen and Heskes, 2011) uses propositional rules to identify invariant features in PAGs under latent confounders or selection bias. By introducing minimal (in)dependence relations and translating them into logical statements, LoCI provides a conceptual framework for deducing PAG edges and directionality. However, its practical utility is limited due to the lack of an implementation. Additionally, LoCI does not address conflict identification or resolution, nor does it incorporate external knowledge into its reasoning process.

Building on LoCI, Ancestral Causal Inference (ACI) (Magliacane et al., 2016) extends logical characterisations to produce PAGs and encodes them in ASP. ACI supports external knowledge integration and benefits from an available implementation, making it more practical than LoCI. However, ACI also lacks mechanisms for transparent conflict resolution in its reasoning process. These gaps reduce its utility in scenarios where stakeholders require interpretable and contestable causal graphs.

Our Causal ABA framework builds upon these advancements by addressing the limitations of both ASPCR and ACI. Using ASP to calculate extensions, Causal ABA dynamically incorporates external knowledge while explicitly identifying and rejecting unacceptable inputs. This ensures both consistency and transparency in output causal graphs, addressing gaps in conflict resolution and interpretability.

[1]Acronym is ours, for ASP Causal Representation, as we use this method as a baseline in Chapter 6.

Unlike ASPCR and ACI, which focus primarily on constraint optimisation, our approach balances scalability, stakeholder engagement, and principled conflict resolution.

Other methods, such as ETIO (Borboudakis and Tsamardinos, 2016), tackle specific challenges like missing data to generate DAGs but similarly fall short in addressing conflicts or promoting transparency in reasoning. Meanwhile, Entner et al. (2013) and Watson and Silva (2022) focus on adjustment set selection or managing confounding variables. While these approaches underscore the adaptability of logic-based methods, they highlight the need for frameworks like ours, that extend their applicability to diverse causal discovery scenarios.

By leveraging the strengths of existing logic-based methods while addressing their limitations, Causal ABA offers a principled and scalable solution for conflict resolution and external knowledge integration. The next section explores how argumentation supports causal discovery and the implications for transparency and accountability in AI systems, contextualising our work in Chapter 6.

3.3.4 Argumentative Methods

Computational argumentation captures reasoning and decision-making processes by constructing arguments and evaluating their acceptability. Beyond its non-monotonic reasoning capabilities, it is inherently transparent and contestable—key attributes of trustworthy AI. An overview of argumentation, particularly ABA, is provided in §2.6. In the context of causal discovery, argumentation can be seen as a specific application of logical reasoning frameworks, which prioritise both consistency and transparency. This section highlights both their role in causal discovery and their broader contributions to Explainable AI (XAI), providing context for our use of argumentation in Chapter 6.

Argumentation for Causal Discovery The first application of argumentation to causal discovery was proposed by Bromberg and Margaritis (2009), marking the introduction of logic-based methods in this domain. Their Argumentative Independence Test (AIT) selects a subset of independence test results performed by the PC algorithm, excluding less trustworthy tests. AIT combines preference-based argumentation (Amgoud and Cayrol, 2002) with deductive reasoning, as reviewed in Besnard and Hunter (2018).

AIT relies on graphoid axioms (Pearl and Paz, 1986) to derive additional independence relations from observed ones. Conflicts arise when inconsistencies occur between data-driven and derived independence relations. These are resolved using heuristics-driven preferences to break bi-directed attacks within the AF. However, as graphoid axioms are incomplete (Studený, 1989), some inconsistencies remain

undetected and unresolved. This limitation constrains the reliability of AIT's outputs.

Our work builds on and extends this foundational idea by introducing the Causal ABA framework in Chapter 6. Specifically, we address key limitations of AIT by:

- Encoding d-separation rules as ABA rules, ensuring a complete representation of DAGs.
- Using stable semantics to reject unacceptable independence tests, ensuring a one-to-one correspondence between accepted tests and the output DAG.
- Allowing the integration of external knowledge, aligning causal discovery outputs with domain expertise and enhancing stakeholder engagement.

Through these advancements, Causal ABA resolves conflicts more comprehensively and represents causal graphs as structured, argument-driven objects. This framework facilitates a sound, interpretable, and contestable causal discovery process, enabling stakeholders to engage with and challenge the causal graphs. These features promote transparency, trust, and alignment with domain expertise, which are critical for explainable and accountable AI systems.

Argumentation for Explainable AI Beyond causal discovery, computational argumentation plays a broader role in XAI, providing structured explanations for reasoning processes. Unlike post-hoc explanation methods, which often offer limited insights into model behaviour, argumentation delivers layered, interactive explanations that stakeholders can explore in detail (Rago et al., 2020).

Cyras et al. (2021) categorise argumentation's contributions to XAI into three primary areas:

- Building intrinsically interpretable models;
- Developing complete post-hoc explainers;
- Constructing approximate post-hoc explainers that balance accuracy with interpretability.

Our work aligns with the first category: the Causal ABA framework in Chapter 6 transparently models the reasoning process underlying causal discovery algorithms. By encoding d-separation rules and incorporating domain knowledge, the framework produces inherently interpretable arguments, representing both statistical evidence and causal relationships. These arguments empower stakeholders to scrutinise and refine causal assumptions, promoting transparency and trustworthiness.

Argumentation has also been applied to models such as Causal Bayesian Networks (BNs) (Rago et al., 2021, 2023), which extend causal graphs by associating probabilities with causal relationships (Pearl,

2009). These applications demonstrate the versatility of argumentation in enhancing the interpretability of complex systems while adhering to formal principles that support trustworthiness.

While these approaches motivate and inform our work, they address different challenges. Specifically, our focus lies in leveraging argumentation to improve causal discovery by systematically resolving inconsistencies, integrating external knowledge, and transparently producing robust causal graphs. This distinctive focus sets our work apart and underscores its contributions to both causal discovery and XAI.

3.3.5 Other Methods

The landscape of causal discovery extends beyond score-, constraint-, and logic-based methods, incorporating approaches that rely on specific functional assumptions about the DGP or explore entirely new paradigms. Among these, **FCM-based methods** represent one of the original categories of causal discovery (Glymour et al., 2019). These methods assume specific noise distributions as well as functional forms for the relationships between variables, enabling guarantees on the recovery of ground truth causal graphs under strong, but testable, assumptions (Bühlmann et al., 2014).

Examples of FCM-based methods include the Causal Additive Model (CAM) (Bühlmann et al., 2014), which employs regularised regression for variable selection and restricted maximum likelihood for causal order estimation. The weights derived from the variable selection are then used for pruning the graph and avoid cycles, method that has inspired our DAG mapping function in Chapter 4. Linear Non-Gaussian Additive Models (LiNGAM) (Shimizu et al., 2006), Post Non-Linear (PNL) models (Zhang and Hyvärinen, 2009), and Additive Noise Models (ANMs) (Hoyer et al., 2008) handle different combinations of linearity, noise distributions and functional forms. While FCM-based methods rely on stronger assumptions than constraint- or score-based methods, their model assumptions can serve to build SEMs used as synthetic data generators. For instance, ANMs are used in Chapter 5 to evaluate our proposed methodology.

Hybrid methods combine elements of independence testing, scoring, and functional assumptions to perform causal discovery. Bayesian approaches such as (Claassen and Heskes, 2012; Triantafilou et al., 2014) model constraints probabilistically, incorporating priors to guide graph estimation and have probabilities on outputs. Max-Min Hill-Climbing (Tsamardinos et al., 2006) and COmbINE (Triantafillou and Tsamardinos, 2015), integrate CITs with scores, aiming at leveraging strengths of both methods while addressing complementary weaknesses. These approaches exemplify the benefits of integrating diverse techniques to address the limitations of individual methods in causal discovery.

Beyond these foundational and hybrid approaches, alternative paradigms have been developed that expand the scope of traditional causal discovery. Leveraging emerging technologies, these methods offer complementary tools and perspectives, exploring novel ways to represent and solve causal discovery tasks. Some of these methods include:

- Permutation-based methods, e.g., (Andrews et al., 2023; Lam et al., 2022), which assess variable relationships through reordering,
- Amortized Inference, e.g., (Lorch et al., 2022), which uses variational autoencoders for scalable causal learning,
- Supervised Learning, e.g., (Dai et al., 2023), which predicts causal relationships using pre-trained models,
- Bayesian methods, e.g., (Rittel and Tschiatschek, 2023), which incorporate structured priors for improved inference,
- Reinforcement Learning, e.g., (Sauter et al., 2023), which learns causal graphs by optimising decision policies,
- Generative Flow Networks, e.g., (Deleu et al., 2022), which explore graph structures using flow-based models, and
- Optimal Transport, e.g., (Tu et al., 2022), which aligns data distributions to infer causal relationships.

While these approaches demonstrate the diversity and innovation in causal discovery, they are not directly aligned with the methodologies or contributions of this thesis. This work focuses on enhancing constraint-based and logic-based methods, providing provable guarantees to foster trust. In addition, it integrates causal discovery with predictive tasks to improve transparency, robustness, and accountability in NN-based systems, aligning with the broader goals of trustworthy AI.

3.4 Summary

This chapter reviewed the foundational literature underpinning the methodologies developed in this thesis, situating the proposed contributions within three key research areas: contestable AI, knowledge

injection, and causal discovery. By contextualising these domains, the chapter identified key challenges and gaps, laying the groundwork for the novel methodologies introduced in subsequent chapters.

The section on contestable AI highlighted the importance of enabling stakeholders to challenge and refine AI decisions. Existing approaches often focus on static explanations or post-hoc analysis, limiting their ability to support dynamic, iterative contestation. Additionally, there is a gap in operationalising contestability, particularly in mechanisms for redress. This thesis addresses these gaps by integrating causal discovery into NNs to create Contestable NNs, which enable interactive contestation and refinement of model outputs.

The discussion on knowledge injection examined its role in informed ML, particularly in improving interpretability, robustness, and contestability through the integration of domain expertise. While existing methods incorporate domain knowledge into ML models, they lack mechanisms for transparent and iterative causal graph integration that enable contestability, while representing both model and real-world mechanisms. This thesis fills this gap by combining causal discovery with expert feedback, enabling dynamic interactions between humans and NN systems to refine causal graphs, improve transparency, and align model behaviour with domain expertise.

The review of causal discovery methods focused on three main approaches: continuous score-based, constraint-based algorithms, and logic-based methods. Continuous score-based can integrate causal discovery with predictive tasks, but lack interpretability and theoretical guarantees. By contrast constraint-based algorithms rely on statistical tests but struggle with noisy data. Logic-based methods provide stronger formal guarantees but often lack mechanism for stakeholder engagement and contestability. This thesis extends constraint-based methods to handle noisy and quasi-unfaithful data, improving robustness. Additionally, it bridges the gap in transparency and guarantees by allowing experts to contest continuous score-based causal discovery methods. Finally, we leverage computational argumentation for conflict resolution and knowledge integration, delivering stronger theoretical guarantees and a methodology designed for contestability.

In summary, this thesis integrates causal discovery and expert feedback into cohesive frameworks to enhance transparency, robustness, and accountability. By addressing gaps in contestability, knowledge injection, and causal discovery, it advances trustworthy AI and its suitability for high-stakes decision-making contexts. The subsequent chapters detail the methodologies developed to achieve these goals, towards more trustworthy AI systems.

Chapter 4

Contestable Neural Networks

In this chapter, we introduce **Contestable NNs**, a methodology to allow stakeholders to challenge NNs through causal graphs and knowledge injection. This work builds on the method introduced in (Russo and Toni, 2023a), while extending it with a formalisation of the contestability process based on the framework proposed in (Leofante et al., 2024). The technical methodology remains consistent with (Russo and Toni, 2023a), but the formalisation provides a conceptual extension.

We organise our approach within a formal **Contestation Framework** (§4.2), leveraging causal discovery and knowledge injection to enhance transparency, accountability and robustness. To initiate the contestation process, we propose a global **explanation** method that employs discovered causal graphs (§4.3). This is followed by the development of two key algorithms. The first algorithm injects expert knowledge into NNs via causal graphs, enabling the **redress** of contested models (§4.4). The second algorithm supports human-AI collaboration during training, allowing practitioners to **contest** predictions using the estimated causal graphs (§4.5). We evaluate these algorithms on real and synthetic tabular datasets, focusing on regression and classification tasks (§4.6), to then discuss the impact for prediction, causal discovery and stakeholder engagement (§4.7).

4.1 Motivation

NNs have demonstrated exceptional performance across a wide range of ML tasks (Goodfellow et al., 2016). However, concerns persist about their ability to accurately encode the DGP and capture the causal structures governing the data. This shortfall is evident in their vulnerability to data perturbations and adversarial attacks (Goodfellow et al., 2015; Szegedy et al., 2014), which challenge their robustness and reliability in high-stakes *automated* decisions, such as credit scoring or parole evaluations (Rudin, 2019).

Three core weaknesses limit the deployment of NNs in these critical domains: a lack of **robustness** in noisy or low-data settings, restricted **transparency** in understanding how input features influence predictions, and, critically, insufficient **contestability** to allow stakeholder challenge. While there has been significant progress in addressing transparency (Ali et al., 2023), and robustness (Huang et al., 2020), the challenge of contestability remains largely unaddressed, with no algorithmic frameworks proposed to operationalise it. To overcome these limitations, this chapter proposes a methodology focused on enhancing all three attributes, aligning with the objectives set out in Chapter 1.

Our methodology leverages knowledge injection to integrate human expertise with data-driven models, addressing the inherent limitations of data-only learning. Knowledge injection has been shown to improve generalisation (Roscher et al., 2020), reduce data requirements (Borghesi et al., 2020), and enhance interpretability. Building on these principles, we propose a novel framework for rendering NNs *contestable* by combining knowledge injection with causal discovery.

Causal graphs are central to our framework, providing interpretable and transparent representations of the NN's learned relationships (**Objective 3**). These graphs enable stakeholders to evaluate, challenge, and iteratively refine the causal assumptions driving model predictions (**Objective 1**), fostering a dynamic form of human-AI collaboration. This focus on contestability ensures that models align with domain knowledge (**Objective 2**), allowing stakeholders to validate and refine learned causal relationships and make interventions that influence model behaviour.

The foundation of our approach is the CASTLE model (Kyono et al., 2020), which uses causal discovery as a regularisation mechanism during NN training. While CASTLE enhances accuracy and generalisation to unseen data through causal discovery (see §2.4), it does not allow stakeholders to modify or refine learned causal relationships. Our method overcomes this limitation by enabling iterative interactions between Subject-Matter Experts (SMEs) and NNs, refining causal graphs with domain expertise. This iterative refinement improves robustness in small-data scenarios and ensures resilience to noise, directly addressing **Objective 5**.

This approach also addresses ethical concerns in automated decision-making systems. Stakeholders can inspect and modify causal representations to uncover and rectify biases, promoting fairness and accountability, in alignment with global principles of trustworthy AI (High-Level Expert Group on AI, 2019; National Institute of Standards and Technology, 2023). For instance, SMEs can identify and remove causal links that unfairly discriminate against protected attributes, such as race or gender, ensuring equitable and legally compliant Automated Decision Systems (ADSs).

To demonstrate the practical implications of our Contestation Framework, we now present a case

study on real data. This study builds on Example 2.1.1 and showcases how causal graphs, derived from data, enable stakeholders to scrutinise and refine the relationships driving model predictions.

4.1.1 Case Study: Predicting Income from Demographic Data

We use the Adult Income Prediction dataset from the UCI Machine Learning Repository (Becker and Kohavi, 1996), which contains demographic information such as race, education, and occupation, to predict whether an individual earns more than $50,000 annually. The adjacency matrix in Figure 4.2 provides the list of variables in the dataset, and further details can be found in the UCI repository.

The prediction of income based on demographic features holds practical significance in domains such as finance, marketing, and social policy, where understanding determinants of income is critical. For instance, financial institutions assess loan eligibility using predictive models, which must comply with regulatory guidelines regarding responsible lending, specifically in regard to affordability (Financial Conduct Authority, 2024).

Unlike causal inference tasks that evaluate the effects of interventions (e.g., "how does improving education influence income?"), our focus here is on predicting income levels from observed demographic attributes, answering the question: *what if we see...?*. This classification task predicts the likelihood that an individual earns more than $50,000 annually.[1] While correlation-based predictions may achieve high accuracy, such models risk relying on spurious associations that undermine robustness, fairness, and ethical decision-making.

To mitigate these risks, our methodology prioritises the extraction and evaluation of causal relationships learned by the model to ensure that its predictions are both meaningful and reliable.

We begin by fitting a CASTLE model to predict income, leveraging causal discovery as an auxiliary task (see §2.4 for details). Given its state-of-the-art performance in predictive tasks compared to mainstream regularisation methods (Kyono et al., 2020), CASTLE serves as the baseline model for this study. The model achieves an Area Under the Curve (AUC) of 0.76 on a held-out test set, demonstrating moderate predictive performance.

However, before deployment, it is essential to verify that the model's learned relationships align with domain knowledge and ethical standards. To this end, our *Contestation Framework* enables stakeholders to evaluate, challenge, and refine these relationships systematically. We introduce this approach in the following section.

[1]This task could also be formulated as a regression problem if the income variable were not thresholded, with no changes required to make it work in our proposed framework.

4.2 The NNs' Contestation Framework

In this section, we propose a structured approach to enabling contestability in NNs. Building upon the analysis by (Leofante et al., 2024), this framework identifies the key stakeholders, the elements of the NN subject to contestation, and the tools and techniques that facilitate effective interaction, including mechanisms for redress.

Contesting and Contested Entities The primary stakeholders in this framework are SMEs, who possess domain-specific expertise essential for evaluating and refining the NN's behaviour. Unlike technical practitioners such as data scientists or engineers, SMEs focus on ensuring that the NN aligns with domain knowledge, ethical principles, and regulatory requirements. Their role extends beyond scrutinising individual predictions to assessing the causal assumptions underpinning the NN's global behaviour. For instance, doctors can evaluate whether medical diagnosis models align with established knowledge, while risk managers and underwriters can scrutinise credit scoring models for compliance with financial risk standards. By involving SMEs, the NN remains grounded in real-world expertise and ethical considerations.

In this framework, SMEs evaluate and refine the NN's predictions, focusing on its overall behaviour rather than individual outputs. This is critical for ADSs performing large-scale classification or regression tasks. Hence, we propose a method for contesting the NN's behaviour as a whole, ensuring it adheres to domain-specific requirements and ethical standards, beyond achieving high predictive accuracy. Through iterative evaluation and targeted feedback, SMEs have the opportunity to enhance the NN's transparency and trustworthiness, while ensuring compliance with relevant regulations.

Note that our experiments in this chapter are aimed at demonstrating the feasibility of this approach, and we do not involve real SMEs in the evaluation. We discuss the potential for real-world applications and user studies investigating collaboration and interaction aspects in Chapter 7.

Contesting Methods and Redress Mechanisms Operationalising contestability requires providing SMEs with interpretable and actionable explanations of the NN's learned behaviour (Yurrita et al., 2023b). This is achieved using causal graphs, which act as global process-centric explanations by summarising the inferred relationships across the dataset (§4.3). These graphs allow SMEs to systematically evaluate the model's causal assumptions, identify questionable or misaligned links, and provide targeted feedback.

Feedback from SMEs is integrated through the *Graph Injection Algorithm* (§4.4), which refines the NN by adjusting its weights to align with expert input. This ensures that the model's predictions rely

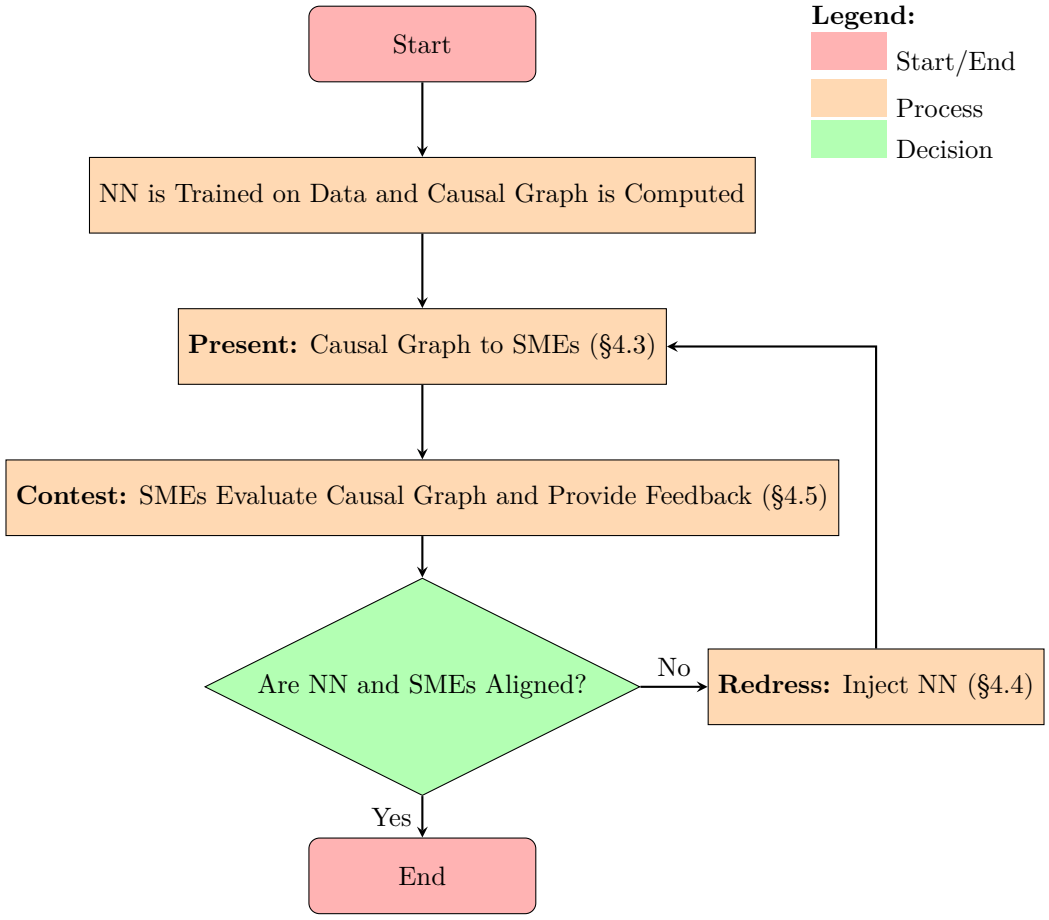


Figure 4.1: Contestation Framework Flowchart showing the iterative process that allows SMEs to contest NNs by challenging the relations in the causal graph that the NN has extracted from data and is using for prediction.

only on validated variables and relationships, effectively closing the contestation loop. By enabling iterative interactions between SMEs and the NN, this process not only resolves identified issues but also improves the model's reliability and robustness in noisy or low-data scenarios (§4.6).

Interactive and Iterative Contestation The contestation process, illustrated in Figure 4.1, involves iterative interactions where SMEs, aided by estimated causal graphs and empowered with the tools to influence them, challenge and refine the NN's causal graph during model training. This iterative loop ensures that the NN evolves to reflect both robust predictive performance and interpretable causal relationships.

The flowchart in Figure 4.1 outlines the components of the framework. The process begins with the NN being trained on data and the generation of a causal graph to be presented to SMEs for evaluation (§4.3). SMEs scrutinise the causal relationships, contesting any edges or directions that conflict with domain knowledge (§4.5). If the feedback reveals misalignments, the graph injection algorithm updates

the model's parameters and causal graph. The revised graph is then re-evaluated, and the process repeats until alignment between the NN and domain expertise is achieved. We begin by detailing the graphical interface that initiates the contestation process.

4.3 A Graphical Interface into a NN's Causal Reading of the Data

In §2.4, we introduced CASTLE, a regularisation method that enhances the generalisation capabilities of FFNNs by leveraging causal discovery as an auxiliary task (Kyono et al., 2020). While CASTLE focuses on predictive performance, our methodology builds upon its framework to provide a graphical interface that visualises the NN's causal interpretation of the data. This interface is a key element of our contestation framework, enabling stakeholders to evaluate and refine the causal relationships inferred by the NN.

Central to this interface is the weighted adjacency matrix $\mathbf{W}_\Theta$, which encodes the strength of relationships between variables in the dataset as learned by the NN, through the optimisation of its weights Θ using the DAG loss. Each entry w_{ik} in $\mathbf{W}_\Theta$ represents the influence of input variable i on output variable k, where $i, k \in \{1, \ldots, d+1\}$, with d input variables and the target variable. For details on our predictive modelling setup and CASTLE, refer to §2.4.

To construct a DAG from $\mathbf{W}_\Theta$, we apply two key steps: thresholding and cycle removal. Thresholding eliminates weaker, potentially spurious connections, while cycle removal ensures the resulting graph is acyclic. These steps are formalised below.

Definition 4.3.1 (Edge Creation Function)**.** *Given $\mathbf{W}_\Theta$ from Definition 2.4.1, a threshold $\tau \geq 0$, and $i, k \in \{1, \ldots, d+1\}$ a set of nodes, the set of edges $\mathcal{E}_\tau(\mathbf{W}_\Theta)$ encoded by $\mathbf{W}_\Theta$ is defined as:*

$$\mathcal{E}_\tau(\mathbf{W}_\Theta) = \{(i, k) \mid w_{ik} > w_{ki} \wedge w_{ik} > \tau\}.$$

The edge creation function identifies directed edges (i, k) where w_{ik} exceeds both w_{ki} and the threshold τ. This process removes 1-step cycles (of length 2) and weaker links. If cycles involving longer paths remain, they are resolved by iteratively removing the weakest edge in each cycle until acyclicity is achieved, as shown in Algorithm 2. The procedure uses Definition 4.3.1 to extract directed edges from the adjacency matrix, followed by cycle finding using Johnson's algorithm (Johnson, 1975), as implemented in the `Networkx` package (Hagberg et al., 2008). The algorithm then iteratively removes the weakest edge in each cycle until the graph is acyclic, making sure that only necessary edges are removed (lines 6-7). The arg min function on line 8 breaks ties lexicographically.

Algorithm 2: DAG Mapping Function

Input: Weighted adjacency matrix $\mathbf{W}_\Theta$, threshold $\tau \geq 0$.

Function: $\Gamma(\mathbf{W}_\Theta, \tau)$:

$\mathcal{E} \leftarrow \mathcal{E}_\tau(\mathbf{W}_\Theta)$ ▷ Extract directed edges using Definition 4.3.1
$\mathcal{G} \leftarrow (\mathbf{V}, \mathcal{E})$
while $\mathcal{G}$ contains cycles **do**
    $\mathbf{C} \leftarrow$ `find_cycles`$(\mathcal{G})$ ▷ Johnson's algorithm
    for $c \in \mathbf{C}$ **do**
        if $\exists (i,k) \in c$ such that $(i,k) \notin \mathcal{E}$ **then**
            continue ▷ Skip if cycle was already broken
        $(i,k)^* = \arg\min_{(i,k)\in c} w_{ik}$
        $\mathcal{E} \leftarrow \mathcal{E} \setminus \{(i,k)^*\}$ ▷ Remove weakest edge
    $\mathcal{G} \leftarrow (\mathbf{V}, \mathcal{E})$
return $\mathcal{G}$

Output: DAG $\mathcal{G}$

Unlike methods such as (Zheng et al., 2018, 2020), which rely solely on thresholding and can allow cycles, our approach explicitly enforces acyclicity using a mechanism inspired by (Bühlmann et al., 2014), which uses penalised regression to calculate the weights that we derive from the NN.

Facilitating Stakeholder Engagement The DAG resulting from Algorithm 2 offers a transparent representation of the NN's causal relationships, forming the basis for stakeholder scrutiny and refinement. SMEs can leverage this interface to challenge causal assumptions, refine specific relationships, and iteratively align the model with domain knowledge.

The threshold τ plays a critical role in this process, controlling the sparsity of the graph. Setting $\tau = 0$ retains all relationships in $\mathbf{W}_\Theta$, while higher values filter out weaker connections. By iteratively adjusting τ, practitioners can initially focus on the most significant relationships, and progressively explore less prominent, but plausible, connections.

This interface bridges complex NN architectures with actionable insights, enabling the contestation process outlined in §4.2. To demonstrate this first step of the methodology, we return to the income prediction case study introduced in §4.1.1.

Example 4.3.2. ***(Extracting a DAG for the Income Prediction Case Study)*** *In Figure 4.2, we show the adjacency matrix generated using CASTLE's weights and Definition 2.4.1 for the Adult dataset (Becker and Kohavi, 1996). Gray cells indicate the exclusion of self-loops, while darker blue shades represent stronger relationships between variables. For clarity, weights below a threshold of* $\tau = 0.08$ *are omitted, simplifying the matrix to highlight the most significant relationships.*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	
(1)			0.0	0.2				0.1					36.1	3.7	21.3	income
(2)													1.6	0.3	11.5	race
(3)								0.1					1.9	0.4	16.3	sex
(4)													0.8	0.1	2.0	age
(5)	0.0												1.1	0.2	4.9	native-country
(6)			0.0										0.6	0.2	4.0	occupation
(7)													1.2	0.3	6.4	workclass
(8)													0.8	0.1	3.6	hours-per-week
(9)													0.5	0.2	5.5	education
(10)	0.0												2.7	0.4	6.9	education-num
(11)	0.1												2.0	0.3	18.1	marital-status
(12)	0.0		0.0	0.1		0.0	0.0	0.1		0.0			2.6	0.5	15.0	relationship
(13)															0.1	capital-gain
(14)													0.2		0.5	capital-loss
(15)																fnlwgt

Figure 4.2: Weighted adjacency matrix $\mathbf{W}_\Theta$ obtained using Definition 2.4.1 from the weights Θ of a Joint NN, fitted on the Adult data (Becker and Kohavi, 1996). Y = income. Darker shades indicate stronger relationships between variables. Weights below $\tau = 0.08$ are blanked for readability. $w_{ik} = 0.0$ indicates weights that are between 0.08 and 0.1.

From this weighted adjacency matrix, a binary version can be derived and mapped into a DAG using Algorithm 2. The resulting DAG includes an arrowhead for every blue cell, irrespective of its shade, capturing the causal relations learned by the NN from the Adult dataset.

The resulting DAG is shown in Figure 4.3. The figure represents a DAG that captures the causal relationships between variables in the dataset, as learned by the NN. For instance, the DAG indicates that the number of years of education *and* marital status *have a direct influence on* income*, while* income *in turn causes financial indicators like* capital gain *and* loss*. However, the NN has also learned anti-causal (spurious) relationships (purple arrows), such as* income *causing* age *and* sex *(i.e., gender).*

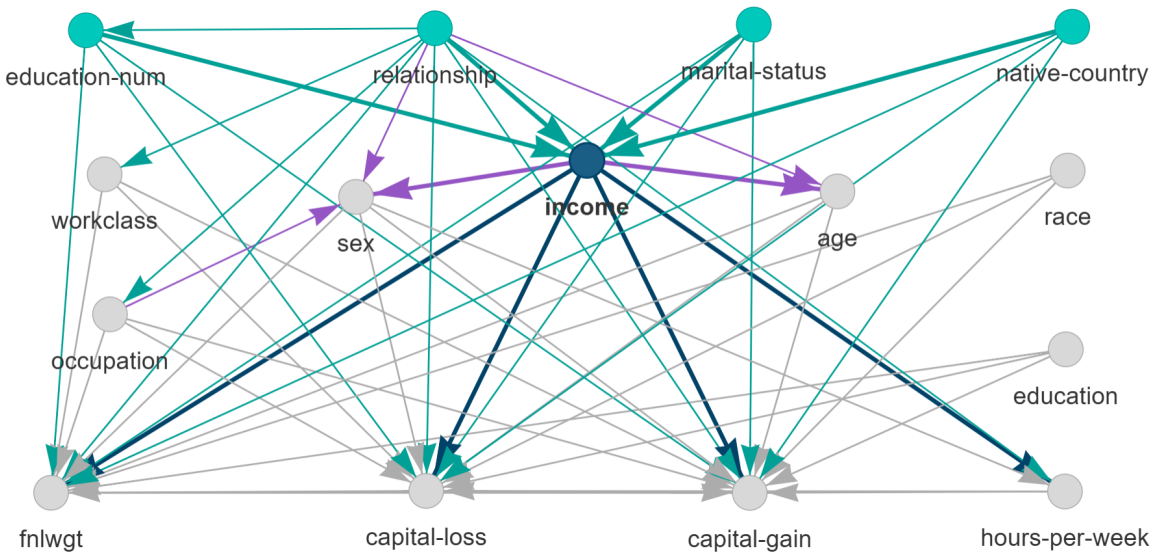


Figure 4.3: DAG obtained from the adjacency matrix in Figure 4.2 using Algorithm 2 with $\tau = 0.08$. The parents of the target variable (*income*) are highlighted in green, the children of the target are in blue while in purple are the spurious (anti-causal) relationships.

Algorithm 3: Inject Causal Graph

Input: Training Data $\mathcal{D}$; NN with M layers, h neurons in the first hidden layer, trained for t steps with final weights Θ_t; max number of steps T; patience $T_s < T$; causal graph $\mathcal{G} = (V, E)$
Function: `inject_graph`$(\mathcal{D}, \Theta_t, T, T_s, \mathcal{G})$:

for $i \in \{1, \ldots, d+1\}$, **for** $k \in \{1, \ldots, d+1\}$ **do**,
  for $j \in \{1, \ldots, h\}$ **do**
    if $i, k \in V$ & $(i, k) \notin E$ **then**
      $\Theta_{1,t+1}^{i,j,k} \leftarrow 0$ ▷ Mask anti-causal relations
    else
      $\mathcal{L}_{\text{best}} \leftarrow \infty$
      $t_s \leftarrow 0$
      while $t < T$ **and** $t_s < T_s$ **do**
        $\Theta_{1,t+1}^{i,j,k} \leftarrow \texttt{update}(\Theta_{1,t}^{i,j,k}, \mathcal{D})$ ▷ Update causal relations
        for $m \in \{1, \ldots, M\}$ **do**
          $\Theta_{m,t+1} \leftarrow \texttt{update}(\Theta_{m,t}, \mathcal{D})$
        $t \leftarrow t + 1$
        if $\mathcal{L}_t < \mathcal{L}_{\text{best}}$ **then**
          $\mathcal{L}_{\text{best}} \leftarrow \mathcal{L}_t$
          $t_s \leftarrow 0$
        else
          $t_s \leftarrow t_s + 1$
$\Theta_{\mathcal{G}} \leftarrow \Theta_t$
return $\Theta_{\mathcal{G}}$

Output: Injected NN with weights $\Theta_{\mathcal{G}}$

As shown in this example, causal graphs provide a transparent representation of the model's inferred causal relationships. By visualising these relationships, practitioners can assess the validity and ethical alignment of the model's predictions, facilitating model refinement and more informed decision-making. Next, we introduce our method to incorporate expert feedback into the NN.

4.4 Redress: The Graph Injection Algorithm

To close the contestation loop and facilitate redress in response to experts' contestations, we need a method to incorporate expert feedback into the NN. This is achieved through the *Graph Injection Algorithm*, which incorporates constraints from a causal graph directly into the NN's weight parameters. We outline the injection process in Algorithm 3, which accepts three primary inputs:

- A training dataset $\mathcal{D}$ with $d+1$ variables, including the target variable, and n samples;
- A Joint NN with $d+1$ stacked subnetworks each predicting one of the variables without using it (see §2.4, Figure 2.1). The NN's weights Θ_t are either randomly initialised or pre-trained on $\mathcal{D}$;

- A causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, with $\mathbf{V} = d + 1, \mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ representing either causal relations about all variables or partial knowledge.

The training dataset $\mathcal{D}$ is, as usual, split into training and validations sets. The NN's weights Θ_t can be either randomly initialised or pre-trained on (a subset of) $\mathcal{D}$. In the pre-trained scenario, our method acts as a *fine-tuning* process, adjusting the NN to comply with the causal constraints specified by the input DAG, while leveraging its prior optimisation for the task. The input DAG serves to constrain the relationships used by the NN, focusing training on accepted relationships.

The NN is trained using backpropagation, represented by the `update` function in the algorithm. This process iteratively updates the NN's weights to minimise the loss function on the training data while adhering to the constraints defined by the input DAG. Training continues for a maximum of T steps or until a patience threshold $T_s < T$ is reached, indicating no improvement in validation loss.

In many practical scenarios, obtaining a complete DAG that describes all relationships among variables is unrealistic. To address this limitation, our approach supports partial causal graphs that constrain only a subset of input variables, enabling the discovery of additional relationships during training. In exploratory settings, where no prior causal knowledge is available, the framework can operate with a *complete* graph, represented as an adjacency matrix with zeros on the diagonal and ones elsewhere. In such cases, the framework relies solely on the DAG loss, which encourages sparsity and acyclicity (see §2.4), to infer causal structures.

From the combination of initialisations and causal knowledge availability we explore two training modalities:

- **Partial Knowledge:** A randomly initialised NN constrained by a PDAG encoding a subset of variable relationships.
- **Full Knowledge:** A pre-trained NN fine-tuned to align with an input DAG.

In a third scenario, we envisage a *complete* graph with a randomly initialised NN, which is equivalent to training a CASTLE model from scratch. We experiment with this scenario to explore the process of extracting an interpretable causal graph from a NN with **no a priori causal knowledge** (see §4.6).

The output of Algorithm 3 is a Joint NN with weights $\Theta_{\mathcal{G}}$, that we call an *Injected* NN. An Injected NN exclusively utilises the relationships specified in the input causal graph $\mathcal{G}$, through a targeted masking scheme that sets the weights in the input layer to 0 for any pair of input neurons (or variables) i, k where $(i, k) \notin \mathbf{E}$. Specifically, for all neurons j in the first hidden layer, we set $\Theta_1^{i,j,k} = 0$, resulting in $w_{ik} = 0$ and effectively eliminating the direct influence of variable i on variable k.

Note that Algorithm 3 handles the presence and absence of directed edges within an input graph to be injected differently. The algorithm masks weights corresponding to the absence of directed edges, thereby enforcing the NN to adhere to the causal constraints that specify the absence of certain causal relations. However, it does not enforce the influence of a variable on another when an edge is present. This is because the NN continues to learn the weights associated with existing edges, as the algorithm does not intervene in their learning process.

Effectively, our algorithm restricts the NN from utilising relationships that are not present in the input graph, while allowing it to learn weights for the existing edges. Consequently, specifying the *anti-causal* relationships (i.e., the absence of edges) plays a more pivotal role in steering the training process than specifying the causal relationships themselves.

Introducing edges that the NN found to have minimal significance during training into the input causal graph of a trained model may lead to limited changes in the model's behaviour. This is likely because the model has already converged to a local minimum in the loss landscape. In contrast, when training begins from scratch, the inclusion of edges in the input graph increases the likelihood that the model will incorporate and utilise these specified edges during the learning process. We discuss potential refinements to the treatment of present edges as part of future work in Chapter 7.

A PDAG can be injected by specifying only a subset of the variables and their relationships. In this case, the algorithm will mask the weights corresponding to the absence of directed edges between the specified variables, while allowing the model to learn the weights associated with the existing edges. Line 3 in Algorithm 3 ensures that the algorithm only masks weights for variables specified in the input graph, while allowing the model to learn the weights associated with the remaining variables. This way, we can accommodate knowledge about a subset of relations between the variables, enabling the discovery of additional relationships during training, while adhering to the specified constraints.

Computationally, the primary distinction between Algorithm 3 and CASTLE's training loop lies in the focused training on weights corresponding to accepted edges, i.e., the causal relationships. This can be conceptualised as a "selective causal" dropout method. While CASTLE induces the NN to prefer using causal parents over children and siblings for its predictions (Kyono et al., 2020), Algorithm 3 enforces this preference strictly. By reconstructing each variable and predicting the target using only its parents, we avoid predicting variables using their children or siblings, whose relations may be unstable (Pearl, 2009).

Our proposed masking scheme encodes responses to causal inquiries. For example, consider the question: *Does education influence income?* If the answer is negative, the corresponding weights are

Algorithm 4: Contest Computed Causal Graph

Input: Training Data $\mathcal{D}$, NN with weights Θ_t trained for t steps; maximum training steps T, patience $T_s < T$, and *Expert Knowledge*.

Function: `contest_graph`$(\mathcal{D}, \Theta_t, T, T_s, \textit{Expert Knowledge})$:

contested $\leftarrow$ True
while *contested* **do**
    $\mathcal{G} \leftarrow \Gamma(\mathbf{W}_{\Theta_t}, \tau)$ ▷ Extract causal graph (Algorithm 2)
    $\mathcal{G}_r, \tau \leftarrow$ `revise_graph`$(\mathcal{G}, \textit{Expert Knowledge})$ ▷ **Present** to SME and revise $\mathcal{G}$
    if $\mathcal{G}_r \neq \mathcal{G}$ **then** ▷ **Contested?**
        $\Theta_t \leftarrow$ `inject_graph`$(\mathcal{D}, \Theta_t, T, T_s, \mathcal{G}_r)$ ▷ **Redress** using Algorithm 3
    else
        contested $\leftarrow$ False
return $\mathcal{G}_r$, Θ_t

Output: Refined DAG $\mathcal{G}_r$ and Injected NN weights Θ_t

set to zero, thereby preventing education from directly influencing the prediction of income in the NN. Importantly, while direct effects are masked, indirect effects can still be captured through the hidden layers, enabling the NN to model complex causal relationships without relying on direct connections that the SMEs decided to exclude.

We hypothesise that the structure of the Joint NN and the masking scheme will naturally guide the NN to prefer the bespoke subnetwork for predicting each variable, with the middle layers playing a reduced role in these predictions. This preference could help mitigate potential issues, such as the NN using indirect relationships to capture spurious correlations. While this hypothesis would limit the issue, the lack of control over indirect relations remains a limitation of the current approach. Addressing this challenge presents a compelling direction for future work, discussed in Chapter 7.

Next, we combine the introduced methods to construct and *present* the NN's causal graph and to *redress* the model's behaviour based on feedback from SMEs, culminating in our contesting algorithm.

4.5 The Contesting Algorithm

This section formalises our contestation framework through the *Contesting Algorithm*, enabling iterative interactions between the NN and SMEs to validate and refine the model's causal understanding of the data, either during or after training. The algorithm, outlined in Algorithm 4, orchestrates the collaborative process by integrating SME feedback to progressively align the NN with domain expertise. A high-level overview of this process is presented in Figure 4.1.

Algorithm 4 extends the functionality of Algorithm 3, replacing the static input causal graph $\mathcal{G}$

with dynamic *Expert Knowledge*. This knowledge, provided by SMEs, reflects domain-specific insights into the relationships among variables and guides the refinement of the causal graph extracted from the NN. The refined graph ensures alignment with SME expertise and adherence to domain-specific ethical and operational requirements.

The process begins by extracting a causal graph $\mathcal{G}$ from the NN using Algorithm 2. This graph reflects the NN's current causal representation, which may be based solely on the training data or include prior causal constraints if the NN has already been injected with an input graph. The `revise_graph` function facilitates stakeholder engagement by presenting $\mathcal{G}$ to the SMEs for evaluation and refinement based on their domain expertise, operating in two stages:

1. **Graph Presentation**: The current causal graph $\mathcal{G}$ is visualised and shared with SMEs, enabling them to scrutinise the estimated causal relationships. SMEs can evaluate individual links, groups of connections, or the overall structure.

2. **Feedback Collection and Revision**: Feedback from SMEs is collected in two forms:

   - *Threshold Adjustments*: SMEs may modify the threshold τ to control the sparsity of $\mathcal{G}$, filtering weaker relationships to focus on the most significant links, or exploring less strong but possibly relevant relations.
   - *Direct Edge Modifications*: Specific edges in $\mathcal{G}$ can be added, removed, or reoriented based on SMEs' domain insights, allowing precise refinements to the causal structure.

   The output is a revised graph $\mathcal{G}_r$ and an updated threshold τ, based on the collected feedback.

This iterative refinement process enables SMEs to tailor the causal graph $\mathcal{G}$ to align with their domain expertise, focusing first on the most influential relationships and gradually exploring weaker ones by adjusting τ. These feedback mechanisms allow SMEs to systematically improve the interpretability and validity of the NN's causal assumptions. The revised graph $\mathcal{G}_r$ serves as input to the `inject_graph` function (Algorithm 3), which updates the NN's weights and causal graph to reflect the proposed refinements.

By embedding SMEs' feedback into the training process, the Contesting Algorithm fosters a dynamic, two-way learning interaction between the NN and its stakeholders. While the NN refines its causal assumptions and predictions based on SME input, the graphical representation of the NN's causal understanding can also provide SMEs with novel insights about potential relationships in the

data. This collaborative process ensures that the model evolves to align with domain-specific expertise while offering stakeholders a deeper understanding of the underlying data structure.

This synergy enhances interpretability, ethical soundness, and robustness, creating a feedback loop where both the model and stakeholders benefit from iterative refinements. Additionally, embedding SME feedback into the model's training process can significantly improve predictive performance, particularly in low-data regimes where expert knowledge compensates for sparse information, further highlighting the value-add of HITL training of ML models.

In the next section, we evaluate the performance of our *Contestable Neural Networks* on real and synthetic datasets, completing our case study on income prediction (§4.1.1) and demonstrating the broader applicability of our framework on other real and synthetic datasets.

4.6 Empirical Evaluation

In this section, we evaluate the *Contestable NN* framework to demonstrate how practitioners can leverage the *Contesting Algorithm* to develop NNs that are both principled and predictive. Our evaluation is structured into two sets of experiments, using both real and synthetic data to assess the benefits of our proposed methods.

Firstly, we discuss the results of our case study on income prediction (§4.1.1) across three scenarios: *No A Priori Knowledge*, *Partial Knowledge*, and *Contesting a Computed Graph*. For the first scenario, we present results from additional real datasets in the financial sector: one more classification and two additional regression tasks. For classification, we use the Adult dataset from the case study and the FICO Home Equity Line of Credit (HELOC) dataset (FICO, 2017) for *credit risk* assessment, predicting loan repayment likelihood. Instead, for regression, we predict house prices using the Boston (Harrison and Rubinfeld, 1978) and California (Pace and Barry, 1997) housing datasets. These results, detailed in §4.6.1, showcase the effectiveness of our injection and contestation algorithms in refining the NN's causal representation and enhancing its predictive performance, even in the absence of prior knowledge.

Secondly, we perform experiments on synthetic data (§4.6.2), demonstrating that prior knowledge can improve both predictive performance and causal discovery, particularly in low-data regimes. These findings align with prior work on causal discovery (Chowdhury et al., 2023; Constantinou et al., 2023), which we evaluate in our prediction setting as well, highlighting the importance of causal knowledge injection in refining NN performance. More details on the datasets and the implementation, including links to code, are provided in Appendix A.1.

4.6.1 Experiments on Real Data

To demonstrate the versatility of our contesting framework in creating transparent and robust NNs, and its capacity to refine their "causal understanding" through human-AI collaboration, we designed three experimental scenarios to evaluate performance under varying levels of prior causal knowledge:

1. **No *A Priori* Knowledge**: Extract a causal graph solely from data using threshold optimisation, without leveraging any pre-existing causal assumptions.

2. **Partial *A Priori* Knowledge**: Integrate predefined causal constraints derived from common-sense assumptions or domain expertise to guide the NN during training.

3. **Contesting a Computed Graph**: Refine a computed causal graph by incorporating the partial knowledge from Scenario 2, demonstrating the effects of expert feedback integration.

These scenarios allow us to systematically investigate how varying levels of causal knowledge and expert intervention influence the NN's ability to model causal relationships and affect predictive performance. In the absence of a true DAG for the real datasets, our quantitative metric is predictive performance, namely MSE for regression and AUC for classification tasks. All results are reported with observed significance levels for a t-test for difference in means (details of the testing procedure are provided in Appendix A.1.1).

Our objective is to demonstrate that predictive performance is not necessarily impacted by knowledge injection, which can, in turn, help to build more transparent and trustworthy NNs.

No *A Priori* Knowledge

In this scenario, we evaluate the performance of our *Contestable NN* framework in the absence of prior causal knowledge. Algorithm 4 is used to iteratively construct potential DAGs by optimising the threshold parameter τ within the Γ_τ function (Algorithm 2). These DAGs are then injected into the NN, and the predictive performance is measured for several thresholds.

To identify optimal thresholds, we systematically generate a set of candidate thresholds τ for each dataset through a binning approach. Specifically, the range of weights in $\mathbf{W}_\Theta$ is divided into n equal intervals, corresponding to n candidate thresholds, with $n = 10$ in our experiments. For each threshold, we evaluate the NN's performance and the sparsity of the resulting DAG. The selection of the "best" DAG balances predictive performance with interpretability by jointly considering the NN's accuracy and the DAG size. For regression tasks, we favour lower MSE, while for classification tasks, higher AUC is prioritised. In cases of comparable performance, sparser graphs (fewer edges) are preferred.

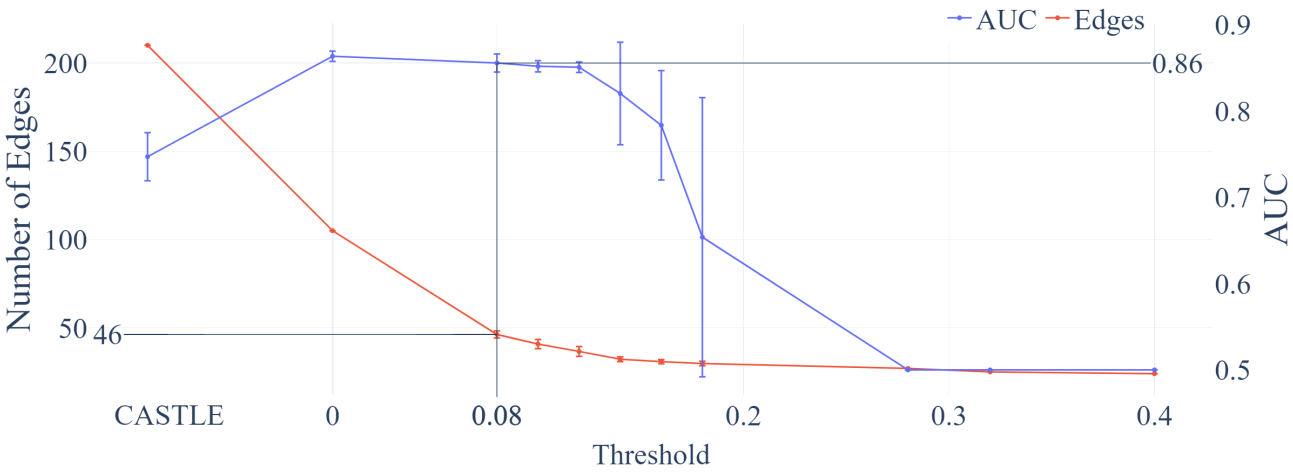


Figure 4.4: Threshold optimization for the Adult dataset. The plot shows the change in AUC for different thresholds τ in the range $[0, 0.4]$. CASTLE, corresponding to the injection of a complete graph, is used as a baseline. $\tau = 0$ corresponds to preventing cycles using Algorithm 2, but not enforcing sparsity. The chosen threshold is $\tau = 0.08$, producing a DAG with 46 edges and AUC=0.86.

Figure 4.4 illustrates the interplay between predictive performance and sparsity for the Adult dataset in our case study. The x-axis represents varying thresholds τ, while the y-axes depict the number of edges (left, red line) and AUC (right, blue line). As expected, increasing the threshold monotonically reduces the number of edges, resulting in sparser graphs. However, changes in AUC reflect dataset-specific trends, highlighting the importance of evaluating both dimensions in tandem. Further comparisons across datasets are provided in §A.1.1, Figure A.1.

As shown in Figure 4.4, for the Adult dataset, $\tau = 0.08$ retains the highest AUC while reducing the number of edges by more than 50% compared to $\tau = 0$. Increasing τ further yields marginal improvements in parsimony (fewer edges) but begins to negatively impact performance.[2]

The CASTLE model, used as a baseline for comparison and positioned to the left of the x-axis, exhibits significantly higher edge counts than all injected models, and a lower AUC than several models with varying thresholds. Further details of this evaluation, along with plots for the other datasets, are provided in Appendix A.1.1.

For this scenario, the contesting process described in Algorithm 4 concludes with a DAG selected through a semi-automatic procedure. The goal is to evaluate whether fixing a plausible—but not human-validated—DAG can enhance predictive performance and how these improvements vary with sample size. To this end, the structure of the NN remains fixed, while datasets of varying sizes are sampled to assess the impact of sample size on the performance of the *Contestable NN*.

[2]For instance, a threshold of $\tau = 0.12$, the second marker to the right of $\tau = 0.08$ in the plot, produced a DAG with 37 edges and AUC = 0.85. This would have been an equally good candidate for further evaluation. A systematic weighting of DAG size and AUC could better guide the choice of τ, but outside the scope of this experiment.

Table 4.1: Experiments with real data. We report mean AUC or MSE ± Standard Deviation (std) for regression and classification, respectively, across different sample sizes of the training data (N) and 5-fold nested cross-validation. When N is larger than the available samples, the cell is left empty. Best results are in bold. Observed significance levels against CASTLE are reported with the following intervals: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. $|V|$ is the number of variables/nodes and $|E|$ the number of edges in the injected DAG for our method and in the graph built using Definition 2.4.1 for CASTLE. *Injected* refers to *No A Priori Knowledge*, *Partial* to *Partial A Priori Knowledge*, and *Refined* to *Contesting a Computed Graph*.

Metric	Data & Method	N=**100**	N=**500**	N=**1K**	N=**2K**	N=**5K**	N=**10K**	N=**20K**
AUC ↑	**Adult** ($\|V\| = 14$)							
	CASTLE ($\|E\| = 210$)	0.67±0.03	0.72±0.04	0.75±0.03	0.74±0.03	0.75±0.03	0.75±0.02	0.76±0.02
	Injected ($\|E\| = 46$)	**0.69.**	**0.74***	**0.76**	**0.77*****	**0.79*****	**0.85*****	**0.86*****
	Partial ($\|E\| = 116$)	0.66	0.71	0.74	**0.76***	**0.76**	**0.76.**	**0.77.**
	Refined ($\|E\| = 30$)	**0.69.**	**0.74***	**0.76**	**0.77*****	**0.79*****	**0.85*****	**0.86*****
	HELOC ($\|V\| = 23$)							
	CASTLE ($\|E\| = 552$)	**0.75±0.02**	**0.79±0.01**	0.78±0.01	**0.79±0.01**	0.79±0.01	**0.80±0.01**	
	Injected ($\|E\| = 85$)	0.74	0.78***	0.78	0.78***	0.79	0.79***	
MSE →	**California** ($\|V\| = 8$)							
	CASTLE ($\|E\| = 72$)	7.05±12.81	2.33±1.39	2.96±4.12	3.86±3.68	4.91±7.41	1.74±1.70	**0.66±0.08**
	Injected ($\|E\| = 31$)	**2.94*****	**2.25**	**1.68****	**1.71*****	**1.51*****	**1.16***	1.02**
	Boston ($\|V\| = 14$)							
	CASTLE ($\|E\| = 182$)	112.0±91.1	21.95±6.84					
	Injected ($\|E\| = 48$)	**86.17*****	**20.45***					

Results Table 4.1 reports the results for all four datasets we consider: Adult, California, Boston, and HELOC. These results demonstrate the impact of our contestable NNs framework on predictive performance and model parsimony across datasets, with comparisons to the CASTLE baseline.

For the Adult dataset, the injected NN consistently achieves higher AUC than CASTLE, with improvements of up to 13% ($N = 20K$). These improvements are statistically significant for most sample sizes and are achieved using only approximately 20% of the relationships that the unconstrained CASTLE network employs ($|E| = 46$ vs $|E| = 210$ for the injected NN and CASTLE, respectively). This reduction matches the reduction in the NN's weights at the input layer.

In the California dataset, the injected NNs significantly outperform CASTLE in MSE across all but the largest sample size and $N = 500$, where the difference in means is not significant. Importantly, the injected NNs achieve these improvements with a 40% reduction in the computed DAG's edges, highlighting the benefits of parsimony. For the Boston dataset, injection reduces edge count by approximately 75%, while the injected NN significantly improves performance compared to CASTLE across both considered sample sizes.

Finally, for the HELOC dataset, the injected NNs achieve AUCs comparable to CASTLE in half the cases, with a maximum reduction of 1%. However, the edge count is drastically reduced, with the

injected NN retaining only 15% of the first-layer weights compared to CASTLE. While this sparsity aligns with principles of *minimality* (Pearl, 2009) and *parsimony* (Vandekerckhove et al., 2015), some beneficial relationships may have been excluded. Future refinement could explore reintroducing selected edges to balance performance and parsimony more effectively.

Overall, simulating the use of Algorithm 4 without prior knowledge produced parsimonious NNs with an average of 75% fewer connections in the input layer. Moreover, predictive performance generally improves significantly or, in the worst case, worsens by at most 1%, regardless of sample size.

As previously mentioned, the strategy for selecting DAGs to inject is purely mechanical in this initial scenario. However, as illustrated in Figure 4.3, the DAG chosen for our case study on income prediction reveals that the NN can learn spurious relationships, such as *income* influencing *age* or *sex*. This underscores the importance of incorporating contestation and expert judgment to guide model validation and refinement.

The adopted strategy is meant to gauge the impact on predictive performance without a qualitative assessment of the validity of domain-specific causal assumptions, a task that SMEs should carry out. We envisage this strategy as a useful starting point in the absence of causal knowledge defined *a priori*. However, in real-life applications, we envisage the use of Algorithm 4 by a panel of SMEs, iteratively assessing intermediate outputs to refine the NN in light of previous experiments, leveraging their experience and knowledge of the modelling task while learning more about it.

Next, we continue our case study and provide examples of how our Contesting Algorithm 4 can work in a human-AI collaboration setting, with SMEs constraining and contesting DAGs computed by the NNs. We run these experiments only on the Adult dataset, as it contains variables such as *race*, *sex*, *age*, and *native-country* that lend themselves to the construction of common-sense causal assumptions by lay users (e.g., *sex* cannot be caused by *age*). This way, we avoid formulating domain-specific assumptions while providing a tangible example application.

Partial *A Priori* Knowledge

In this scenario, we consider the case where some causal knowledge is available *a priori*, and we aim at assessing the impact of injecting this knowledge into the NN on predictive performance. We build a simple yet intuitive partial set of constraints ($\mathcal{G}_p$) for the Adult dataset. The graph $\mathcal{G}_p$ is used to constrain the NN's learning from the start, adhering to the following assumptions (reflected in the adjacency matrix in Figure 4.5):

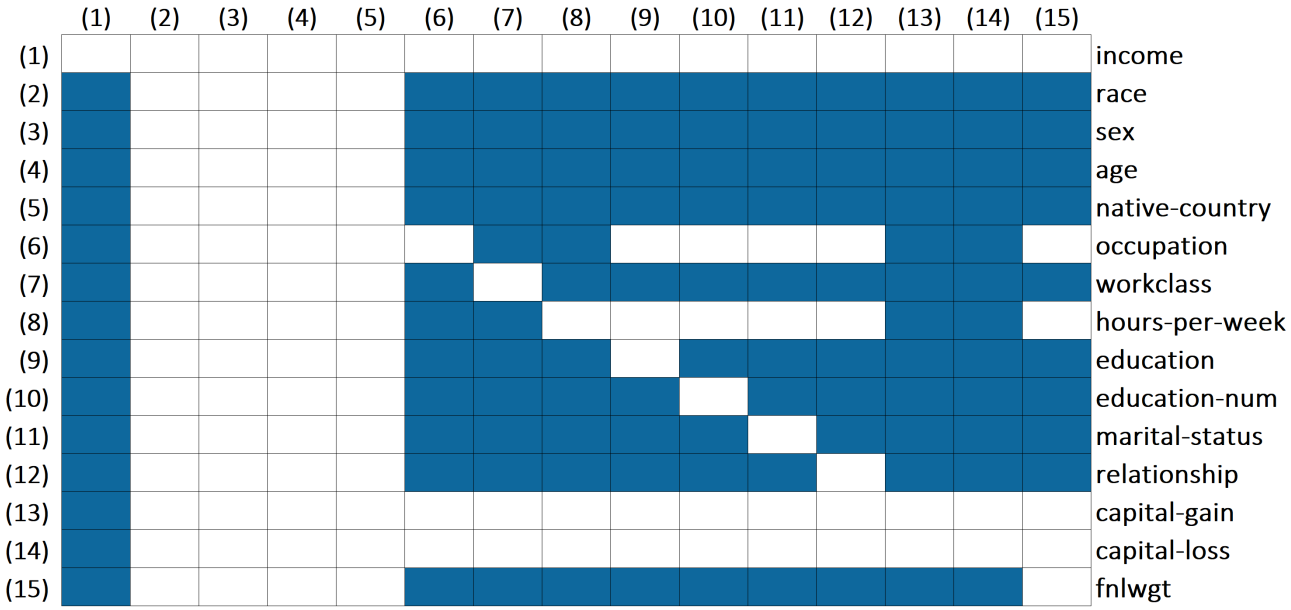


Figure 4.5: Adjacency matrix $\mathbf{W}$, representing partial causal knowledge for the Adult dataset. Blue represents edges ($w_{ik} = 1$); missing edges are shown in white (hard constraints, $w_{ik} = 0$).

- *race, sex, age, native-country* cannot be caused by any variable, i.e., these variables cannot have incoming edges. As a result, columns 2 to 5 of the adjacency matrix in Figure 4.5 are blanked ($w_{ik} = 0$).
- *occupation* and *hours-per-week* cannot cause *fnlwgt (demographics index)*, *education*, *education-num*, *relationship*, and *marital-status*; the respective cells in Figure 4.5 are blanked.
- The target variable ($income > USD50K$) cannot cause any variable, and *capital-gain* and *capital-loss* can only affect the target. Thus, rows 1, 13, and 14 in Figure 4.5 are blanked.

Note that some of these assumptions could easily be challenged, for example, by arguing that the target variable income *can* cause variables such as *capital-gain* and *capital-loss*. We adopt these assumptions solely to illustrate their effects on the adjacency matrix and to simulate a scenario in which the modeller tests whether the algorithm identifies relationships in the data that aid the prediction task. As discussed in (Pearl, 2009; Pearl and Mackenzie, 2018), the ability to extract and enforce such assumptions, or simply discuss them, has the potential to make models more transparent, robust, and representative of the world's causal mechanisms, ultimately rendering them more trustworthy.

We report the AUCs of the NNs injected with the graph in Figure 4.5, and trained on dataset of varying sizes, in Table 4.1, under the label ***Partial***. The "causal constraints" result in performances that are generally not significantly different from CASTLE, but with a computed DAG that has about half the edges of the unconstrained NN resulting from fitting CASTLE ($|E| = 116$ vs $|E| = 210$).

Overall, we obtain a sparser NN that adheres to common-sense causal knowledge based on our assumptions, making its recommendations more understandable and trustworthy while maintaining performance comparable to a regularised, but unconstrained NN.

Contesting a Computed DAG

This final experiment on real data demonstrates the process of contesting and refining the causal graph computed and used by the NN, as facilitated by Algorithm 4. Starting without prior knowledge about the problem, we use the same threshold optimisation strategy from the *No A Priori Knowledge* scenario. In the initial runs of Algorithm 4, the `revise_graph` function adjusts only the threshold τ, terminating with selecting $\tau = 0.08$.

With $\tau = 0.08$, a DAG is extracted and presented to the SMEs, as shown in Figure 4.3. In this DAG, the target variable ($income > 50K$) erroneously appears to cause several variables, including *sex* and *age* (see purple arrows). This violates common-sense and causal reasoning principles.

The contestation process focuses on addressing these specific edges, and the `revise_graph` function produces a refined DAG ($\mathcal{G}_r$) by removing the spurious purple edges. This refined DAG is then injected back into the NN, resulting in the predictive performance detailed in Table 4.1, under the label ***Refined***. Importantly, the refined DAG aligns with common sense and causal principles and maintains the same predictive performance as the injected NN using nonsensical relationships.

Additionally, when compared to the unconstrained CASTLE models, the refined NN is seven times smaller in the input layer, and achieves up to 13% higher predictive accuracy across varying sample sizes. This highlights the dual benefits of improved performance while significantly reducing model complexity, aiding interpretability.

This case study demonstrates how our framework allows causal knowledge to be contested and injected within the NN's structure. By leveraging partial knowledge or contesting nonsensical computed DAGs, the framework can improve predictive performance and interpretability. Additionally, the contesting process empowers stakeholders to identify and challenge incorrect or unethical relationships learned by the model, fostering greater transparency, accountability, and ethical compliance.

In the next section, we use synthetic data to showcase another advantage of injecting causal knowledge: the improved ability of NNs to retrieve the ground-truth causal graph underlying the DGP.

4.6.2 Experiments on Synthetic Data

To further validate the findings from our real data experiments, we assess the effectiveness of our proposed method using synthetic data. This set of experiments is designed to address the following questions:

Q1. Does knowledge injection by our algorithms improve predictive performance?

Q2. Can we reconstruct a DAG, known to be underpinning the DGP, using Algorithm 4?

Q3. How well can Algorithm 4 fill the gaps in an input graph contributing only partial knowledge?

Q4. How does knowledge injection performance change across different data size regimes?

Q5. How resilient are our algorithms to noise?

To answer these questions, we generate synthetic data from known DAGs as follows.

Synthetic Data Generation We generate synthetic data adhering to a series of randomly generated DAGs of varying sizes, following the methodology described in (Kyono et al., 2020). This process involves creating Erdös-Rényi (ER) graphs (Erdős and Rényi, 1959), which have uniform probability of edge existence, and then transforming them into DAGs by removing cycles. The DAGs are then used to generate data by initially assigning Gaussian noise to all variables and then, in topological order, adding the sigmoid of parent values to their corresponding children.[3]

The synthetic DAGs and corresponding data vary along three main dimensions: the number of nodes in $\mathcal{G}$ ($|V| \in \{10, 20, 50\}$), the number of edges ($|E| = |V| \times e$, where $e \in \{1, 2, 5\}$), and the sample size ($N = |V| \times s$, where $s \in \{50, 100, 200, 300, 500\}$). We refer to s as the *proportional sample size*.

To ensure comparability, the data is standardised so that all variables have a mean of 0 and a std of 1. As noted by Reisach et al. (2021), this standardisation can lead to conservative estimates of reconstruction accuracy, as it removes information that continuous score-based causal discovery methods might otherwise use to infer causal ordering.

Evaluation Metrics We use the average and std of *MSE* to evaluate *predictive performance* (Q1, Q4, Q5). To evaluate *reconstruction accuracy* (Q2, Q3, Q5), we analyse the distribution of the percentage of estimated edges that match the true DAG. Each scenario, involving a combination of $|V|$, e, and s, is run 10 times. Results are presented as boxplots in Figure 4.6 and compared with the baselines detailed below. As in the real data experiments, statistical significance is assessed using a two-sample t-test.

[3]For details, refer to Appendix B.1 of (Kyono et al., 2020).

Baselines CASTLE's predictive performance has been extensively compared with leading regularisation methods for NNs (Zhang et al., 2021), demonstrating superior results in both classification and regression tasks across synthetic and real datasets. Thus, for predictive performance, our baseline is a state-of-the-art regularised NN: CASTLE.

For reconstruction accuracy, however, CASTLE alone is insufficient as a baseline. Instead, we introduce a derived baseline called *CASTLE+*, built upon the adjacency matrix extracted from CASTLE's weights. Specifically, using CASTLE's weights Θ, we compute the adjacency matrix $\mathbf{W}_\Theta$ according to Definition 2.4.1. This adjacency matrix is then converted into a DAG $\mathcal{G} = \Gamma_\tau(\mathbf{W}_\Theta)$ using Algorithm 2, with an appropriate threshold τ.

The key differences between CASTLE+ and our proposed Algorithm 3 are as follows:

- **CASTLE+:** The NN is trained *unconstrained*, and the causal graph is only extracted *post-training* by applying the adjacency matrix definition (Definition 2.4.1) and transforming it into a DAG with a selected threshold τ.

- **Injected NNs:** The NN is trained *with constraints* provided by Algorithm 3. A PDAG, reflecting partial causal knowledge, is used as input. This PDAG masks a proportion of anti-causal relationships in the true graph, ensuring that only non-masked weights are optimised during training.

Effectively, CASTLE+ can be interpreted as using Algorithm 3 with a complete graph in input, meaning all causal directions are initially considered valid. In both CASTLE+ and our method, the threshold τ is selected to minimise mismatches in reconstruction accuracy.

This comparison allows us to evaluate the advantages of incorporating causal constraints during training, as opposed to inferring causal graphs only after the model is trained.

Simple DGP Using Algorithm 4, we represent the *Expert Knowledge* as an input graph $\mathcal{G}$ encapsulating partial *a priori* causal knowledge among a subset of the input variables.

In the experiments shown in Figure 4.6, we inject a 20% random sample of the total edges in the true DAG; experiments with 10% and 50% of true edges injected are reported in Appendix A.1.2, showing diminishing returns for increased injected knowledge. Once the known edges are selected, the corresponding edges in the adjacency matrix $\mathbf{W}$ are set to 0 in the opposite direction, e.g., if $X_i \to X_j$ is known, then $(i, j) \in E$ and $w_{ji} = 0$.

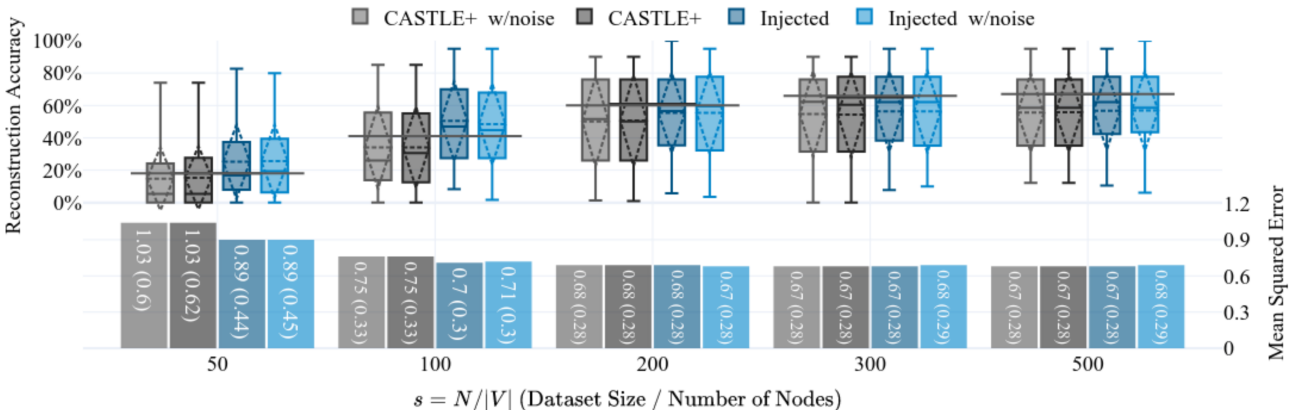


Figure 4.6: Reconstruction Accuracy and MSE for varying proportional sample size $s = N/|V| \in \{50, 100, 200, 300, 500\}$ and number of nodes $|V| \in \{10, 20, 50\}$ in the causal DAG. Darker (grey or blue) colours refer to the no-noise scenario, while lighter colours indicate the noisy DGP scenario. Each combination of $|V|$, s, and $e = |E|/|V| \in \{1, 2, 5\}$ is run 10 times. Boxplots (left y-axis) show Min/Max/Median (solid lines) and Mean/std (dashed lines) for reconstruction accuracy. Bottom bars (right y-axis) show the MSE (std). Solid horizontal lines re-base the CASTLE+ mean to account for the advantage that our Injection methodology knows 20% of the edges. If the mean of *Injected* is above/below the horizontal lines, the average increase in reconstruction accuracy is more/less than proportional to the amount injected.

The results for this scenario, presented in darker colours in Figure 4.6, show two metrics: predictive performance (MSE) represented by the bottom bars, and reconstruction accuracy represented by the boxplots. The results vary across the proportional sample size s. Analysis of changes in other dimensions ($e = |E|/|V|$ and $|V|$) can be found in Appendices A.1.2 and A.1.2.

From Figure 4.6, we observe that injection significantly improves predictive performance in low-data regimes, with gains of up to 15% for sample sizes $s < 200$ (statistically significant for $s = 50$ and $s = 100$ at the 0.05 level: $t(178) = 2.226, p = 0.027$ and $t(178) = 2.464, p = 0.015$, respectively). In contrast, no statistically significant differences in predictive performance are observed for larger sample sizes. The largest improvements in reconstruction accuracy also occur in low-data scenarios. Notably, the number of correct edges increases beyond the proportion of edges injected by up to 10%, as shown by boxplots lying above the grey longitudinal lines. However, for higher values of s, no statistically significant differences in reconstruction accuracy are detected.

These results answer Q1 through Q4: Algorithm 3 improves DAG reconstruction (Q2) and fills gaps in partial causal knowledge (Q3) without decreasing prediction performance (Q1), but primarily for low-data regimes (Q4).

Noisy DGP To assess the robustness of Algorithm 3 to noise (Q5), we add variables to the training data that amount to 20% of the number of nodes in the DAG used to generate the data. These additional noisy variables are sampled from a standard normal distribution and are disconnected from

the other variables in the data. The results are presented in Figure 4.6, comparing the no-noise and noise scenarios. As shown in the bottom bars, the MSE for the target variable Y remains effectively the same across different proportional data sizes (with no significant differences in mean). Similarly, reconstruction accuracy (top boxplots) shows no significant differences in means. This aligns with findings in (Kyono et al., 2020) for prediction. We conclude that our algorithm is resilient to noise in both reconstruction and predictive performance (Q5).

In this section, we demonstrated that Contestable NNs achieve significant improvements in transparency without sacrificing predictive performance. By reducing parameter size, improving causal graph reconstruction, and empowering stakeholders to identify and challenge problematic causal relations, our framework aligns predictive performance with interpretability and accountability.

We now summarise the key contributions and outcomes of this work, positioning them within the broader context of trustworthy AI.

4.7 Summary

This work introduces a principled framework to make NNs contestable through the integration of expert knowledge, facilitated by causal graphs. The proposed methodology enables SMEs to challenge NN models based on causal assumptions discovered from the data, addressing key challenges in robustness, transparency, and accountability for AI systems.

Two algorithms were developed to achieve this goal. The first incorporates expert feedback into NN training by modifying NNs' weights according to causal graphs representing them. The second algorithm uses computed causal graphs to elicit stakeholder input, allowing SMEs to refine assumptions and inject domain expertise into the NN. Together, these algorithms address **Objective 1** by making models accessible and contestable, while also enhancing interpretability through causal graphs, in line with **Objective 3** and ensuring formal consistency between data-driven and expert-informed causal representations, as per **Objective 2**.

Empirical results on financial datasets demonstrate that these methods yield parsimonious NNs, which are more interpretable and easier to debug, while either significantly improving or minimally reducing predictive performance. Additionally, synthetic data experiments reveal that knowledge injection enhances causal discovery in low-data scenarios and under noisy conditions, contributing to **Objective 5** by strengthening model robustness.

The broader impact of this work lies in fostering trustworthiness in AI. The framework equips NNs

with mechanisms for causal graph integration and stakeholder interaction, enabling exploration and refinement of the relationships underlying model predictions. This capability leads to models that align more closely with domain knowledge and stakeholder priorities.

Future research could explore interactions between latent causal effects and hidden NN layers, refine uncertainty quantification using Bayesian learning, and incorporate multiple expert perspectives into consistent causal graphs. These directions, further discussed in Chapter 7, promise to enhance the robustness and applicability of contestable NNs in high-stakes decision-making contexts.

In conclusion, this chapter establishes a foundation for creating interpretable, accountable, and effective NN systems that support high-stakes decision-making while aligning with stakeholder values. By integrating expert input, we enable stakeholders to validate and refine causal relationships discovered during NN training, fostering contestability and trust.

However, the reliance on stochastic methods presents a notable limitation: the lack of formal guarantees on the causal structures extracted by the NNs. Additionally, expecting SMEs to shoulder the entire burden of validation is impractical, particularly in complex systems. To overcome these challenges, we now shift focus to causal discovery methods that offer formal guarantees. These methods provide a principled foundation for extending both the applicability of Contestable NNs, and more broadly, the construction of causal models.

Chapter 5

Constraint-based Causal Discovery with Shapley Values

In this chapter, we introduce a novel constraint-based causal discovery method that leverages Shapley values to improve the identification of causal graphs from observational data, building on the approach proposed in (Russo and Toni, 2023b). We begin by outlining the motivation for enhancing current causal discovery methods, in line with our objectives (§5.1). Next, we propose an innovative **Shapley-based decision rule** for orienting v-structures in causal graphs (§5.2). Building on this, we introduce the **Shapley-PC** algorithm (§5.3), which integrates our decision rule into the PC algorithm to enhance its robustness in finite-sample settings. We also examine the theoretical properties and guarantees of Shapley-PC, highlighting its contribution to advancing constraint-based methods. Finally, we validate the effectiveness of Shapley-PC through an extensive empirical evaluation, showcasing its performance improvements over existing algorithms (§5.4). We conclude with a discussion of the broader implications of our approach for causal discovery and its alignment with the objectives of this thesis (§5.5).

5.1 Motivation

In Chapter 4, we demonstrated how integrating expert feedback into causal discovery enhances Neural Networks' interpretability and robustness. This flexible approach combines data-driven discovery with expert input to foster transparency, stakeholder engagement, and performance. Additionally, our experimental results revealed that having partial knowledge about the true causal graph can significantly aid discovery in low-data regimes. However, the stochastic nature of NNs prevents formal guarantees on structure recovery—an essential requirement in high-stakes decision-making.

However, requiring SMEs to validate complex systems can be impractical, highlighting the need for trustworthy data-driven methods to complement expert knowledge. Beyond prediction with Contestable NNs, causal graphs form the backbone of causal reasoning, and are required for robust causal inference (Pearl, 2009; Peters et al., 2017). While domain experts can provide valuable assumptions about causal relationships, eliciting complete and correct information remains a challenge in complex systems. Robust causal discovery algorithms offer a principled starting point for identifying these relationships, which can then be refined with expert input (Spirtes et al., 2000).

In this chapter, we focus on constraint-based methods, such as the PC algorithm (Spirtes et al., 2000), which provide theoretical guarantees for identifying causal structures—a critical requirement for trustworthy AI systems. These methods are modular and transparent, making them well-suited for fostering the understanding of the causal discovery process. However, despite their soundness and completeness guarantees, constraint-based methods are sensitive to finite-sample errors in CITs, which can propagate multiple edge orientation errors (Ramsey et al., 2006; Tsagris, 2019).

State-of-the-art PC-based methods, such as PC-Max (Ramsey, 2016), address these challenges by prioritising v-structures based on statistical significance (see §3.3). Yet, they remain vulnerable to errors, particularly in low-data regimes, underscoring the need for further improvements in robustness.

To address this limitation, we propose a novel approach that integrates Shapley values (Shapley, 1953) into the PC algorithm to enhance v-structure identification. Shapley values (covered in §2.5) quantify the contribution of each variable to an observed outcome, making them ideal for analysing CIT results. While Shapley values have been widely applied in fields such as economics (Ichiishi, 1983) and machine learning (Frye et al., 2020; Heskes et al., 2020; Lundberg and Lee, 2017; Teneggi et al., 2023), their use in causal discovery remains unexplored.

By extracting additional information from statistical tests, our proposed Shapley Independence Values (SIVs) mitigate finite-sample errors, improving the reliability of v-structure orientations within the PC-Stable algorithm. This enhancement strengthens the robustness of causal discovery (**Objective 6**) while producing results that are inherently transparent (**Objective 4**) due to the interpretability of Shapley values. Such transparency also provides a foundation for contestability, as it enables stakeholders to better understand and scrutinise the causal relationships identified by the algorithm.

While Contestable NNs integrate expert feedback with continuous score-based causal discovery to improve predictions and alignment to social standards, here we focus on extracting reliable causal assumptions from data. These methods serve distinct but complementary roles, providing stakeholders with transparent and robust tools for validating causal assumptions and fostering trust in AI systems.

5.2 Shapley Values to Detect Colliders

In this section, we propose a novel strategy for detecting colliders and orienting v-structures in causal graphs using Shapley values. By leveraging the Shapley value of a candidate collider, we assess its contribution to the observed association between the two disconnected variables in an UT. This approach leads to the definition of *Shapley Independence Values* (SIVs) (§5.2.1), which quantify how the conditional independence between two variables changes when a potential collider is included in the conditioning set. This mechanism provides a robust and generalisable decision rule for v-structure orientation, applicable to any causal discovery algorithm that relies on a skeleton and an association measure between variables (§5.2.2).

5.2.1 Shapley *Independence* Values (SIVs)

As discussed in §2.5, Shapley values quantify the contribution of individual players to the outcome of a cooperative game. Here, we introduce the *Independence Game*, which consists of three fundamental components: **players**, **coalitions**, and a **value function**. In this game, the players represent variables adjacent to disconnected nodes in a UT, coalitions correspond to subsets of these players forming potential conditioning sets, and the value of the game is the conditional independence between the disconnected nodes given a coalition.

The Shapley value of a variable within this game, which we term the *Shapley Independence Value* (SIV), quantifies its marginal contribution to the independence between the disconnected nodes across all possible coalitions. Below, we formally define the components of the independence game and the resulting SIVs.

Players and Coalitions Setup In the Independence Game, the players are variables that have the potential to influence the independence between two other variables, when included in a conditioning set. These players are selected based on their adjacency to the disconnected nodes in a UT, as follows:

Definition 5.2.1 (Players)**.** *Let $\mathcal{C}$ be the skeleton of a DAG, and $X_i - X_j - X_k$ a UT in $\mathcal{C}$ such that $X_i, X_j, X_k \in \mathbf{V}$. Then, the set of players $\mathbf{P}$ is defined as:*

$$\mathbf{P} = \{adj(\mathcal{C}, X_i) \cup adj(\mathcal{C}, X_k)\} \setminus \{X_i, X_k\}.$$

Coalitions, as per the original Shapley value framework (§2.5), are all possible subsets (powerset) of the players. In our Independence Game, these subsets correspond to potential conditioning sets used

to evaluate the conditional independence between the disconnected nodes X_i and X_k. By leveraging the structure of the skeleton, we focus only on subsets of the adjacency sets of either X_i or X_k:

Definition 5.2.2 (Coalitions)**.** *Let $\mathcal{C}$ be the skeleton of a DAG, and $X_i - X_j - X_k$ a UT in $\mathcal{C}$ such that $X_i, X_j, X_k \in \mathbf{V}$. Then, the set of coalitions $\mathbf{C}$ is defined as:*

$$\mathbf{C} = \{\mathbf{S} \mid \mathbf{S} \subseteq adj(\mathcal{C}, X_i) \setminus \{X_j\} \vee \mathbf{S} \subseteq adj(\mathcal{C}, X_k) \setminus \{X_j\}\}.$$

This targeted definition encourages computational efficiency and aligns with the theoretical guarantees provided by constraint-based causal discovery methods. Spirtes et al. (2000, Lemma 5.1.3) prove that if X_i and X_k are not adjacent, there exists a set of vertices $\mathbf{S}$ that either satisfies $\mathbf{S} \subseteq adj(X_i, \mathcal{G}) \setminus \{X_k\}$ or $\mathbf{S} \subseteq adj(X_k, \mathcal{G}) \setminus \{X_i\}$, such that $\mathbf{S}$ d-separates X_i and X_k in $\mathcal{G}$. This property ensures that the adjacency sets of X_i and X_k are sufficient to determine the conditional independence between these variables given any subset of their respective adjacency sets. By restricting coalitions to subsets of adjacency sets, Definition 5.2.2 improves tractability in the computation of Shapley values, while adhering to the theoretical guarantees of the PC algorithm.

Value Function The next step in defining the framework for our Independence Game is specifying the value function. The value function quantifies the conditional independence between two variables when a third variable is included in the conditioning set. To investigate the properties of our method under ideal conditions, we begin with an assumption of perfect independence information. This is formalised through the concept of a *Perfect Conditional Independence Test (Perfect CIT)*, which acts as an oracle, providing accurate results for conditional independence relationships.

Definition 5.2.3 (Perfect CIT)**.** *For any $X_i, X_j \in \mathbf{V}$ and $\mathbf{S} \subseteq \mathbf{V} \setminus \{X_i, X_j\}$, the* Perfect CIT I_∞ *is defined as:*

$$I_\infty(X_i, X_j \mid \mathbf{S}) = \begin{cases} 1 & \text{if } X_i \perp\!\!\!\perp X_j \mid \mathbf{S}, \\ 0 & \text{otherwise.} \end{cases}$$

This idealised test provides a binary output, indicating whether two variables are conditionally independent given a set of other variables. By rooting the value function in Perfect CITs, we can isolate the theoretical contributions of a candidate collider under ideal conditions, enabling precise analysis of SIVs' properties. Given all the components of the Independence Game, we can now define the SIVs.

Definition 5.2.4 (SIV)**.** *Let $\mathcal{C}$ be the skeleton of a DAG, $X_i - X_j - X_k$ be a UT in $\mathcal{C}$, with $X_i, X_j, X_k \in \mathbf{V}$, $n = |\mathbf{P}|$ as per Definition 5.2.1, $\mathbf{C}$ as per Definition 5.2.2, and $I_\infty(\cdot)$ a Perfect CIT as per Definition 5.2.3. The* SIV *ϕ_{I_∞} of X_j in the UT $X_i - X_j - X_k$ is given by:*

$$\phi_{I_\infty}(X_j, \{X_i, X_k\}) = \sum_{\mathbf{S} \in \mathbf{C}} w_{\mathbf{S}}^n [I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) - I_\infty(X_i, X_k \mid \mathbf{S})],$$

where $w_{\mathbf{S}}^n = \frac{|\mathbf{S}|!(n-|\mathbf{S}|-1)!}{n!}$ is the Shapley weighting factor as defined in Definition 2.5.1.

The SIV quantifies the unique contribution of the candidate collider X_j to the conditional independence between X_i and X_k. A negative SIV signifies that X_j induces conditional dependence between X_i and X_k, identifying X_j as a collider. With this formalisation, we now state a key theoretical result regarding the behaviour of SIVs under perfect conditional independence information.

Lemma 5.2.5. *Given a skeleton $\mathcal{C}$, a UT $X_i - X_j - X_k \in \mathcal{C}$, $X_i, X_j, X_k \in \mathbf{V}$, and a perfect CIT I_∞, the SIV of variable X_j $\phi_{I_\infty}(X_j, \{X_i, X_k\}) < 0$ if and only if X_j is a collider for X_i and X_k.*

Proof. Assume a perfect CIT I_∞ (Definition 5.2.3). The marginal contribution of $X_j \in \mathbf{V}$ for a conditioning set $\mathbf{S} \in \mathbf{C}$ (Definition 5.2.2) is: $\Delta_{X_j}(\mathbf{S}) = I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) - I_\infty(X_i, X_k \mid \mathbf{S})$.

($\Rightarrow$) Assume X_j is indeed a collider for X_i, X_k and consider the cases based on whether $\mathbf{S}$ blocks paths other than the one passing through X_j:

Case 1 ($\mathbf{S}$ blocks all other paths)

Conditioning on X_j opens a previously blocked path, inducing dependence:

$$I_\infty(X_i, X_k \mid \mathbf{S}) = 1, \quad I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) = 0, \quad \Delta_{X_j}(\mathbf{S}) = -1.$$

Case 2 ($\mathbf{S}$ does not block all other paths)

Conditioning on X_j opens an additional path, not altering dependence:

$$I_\infty(X_i, X_k \mid \mathbf{S}) = I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) = 0, \quad \Delta_{X_j}(\mathbf{S}) = 0.$$

Given the UT configuration, $X_i \notin \text{adj}(X_k)$, and adjacency faithfulness (Definition 2.2.7) ensures that at least one $\mathbf{S}$ blocks all paths. Hence:

$$\phi_{I_\infty}(X_j, \{X_i, X_k\}) = \sum_{\mathbf{S} \in \mathbf{C}} w_{\mathbf{S}}^n \Delta_{X_j}(\mathbf{S}) < 0.$$

($\Leftarrow$) Assume, instead, that X_j is not a collider for X_i, X_k, and again by case distinction:

Case 1 ($\mathbf{S}$ blocks all other paths)

Conditioning on X_j blocks the only open path, inducing independence:

$$I_\infty(X_i, X_k \mid \mathbf{S}) = 0, \quad I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) = 1, \quad \Delta_{X_j}(\mathbf{S}) = 1.$$

Case 2 ($\mathbf{S}$ does not block all other paths)

Conditioning on X_j blocks one of the open paths, but does not alter dependence:

$$I_\infty(X_i, X_k \mid \mathbf{S}) = I_\infty(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) = 0, \quad \Delta_{X_j}(\mathbf{S}) = 0.$$

By adjacency faithfulness:

$$\phi_{I_\infty}(X_j, \{X_i, X_k\}) = \sum_{\mathbf{S} \in \mathbf{C}} w_{\mathbf{S}}^n \Delta_{X_j}(\mathbf{S}) > 0.$$

Therefore, $\phi_{I_\infty}(X_j, \{X_i, X_k\}) < 0$ if and only if X_j is a collider for X_i and X_k. ■

Beyond the setting of perfect independence information, SIVs do not always guarantee the correct identification of colliders due to inaccuracies in CITs. However, a negative SIV reflects the relative contribution of a candidate collider to the conditional independence between two variables, offering a principled basis for comparing multiple candidates systematically. While our theoretical analysis of SIVs focuses on the asymptotic setting, where the weighting scheme of Shapley values does not influence results, the framework defines and aggregates contributions in a principled manner.

In the next section, we introduce our Shapley-based decision rule for orienting v-structures in causal graphs, leveraging SIVs to systematically identify colliders.

5.2.2 Shapley-based Decision Rule

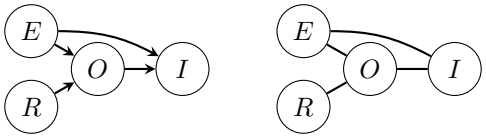
Lemma 5.2.5 established that a collider in a UT exhibits a negative SIV, reflecting its role in contributing to the conditional dependence between two other variables. Based on this insight, we propose the following *Shapley-based decision rule*:

> For any UT $X_i - X_j - X_k$, declare X_j a collider if its SIV $\phi_I(X_j, \{X_i, X_k\}) < 0$.

We illustrate this rule with two examples: one in an idealised setting with perfect independence information and another under realistic conditions with potential inaccuracies. These examples use the

causal graph from Example 2.1.1, represented here with variables *Education* (E), *Race* (R), *Occupation* (O), and *Income* (I).

Example 5.2.6 (Collider Decision with Perfect Information)**.** *Consider again Example 2.1.1; we show the true DAG (left) and its undirected skeleton $\mathcal{C}$ (right) again for reference:*



Our goal is to determine whether O is a collider in the unshielded triple $E - O - R$. Based on the skeleton's adjacency sets, $adj(\mathcal{C}, E) = \{O, I\}$ and $adj(\mathcal{C}, R) = \{O\}$. The following test results are available (e.g., for $\alpha = 0.05$):

- $I_\infty(E, R) = 1 \geq \alpha$ *($E \perp\!\!\!\perp R$),*
- $I_\infty(E, R \mid O) = 0 < \alpha$ *($E \not\perp\!\!\!\perp R \mid O$),*
- $I_\infty(E, R \mid I) = 0 < \alpha$ *($E \not\perp\!\!\!\perp R \mid I$),*
- $I_\infty(E, R \mid \{O, I\}) = 0 < \alpha$ *($E \not\perp\!\!\!\perp R \mid \{O, I\}$).*

Given $n = 2$, $w_{\mathbf{S}}^n = 0.5$, the SIV of O is computed as:

$$\begin{aligned}\phi_{I_\infty}(O, \{E, R\}) &= 0.5[I_\infty(E, R \mid O) - I_\infty(E, R)] + 0.5[I_\infty(E, R \mid \{O, I\}) - I_\infty(E, R \mid I)] \\ &= 0.5(0 - 1) + 0.5(0 - 0) \\ &= -0.5.\end{aligned}$$

Since $\phi_{I_\infty}(O, \{E, R\}) < 0$, O reduces the conditional independence between E and R, correctly identifying O as a collider and orienting the v-structure as $E \to O \leftarrow R$.

With correct tests, as shown in Example 5.2.6, the decision rules of all CPC, MPC, PC-Max, and Shapley-PC algorithms correctly infer the v-structure from the marginal independence $E \perp\!\!\!\perp R$ and the conditional dependencies between E and R given all subsets of other variables. This demonstrates that under ideal conditions, these methods are all capable of identifying colliders accurately.

However, in more realistic settings where test results are noisy or inconsistent, some algorithms may fail to orient v-structures correctly. In contrast, the Shapley-PC algorithm, by aggregating information across tests, retains its robustness to such inaccuracies. This capability is illustrated next.

Errors in Test Results In real-world scenarios, test results are often subject to noise and inconsistencies due to finite sample sizes, high-dimensional data, or other limitations (Xia et al., 2024). For example, as the size of the conditioning set increases, the corresponding data slices may shrink, leading to less reliable estimates of conditional independence. This is particularly relevant in constraint-based causal discovery, where accurate orientations depends heavily on reliable CITs (Shah and Peters, 2018).

Building on established interpretations of p-values (Hung et al., 1997; Raghu et al., 2018; Ramsey, 2016), we assume that a higher p-value indicates a greater likelihood of independence. Consequently, a lower $\phi_I(X_j, \{X_i, X_k\})$ suggests a reduced contribution of X_j to the conditional independence between X_i and X_k, increasing the likelihood that X_j is a collider. By leveraging p-values as proxies for the strength of association—as common in constraint-based causal discovery (Tsamardinos et al., 2006)—our decision rule enables more fine-grained evaluations of independence.

Furthermore, the Shapley-based approach aggregates information across multiple tests, mitigating errors from individual inconsistencies and providing a robust mechanism for collider identification. This stands in contrast to methods like PC-Max (Ramsey, 2016), which rely on a single test result for v-structure orientation and are thus more vulnerable to noisy conditions.

The following example illustrates how our Shapley-based decision rule can identify colliders even when faced with noisy or inconsistent test results.

Example 5.2.7 (Collider Decisions with Noisy Information)**.** *Consider the causal graph from Example 5.2.6, and assume the following test results for* $\alpha = 0.05$*:*

- $I(E, R) = 0.7 \geq \alpha$ *(thus* $E \perp\!\!\!\perp R$*),*
- $I(E, R \mid O) = 0.01 < \alpha$ *(thus* $E \not\perp\!\!\!\perp R \mid O$*),*
- $I(E, R \mid I) = 0.1 \geq \alpha$ *(thus* $E \perp\!\!\!\perp R \mid I$*),*
- $I(E, R \mid \{O, I\}) = 0.75 \geq \alpha$ *(thus* $E \perp\!\!\!\perp R \mid \{O, I\}$*).*

These results reflect realistic inconsistencies. For instance, $I(E, R \mid I)$ *is just above the threshold* α *while it should be 0;* $I(E, R \mid \{O, I\})$ *is instead entirely incorrect, possibly due to data fragmentation from larger conditioning sets.*

The SIV for O *is computed as* $\phi_I(O, \{E, R\}) = -0.03$*. Despite noisy test results, the negative SIV correctly identifies* O *as a collider. Other decision rules fail to orient the v-structure:*

- ***MPC and CPC*** *do not orient the v-structure due to the inconsistency between* $E \not\perp\!\!\!\perp R \mid O$ *and* $E \perp\!\!\!\perp R \mid \{O, I\}$*.*

- ***PC-Max*** *does not orient the v-structure because the maximum p-value test includes O. If instead $I(E, R \mid \{O, I\}) < 0.7$, PC-Max would have also identified the marginal independence $I(E, R) = 0.7$ as the maximum p-value and correctly oriented the v-structure.*

This example demonstrates the reduced reliance of our proposed decision rule to single (possibly erroneous) tests. However, the Shapley-based decision rule is not immune to errors in test results. We illustrate this in the following example, where the Shapley-PC algorithm fails to orient a chain structure due to inconsistencies in the conditional independence tests.

Example 5.2.8 (Chain Decisions with Noisy Information)**.** *Suppose now that the same triple $E - O - R$ considered in the previous example was not a v-structure, but instead a chain $E \to O \to R$. Under perfect conditional independence tests, we would observe:*

$$E \not\perp\!\!\!\perp R, \quad E \perp\!\!\!\perp R \mid O, \quad E \not\perp\!\!\!\perp R \mid I, \quad E \perp\!\!\!\perp R \mid \{O, I\}$$

However, if again two tests were incorrect such that $I(E, R) = 0.8$, $I(E, R \mid O) = 0 = I(E, R \mid I)$ and $I(E, R \mid \{O, I\}) = 0.7$, then:

- ***Shapley-PC*** *would calculate a negative SIV for O ($\phi_I = -0.05$), wrongly deeming it a collider.*
- ***PC-Max*** *would pick up the wrong signal, as the highest p-value is $p = 0.8$ (a wrong test).*
- ***CPC and MPC*** *would not orient the structure due to the inconsistency between the tests involving O and $\{O, I\}$.*

This highlights the conservativeness of the decision rules of CPC and MPC as a desirable trait, preventing false positives when orienting v-structures. Their stricter decision rules, which require consistency across multiple conditioning sets, act as safeguards in cases where tests are unreliable.

Overall, these examples highlight the strengths and limitations of various decision rules, illustrating how each balances conservatism and informativeness while leveraging data through CITs. The Shapley-based decision rule demonstrates its ability to aggregate information across tests, enabling it to correctly identify colliders even under noisy conditions. This approach is particularly advantageous as it mitigates the influence of individual inconsistencies in conditional independence tests.

In this section, we introduced the Shapley-based decision rule for v-structure orientation and illustrated its ability to identify colliders even with noisy test results. By comparing its performance against other decision rules, we demonstrated its robustness and flexibility. In the next section, we embed our Shapley-based decision rule into the PC algorithm, resulting in the Shapley-PC algorithm.

Algorithm 5: Shapley-PC

Input: $I(X_i, X_j \mid \mathbf{S})\ \forall\ X_i, X_j \in \mathbf{V}, \mathbf{S} \subseteq \mathbf{V} \setminus (X_i, X_j)$; α

Step 1: Adjacency Search (Colombo and Maathuis, 2014)

$\mathcal{C} \coloneqq \langle \mathbf{V}, \mathbf{E} \rangle, \mathbf{E} = \mathbf{V} \times \mathbf{V}$ ▷ Complete Graph over $\mathbf{V}$
$\mathbf{C_S} = \emptyset$
for $X_i \in \mathcal{C}$ **do**
  for $X_j \in \text{adj}(\mathcal{C}, X_i)$ **do**
    $l \coloneqq 0$
    for $\mathbf{S} \subseteq \text{adj}(\mathcal{C}, X_i)$ with $|\mathbf{S}| = l$ **do**
      $\mathbf{C_S} = \mathbf{C_S} \cup I(X_i, X_j \mid \mathbf{S})$ ▷ Store CIT Results
    $l = l + 1$
    for $\mathbf{S} \in \mathbf{C_S}$ **do**
      if $I(X_i, X_j \mid \mathbf{S}) \geq \alpha$ **then**
        Remove edge from $\mathcal{C}$ ▷ Remove only after all tests for size l
return $\mathcal{C}, \mathbf{C_S}$ ▷ Skeleton, and CIT Results

Step 2: Orient v-structures (Our Decision Rule)

for $X_i - X_j - X_k \in \mathcal{C}$ **do**
  for $\mathbf{S} \notin \mathbf{C_S} \cap \mathbf{C}$ **do** ▷ Test all coalitions in Definition 5.2.2
    $\mathbf{C_S} = \mathbf{C_S} \cup I(X_i, X_j \mid \mathbf{S})$
  if $\phi_I(X_j, \{X_i, X_k\}) < 0$ **then**
    if $X_i - X_j - X_k$ not fully directed **then**
      if do not add a cycle or bi-directed edge **then**
        orient: $X_i \rightarrow X_j \leftarrow X_k$
return $\mathcal{C}$ ▷ PDAG

Step 3: Pattern Completion (Meek, 1995)

Apply Meek's rules to $\mathcal{C}$ until no more edges can be oriented
return $\mathcal{C}$ ▷ CPDAG

5.3 Shapley-PC Algorithm

We introduce our end-to-end causal discovery algorithm, *Shapley-PC*, which integrates the Shapley-based decision rule outlined in §5.2. This enhancement builds on the foundation of the PC-Stable algorithm by leveraging Shapley values to improve collider identification.

The *Shapley-PC* algorithm, presented in Algorithm 5, retains the three main steps of the PC algorithm: adjacency search, v-structure orientation, and pattern completion (cf. §2.3, Algorithm 1). At each step, the algorithm progressively refines the causal graph, with the Shapley-based decision rule playing a critical role in Step 2 by accurately identifying colliders.

Step 1: Adjacency Search The algorithm begins by conducting an adjacency search similar to the original PC algorithm (Spirtes et al., 2000), resulting in an undirected skeleton $\mathcal{C}$ over the set of nodes $\mathbf{V}$. As discussed in §2.3, we use Stable-PC (Colombo and Maathuis, 2014) for this phase, as this modified adjacency search ensures order independence. At a high level, we start from a complete

graph (line 2) to then test each pair of variables X_i and X_j connected in $\mathcal{C}$, evaluating their conditional independence given subsets $\mathbf{S}$ of adjacent nodes. If $I(X_i, X_j \mid \mathbf{S}) \geq \alpha$, we remove the edge between X_i and X_j in $\mathcal{C}$ (lines 3-11).

Instead of storing the separating sets, we store the p-values of the independence tests for each pair of variables (line 7), which are then used to calculate the SIVs in the next step. To minimise computational complexity, the size of the conditioning sets is incrementally expanded (line 8). The key difference in this step, as compared to the original PC algorithm, is that edges are removed only after all tests for a given size of the conditioning set have been performed (line 9-11). This ensures the algorithm is order-independent (Colombo and Maathuis, 2014).

Step 2: Orienting v-structures Once the skeleton $\mathcal{C}$ is constructed, we identify UTs $X_i - X_j - X_k \in \mathcal{C}$. For each such UT, like CPC, MPC, and PC-Max, we perform additional tests beyond those used in the adjacency search. Specifically, we test the independence between X_i and X_k given all the coalitions from Definition 5.2.2 if they have not already been tested in the adjacency search (lines 14-15). We then calculate the SIVs for the candidate collider X_j using Equation 5.2.4.

If X_j has a negative SIV $\phi_I(X_j, \{X_i, X_k\})$, we declare it a collider and orient the triple as $X_i \rightarrow X_j \leftarrow X_k$ (lines 16 and 19). This step directly implements our Shapley-based decision rule defined in §5.2. Additionally, we introduce two extra checks: at line 17 we ensure we are not overwriting previously oriented triples (Ramsey, 2016), and at line 18 we check that no bi-directed edges or cycles are introduced (Ramsey, 2016; Tsagris, 2019).

Step 3: Pattern Completion Finally, we apply Meek's orientation rules (Meek, 1995) to iteratively orient additional edges in the graph until none of the rules can be applied any more. This step produces the final CPDAG representing the MEC of the true causal graph.

Compared to our reference versions of the PC algorithm in the literature, Shapley-PC introduces key refinements in the v-structure orientation step. Unlike CPC (Ramsey et al., 2006) and MPC (Colombo and Maathuis, 2014), which rely on dichotomous independence tests, our approach uses a continuous characterisation of the degree of independence. Additionally, we incorporate checks for cycles and bi-directed edges to avoid creating invalid DAGs. While PC-Max (Ramsey, 2016) also uses p-values to detect colliders and checks for bi-directed edges, it relies on the test with the maximum p-value, making it more prone to mistakes than our proposed method.

5.3.1 Theoretical Guarantees

Shapley-PC retains the theoretical guarantees of the original PC algorithm: soundness, completeness, and asymptotic consistency. These properties ensure that the algorithm correctly identifies the MEC of the true causal graph under faithfulness and perfect conditional independence information.

Theorem 5.3.1 (Correctness of Shapley-PC)**.** *Let $\mathbb{P}(\mathbf{V})$ be a joint distribution faithful to a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and assume access to perfect conditional independence information for all pairs $(X_i, X_j) \in \mathbf{V}$ given subsets $\mathbf{S} \subseteq \mathbf{V} \setminus \{X_i, X_j\}$. Then the output of Shapley-PC is the CPDAG representing the MEC of $\mathcal{G}$.*

Proof. The proof follows straightforwardly from Lemma 5.2.5 and two additional results in the literature. Assume faithfulness and perfect CITs, then:

- **Step 1**: The skeleton is guaranteed to be correct (Colombo and Maathuis, 2014, Thm. 2).
- **Step 2**: Given a correct skeleton, by Lemma 5.2.5, $\phi_I(X_j, \{X_i, X_k\}) < 0$ correctly identifies colliders in v-structures.
- **Step 3**: Given a PDAG, Meek's rules are sound and complete (Meek, 1995, Thm. 2 and 3).

Thus, Shapley-PC outputs the correct CPDAG. ■

High-Dimensional Consistency Shapley-PC inherits the high-dimensional consistency guarantees of the original PC algorithm, which apply under specific assumptions about sparsity and data properties. These guarantees hold for sparse graphs and certain distributional assumptions, such as multivariate Gaussian distributions (Kalisch and Bühlmann, 2007) or Gaussian copulas (Harris and Drton, 2013).

As outlined in (Colombo and Maathuis, 2014, 4.5), high-dimensional consistency applies when the number of observed variables d, and hence the size of the DAG $\mathcal{G}$, grow with the sample size N, such that $d_N = O(N^a)$ for some $0 \leq a < \infty$. Let $d = d_N$, $\mathbf{V} = \mathbf{V}_N = (X_{N,1}, \ldots, X_{N,d_N})$, and $\mathcal{G} = \mathcal{G}_N$. The corresponding CPDAG is denoted by $\mathcal{C}_N$, while $\hat{\mathcal{C}}_n(\alpha_N)$ represents the estimated CPDAG at significance level α_N.

The consistency result ensures that, under sparsity and suitable data properties, there exists a sequence $\alpha_N \to 0\,(N \to \infty)$ such that:

$$\mathbb{P}(\hat{\mathcal{C}}_N(\alpha_N) = \mathcal{C}_N) \to 1, \quad (N \to \infty). \tag{5.1}$$

This result is a property of the original PC algorithm (Kalisch and Bühlmann, 2007) and is referenced here to establish that Shapley-PC retains the same guarantees under the same assumptions.

The sparsity assumption necessary for these guarantees is minimal, requiring only that the neighbourhoods in the DAG are of lower order than the sample size (Kalisch and Bühlmann, 2007). This ensures that the size of the conditioning sets used in CITs is computationally manageable and statistically reliable. Shapley-PC adheres to these requirements by limiting the size of conditioning sets in CITs to the degree of the graph, aligning with the sparsity assumption.

In summary, while these assumptions are necessary for consistency, they are not unique to Shapley-PC but are inherited directly from the original PC algorithm. Consequently, Shapley-PC retains the high-dimensional consistency guarantees of PC under appropriate data conditions.

5.3.2 Additional Properties

This section examines additional key properties of the Shapley-PC algorithm, focusing on its finite-sample behaviour and computational complexity, highlighting both its strengths and limitations.

Order-Independence Shapley-PC, like CPC and MPC, is order-independent. This means that the orientation of v-structures is unaffected by the order in which variables are considered during the execution of the algorithm. While theoretical guarantees under perfect information or infinite samples are not affected by order-dependence, practical inconsistencies can arise with sample data (Colombo and Maathuis, 2014). By aggregating information from all adjacent nodes before determining causal directions, Shapley-PC ensures consistency and avoids the pitfalls of order-dependent methods.

Reduced Dependency on Significance Threshold The significance threshold α, affecting the outcomes of CITs, is a key parameter in constraint-based causal discovery (Colombo and Maathuis, 2014). Selecting an appropriate α can be challenging, as highlighted in statistical literature (Mudge et al., 2012). Shapley-PC reduces reliance on α during v-structure orientation by aggregating information across multiple tests using SIVs. This approach mitigates the effects of false positives or negatives from individual tests, offering a more robust alternative to methods like CPC, PC-Max, and MPC, which rely on single test results or majority votes.

Computed DAGs Validity Shapley-PC is robust to noisy data also due to additional checks in the decision rule. Specifically, it ensures that no cycles or bi-directed edges are introduced, upholding the acyclicity assumption. This property, shared with methods like (Tsagris, 2019) for the acyclicity checks and inspired by (Ramsey, 2016) for bi-directed edges, enhances robustness under noisy conditions.

Limitations of Skeleton Estimation The computation of $\phi_I(X_j, \{X_i, X_k\})$ depends on the adjacency sets of X_i and X_k. Errors during skeleton estimation in Step 1 of Algorithm 5—caused by violations of adjacency faithfulness (Definition 2.2.7)—can exclude relevant variables from the conditioning sets used in Step 2. This limitation, shared by all PC-based methods, underscores the importance of accurate skeleton estimation for reliable v-structure orientation and graph estimation.

Limitations of p-values While p-values are widely used as a measure of association in causal discovery (Tsamardinos et al., 2006), they have intrinsic limitations. Primarily, p-values provide a basis for rejecting the null hypothesis of conditional independence but do not directly quantify the *strength* of association. Instead, they measure evidence against independence (Casella and Berger, 2002, 8.5.2). This asymmetry introduces susceptibility to spurious results, particularly when high p-values incorrectly confirm independence.

The advantage of using p-values in our Shapley-based decision rule lies in their comparability under the alternative hypothesis (dependence), which allows us to avoid optimising α as a hyperparameter. However, under the null hypothesis (independence), p-values are uniformly distributed (see §2.1), meaning that a high p-value may incorrectly signal independence. Given a significance level α, such misclassifications may occur in up to an α proportion of the number tests, introducing noise into the orientation process.

Nonetheless, for this chapter, we use p-values as a pragmatic and accessible choice for evaluating SIVs, given their established use in constraint-based methods. This facilitates benchmarking against our baselines and isolating the contribution of our decision rule to the change in performance. We discuss potential alternatives to p-values in Chapter 7.

Complexity

The computational complexity of Shapley-PC, like other PC-based methods, is primarily influenced by the graph's sparsity and the cost of CITs. As highlighted by Spirtes et al. (2000) and Kalisch and Bühlmann (2007), it is hard to define tight complexity bounds, given the interplay between structure and amount of CITs to perform. However, we analyse the complexity of each step in Shapley-PC to provide upper bounds that can help the comparison between the different variants considered, and make sense of the empirical results we perform in §5.4.

Step 1: Skeleton Estimation Within this step, the algorithm considers all pairs of variables (X_i, X_j) and tests for independence conditioned on subsets of other variables. For a graph with $n = |\mathbf{V}|$

nodes, there are $\binom{n}{2} = O(n^2)$ pairs to consider. For each pair, subsets of neighbours are considered as conditioning sets. Let k denote the maximum degree of any node in the graph. The number of such subsets grows combinatorially as $\sum_{i=0}^{k} \binom{n-1}{i}$. Thus, following the analysis in (Spirtes et al., 2000, §5.4.2.1), the number of CITs in Step 1 is bounded by:

$$2\binom{n}{2}\sum_{i=0}^{k}\binom{n-1}{i} \approx O\left(n^2(n-1)^k\right),$$

where $O((n-1)^k)$ accounts for the exponential growth in subsets, assuming $k \ll n$.

Step 2: V-Structure Orientation For each node X_j there are $\binom{k}{2} = O(k^2)$ pairs of neighbours to form a UT. Since there are n nodes in total, the maximum number of UTs across the entire graph is $\frac{n \times k \times (k-1)}{2} = O(nk^2)$. For each UT $X_i - X_j - X_k$, there are $k-1$ possible choices for both X_i and X_k, as each can be connected to up to $k-1$ other neighbours excluding X_j. Consequently, for each UT, the algorithm tests all subsets of these neighbours for independence, leading to 2^k possible subsets per UT. Therefore, the total number of CITs introduced in Step 2 is bounded by:

$$O(nk^2 \cdot 2^k)$$

SIVs Calculating Shapley values is generally expensive due to the need to evaluate all subsets of players (Lundberg and Lee, 2017). For m players, the number of subsets is 2^m. However, in Shapley-PC, SIVs are computed *exactly* only for the UTs identified in Step 1. This targeted computation reduces the number of Shapley calculations significantly. Crucially, the computation of SIVs does not involve additional CITs; instead, it aggregates results from existing tests in Step 2 using lightweight operations like summation and weighting. Consequently, the cost of SIV computation is negligible compared to CITs, leaving the overall complexity of Shapley-PC unchanged.

Cycle Checking Shapley-PC includes an additional step to ensure that no cycles are introduced during edge orientation (line 9 of Algorithm 5). For sparse graphs, where $k \ll n$ and the number of edges $e = O(nk)$, cycle detection using Depth-First Search (DFS) scales as $O(n+e)$, which simplifies to $O(n)$ under the sparsity assumption. Over $O(e)$ edge updates, the total cost of cycle checking becomes $O(n^2)$. Notably, this step serves as a conservative safeguard to ensure acyclicity and operates independently of the rest of the algorithm. In practice, this check could be performed at the end only, reducing the cost to $O(n)$, or removed if the acyclicity is assumed.

Step 3: Pattern Completion This step is computationally lightweight compared to the previous steps, as it only involves a series of local checks to orient edges if specific configurations of three and four edges are found (Meek, 1995).

Combined Complexity Combining the costs of Step 1, Step 2, and cycle checking, the overall complexity of Shapley-PC for sparse graphs is:

$$O\left(n^2(n-1)^k + nk^2 \cdot 2^k + n^2\right)$$

As highlighted by Ramsey (2016), the computational cost of the PC algorithm is dominated by Step 1, where the majority of CITs are performed. Shapley-PC inherits this trait, as it relies on the same adjacency search process to construct the skeleton. Step 2 introduces additional costs to perform the extra tests necessary to check orientation faithfulness (Ramsey et al., 2006), which, while non-negligible, are asymptotically smaller than the exponential growth driven by Step 1 in most sparse graph scenarios.

Crucially, our Shapley value computations (SIVs) and the (optional) cycle-checking step contribute negligible overhead within Step 2. Consequently, while Shapley-PC incurs slightly higher complexity than the original PC algorithm, its overall cost remains comparable to other variants, such as CPC, MPC, and PC-Max, while providing enhanced robustness in v-structure orientation. We evaluate its empirical performance in the next section.

Practical Implications While the theoretical bounds describe worst-case scenarios and assume $k \ll n$, many graphs in practice are sparse, with relatively small k. This significantly reduces the number of subsets tested in Steps 1 and 2. Furthermore, the upper bounds assume no two variables are independent for conditioning sets of size less than k, which is unlikely in real-world datasets (Spirtes et al., 2000, §5.4.2.1). The average number of CITs is therefore much lower than the theoretical maximum.

Additionally, parallel computation techniques (Ramsey, 2016) can distribute the computation of CITs, further improving runtime efficiency. Note also that CITs can significantly vary in complexity (Zhang et al., 2011), and the actual runtime of Shapley-PC will depend on the specific type and implementation of CITs used.

In the next section, we evaluate the performance of Shapley-PC against our reference PC-based algorithms using synthetic and pseudo-real data. Specifically, we assess its ability to identify v-structures, recover causal directions, and compare its computational efficiency (Table 5.2).

5.4 Empirical Evaluation

In this section, we conduct a comprehensive simulation study to evaluate Shapley-PC against reference PC-based algorithms, including PC-Stable, CPC, MPC, and PC-Max (see §2.3 for background). The evaluation is performed on synthetic data generated from diverse graph types, sizes, and noise distributions. Additionally, we compare Shapley-PC against these baselines on pseudo-real data generated from BNs publicly available in the `bnlearn` package (Scutari, 2014).

All methods use Fisher's Z test (Fisher, 1970) as the CIT. Following Ramsey (2016), the significance threshold α is adjusted based on the number of nodes in the graph: $\alpha = 0.1, 0.05, 0.01$ for $|\mathbf{V}| = 10, 20, 50$, respectively. Full implementation details and code for reproducing the experiments are provided in §B.1 of the Appendix.

Data Generating Process (DGP) We designed our synthetic data experiments to rigorously stress-test Shapley-PC under challenging conditions, where the data are "close-to-unfaithful to the true graph" (Ramsey et al., 2006). Such scenarios, characterised by *weak links* between variables, pose significant challenges for causal discovery algorithms (Robins et al., 2003; Zhang and Spirtes, 2003). Weak links refer to edges in the true graph that are difficult to detect due to low signal-to-noise ratios, resulting in inconsistent separating sets in CITs and violations of the orientation faithfulness assumption (Definition 2.2.6). These experiments enable us to evaluate the robustness of Shapley-PC to noisy and finite-sample data.

For each experiment, we generate 10 random graphs by varying three key parameters:

- **Graph Types:** two widely used random graph models are employed:
    1. ER graphs (Erdős and Rényi, 1959): Edges are added between node pairs with a fixed probability, producing relatively uniform degree distributions.
    2. Scale-Free (SF) graphs (Barabási and Albert, 1999): These graphs feature a few highly connected hub nodes alongside many nodes with lower connectivity.
- **Graph Sizes:** we consider three graph sizes, with the number of nodes set to $|\mathbf{V}| \in \{10, 20, 50\}$.
- **Graph Densities:** three levels of graph density are evaluated, $d \in \{1, 2, 4\}$, where the total number of edges is $|\mathbf{E}| = |\mathbf{V}| \times d$.

All generated DAGs have a maximum degree of 10 to ensure sparsity and computational feasibility.

For each ground truth DAG, data are simulated using an Additive Noise Model (ANM) (Hoyer et al., 2008) defined as:

$$X_j = f_j(\text{pa}(\mathcal{G}, X_j)) + u_j, \quad \forall j \in \{1, \ldots, |\mathbf{V}|\},$$

where f_j is a linear function of the parent variables $\text{pa}(\mathcal{G}, X_j)$, and u_j represents additive noise. The weights of the linear functions are sampled randomly from a uniform distribution with parameters $[-1.5, -0.5] \cup [0.5, 1.5]$ for 95% of the effects, while the remaining 5% are sampled from $[-0.001, 0.001]$, to simulate the presence of weak edges (Ramsey et al., 2006). We vary the noise distributions, considering Gaussian, Exponential, Gumbel, and Uniform distributions.

To assess how sample size impacts performance, we vary the number of samples (N) relative to the number of nodes ($|\mathbf{V}|$). Specifically, we evaluate $N = s \times |\mathbf{V}|$, where $s \in \{100, 500, 1000\}$ is the proportional sample size. This setup allows us to explore the data-efficiency of the algorithms.

Experimental Design For each combination of parameters (graph type, size, density, noise distribution, and sample size), we generate 10 datasets. Compared to prior work (Ramsey et al., 2006), we reduced the range of graph sizes and densities to allow for a deeper exploration of graph types, noise distributions, and sample size effects. This design ensures a comprehensive evaluation of the robustness, scalability, and accuracy of Shapley-PC and baseline algorithms. Full details of the DGP are provided in §B.1.2 of the Appendix.

Evaluation Metrics In line with (Ramsey, 2016; Ramsey et al., 2006), we evaluate the performance of Shapley-PC against our baselines along two dimensions:

- Ability to **identify v-structures.**
- Ability to recover the **direction of causal arrows.**

To these ends, we use two primary metrics: V-structure F1 (V-F1) and ArrowHead F1 (AH-F1). V-F1 specifically evaluates the ability to identify v-structures, which is the main focus of our algorithm, while AH-F1 measures how well the methods recover the direction of all arrows in the graph. As a reminder, the F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. Specifically:

- **Precision:** The ratio of correct orientations among all estimated directed edges.
- **Recall:** The ratio of correctly oriented edges among all true edges.
- **F1 Score:** The harmonic mean of precision and recall, balancing these two aspects.

Table 5.1: ArrowHead (AH) and V-structure (V) F1 Scores ± std for ERd and SFd graphs of nodes $|\mathbf{V}| \in \{10, 50\}$. d is the number of edges per node in the true DAG. Bold if significantly different from the runner-up according to a t-test (see Appendix §B.1.2 for details).
Observed significance intervals: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

		Method	ER2	ER4	SF2	SF4
|V|=10	AH-F1	PC-Stable	0.36±0.28	0.15±0.16	0.67±0.23	0.32±0.27
		CPC	0.42±0.26	0.15±0.17	0.74±0.13	0.34±0.23
		MPC	0.42±0.26	0.15±0.17	0.74±0.13	0.36±0.23
		PC-Max	0.5±0.22	0.07±0.12	0.73±0.2	0.39±0.26
		Shapley-PC	**0.63±0.16 ** **	**0.25±0.16 ** **	**0.82±0.08 ** **	**0.54±0.16 ** **
	V-F1	PC-Stable	0.46±0.36	0.22±0.32	0.81±0.29	0.49±0.41
		CPC	0.59±0.33	0.23±0.33	0.88±0.18	0.54±0.36
		MPC	0.58±0.34	0.22±0.33	0.88±0.18	0.59±0.36
		PC-Max	0.67±0.3	0.09±0.24	0.9±0.22	0.59±0.4
		Shapley-PC	**0.86±0.16 *****	0.36±0.42	**0.99±0.02 ***	**0.84±0.23 ** **
|V|=50	AH-F1	PC-Stable	0.25±0.3	0.04±0.09	0.63±0.37	0.25±0.35
		CPC	0.4±0.36	0.06±0.1	0.53±0.44	0.4±0.39
		MPC	0.35±0.34	0.04±0.09	0.51±0.44	0.36±0.39
		PC-Max	0.63±0.27	0.05±0.11	0.56±0.44	0.59±0.37
		Shapley-PC	**0.75±0.06 ** **	**0.19±0.15 *****	**0.9±0.04 *****	**0.83±0.07 *****
	V-F1	PC-Stable	0.31±0.37	0.08±0.17	0.67±0.4	0.28±0.4
		CPC	0.5±0.44	0.14±0.24	0.58±0.48	0.46±0.45
		MPC	0.42±0.41	0.08±0.18	0.57±0.49	0.42±0.45
		PC-Max	0.81±0.34	0.12±0.27	0.61±0.48	0.69±0.43
		Shapley-PC	**0.98±0.03 ** **	**0.48±0.36 *****	**1.0±0.0 *****	**0.99±0.03 *****

All metrics are computed on the output CPDAGs. Details about the metrics are provided in the Appendix (§B.1.1), alongside results breakdowns of F1 into precision and recall (§B.1.2).

Results We present the performance of Shapley-PC and baseline algorithms for ER and SF graphs with 10 and 50 nodes in Table 5.1. Additional results are provided in the Appendix (§B.1.2): for $|\mathbf{V}| = 20$ and $s \in \{100, 500\}$, as they corroborate those in Table 5.1, and for $d = 1$ where no significant variations across methods are observed due to the near-optimal performance.

As visible in Table 5.1, Shapley-PC outperforms other PC-based algorithms for both ER and SF graphs with densities $d \in \{2, 4\}$. The improvement in v-structure identification (V-F1) is up to 6x for ER4 graphs and 3.5x for SF4 graphs and more than 3x higher ArrowHead F1 scores (ER2) when compared to PC-Stable.

Pairwise t-tests highlight the best results in bold if significantly different from the runner-up ($\alpha = 0.05$). Stars indicate the bin where the p-value belongs (details in Appendix §B.1.2).

Performance varies across graph types, densities, and sizes:

Table 5.2: Runtime comparison for the experiments presented in Table 5.1: median elapsed time in seconds for ER and SF graphs with nodes $|\mathbf{V}| \in \{10, 20, 50\}$.

	ER			SF		
$\|\mathbf{V}\|$	**10**	**20**	**50**	**10**	**20**	**50**
PC-Stable	0.1	0.6	4.7	0.1	0.6	2.4
CPC	0.1	0.7	5.5	0.1	0.8	4.3
MPC	0.1	0.7	5.4	0.1	0.8	4.4
PC-Max	0.1	0.7	5.5	0.1	0.8	4.5
Shapley-PC	0.1	0.7	6.3	0.1	0.8	5.2

- **Graph Type:** ER graphs are generally more challenging to retrieve than SF graphs, possibly due to the uniform degree distribution and lack of highly connected hub nodes which can simplify the dependence structure around fewer variables.

- **Density Impact:** Increasing density in ER graphs has a larger negative impact on performance compared to SF graphs, as evidenced by a steeper drop in performance from $d = 2$ to $d = 4$.

- **Graph Size:** For the same density d, larger graphs result in improved performance. For example, $d = 4$ corresponds to 40 edges in a 10-node graph, approaching the acyclic limit $\frac{|\mathbf{V}|(|\mathbf{V}|-1)}{2} = 45$. In contrast, for a 50-node graph, 200 edges represent only about 15% of the maximum edge capacity, making the graph sparser.

PC-based methods generally perform best on sparse graphs (Kalisch and Bühlmann, 2007). However, Shapley-PC demonstrates improved performance also on denser graphs, possibly due to its robustness in aggregating information across multiple tests and reducing reliance on single, potentially erroneous tests for orientation decisions.

The runtime comparison for Shapley-PC and baselines is summarised in Table 5.2. The table presents the median elapsed time in seconds for experiments on both ER and SF graphs, with varying node set sizes ($|\mathbf{V}| \in \{10, 20, 50\}$). As expected, runtimes increase with graph size across all algorithms, reflecting the growing computational burden of evaluating CITs as node and edge counts rise.

For smaller node sets ($|\mathbf{V}| = 10$ and $|\mathbf{V}| = 20$), runtime differences across algorithms are minimal. However, for $|\mathbf{V}| = 50$, a pronounced gap emerges between PC-Stable and the other algorithms, particularly for SF graphs. This highlights the impact of the maximum degree (k) on runtime, as SF graphs, with their high-degree hubs, necessitate more extensive evaluations of subsets during Step 2.

Shapley-PC consistently exhibits runtimes approximately 15% higher than PC-Max for $|\mathbf{V}| = 50$. This difference is primarily attributable to the additional cycle-checking procedure, which scales with

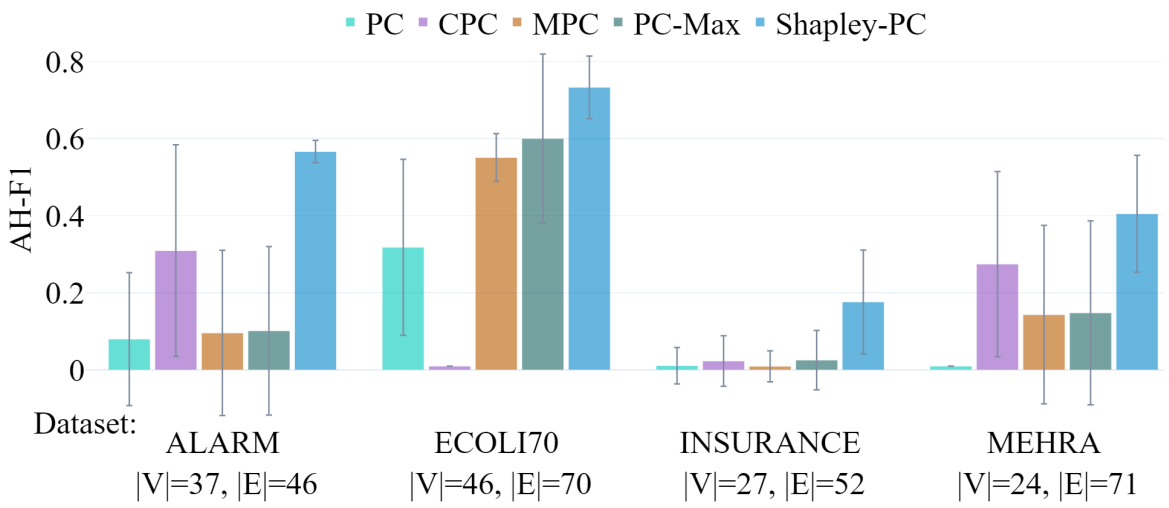


Figure 5.1: Mean and standard deviation ArrowHead F1 score for pseudo-real BNs from the `bnlearn` repository.

the number of nodes ($O(n^2)$ for sparse graphs). While the computational cost of cycle checking remains minor compared to CITs, its role is crucial in ensuring acyclic graph outputs, particularly for larger, more complex graphs.

Overall, the results underscore the influence of graph structure and degree distribution on runtime performance. For ER graphs, the uniform degree distribution limits k, leading to fewer subsets and shorter runtimes. In contrast, SF graphs demand higher computational effort due to their hub-dominated structure. Despite these challenges, Shapley-PC maintains competitive performance and provides enhanced reliability, reinforcing its suitability for causal discovery in diverse graph settings.

Pseudo-Real Data To further validate our findings, we test Shapley-PC on datasets from the `bnlearn` repository, which represent a broad range of real-world structures. These datasets are sampled from BNs with fixed conditional probability tables and categorised into discrete, Gaussian, and Conditional Linear Gaussian BNs. We sample 50,000 examples with 10 different random seeds.

The results, summarised in Figure 5.1, show the AH-F1 scores along with standard deviations. Shapley-PC consistently outperforms the other methods across all categories, achieving the highest F1 scores. On discrete datasets (Alarm, Insurance), Shapley-PC significantly outperforms competitors at the 0.05 significance level. For continuous data (Ecoli70), it is on par with PC-Max, and better than the other methods. Finally, on mixed datasets (Mehra), Shapley-PC achieves results comparable to CPC while outperforming all other methods. Details on the datasets, results for V-F1, and additional datasets from `bnlearn` are provided in the Appendix, §B.1.3.

Overall, the results demonstrate the effectiveness of Shapley-PC in recovering causal structures from finite samples, even in challenging scenarios where traditional PC-based methods may struggle. The algorithm consistently outperforms its predecessors in identifying v-structures and recovering causal directions, showcasing its robustness across diverse graph types and data distributions.

While Shapley-PC introduces some computational overhead compared to simpler algorithms like PC-Stable, it achieves competitive runtimes relative to improved variants such as PC-Max and CPC. The added cost, driven by features like cycle checking, ensures reliable and valid graph outputs. With 50-variable datasets processed in under 7 seconds, the runtime differences are minor and do not hinder scalability or real-world applicability. These findings highlight the practicality and robustness of Shapley-PC, making it a valuable tool for causal discovery in diverse settings.

5.5 Summary

In this chapter, we introduced the *Shapley-PC* algorithm, a novel causal discovery method that incorporates a *Shapley-based decision rule* to improve the robustness of constraint-based causal discovery. By leveraging Shapley values, our decision rule systematically evaluates the contributions of adjacent variables to conditional independence relations, addressing key challenges in finite-sample and quasi-unfaithful settings. Integrated into the PC-Stable framework (Colombo and Maathuis, 2014), Shapley-PC retains the theoretical guarantees of soundness, completeness, and consistency offered by the original PC algorithm (Spirtes et al., 2000) while significantly enhancing its robustness in orienting v-structures. These advancements directly address our **Objective 6** by mitigating statistical errors and enabling the generation of more accurate causal graphs.

The empirical evaluation conducted in this chapter demonstrates that Shapley-PC consistently outperforms its PC-based predecessors in both v-structure identification and overall causal direction recovery. By reducing susceptibility to finite-sample errors, the algorithm generates more reliable causal graphs, facilitating the creation of robust causal models. This improvement positions Shapley-PC as a key contribution to the advancement of constraint-based methods, with significant potential for underpinning transparent and trustworthy AI systems.

In addition to robustness, the Shapley-based decision rule introduces a principled mechanism for analysing inconsistent tests within v-structure orientation. By systematically evaluating statistical associations, it offers a transparent and interpretable approach, addressing **Objective 4**. Furthermore, by producing reliable causal graphs whose reasoning for v-structure orientation is accessible for

evaluation and refinement, the algorithm aligns with **Objective 1**, with potential to supporting domain experts in validating and contesting causal graphs.

There are several promising directions for enhancing Shapley-PC. Extending its principles to the skeleton estimation phase could reduce errors in adjacency determination, while incorporating Shapley values into pattern completion could address remaining ambiguities in edge orientation (Tsagris, 2019). Additionally, integrating domain expert feedback into the workflow would further strengthen human-AI collaboration, enabling iterative refinement of causal discoveries. These advancements would broaden the applicability of Shapley-PC in high-stakes decision-making contexts, as outlined in Chapter 7.

Despite its strengths, Shapley-PC does not guarantee consistency between its output DAG and the results of independence tests—an inherent limitation of its efficient variable selection (§2.3). Addressing such inconsistencies requires a systematic approach to resolving conflicts in test results, further enhancing the robustness of the computed causal graphs. Besides, while constraint-based methods are inherently interpretable due to their reliance on CITs and rules, they lack a formal mechanism to establish how various tests from data relate to causal assumptions in a unified framework.

To address these limitations, in the next chapter, we propose representing causal discovery as a structured set of rules and assumptions within an argumentation framework. This leads to *Causal ABA*, a novel method able to refine Shapley-PC's outputs, ensuring consistency, while incorporating expert knowledge to promote accountability and robustness.

Chapter 6

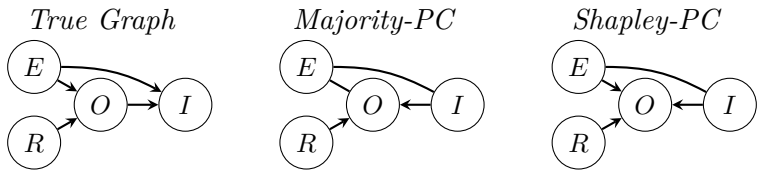
Argumentative Causal Discovery

In this chapter, building on (Russo, 2023; Russo et al., 2024), we introduce **Causal ABA**, a novel argumentation framework that accepts both statistical and causal arguments, guaranteeing robust causal discovery. After motivating the need for consistency guarantees in causal discovery (§6.1), we formalise causal graph properties within the Causal ABA framework (§6.2). This methodology systematically integrates statistical evidence and expert knowledge, ensuring that the resulting causal graphs are both interpretable and formally consistent with the data. We establish a formal correspondence between the extensions of Causal ABA and DAGs, along with their implied (in)dependencies via d-separation. We then present the **ABA-PC algorithm** (§6.3), an instantiation of Causal ABA, which uses the statistical tests from Shapley-PC to construct causal graphs. In our empirical evaluation (§6.4), we compare ABA-PC with Shapley-PC and other state-of-the-art methods, demonstrating its ability to reliably recover causal graphs. Finally, we reflect on how these methods address our objectives and contribute to advancing causal discovery (§6.5).

6.1 Motivation

In Chapter 5, we explored how finite-sample limitations impact causal graph recovery, even for algorithms that are sound and complete in the infinite-data setting. While Shapley-PC mitigates statistical errors, a key challenge remains: inconsistency between independence relations derived from data and the CPDAG output by constraint-based algorithms such as PC (Tsagris, 2019). These algorithms interpret test results locally, without enforcing global coherence across the full set of test performed, leading to causal graphs that can deviate from both the ground truth and the observed data. To illustrate this issue, we present a motivating example.

Example 6.1.1. *Consider again our Income (I) prediction Example 2.1.1. After collecting data, we conduct CITs that correctly report* Education *(E) and* Race *(R) as independent ($E \perp\!\!\!\perp R$). However, the tests also indicate independence when conditioned on* Occupation *(O) ($E \perp\!\!\!\perp R \mid \{O\}$), which contradicts intuition. Given the assumed graph,* Occupation *must depend on either* Education *or* Race*. Hence, knowing one should inform the other, making them conditionally dependent. Below, we depict the ground truth causal graph (left) and the output of Majority-PC (Colombo and Maathuis, 2014) (middle) and Shapley-PC (Chapter 5) (right), when run on a dataset sampled from a Bayesian Network with the ground truth structure.*



Since the conditional independence test incorrectly reports $E \perp\!\!\!\perp R \mid \{O\}$, it becomes impossible to retrieve the ground truth while satisfying all reported independence relations. In fact, there may be, as in this example, no graph that faithfully represents all test results.

To resolve the conflicts illustrated in the example, we propose a structured argumentative approach based on ABA. By integrating statistical outputs and domain knowledge into an ABA framework, our method identifies all DAGs consistent with the provided causal and statistical relationships. The workflow for this approach is summarised in Figure 6.1, which highlights how statistical methods and expert domain knowledge interact to produce interpretable causal graphs.

The *Causal ABA algorithm* addresses **Objective 2** by ensuring formal consistency between statistical evidence and domain expertise. It supports **Objective 4** through the intrinsic transparency of ABAFs and their ability to support interaction and interpretability, by providing arguments in the form of causal directions, (in)dependencies or combinations thereof, deriving from inference rules.

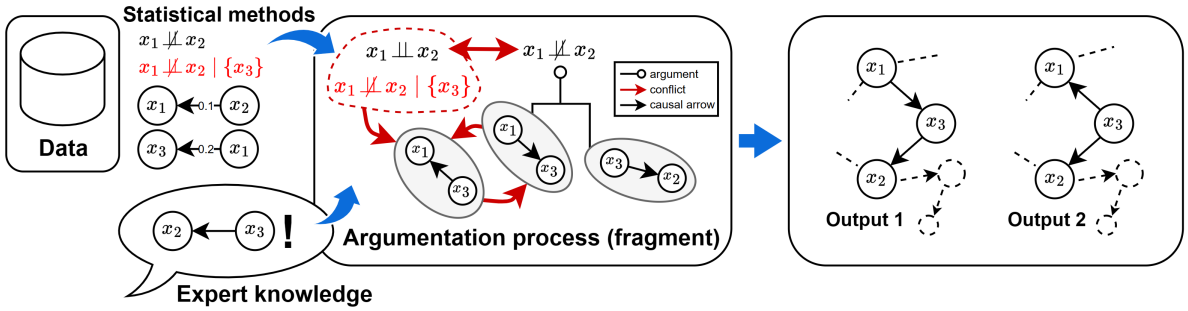


Figure 6.1: Overview of the workflow of our *Causal ABA algorithm*, which combines statistical methods and expert domain knowledge with non-monotonic reasoning to output causal graphs consistent with reported causal and statistical relationships.

Additionally, Causal ABA tackles **Objective 6** by improving robustness in handling noisy or conflicting data with formal consistency guarantees. Finally, as argued in (Leofante et al., 2024), the inherent transparency of Causal ABA, and its structured approach to resolving conflicts, provides a "contestability-ready" approach to causal discovery, aligned with **Objective 1**. Through this integration of formal and statistical methods with expert knowledge, we aim to advance the trustworthiness and applicability of causal discovery in complex, high-stakes domains, such as finance or healthcare.

6.2 Causal ABA

We formalise causal graphs using ABA. Recall that an ABAF is a tuple $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$, where $(\mathcal{L}, \mathcal{R})$ is a deductive system, $\mathcal{L}$ is a language, and $\mathcal{R}$ is a set of inference rules. We omit explicit mention of the language $\mathcal{L}$ for simplicity. The set $\mathcal{A}$ contains assumptions, and $^-$ is a contrary function mapping each assumption $a \in \mathcal{A}$ to a contrary $\overline{a} \in \mathcal{L}$. Each assumption a has a unique contrary a_c, and for clarity, we write $\overline{a}$ instead of a_c when it does not cause confusion. See §2.6 for background.

In this chapter, we present results using several argumentation semantics, including conflict-free, admissible, grounded, complete, preferred, and stable ABA semantics (abbreviated *cf*, *ad*, *gr*, *co*, *pr*, *stb*, respectively). The stable semantics will play a crucial role in our approach since we theoretically show one-to-one correspondence between extensions and DAGs, and we used it in our experiments, leveraging its correspondence to stable models in ASP. Throughout the chapter, we assume a fixed, arbitrary set of variables $\mathbf{V}$ with $|\mathbf{V}| = d$.

As in the previous chapter, we assume sufficiency (no latent variables) and acyclicity. To start with, our goal is to capture the set of all DAGs with d nodes. To achieve this, we first represent DAGs within an ABAF, through assumptions and inference rules.

6.2.1 DAGs Representation

We formalise the graph-theoretic properties needed to represent DAGs, since our intended outcome—the extensions of the ABA framework—are graphs. In this setup, the assumptions of our ABAF are arrows:

$$\mathcal{A}_{arr} = \{arr_{xy} \mid x, y \in \mathbf{V}, x \neq y\}.$$

To ensure acyclicity, we define the following ABAF:

Definition 6.2.1. *Let $D_{dag} = (\mathcal{A}_{dag}, \mathcal{R}_{dag}, ^-)$ be an ABAF, where*

$$\mathcal{A}_{dag} = \mathcal{A}_{arr} \cup \{noe_{xy} \mid x \neq y, \{x, y\} \subseteq \mathbf{V}\}.$$

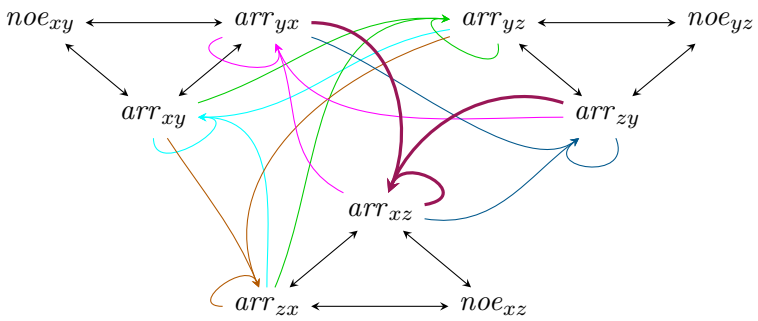
And the set $\mathcal{R}_{dag}$ of inference rules contains the following:

1. $\overline{a} \leftarrow b$, *where* $a \neq b$, $a, b \in \{arr_{xy}, arr_{yx}, noe_{xy}\}$, *and* $x, y \in \mathbf{V}$;
2. $\overline{arr_{x_i x_{i+1}}} \leftarrow arr_{x_1 x_2}, \ldots, arr_{x_{k-1} x_k}$ *for any sequence* $x_1 \ldots x_k$ *with* $x_1 = x_k$, *for all* $1 \leq i < k$.

Here, noe_{xy} represents the absence of an edge between x and y. Note that we only need to define one atom noe_{xy} for each pair of variables x, y. The first set of rules enforces a choice between noe_{xy}, arr_{xy}, and arr_{yx}. The second set ensures that no extension contains a cycle.

Example 6.2.2. *Consider $\mathbf{V} = \{x, y, z\}$. The corresponding ABAF D_{dag} contains 9 assumptions: for each pair of variables $u, v \in \mathbf{V}$, we have arr_{uv}, arr_{vu}, and noe_{uv}.*

There are two cyclic sequences of length > 2: (1) $c_1 = xyzx$ and (2) $c_2 = xzyx$. Both cycles attack every arrow they contain. The attack structure of the ABAF is depicted below.



The figure shows the ABA representation of a DAG, not a DAG itself. In this representation, the nodes correspond to assumptions, and the arrows signify attacks rather than causal relationships. This accounts for the presence of cycles and double-headed arrows in the diagram. The double-headed arrows represent symmetric attacks between assumptions, as produced by the first set of rules in Definition 6.2.1. The colourful arcs represent collective attacks against cycles, deriving from the second set of rules. For example, the thick, purple arrows pointing to arr_{xz} represent the attack from the set $\{arr_{yx}, arr_{xz}, arr_{zy}\}$ to the assumption arr_{xz} based on the derivation $\{arr_{yx}, arr_{xz}, arr_{zy}\} \vdash \overline{arr_{xz}}$.

We show that D_{dag} correctly captures the set of all DAGs of fixed size d. This correspondence holds for all (cf, ad, co, pr, stb), except the grounded semantics. Below, the assumption arr_{xy} represents the arrow (x, y).

Theorem 6.2.3. $\{(\mathbf{V}, S \cap \mathcal{A}_{arr}) \mid S \in \sigma(D_{dag})\} = \{\mathcal{G} \mid \mathcal{G} \text{ is a DAG}\}$ *for* $\sigma \in \{cf, ad, co, pr, stb\}$.

Proof. By the relations between argumentation semantics reported in (Cyras et al., 2017), i.e., $cf(D) \supseteq ad(D) \supseteq co(D) \supseteq pr(D) \supseteq stb(D)$, it suffices to prove the statement for conflict-free and stable semantics.

- $(\sigma = cf)$

  ($\subseteq$) Let $S \in cf(D_{dag})$. By definition, S cannot contain cycles; hence, S corresponds to a DAG.

  ($\supseteq$) Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG. Let $S = \{arr_{x,y} \mid (x, y) \in \mathbf{E}\}$. By acyclicity of $\mathcal{G}$, we obtain that S is conflict-free.

- $(\sigma = stb)$

  ($\subseteq$) Let $S \in stb(D_{dag})$. By definition, S cannot contain cycles; hence, S corresponds to a DAG.

  ($\supseteq$) Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG. Let $S = \{arr_{x,y} \mid (x, y) \in \mathbf{E}\} \cup \{noe_{xy} \mid (x, y), (y, x) \notin \mathbf{E}\}$. Note that S contains exactly one of arr_{xy}, arr_{yx}, noe_{xy} (since $\mathcal{G}$ does not contain a bi-directed edge and S contains noe_{xy} only if x and y are not linked in $\mathcal{G}$). Therefore, S attacks each assumption which is not contained in S. Moreover, S does not contain cycles by acyclicity of $\mathcal{G}$. Therefore, S is conflict-free. ■

Note that the grounded extension corresponds to the fully disconnected graph $\mathcal{G} = (\mathbf{V}, \emptyset)$ since the empty set is complete.

The correspondence between DAGs and the extensions of the ABAF is one-to-many for complete, admissible, and conflict-free sets. A single acyclic graph corresponds to several complete extensions. Specifically, the absence of an edge between two variables x, y can be achieved either by accepting noe_{xy} or by accepting none of noe_{xy}, arr_{xy}, or arr_{yx} in the extension.

Example 6.2.4. *In the ABAF from Example 6.2.2, the fully disconnected graph* $(\mathbf{V}, \emptyset)$ *corresponds to* 2^3 *complete extensions; i.e., each subset of* $\{noe_{xy}, noe_{yz}, noe_{zx}\}$ *represents one extension.*

For preferred and stable semantics, the correspondence between DAGs and extensions is one-to-one. In fact, the semantics coincide in D_{dag}, as formalised below.

Lemma 6.2.5. $\sigma(D_{dag}) = \tau(D_{dag})$ *for* $\sigma, \tau \in \{pr, stb\}$.

Proof. Since $pr(D) \supseteq stb(D)$, it suffices to show $pr(D_{dag}) \subseteq stb(D_{dag})$. Let $S \in pr(D_{dag})$.

To prove the equivalence, exactly one of noe_{xy}, arr_{xy}, or arr_{yx} must be in S, for each pair of variables $x, y \in \mathbf{V}$. Towards a contradiction, suppose that there exists a pair $x, y \in \mathbf{V}$ such that:

$$S \cap \{noe_{xy}, arr_{xy}, arr_{yx}\} = \emptyset.$$

Now consider the set $S' = S \cup \{noe_{xy}\}$. S' is conflict free because noe_{xy} is only attacked by arr_{xy}, arr_{yx}, neither of which are in S'; and noe_{xy} defends itself against these attacks, ensuring $S' \in co(D_{dag})$. Since $S \subseteq S'$, this contradicts $S \in pr(D_{dag})$ ($\subseteq$-maximal complete set). Therefore, for each pair $x, y \in \mathbf{V}$, exactly one of noe_{xy}, arr_{xy}, or arr_{yx} is contained in S. Thus, $S \in stb(D_{dag})$, and we conclude $pr(D_{dag}) = stb(D_{dag})$. ■

Corollary 6.2.6. *Let $\sigma \in \{pr, stb\}$. Each set $S \in \sigma(D_{dag})$ corresponds to a unique DAG $\mathcal{G}$ if and only if $\mathcal{G}$ corresponds to S.*

This corollary establishes a one-to-one correspondence between DAGs and the extensions of the ABAF, enabling us to encode the structure of DAGs within the AF. Building on this, we incorporate independence assumptions to connect DAGs with conditional independence relations, through d-separation (Definition 2.1.5).

6.2.2 Encoding d-Separation

Having encoded DAGs in the language of ABA and established their unique correspondence with the (preferred and stable) extensions of our ABAF, we now link DAGs with data via d-separation. The first step in representing d-separation is to extend our ABAF with independence statements. We add the following set of assumptions:

$$\mathcal{A}_{ind} = \{(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \mid \mathbf{Z} \subseteq \mathbf{V}, x, y \in \mathbf{V} \setminus \mathbf{Z}\}.$$

Conditional independence $x \perp\!\!\!\perp y \,|\, \mathbf{Z}$ is violated if the variables x and y are d-connected, given the conditioning set $\mathbf{Z}$. We want to formalise:

$$\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}} \text{ if there exists a } \mathbf{Z}\text{-active path between } x \text{ and } y.$$

By Definition 2.1.4, a $\mathbf{Z}$-active path is a path where the variables in $\mathbf{Z}$ do not block the connection between x and y.

To express d-separation in ABA, we formalise directed paths using the following rules, which are incorporated into the set $\mathcal{R}_{path}$:

$$dpath_{xy} \leftarrow arr_{xy} \qquad dpath_{xz} \leftarrow dpath_{xy}, arr_{yz}$$
$$e_{xy} \leftarrow arr_{xy} \qquad e_{xy} \leftarrow arr_{yx} \qquad \overline{noe_{xy}} \leftarrow e_{xy}$$

Here, e_{xy} represents the existence of an edge between x and y.

Next, we introduce *collider-trees*, which generalise the notion of paths by adding branches from collider nodes.

Definition 6.2.7. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and let $x, y \in \mathbf{V}$. An x-y-*collider-tree *t is a subgraph of $\mathcal{G}$ satisfying the following:*

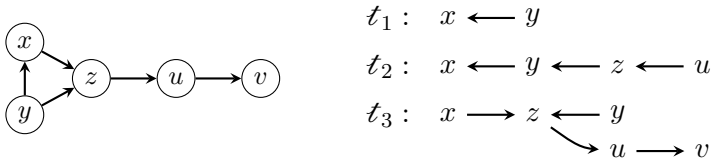
(a) t contains an x-y-path p_t;

(b) For any $u \in t$ not on p_t, there exists a collider $v \in p_t$ such that u is a descendant of v.

A path from a collider $c \in p_t$ to a leaf $l \notin \{x, y\}$ is called a branch *of t. For a set of variables $\mathbf{Z} \subseteq \mathbf{V}$, we call t* $\mathbf{Z}$-active *if p_t is active with respect to $t \cup (\mathbf{V}, \emptyset)$.*

In the remainder of the chapter, we refer to an 'x-y-collider-tree' simply as a collider-tree whenever it does not cause confusion. Let us illustrate the notion of collider-tree with an example.

Example 6.2.8. *Consider a causal graph $\mathcal{G}$ with five variables x, y, z, u, v, as depicted below (left), and some collider-trees (right):*



The collider-trees t_1 and t_2 are active with respect to $\emptyset$, as they have no colliders blocking the paths. The tree t_3 is active for sets containing z, u, and/or v.

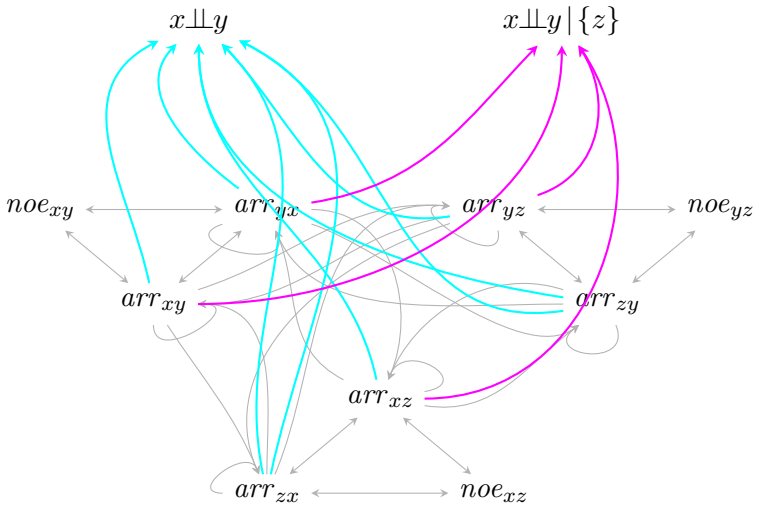
We can now define the *Causal ABA* framework as follows:

Definition 6.2.9. *Let the Causal ABAF $D_{ds} = (\mathcal{A}_{ds}, \mathcal{R}_{ds}, {}^-)$ include assumptions and rules such that*

$$\mathcal{A}_{ds} = \mathcal{A}_{dag} \cup \mathcal{A}_{ind}, \quad \text{and} \quad \mathcal{R}_{ds} = \mathcal{R}_{dag} \cup \mathcal{R}_{path} \cup \mathcal{R}_{act},$$

where $\mathcal{R}_{act}$ *contains rules* $(\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}} \leftarrow t)$ *for each* $\mathbf{Z}$*-active* x*-*y*-collider-tree* t *with* $x \neq y$*, and* $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$*.*

Example 6.2.10. *Let us go back to Example 6.2.2 with* $\mathbf{V} = \{x, y, z\}$*. Extending our ABAF with six independence assumptions and their corresponding contraries, we depict the attack structure below for the pair* x, y*:*



Here, the assumption $(x \perp\!\!\!\perp y)$ *is attacked by* arr_{xy}*,* arr_{yx}*, and all* x*-*y*-paths with inner node* z *except for the collider, which blocks the path between* x *and* y*. Instead the assumption* $(x \perp\!\!\!\perp y \,|\, z)$ *is attacked by* arr_{xy}*,* arr_{yx}*, and the collider* $\{arr_{xz}, arr_{yz}\}$*, as conditioning on the collider* z *activates the path. Attacks for the other pairs are analogous.*

The formalisation of d-separation in the ABAF allows us to establish a correspondence between the extensions of the ABAF and the conditional independences in the graph. We show that the stable extensions of D_{ds} include only the conditional independences implied by the graph. Let us first define the graph corresponding to a set of assumptions S.

Definition 6.2.11. *For a set of assumptions* S*, we denote its corresponding graph* $\mathcal{G}$ *as* $\mathcal{G}(S) = S \cap \mathcal{A}_{arr}$*.*

Theorem 6.2.12. *Let* $\sigma \in \{pr, stb\}$*,* $S \in \sigma(D_{ds})$*,* $x, y \in \mathbf{V}$*,* $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$*. Then:*

$$(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S \iff (x \perp\!\!\!\perp_{\mathcal{G}} y \mid \mathbf{Z})$$

Proof. By Lemma 6.2.5, it suffices to prove the theorem for stable semantics. This is because the attack structure of the ABAF D_{dag} was not affected by the introduction of independence assumptions, which can only be attacked by the graphical elements of the ABAF.

(⇒) Suppose $(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S$. Towards a contradiction, assume that $\mathcal{G}(S)$ implies $x \not\perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}$. By definition of d-connection, there exists a $\mathbf{Z}$-active path $\wp$ between x and y in $\mathcal{G}(S)$. Consequently, there exists a corresponding $\mathbf{Z}$-active x-y-collider-tree t, where $\wp$ serves as the connecting path (i.e., $\wp = \wp_t$).

By Definition 6.2.11, every arrow (a, b) in $\wp$ must satisfy $arr_{ab} \in S$. This means that $S \vdash \overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}$, which contradicts the conflict-freeness of S. Hence, no such $\mathbf{Z}$-active path can exist, and we conclude $x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}$.

(⇐) Suppose $\mathcal{G}(S)$ implies $x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}$. To prove that $(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S$, we need to verify that any set of assumptions T, with $T \vdash \overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}$, is attacked by S.

For $T \vdash \overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}$, there must exist a $\mathbf{Z}$-active x-y-collider-tree t, and a path $\wp_t$ connecting x and y in t such that every arrow (a, b) in $\wp_t$ must satisfy $arr_{ab} \in \mathcal{G}(T)$. Instead, at least one arrow in t must not be in S, otherwise $\mathcal{G}(S)$ would imply $x \not\perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}$, contradicting $x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}$.

Let (a, b) be an arrow in $\wp_t$ such that $arr_{ab} \notin S$. Since $S \in stb(D_{ds})$, the stability of S ensures that: $\{arr_{ab}, arr_{ba}, noe_{ab}\} \cap S \neq \emptyset$. Given $arr_{ab} \notin S$, it follows that either $noe_{ab} \in S$ or $arr_{ba} \in S$. This means that S attacks the assumption $arr_{ab} \in T$, thereby invalidating $\wp_t$ as a $\mathbf{Z}$-active path.

Since t was arbitrary, all sets of assumptions T that derive $\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}$ are attacked by S. Therefore, $(x \perp\!\!\!\perp y \,|\, \mathbf{Z})$ is defended by S, and we conclude that $(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S$. ■

Having established the correspondence for stable and preferred semantics, we now note that we cannot guarantee the correspondence for complete semantics.

Example 6.2.13. *In the ABAF from Example 6.2.10, $S = \emptyset$ is complete because D_{ds} does not contain any assumptions that are unattacked. The corresponding graph is $\mathcal{G} = (\mathbf{V}, \emptyset)$ (cf. Example 6.2.4). In $\mathcal{G}$, all pairs of variables are d-separated; however, S does not contain any independence statements.*

This example illustrates that, without $\subseteq$-maximality, we lose the reverse direction of the correspondence. However, the forward direction is preserved, as shown in the following proposition.

Proposition 6.2.14. *Let $S \in co(D_{ds})$, $x, y \in \mathbf{V}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$. Then*

$$(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S \Longrightarrow (x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}).$$

Proof. The proof follows the same reasoning as the forward direction in the proof of Theorem 6.2.12. Suppose $(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S$. Towards a contradiction, suppose $\mathcal{G}(S)$ implies $x \not\perp\!\!\!\perp_{\mathcal{G}} y \mid \mathbf{Z}$. Then there exists a

$\mathbf{Z}$-active path $\not{p}$ between x and y in $\mathcal{G}(S)$. Consequently, there is a $\mathbf{Z}$-active x-y-collider-tree $\not{t}$ where $\not{p}$ is the connecting path between x and y (i.e., $\not{p} = \not{p}_{\not{t}}$). By the definition of $\mathcal{G}(S)$, each directed edge in $\mathcal{G}(S)$, and hence in $\not{p}$, is contained in S. Hence, $S \vdash \overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}$, contradicting the conflict-freeness of S as a complete extension. ■

Thus, if a complete extension contains an independence statement, that independence must be reflected in the corresponding graph.

6.2.3 Integrating External Knowledge

Our proposed ABAF can be integrated into any causal discovery pipeline to ensure that the discovered graph provably corresponds to the relations found in the data. To achieve the integration of information from external sources—whether derived from statistical methods or expert knowledge— we treat them as *facts*, that are assumed to be true.[1]

So far, we have introduced an ABAF that faithfully captures d-separations in causal graphs. We have shown that an independence statement $(x \perp\!\!\!\perp y \,|\, \mathbf{Z})$ is contained in an extension S if and only if it is consistent with the graph corresponding to S (using the arrow assumptions). This correspondence also extends to dependencies in the graph: x and y are dependent given $\mathbf{Z}$ if and only if $(x \not\perp\!\!\!\perp y \,|\, \mathbf{Z}) \notin S$.

For an ABAF $D = (\mathcal{A}, \mathcal{R}, ^-)$ and a rule r, we write $D \cup \{r\} = (\mathcal{A}, \mathcal{R} \cup \{r\}, ^-)$ to represent the extension of D with r.

Example 6.2.15. *Consider the three-variable scenario from Example 6.2.10. Suppose we learn that x and y are marginally independent. We incorporate this information by adding the rule:*

$$(x \perp\!\!\!\perp y \leftarrow).$$

This rule ensures that each extension contains $(x \perp\!\!\!\perp y)$. Since extensions must be closed and conflict free, no active path between x and y can be accepted.

We can similarly incorporate specific causal relations into the ABAF, in the form of directed and undirected edges. In doing so we ensure that the graph inferred from the data is consistent with the external knowledge, while mantaining the correspondence between the graph and the extensions of the ABAF. We formalise this as follows:

[1] This approach assumes the reported data is sufficiently accurate. In our final system (§6.3), we assign weights to the reported independence statements to account for statistical errors.

Proposition 6.2.16. *Let* $x, y, a, b \in \mathbf{V}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$, $\mathbf{W} \subseteq \mathbf{V} \setminus \{a, b\}$, $\sigma \in \{pr, stb\}$, *and* $r \in \{(x \perp\!\!\!\perp y \,|\, \mathbf{Z} \leftarrow), (arr_{xy} \leftarrow)\}$. *For each* $S \in \sigma(D_{ds} \cup \{r\})$, *it holds that:*

$$(a \perp\!\!\!\perp b \,|\, \mathbf{W}) \in S \iff (a \perp\!\!\!\perp_{\mathcal{G}} b \,|\, \mathbf{W}).$$

Proof. By Lemma 6.2.5, it suffices to prove the proposition for stable semantics. Since S is conflict-free and closed under the rules of $D_{ds} \cup \{r\}$, the addition of r ensures that S consistently includes the assumption $x \perp\!\!\!\perp y \,|\, \mathbf{Z}$ or arr_{xy}, as required. The result then follows directly from Theorem 6.2.12. ■

However, incorporating external facts has implications for the structure of the ABAF. Specifically, the ABAF is no longer flat, as independence and arrow literals may now appear in the head of rules.

Now, what happens if we incorporate any (in)dependence or causal relations from an external source? Suppose we discover that x and y are marginally dependent. Adding the rule $(\overline{x \perp\!\!\!\perp y} \leftarrow)$ successfully renders $(x \perp\!\!\!\perp y)$ false; however, this introduces a limitation. We lose the correspondence between the (in)dependence statements and the graph of a given extension: adding the contrary of $(x \perp\!\!\!\perp y)$ does not, in the framework presented so far, prevent us from simultaneously accepting one of the arrows arr_{xy} or arr_{yx}.

To address this, we need to generalise the framework and ensure that our ABAF remains sound when dependencies are added as facts.

Blocked Paths

The ABAF D_{ds} correctly captures that an active path implies dependence. To guarantee soundness, we also need to formalise the opposite direction: independence between two nodes x and y implies that all paths linking them are blocked. To achieve this, we introduce new assumptions:

$$\mathcal{A}_{bp} = \{bp_{\not p|\mathbf{Z}} \mid \not p \text{ is an x-y-path}, \mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}\}$$

with contraries defined as:

$$\overline{bp_{\not p|\mathbf{Z}}} = ap_{\not p|\mathbf{Z}}.$$

Two new sets of rules are required: the first to require that independence between x and y given $\mathbf{Z}$ implies that all paths between them must be blocked; the second to specify when a path is $\mathbf{Z}$-active.

Definition 6.2.17. *For $x, y \in \mathbf{V}$, $x \neq y$, and $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$, we define $\mathcal{R}_{ext} = \mathcal{R}_{ds} \cup \mathcal{R}_{xy\mathbf{Z}}$, where $\mathcal{R}_{xy\mathbf{Z}}$ contains:*

- $x \perp\!\!\!\perp y \,|\, \mathbf{Z} \leftarrow bp_{\mathcal{p}_1|\mathbf{Z}}, \ldots, bp_{\mathcal{p}_k|\mathbf{Z}}$, *where $\mathcal{p}_1, \ldots, \mathcal{p}_k$ denote all paths between x and y;*
- $ap_{\mathcal{p}|\mathbf{Z}} \leftarrow \mathcal{p}_t$, *for each $\mathbf{Z}$-active x-y-collider-tree t with underlying x-y-path $\mathcal{p}_t$.*

Let us illustrate the new rules with an example.

Example 6.2.18. *Consider Example 6.2.10, but now $(x \not\perp\!\!\!\perp y \,|\, z)$ is observed. Adding $(\overline{x \perp\!\!\!\perp y \,|\, z} \leftarrow)$ prevents all $bp_{\mathcal{p}|\{z\}}$ assumptions from being accepted simultaneously. Since each extension S is closed, the assumption $(x \perp\!\!\!\perp y \,|\, z)$ cannot hold, since it leads to a conflict.*

At least one of the $bp_{\mathcal{p}|\{z\}}$ assumptions is attacked, meaning that some path between x and y must be active. In fact, the following paths are $\{z\}$-active:

$$\mathcal{p}_1 \colon x \longrightarrow y \qquad \mathcal{p}_2 \colon x \longleftarrow y \qquad \mathcal{p}_3 \colon x \longrightarrow z \longleftarrow y$$

As shown in Example 6.2.10, these paths attack $(x \perp\!\!\!\perp y \,|\, z)$. According to Definition 6.2.17, each path $\mathcal{p}_i$ derives $ap_{\mathcal{p}_i|\mathbf{Z}}$, which attacks $bp_{\mathcal{p}_i|\mathbf{Z}}$. Thus, each extension S must contain at least one of these paths, ensuring consistency with the observed dependency between x and y.

Definition 6.2.19. *For $x, y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$, we define the extended Causal ABAF $D^{xy\mathbf{Z}}_{csl} = (\mathcal{A}_{csl}, \mathcal{R}_{csl}, {}^-)$, where:*

$$\mathcal{A}_{csl} = \mathcal{A}_{ds} \cup \mathcal{A}_{bp},$$

$$\mathcal{R}_{csl} = \mathcal{R}_{ext} \cup \{\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}} \leftarrow\}.$$

The extended Causal ABAF is both sound and complete, as shown in Theorem 6.2.20.

Theorem 6.2.20. *Let $x, y, a, b \in \mathbf{V}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$, $\mathbf{W} \subseteq \mathbf{V} \setminus \{a, b\}$, $\sigma \in \{pr, stb\}$, and $S \in \sigma(D^{xy\mathbf{Z}}_{csl})$. Then:*

$$(a \perp\!\!\!\perp b \,|\, \mathbf{W}) \in S \iff (a \perp\!\!\!\perp_{\mathcal{G}} b \,|\, \mathbf{W}).$$

Proof. Let $r = \overline{x \perp\!\!\!\perp y \mid \mathbf{Z}} \leftarrow$ and $\sigma = stb$. The proof for $\sigma = pr$ is analogous.

($\Rightarrow$) Following the same reasoning as in Proposition 6.2.16, since S is conflict-free and closed under the rules of $D^{xy\mathbf{Z}}_{csl} \cup \{r\}$, the addition of r ensures that S consistently excludes the assumption $x \perp\!\!\!\perp y \,|\, \mathbf{Z}$, since its contrary must be included. Then the result follows directly from Theorem 6.2.12.

($\Leftarrow$) For the reverse direction, suppose $\mathcal{G}(S)$ implies $a \perp\!\!\!\perp_{\mathcal{G}} b \mid \mathbf{W}$. We proceed by case distinction.

Case 1 ($a = x$, $b = y$, $\mathbf{W} = \mathbf{Z}$)

Towards a contradiction, suppose $x \perp\!\!\!\perp y \mid \mathbf{Z} \notin S$.

Since S is closed, there must exist an x-y-path $\mathscr{p}$ such that $bp_{\mathscr{p}|\mathbf{Z}} \notin S$, otherwise, $S \vdash x \perp\!\!\!\perp y \mid \mathbf{Z}$, generating a contradiction.

As $bp_{\mathscr{p}|\mathbf{Z}} \notin S$, it follows that S attacks this assumption, by stability of S. Specifically, this happens if S derives $ap_{\mathscr{p}|\mathbf{Z}}$. By definition of D_{csl}, such a derivation occurs when S contains a corresponding $\mathbf{Z}$-active x-y-collider-tree $\mathscr{t}$.

Thus, x and y are d-connected given $\mathbf{Z}$, which contradicts that $\mathcal{G}(S)$ implies $x \perp\!\!\!\perp_{\mathcal{G}} y \mid \mathbf{Z}$.

Case 2 ($x \neq a$ or $y \neq b$ or $\mathbf{Z} \neq \mathbf{W}$)

The proof proceeds as in Theorem 6.2.12. To prove that $(a \perp\!\!\!\perp b \mid \mathbf{W}) \in S$, we need to verify that any T such that $T \vdash \overline{a \perp\!\!\!\perp b \mid \mathbf{W}}$ is attacked.

For any T such that $T \vdash \overline{a \perp\!\!\!\perp b \mid \mathbf{W}}$, there must exist a $\mathbf{W}$-active a-b-collider-tree $\mathscr{t}$, and a path $\mathscr{p}_{\mathscr{t}}$ connecting a and b in $\mathscr{t}$, such that every arrow (u, v) in $\mathscr{p}_{\mathscr{t}}$ satisfies $arr_{uv} \in \mathcal{G}(T)$. At least one arrow in $\mathscr{p}_{\mathscr{t}}$ must not be in S, otherwise $\mathcal{G}(S)$ would imply $a \not\perp\!\!\!\perp_{\mathcal{G}} b \mid \mathbf{W}$, contradicting $a \perp\!\!\!\perp_{\mathcal{G}} b \mid \mathbf{W}$.

Let (u, v) be an arrow in $\mathscr{p}_{\mathscr{t}}$ such that $arr_{uv} \notin S$. Since $S \in stb(D_{ds})$, the stability of S ensures that: $\{arr_{uv}, arr_{vu}, noe_{uv}\} \cap S \neq \emptyset$. Given $arr_{uv} \notin S$, it follows that either $noe_{uv} \in S$ or $arr_{vu} \in S$. This means that S attacks the assumption $arr_{uv} \in T$, thereby invalidating $\mathscr{p}_{\mathscr{t}}$ as a $\mathbf{W}$-active path.

Since $\mathscr{t}$ was arbitrary, all sets of assumptions T that derive $\overline{a \perp\!\!\!\perp b \mid \mathbf{W}}$ are attacked by S. Therefore, $(a \perp\!\!\!\perp b \mid \mathbf{W}) \in S$, as required. ■

Together, Proposition 6.2.16 and Theorem 6.2.20 guarantee that external knowledge can be integrated in a faithful way. These results ensure that the extended specification allows us to incorporate (in)dependence facts and arrows while maintaining the consistency of the Causal ABAF.

Independence facts and arrows can be added directly without requiring additional modifications to the framework. However, when incorporating dependence facts, additional rules, as specified in Definition 6.2.17, are necessary to preserve soundness and completeness.

We note that, by Theorem 6.2.20, it is sufficient to add rules only for the dependence fact we want to introduce. For instance, when adding the fact $(\overline{x \perp\!\!\!\perp y \mid \mathbf{Z}} \leftarrow)$, only the rules in Definition 6.2.17 for x, y, and $\mathbf{Z}$ need to be included. This plays an important role in the efficiency of the framework, as we can limit the number of rules added to the ABAF.

We define the extended Causal ABA framework $D^{\mathbf{T}}_{csl}$ as follows:

$$D^{\mathbf{T}}_{csl} = D_{ds} \cup \bigcup_{(\overline{x \perp\!\!\!\perp y | \mathbf{Z}} \leftarrow) \in \mathbf{T}} D^{xy\mathbf{Z}}_{csl},$$

where $\mathbf{T}$ is a set of (in)dependence and arrow facts, but $D^{xy\mathbf{Z}}_{csl}$ is updated only for the dependence facts in $\mathbf{T}$.

Corollary 6.2.21. *Let $\mathbf{T}$ be a set of (in)dependence and arrow facts, $\sigma \in \{pr, stb\}$, $x, y \in \mathbf{V}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{x, y\}$, and $S \in \sigma(D^{\mathbf{T}}_{csl})$. Then:*

$$(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S \iff (x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}).$$

We remark that the ABAF $D^{\mathbf{T}}_{csl}$ is potentially *non-flat* because assumptions can now be derived. Specifically, independence, *arr*, and *bp* assumptions may appear in the head of rules. The flexibility of the extended ABAF allows the integration of external knowledge, such as statistical results or expert knowledge, while maintaining the correspondence between the ABAF and the causal graph.

However, the non-flatness of the presented ABAF complicates the computation of extensions. We address this issue in the next section, where we present an instance, as well as an implementation, of the extended Causal ABAF.

6.3 Implementation

In this section, we present an instance of our *Causal ABA algorithm*, which combines the Causal ABAF presented in the previous section with heuristic approaches to select the independence facts provided as inputs. The pseudocode for our proposed method is in Algorithm 6, which is structured as follows:

1. The algorithm's core function, `causalaba` (lines 9 and 12), constructs the Causal ABAF $D^{\mathbf{T}}_{csl}$ using the number of nodes (d) and a set of facts ($\mathbf{T}$). This framework, detailed in §6.2, is implemented with ASP encoding in `clingo` (Gebser et al., 2019), where stable extensions represent candidate DAGs. The ASP implementation is detailed in §6.3.1.

2. The algorithm takes as input the graph size d, a significance threshold α, and a set of (in)dependence facts ($\mathcal{I}$). The process of extracting these input facts is described in §6.3.2.

3. To address potential inconsistencies caused by errors in statistical tests, we incorporate a *fact*

Algorithm 6: Causal ABA (with independence facts)

Input: $\mathcal{I}, \alpha, |\mathbf{V}| = d$

1: $\mathbf{T} \leftarrow [\,]$
2: **for** $p = I(x, y \mid \mathbf{S}) \in \mathcal{I}$ **do**
3: $\quad s \leftarrow |\mathbf{S}|$
4: $\quad$ **if** $p > \alpha$ **then**
5: $\quad\quad \mathbf{T} \leftarrow \mathbf{T} + [(\texttt{indep}(x, y, \mathbf{S}), \mathcal{S}(p, \alpha, s, d))]$
6: $\quad$ **else**
7: $\quad\quad \mathbf{T} \leftarrow \mathbf{T} + [(\texttt{dep}(x, y, \mathbf{S}), \mathcal{S}(p, \alpha, s, d))]$
8: $\mathbf{T} \leftarrow \text{sort}(\mathbf{T}, \mathcal{S})$ ▷ Sort elements of $\mathbf{T}$ by strength $\mathcal{S}$
9: $\mathbf{M} = \texttt{causalaba}(d, \mathbf{T})$
10: **while** $\mathbf{M} = \emptyset$ **do**
11: $\quad \mathbf{T} \leftarrow \mathbf{T}[2 \ldots |\mathbf{T}|]$ ▷ Drop fact with lowest $\mathcal{S}$
12: $\quad \mathbf{M} = \texttt{causalaba}(d, \mathbf{T})$
13: **for** $\mathcal{G} \in \mathbf{M}$ **do**
14: $\quad \mathcal{S}_{\mathcal{G}} = 0$
15: $\quad$ **for** $p = I(x, y \mid \mathbf{S}) \in \mathcal{I}$ **do**
16: $\quad\quad s \leftarrow |\mathbf{S}|$
17: $\quad\quad$ **if** $(p > \alpha \wedge x \perp\!\!\!\perp_{\mathcal{G}} y \mid \mathbf{S}) \vee (p \leq \alpha \wedge x \not\perp\!\!\!\perp_{\mathcal{G}} y \mid \mathbf{S})$ **then**
18: $\quad\quad\quad \mathcal{S}_{\mathcal{G}} \leftarrow \mathcal{S}_{\mathcal{G}} + \mathcal{S}(p, \alpha, s, d)$
19: $\quad\quad$ **else**
20: $\quad\quad\quad \mathcal{S}_{\mathcal{G}} \leftarrow \mathcal{S}_{\mathcal{G}} - \mathcal{S}(p, \alpha, s, d)$
21: $\mathcal{G} \leftarrow \text{argmax}(\mathcal{G} \in \mathbf{M}, \mathcal{S}_{\mathcal{G}})$ ▷ Select $\mathcal{G}$ with max $\mathcal{S}_{\mathcal{G}}$
return $\mathcal{G}$

selection mechanism. This mechanism assigns weights to facts (lines 2–7) and uses these weights to optimise candidate extensions (via weak constraints in `causalaba`) and rank the resulting DAGs (lines 10–20). We detail this process in §6.3.3.

Algorithm 6 is a sound procedure for extracting DAGs given a consistent set of independence facts, as formalised in the following proposition:

Proposition 6.3.1. *Let* $\mathbf{V}$ *be a set of variables,* $\mathcal{I}$ *a set of (in)dependence tests,* α *a significance threshold and* $\mathbf{T}$ *the output of the evaluation of* $\mathcal{I}$ *wrt* α *(lines 1-7 of Algorithm 6). Then, if* $\mathbf{T}$ *is compatible with a (set of) MEC(s), Algorithm 6 outputs a DAG consistent with* $\mathbf{T}$*.*

Proof. By Corollary 6.2.6, each stable extension $S \in stb(D_{dag})$ corresponds uniquely to a DAG $\mathcal{G}(S)$, and vice versa. Algorithm 6 constructs the Causal ABAF $D^{\mathbf{T}}_{csl}$ using the facts $\mathbf{T}$ derived from the evaluation of $\mathcal{I}$ with threshold α. Specifically:

$$p > \alpha \implies (x \perp\!\!\!\perp y \,|\, \mathbf{Z} \leftarrow) \quad \text{and} \quad p \leq \alpha \implies (\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}} \leftarrow).$$

Independence Facts:

For each evaluated $(x \perp\!\!\!\perp y \,|\, \mathbf{Z} \leftarrow) \in \mathbf{T}$, Proposition 6.2.16 ensures that:

$$(x \perp\!\!\!\perp y \,|\, \mathbf{Z}) \in S \iff x \perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}.$$

Dependence Facts:

For each evaluated $(\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}} \leftarrow) \in \mathbf{T}$, Theorem 6.2.20 guarantees that:

$$(\overline{x \perp\!\!\!\perp y \,|\, \mathbf{Z}}) \in S \iff x \not\perp\!\!\!\perp_{\mathcal{G}} y \,|\, \mathbf{Z}.$$

Therefore, for every $S \in stb(D^{\mathbf{T}}_{csl})$, the corresponding graph $\mathcal{G}(S)$ satisfies all (in)dependence relations encoded in $\mathbf{T}$, which are consistent with $\mathcal{I}$. Thus, Algorithm 6 outputs a DAG consistent with $\mathcal{I}$, as required. ■

The guarantees provided by Proposition 6.2.16 and Theorem 6.2.20 ensure that the output of Algorithm 6 is a DAG consistent with the evaluated (in)dependence facts. Specifically, each stable extension of the Causal ABAF corresponds uniquely to a candidate DAG, representing an acceptable set of assumptions that align with (a subset of) the input relations.

While the algorithm is sound and complete under perfect data and tests, practical challenges arise with finite samples and statistical noise. In the next section, we describe the implementation of this procedure using ASP.

6.3.1 Encoding Causal ABA in ASP

Stable ABA semantics and stable semantics for LPs are closely related (Caminada et al., 2015; Schulz and Toni, 2015). Importantly, this correspondence has been extended to non-flat instances in (Rapberger et al., 2024).

In standard ABA-LP translations, assumptions are associated with their default negated contraries: an assumption $a \in \mathcal{A}$ with contrary $a_c \in \mathcal{L}$ corresponds to the default negated literal *not* a_c. These translations generally consider atomic logical languages. To leverage the full power of ASP, we make adjustments to standard translations while ensuring consistency with the ABA model. For instance, the contrary of an arrow is encoded as the opposite direction of the arrow, and we employ more descriptive contrary names for clarity.

Following our approach to constructing the Causal ABAF, we represent the framework in ASP by:

1. Encoding DAGs: Each stable model (or answer set) corresponds to a DAG.
2. Encoding d-separation: Nodes x and y are independent given $\mathbf{Z}$ *if and only if* x and y are not connected via a $\mathbf{Z}$-active path.

We briefly outline the standard translation from ABA to LP. Assume an ABAF $D = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^-)$ with precisely one contrary for each assumption. For an atom $p \in \mathcal{L}$, we define:

$$rep(p) = \begin{cases} not\ \overline{p}, & \text{if } p \in \mathcal{A}, \\ \overline{a}, & \text{if } p = \overline{a} \in \overline{\mathcal{A}}. \end{cases}$$

We extend this operator element-wise to ABA rules:

$$rep(r) = rep(head(r)) \leftarrow \{rep(p) \mid p \in body(r)\}.$$

For an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$, we define the *associated LP*:

$$P_D = \{rep(r) \mid r \in \mathcal{R}\}.$$

Following the standard ABA to LP translation, each ABA atom arr_{xy} is translated to $not\ \overline{arr_{xy}}$. With a slight deviation from the Causal ABAF, to enhance the clarity of the LP, we identify "not arr_{yx}" simply with **arrow**(x,y) and "not noe_{xy}" with **edge**(x,y). Listing 1 contains our DAG encoding, where the predicate **var(·)** indicates variables.

In Line 1, the program enforces the choice between an arrow or the absence of an arrow; the remaining lines enforce acyclicity. The code faithfully captures the basis of our Causal ABAF: each answer set of Module Π_{dag} corresponds to precisely one DAG of size $|\mathbf{V}|$ for a given set $\mathbf{V}$ of variables.

To link data and DAGs, we encode the d-separation criterion. Using ASP, we represent sets (k-th

Listing 1: Module Π_{dag}

```
1 arrow(X,Y)|arrow(Y,X)|not edge(X,Y) ← var(X),var(Y), X!=Y.
2 edge(X,Y)←  arrow(X,Y), X!=Y, var(X), var(Y).
3 edge(X,Y)←  edge(Y,X), X!=Y, var(X), var(Y).
4 ← arrow(X,Y), not edge(X,Y), var(X), var(Y).
5 ← edge(X,Y), not arrow(X,Y), not arrow(Y,X), var(X), var(Y).
6 ← arrow(Y,X), arrow(X,Y), X!=Y, var(X), var(Y).
7 dpath(X,Y)←  arrow(X,Y), X!=Y, var(X), var(Y).
8 dpath(X,Y)←  arrow(X,Z), dpath(Z,Y), var(X).
9 ← dpath(X,X), var(X).
```

set $S \subseteq \mathbf{V}$) with predicates **in**(k,x). Module Π_{col} in Listing 2 encodes colliders and collider descendants, leveraging natural specifications of the **arrow** and **dpath** (directed path) predicates.

Listing 2: Module Π_{col}

```
1 collider(Y,X,Z) ← arrow(X,Y),arrow(Z,Y),X!=Y,var(X),var(Y),var(Z).
2 coll_desc(N,Y,X,Z) ← collider(Y,X,Z),dpath(Y,N).
3 nb(N,X,Y,S) ← in(N,S), collider(N,X,Y).
4 nb(N,X,Y,S) ← not in(N,S),coll_desc(Z,N,X,Y),in(Z,S),var(N),var(X),var(Y).
5 nb(N,X,Y,S) ← not in(N,S),not collider(N,X,Y),var(N),var(X),var(Y),set(S),N!=X,N!=Y,X!=Y.
```

Next, we introduce *non-blocking* nodes. A node $n \notin \{x, y\}$ in an x-y-path is non-blocking, given $\mathbf{S}$, if:

- n is a collider and $n \in \mathbf{S}$, or
- n is a collider and a descendant of n, $z \in \mathbf{S}$, or
- n is not a collider and $n \notin \mathbf{S}$.

Lines 3-5 in Module Π_{col} in Listing 2 encode these rules.

For each pair $x, y \in \mathbf{V}$ and each set $S \setminus \{x, y\}$, we add rules $\Pi_{ap}(\mathit{p}, (v_i)_{i \leq k})$ as specified in Listing 3 to ensure that if x and y are connected via a $\mathbf{Z}$-active path, then they are dependent.

Listing 3: Module $\Pi_{ap}(\mathit{p}, (v_i)_{i \leq k})$

```
1  ap(v_1,v_k,p,S) ← (arrow(v_i,v_{i+1}))_{i<k}, not in(v_1,S), not in(v_k,S), set(S), (nb(v_i,v_{i-1},v_{i+1},S))_{1<i<k}.
2  dep(v_1,v_k,S) ← ap(v_1,v_k,p,S).
```

Listing 4: Module $\Pi_{bp}(x, y, (\mathit{p}_i)_{i \leq k})$

```
1 indep(x,y,S) ← (not ap(x,y,p_i,S))_{i≤k}, not in(x,S), not in(y,S), set(S).
```

To capture the reverse direction of the correspondence (if dependency, then $\mathbf{Z}$-active path), we use Module $\Pi_{bp}(x, y, (\mathit{p}_i)_{i \leq k})$ in Listing 4. These rules ensure that the absence of an active path implies independence. $(\mathit{p}_i)_{i \leq k}$ denotes the list of all paths between x and y. The Module Π_{bp} in Listing 4 encodes the blocked path rules defined in Definition 6.2.19. The **indep(·)** and **dep(·)** predicates take two variables x, y, and a set $\mathbf{S}$ of variables as arguments.

The number of paths between two variables can grow exponentially (up to $\lfloor (d-2)!e \rfloor$), posing a significant computational challenge. To mitigate this, we leverage the observation that fixing independence facts effectively removes edges between nodes. This observation is commonly utilised in the adjacency search phase of constraint-based algorithms (Spirtes et al., 2000) (see §2.3). When fixing independence facts ($a \perp\!\!\!\perp b \,|\, \mathbf{W} \leftarrow$), we restrict our analysis to paths in the skeleton that exclude the edge (a, b).

By Theorem 6.2.20, fixing dependence facts requires only the addition of the corresponding blocked path rules. Specifically, when adding the fact $(a \not\perp\!\!\!\perp b \,|\, \mathbf{W})$, it suffices to introduce the rules $\Pi_{ap}(\not{p}, (v_i)_{i\leq k})$ and $\Pi_{bp}(a, b, (\not{p}_i)_{i\leq k})$ to ensure correctness.

6.3.2 Sourcing Facts: The ABA-PC Algorithm

As outlined in §6.2, the proposed ABAF efficiently generates all DAGs compatible with a set of fixed facts, which can represent both conditional/marginal (in)dependencies and (un)directed causal relations (arrows and edges). In the instantiation of Causal ABA as ABA-PC we consider here, we input a set of (in)dependence facts, weighting them according to their p-values. This weighting scheme prioritises the most statistically robust relationships, ensuring a principled integration of data into the causal discovery process.

A DAG with d nodes is fully characterised by $\frac{1}{2}d(d-1)2^{d-2}$ (in)dependence relations, leading to exponential growth in the number of relations as the number of nodes increases. Consequently, it is computationally infeasible to carry out all possible tests, as implemented in ASPCR (Hyttinen et al., 2014). To address this computational challenge, several efficient strategies have been proposed in the Causal Discovery literature (Colombo and Maathuis, 2014; Spirtes et al., 2000; Tsamardinos et al., 2006) (see §3.3).

Approaches like those in (Colombo and Maathuis, 2014; Spirtes et al., 2000; Tsamardinos et al., 2006) rely on conditional independence tests such as Fisher Z (Fisher, 1970). Alternatively, score-based methods like (Chickering, 2002a; Ramsey et al., 2017) employ statistical metrics to assess the value of adding or removing an arrow, returning arrow weights as a measure of fit.

In this instantiation, we utilise the proposed Shapley-PC from Chapter 5 to source facts. As detailed in §3.3, the independence tests carried out by Majority-PC (Colombo and Maathuis, 2014), Conservative-PC (Ramsey et al., 2006), PC-Max (Ramsey, 2016) and our Shapley-PC are the same; we will refer only to Shapley-PC for conciseness. Under the assumptions of sufficiency (no unmeasured confounders), faithfulness (data represents a DAG), and perfect conditional information (no estimation error), CPC, MPC and Shapley-PC are provably capable of recovering the ground-truth CPDAG from data. The following example illustrates the use of Shapley-PC in our motivating example.

Example 6.3.2 (Collider Decision with Noisy Information)**.** *We run the Shapley-PC algorithm on the*

scenario described in Example 6.1.1, performing 23 out of 24 possible tests, including the following:

$$E \perp\!\!\!\perp R \mid \{I\} \qquad R \perp\!\!\!\perp I \mid \{E\} \qquad E \perp\!\!\!\perp R$$
$$E \perp\!\!\!\perp R \mid \{O\} \qquad R \not\perp\!\!\!\perp I \mid \{E, O\} \qquad E \perp\!\!\!\perp R \mid \{O, I\}$$

Among these, only $E \perp\!\!\!\perp R$ is correct. The sole other valid d-separation in $\mathcal{G}$, $R \perp\!\!\!\perp_{\mathcal{G}} I \mid \{E, O\}$, is mistakenly classified as a dependence. The remaining 17 tests correctly yield dependencies.

Based on these erroneous results, Shapley-PC produces the graph shown in Example 6.1.1 (right), which deviates from the ground truth. Crucially, this graph fails to capture the independence relations listed above. In fact, no single graph can satisfy all the test results, as it is impossible for E and R to be independent given any set (implying they are disconnected in the graph) while simultaneously ensuring that E and I, as well as R and I, are dependent (indicating connectivity). The dependencies imply a path between E and R, resulting in a contradiction.

Algorithm 6 is flexible regarding the source of facts. For example, it could leverage the tests performed by (Tsamardinos et al., 2006) or, with minor modifications, incorporate arrow weights from score-based methods such as FGS (Ramsey et al., 2017).

6.3.3 Weighting Facts

In this subsection, we outline our strategy for weighting the results of independence tests based on their p-values and the size of their conditioning sets. These weights are used as weak constraints to rank both facts and extensions. As a result of using stable semantics to evaluate our ABAF, wrong tests can render empty extensions if they contradict another (set of) test(s). Our goal is to exclude such erroneous tests that create conflicts and prevent the framework from producing a valid extension.

To rank p-values from independence tests, we use a simple heuristic that is independent of whether the p-value falls below or above the significance threshold α. The ranking is defined by the following normalising function:

$$\gamma(p, \alpha) = \begin{cases} 2p\alpha - 1 & \text{if } p < \alpha, \\ \frac{2\alpha - p - 1}{2(\alpha - 1)} & \text{otherwise.} \end{cases}$$

The plot of γ over the p-value interval, for three commonly chosen levels of α, is shown in Figure 6.2.

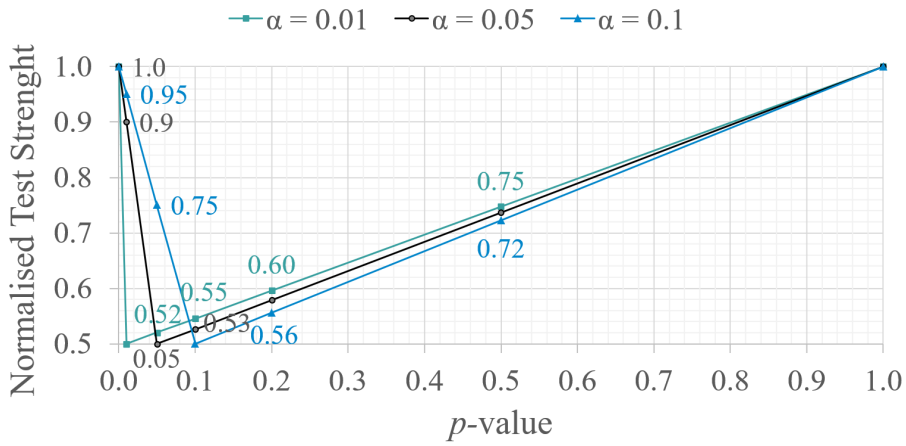


Figure 6.2: Plot of the normalising function $\gamma(p, \alpha)$ for different significance levels α.

The function γ follows the intuition that uncertainty is highest around the significance threshold $p = \alpha$ (Berger, 2003; Sellke et al., 2001), which corresponds to the lowest value of $\gamma = 0.5$. The further the p-value moves from α, the more confident we become about whether the test indicates dependence or independence. The final weight of the (in)dependence facts is determined by scaling the output of γ by a factor penalising larger conditioning set sizes $\mathbf{S}$:

$$\mathcal{S}(p, \alpha, s, d) = \frac{(1 - s)}{(d - 2)} \gamma(p, \alpha), \tag{6.1}$$

where $s = |\mathbf{S}|$ is the size of the conditioning set, and $d = |\mathbf{V}|$ is the total number of variables in $\mathcal{G}$.

The scaling factor reflects the intuition that the accuracy of independence tests decreases as the size of the conditioning set increases (Sellke et al., 2001), due to the need for more precise estimation of frequencies when conditioning on additional variables.

Using these weights $\mathcal{S}$, we rank the tests performed by the Shapley-PC. Our strategy is simple:

> Exclude an incremental number of the lowest-ranking tests until the returned stable extension is not empty.

This approach systematically removes tests that are more likely to be erroneous, ensuring that the ABAF remains consistent and outputs valid extensions. Let us illustrate our strategy with an example.

Example 6.3.3. *Consider again Example 6.1.1. The results of the independence tests from Shapley-PC (using Fisher's Z (Fisher, 1970) and $\alpha = 0.05$) have the following p-values. We show the same subset*

of the 23 tests carried out by the algorithm, as in Example 6.3.2:

$E \perp\!\!\!\perp R$	$p = 0.45$	$\mathcal{S} = 0.71$
$E \perp\!\!\!\perp R \mid \{I\}$	$p = 0.52$	$\mathcal{S} = 0.37$
$E \perp\!\!\!\perp R \mid \{O\}$	$p = 0.33$	$\mathcal{S} = 0.32$
$R \perp\!\!\!\perp I \mid \{E\}$	$p = 0.05$	$\mathcal{S} = 0.25$
$E \perp\!\!\!\perp R \mid \{O, I\}$	$p = 0.39$	$\mathcal{S} = 0.00$
$R \not\perp\!\!\!\perp I \mid \{E, O\}$	$p = 0.03$	$\mathcal{S} = 0.00$

We apply Equation 6.1 to calculate $\mathcal{S}$. Ranking the tests by $\mathcal{S}$, as shown above, the highest-ranking test is the correct one. However, fixing all the tests results in no solution.

We start excluding the test with the lowest strength and progressively exclude more until we find a model. In this example, the correct DAG is obtained by excluding 9 of the 23 tests performed, including the bottom five of the list above, while retaining the 14 strongest tests.

In this example, we obtain exactly one DAG by excluding approximately 40% of the tests performed by Shapley-PC. If multiple models are returned, we score each of them using the approach outlined in Algorithm 6, lines 13–21. Additionally, we encode the (in)dependence facts as *weak constraints*, treated as optimisation statements within ASP (Gebser et al., 2011), to prioritise optimal extensions. Our weighting method resembles the function proposed by Bromberg and Margaritis (2009), with two notable differences:

1. We re-base the strength around 0.5 instead of $1 - \alpha$, allowing for finer discrimination.

2. We incorporate the size of the conditioning set irrespective of the test's result, rather than considering it only in the case of dependence. This ensures that we account for the reduced accuracy of independence tests with larger conditioning sets, and avoid trusting that p-values reflect the probability of Type I errors even when conditioning on multiple variables.

We remark that classical independence tests are inherently asymmetric. Dependence can only be inferred when the null hypothesis is rejected ($p < \alpha$), with an expected Type I error rate corresponding to α. Conversely, when $p \geq \alpha$, no strict inference can be made, as p-values are uniformly distributed in $[0, 1]$ under the null hypothesis ($\mathcal{H}_0$). However, higher p-values are less likely to occur under the alternative hypothesis of dependence (Hung et al., 1997), which justifies the use of γ to rank p-values.

As highlighted by Bromberg and Margaritis (2009) and Hyttinen et al. (2014), directly using p-values as strength is common but neither sound nor consistent. Transforming p-values into probability estimates (Claassen and Heskes, 2012; Jabbari et al., 2017; Triantafilou et al., 2014) could address this issue. However, we leave this as an interesting avenue for future research.

We now turn to the empirical evaluation of ABA-PC, comparing it against established causal discovery methods on pseudo-real datasets.

6.4 Empirical Evaluation

The goal of this evaluation is to assess the overall performance of ABA-PC in learning causal graphs from data, gauging its ability to process data supported by its strong consistency guarantees. We compare ABA-PC against established baselines using datasets from the `bnlearn` repository (Scutari, 2014): Asia, Cancer, Earthquake, and Survey. These datasets represent real-world decision-making challenges across medical, legal, and policy domains (see Appendix §C.1.1 for details).

Unlike Chapter 5, which focused on evaluating decision rules for v-structure orientation and graph retrieval, this chapter broadens the scope to evaluate algorithms' overall performance in causal discovery tasks. Note that, compared to Chapter 5, we used smaller datasets due to the current implementation of ABA-PC, which can handle datasets with up to 10 variables. Future work to improve scalability is discussed in Chapter 7.

Baselines We compare ABA-PC against three baselines, representing diverse causal discovery approaches (see §3.3 for an overview):

- **Shapley-PC (SPC)**: Our proposed constraint-based algorithm from Chapter 5, that uses the same independence tests results that feed into Causal ABAF as input facts.
- **Fast Greedy Search (FGS)** (Ramsey et al., 2017): A score-based algorithm using the BIC to measure fit between graphs and data, and to guide a greedy search.
- **NOTEARS-MLP** (Zheng et al., 2020): A continuous score-based method that leverages NNs to learn causal structures, granting good flexibility to capture non-linearity in data.

To ground results, we also use a **Random baseline** as in (Lachapelle et al., 2020), which consists of randomly sampled graphs with the correct dimensions ($\mathbf{V}$, $\mathbf{E}$). Additional baselines, including MPC (Colombo and Maathuis, 2014) and ASPCR (Hyttinen et al., 2014), are presented in the Appendix

(§C.1.3), as the methods do not outperform the ones presented herein. While a direct comparison with the only other argumentation-based approach to causal discovery (Bromberg and Margaritis, 2009) would have been ideal, no implementation is available.

Evaluation Metrics To quantify performance, we use Structural Interventional Distance (SID) (Peters and Bühlmann, 2015), which measures the discrepancy between the estimated and true CPDAGs in terms of interventional distributions. SID evaluates how errors in causal graph estimation translate into downstream errors in causal inference, making it a meaningful proxy for real-world impact and an important metric in causal discovery evaluation (Reisach et al., 2021). Specifically, it quantifies the number of falsely inferred intervention effects. Since SID is sensitive to graph size, we normalise it (NSID) by dividing by the number of edges in the true DAG, enabling comparisons across datasets with varying numbers of variables.

As discussed in §2.1, CPDAGs represent MECs, which define the upper bound of what can be inferred from a set of independence relations. To ensure consistency in evaluation, we compute SID on CPDAGs for all methods. For ABA-PC and NOTEARS-MLP, which output DAGs, we transform these into CPDAGs by isolating skeletons and v-structures. This transformation ensures that the evaluation is conducted on a comparable representation. Nonetheless, comparisons on DAGs are provided in Appendix §C.1.3.

NSID values can exceed 100% because SID counts discrepancies at the level of intervention effects rather than edges. A single edge misrepresentation in the graph can result in multiple incorrect interventional implications due to the transitive nature of causal relationships. Details on metric calculations are in Appendix §C.1.2, and additional results based on other commonly used metrics such as Structural Hamming Distance (SHD), F1 score, precision, and recall are in Appendix §C.1.3.

Results The results of our experiments are presented in Figure 6.3. Best and worst-case SID are reported for each dataset, where lower NSID values indicate better performance. The number of edges and node is displayed below the dataset name on the x-axis.

ABA-PC achieves strong overall performance, ranking among the top-performing methods in most of the experiments. In the worst-case scenario, ABA-PC achieves the lowest values on the Earthquake, Survey, and Asia datasets, significantly outperforming all other baselines, including Shapley-PC, on Earthquake and Asia (as confirmed by t-tests for mean differences; see §C.1.4). On the Survey dataset, ABA-PC matches the performance of NOTEARS-MLP, while Shapley-PC remains competitive, performing closely to the top-performing methods.

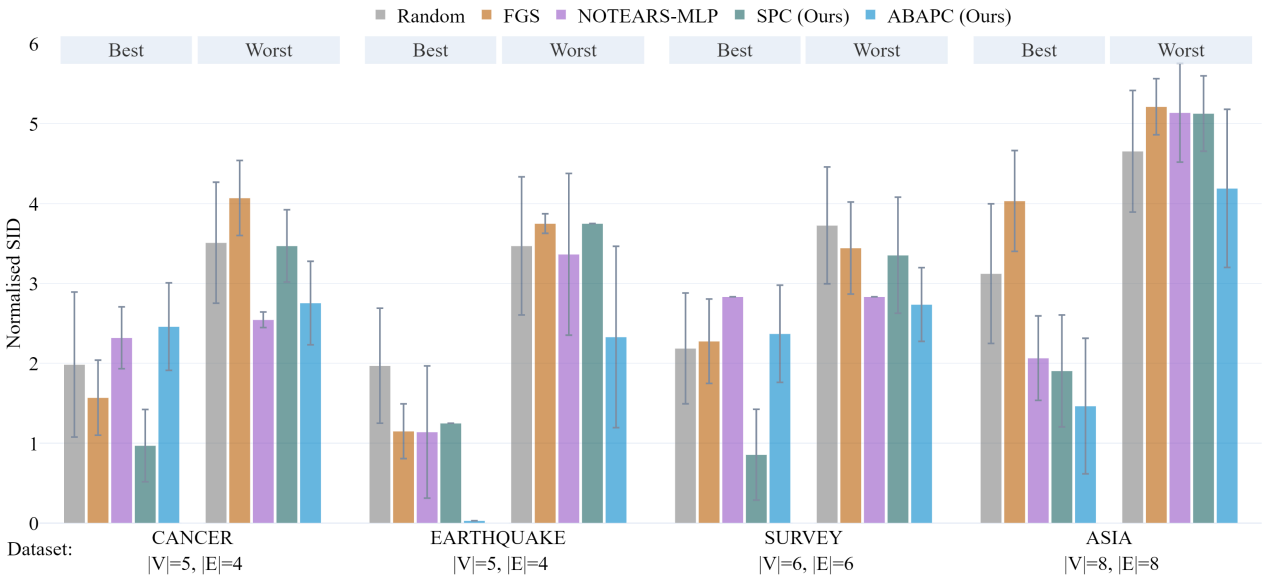


Figure 6.3: Normalised Structural Interventional Distance (NSID) for four datasets from the `bnlearn` repository. Lower values are better. Low (resp. High) represents the SID for the best (resp. worst) DAG within the estimated CPDAGs.

In the best-case scenario, ABA-PC again demonstrates strong performance. It achieves the lowest NSID values on the Earthquake and Asia datasets, significantly outperforming all baselines with reductions of up to 100% for Earthquake, where it makes no mistakes. On the Cancer datasets, NOTEARS-MLP perform slightly better, but ABA-PC remains competitive, with performance close to the leading methods. Shapley-PC also shows solid results in the Best SID case, particularly on the Cancer and Survey datasets, where it significantly outperforms all baselines.

Overall, ABA-PC demonstrates consistent performance across diverse datasets, particularly excelling in the worst-case scenarios where it estimates the ground-truth graph more effectively than the other methods for three of the datasets. Shapley-PC remains a strong baseline, especially in the best-case scenarios. Together, these results highlight the improvements introduced by Causal ABA, while reaffirming the strengths of Shapley-PC.

We note that, even if the empirical results do not uniformly see ABA-PC outperforming all baselines, the method's stronger consistency guarantees and robustness to worst-case scenarios make it a valuable addition to the causal discovery toolbox. Additionally, the ability to refine and contest causal graphs through argumentation provides a unique and valuable feature that is not present in other methods.

Scalability In Figure 6.4 we show the runtime (log scale) as a function of the number of nodes, recorded for the experiments in Figure 6.3 over 50 repetitions per dataset. While ABA-PC achieves promising results in terms of accuracy, it does not scale when increasing the number of nodes.

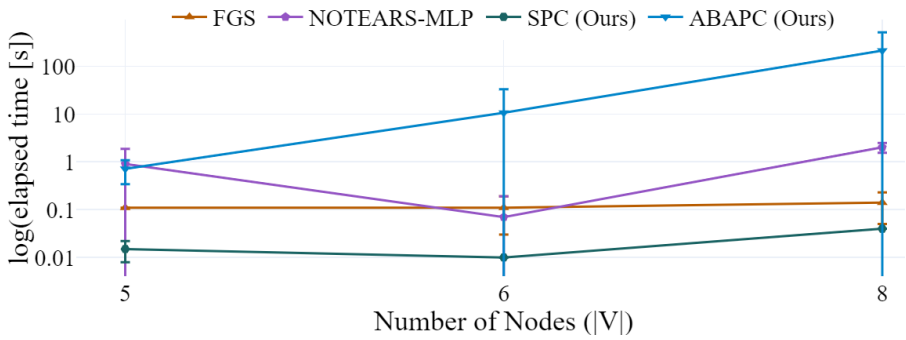


Figure 6.4: Mean and std of log elapsed time by number of nodes, averaged over 50 runs.

Improving the computational efficiency of ABA-PC without sacrificing its stronger consistency guarantees is a critical direction for future work. Nevertheless, the algorithm's ability to achieve superior accuracy and consistent causal graphs—particularly in worst-case scenarios—highlights its value as a robust and theoretically sound approach to causal discovery.

6.5 Summary

In this chapter, we introduced *Causal ABA*, a novel argumentation-based framework for causal discovery that systematically addresses inconsistencies in data while ensuring formal consistency between statistical evidence and domain expertise. By leveraging the non-monotonic reasoning capabilities of ABA, *Causal ABA* formalises the properties of causal graphs, offering a principled approach to resolving conflicts in independence test results and incorporating external knowledge.

A key theoretical contribution of *Causal ABA* is its formal correspondence (**Objective 2**) between stable and preferred extensions of ABA and causal graphs, enabling rigorous reasoning about causal relationships in a structured framework. This general framework provides the foundation for *ABA-PC*, a practical instantiation of *Causal ABA* that uses the Shapley-PC algorithm (Chapter 5) as its underlying statistical engine. By embedding statistical outputs within an argumentative structure, *ABA-PC* produces transparent causal graphs (**Objective 4**) that stakeholders can refine and contest (**Objective 1**). Moreover, the structured reasoning of *Causal ABA* enhances robustness **Objective 6** by resolving ambiguities and conflicts.

Empirical evaluations demonstrated that *Causal ABA*, instantiated through *ABA-PC*, can outperform existing methods in terms of generating consistent and accurate causal graphs. Unlike approaches such as Majority-PC (Colombo and Maathuis, 2014) or our Shapley-PC, which often tolerate ambiguities to manage inconsistencies, *Causal ABA* resolves such conflicts systematically, ensuring that the

resulting causal graphs are both grounded in evidence and free from contradictions.

Looking ahead, several promising research directions emerge. Enhancing the scalability of *Causal ABA* is a key priority, particularly through optimised processing of extensions and incremental solvers for large-scale graphs. Extending the framework to handle latent variables, feedback loops, and alternative ABA semantics would further increase its flexibility and applicability. Finally, integrating domain expertise interactively could strengthen alignment with stakeholder values, fostering greater trustworthiness for high-stakes decision-making. We expand on these directions in Chapter 7.

In conclusion, *Causal ABA* represents a significant theoretical and practical advancement in causal discovery. By bridging statistical evidence and domain knowledge through a robust argumentation framework, it lays the groundwork for creating interpretable, consistent, and contestable causal graphs. This foundational contribution provides a stepping stone for the development of trustworthy AI systems capable of supporting complex real-world decisions.

Chapter 7

Conclusion

This chapter concludes the thesis by summarising its key contributions (§7.1), outlining potential directions for future work (§7.2), and discussing the broader implications of the research findings (§7.3). The methodologies developed in this work tackle critical challenges in AI systems, emphasising robustness, transparency, and accountability while aligning with international principles of trustworthy AI. By integrating causal discovery, expert knowledge, and computational argumentation, this research advances AI systems from predictive tools to robust and interpretable decision-support systems designed to tackle real-world challenges. These advancements have significant implications for high-stakes domains such as healthcare, finance, and public policy, providing a foundation for broader adoption and further innovation. The following sections detail the thesis' contributions and propose specific opportunities for expanding its scope and impact.

7.1 Contributions

This thesis addresses fundamental challenges in trustworthy AI, specifically robustness, transparency, and accountability, through innovative methodologies that integrate causal discovery, ML, game-theory, computational argumentation, and knowledge injection. By enhancing the trustworthiness of causal discovery to the benefit of both prediction and causal inference, this work advances the design of AI systems aligning data-driven insights with expert knowledge. The thesis introduces game-theoretic and argumentation-based approaches to enhance the robustness and transparency of causal discovery, alongside frameworks for making NNs contestable, bridging predictive accuracy with interpretability and accountability. We evaluate the thesis' contributions in relation to the objectives set out in Chapter 1.

Contestable Neural Networks The first contribution of this thesis is the introduction of Contestable NNs, a principled framework to explore and challenge NNs. Unlike traditional "black-box" ML models, this approach directly involves Subject-Matter Experts (SMEs) during or after model training, through interfaces designed to scrutinise the model's discovered relationships and injecting domain-specific causal knowledge. Through the contestation process, stakeholders gain practical tools to interrogate and influence predictive models. This contribution fills critical gaps in transparency and accountability, advancing the integration of NNs into high-stakes decision-making processes.

In this context, causal graphs act as the central representation of the model's "understanding" of the DGP, enabling stakeholders to contest predictions and align the model with expert insights. This work transforms an existing NN architecture by embedding expert feedback directly into its causal graph representation, moving beyond conventional weight regularisation. By incorporating causal discovery as a foundational step in training, the approach provides a transparent and interpretable framework that strengthens both predictive performance and stakeholders' comprehension of the underlying phenomenon.

As detailed in Chapter 4, this methodology enhances transparency through a graphical interface and induced model parsimony, while enhancing predictive performance and fostering active stakeholder engagement through knowledge injection. These advancements fulfil **Objective 3 (transparency)**, **Objective 1 (contestability)**, and **Objective 2 (formal consistency)**. Empirical evaluations further demonstrate robustness to noisy and small datasets, directly contributing to **Objective 5 (robustness)**.

Shapley-based Casual Discovery The second contribution is the development of a Shapley-based approach to causal discovery, introducing a novel decision rule for causal orientation in constraint-based algorithms. Central to this contribution is the study of Shapley Independence Values (SIVs), which we establish as theoretically relevant for the accurate identification of v-structures—a critical step in causal graph orientation.

The Shapley-based decision rule is integrated into the PC-Stable algorithm, a sound and efficient constraint-based method for causal discovery, resulting in the Shapley-PC algorithm. By leveraging the game-theoretic concept of Shapley values to quantify and weight independence tests, the algorithm mitigates errors in statistical tests, significantly improving the reliability of causal discovery, even under noisy or quasi-unfaithful conditions.

Crucially, the Shapley-PC algorithm retains the soundness, completeness, and consistency guarantees of the original PC framework while delivering superior empirical performance in v-structure identification and causal direction recovery. Its robustness is validated through extensive experimentation on synthetic and pseudo-real datasets, where it consistently outperforms existing methods, in scenarios characterised by finite-sample noise or quasi-unfaithfulness. These results highlight Shapley-PC's ability to perform reliably under conditions that challenge traditional causal discovery methods.

As demonstrated in Chapter 5, these advancements contribute to **Objective 6 (robustness)** by improving reliability in noisy settings where common assumptions might fail. Additionally, the integration of Shapley values offers a transparent rationale for algorithmic decisions about causal directions, aligning with **Objective 4 (transparency)** and paving the way for future contestable causal discovery systems that support **Objective 1 (contestability)**.

Causal ABA The final contribution of this thesis is the creation of Causal ABA, a novel approach that combines Assumption-Based Argumentation (ABA) with causal discovery methods. This framework provides a structured mechanism to reconcile inconsistencies and align causal graphs with both statistical evidence and expert knowledge.

The core contribution of Causal ABA is its ability to represent both causal and statistical hypotheses as arguments within an ABA framework. Conflicts between these arguments, such as those arising from noisy data or statistical inaccuracies, are resolved through a formal reasoning mechanism that ensures consistency and transparency. This approach effectively manages noisy datasets while integrating diverse sources of information, tackling key limitations of traditional methods.

We instantiate Causal ABA with the statistical tests generated by our Shapley-PC algorithm, resulting in ABA-PC. Empirical evaluations demonstrate that ABA-PC can improve the accuracy of causal discovery while guaranteeing outputs that identify the most reliable statistical relationships and the graph(s) which are consistent with them.

As demonstrated in Chapter 6, Causal ABA systematically resolves conflicts in the input data, ensuring outputs that have consistent statistical and causal arguments, thereby achieving **Objective 2 (formal consistency)**. It improves reliability by excluding input arguments prone to errors, contributing directly to **Objective 6 (robustness)**. Furthermore, the framework's transparent argumentation-based methodology has potential to allow stakeholders to engage with, and refine, the resulting causal graphs, fulfilling **Objective 4 (transparency)** and fostering **Objective 1 (contestability)**.

By introducing Contestable NNs, Shapley-PC and Causal ABA for prediction and causal discovery, this research equips stakeholders with tools to actively engage with AI systems, bridging the gap between predictive accuracy, robust causal inference and interpretability. These methodologies provide practical solutions for high-stakes decision-making, while aligning with ethical standards and stakeholder needs. Moreover, the findings of this work highlight opportunities to further develop and adapt these innovations to real-world contexts.

7.2 Future Work

Building on the contributions of this thesis, several avenues for future research emerge, aimed at enhancing scalability, expanding real-world applicability, and deepening integration with emerging technologies. These directions not only stem from the findings of this research but are also inspired by the evolving challenges of advancing trustworthy AI systems.

7.2.1 Contestable Neural Networks

A promising direction of work involves the further development of Contestable NNs. While this thesis demonstrates the feasibility of integrating causal graph injection into NN training, some **methodological aspects** would benefit from further investigation. One key direction is to explore the indirect causal effects that may emerge within the hidden layers of injected NNs. Such effects, including latent interactions and spurious correlations, could obfuscate the true causal relationships under analysis. Adapting techniques like GranDAG (Lachapelle et al., 2020) could help identify and mitigate these effect through path analysis, thereby improving the interpretability and reliability of contestable NNs. In addition, incorporating the Gumbel-Softmax trick from (Ng et al., 2022) alongside our thresholding approach could enhance the process of identifying the most significant causal relationships within the model. While the threshold would remain valuable for interpretability, the Gumbel-Softmax approach could serve as a principled and automated starting point for this analysis.

Incorporating Bayesian learning into the NN weights estimation presents another promising opportunity to **advance human-AI collaboration**. Bayesian techniques, such as those proposed by Gal and Ghahramani (2016), would enable the model to quantify uncertainty in both its predictions and the extracted causal assumptions. This feature would allow stakeholders to assess the reliability of estimated causal structures and refine injected knowledge accordingly, thereby strengthening contestability and trustworthiness.

Moreover, incorporating mechanisms for aggregating diverse expert perspectives into a consistent causal graph would enhance the usability of contestable NNs in collaborative decision-making scenarios, such as policy development or interdisciplinary research. Approaches like our proposed Causal ABA, the one proposed by (Alrajeh et al., 2020) for causal models, or by (Ioannou and Michael, 2024) for dialogue arbitration could serve as a foundation for such advancements.

Expanding the **debugging capabilities** of contestable NNs is also critical. High-stakes decision-making models require detailed explanations and robust interaction mechanisms to ensure reliability and fairness (Miller, 2023). Future research could explore the design of intuitive interfaces and interaction modalities that enable SMEs to effectively interrogate, contest, and refine model outputs.

An *Evaluative AI* framework, as proposed by Miller (2023), where various options are analysed to elicit expert feedback, could be adapted to build effectiveness and unbiased interfaces for contestable NNs. Evaluating these interfaces in terms of ease of use, effectiveness, and their ability to foster stakeholder trust would provide valuable insights for deploying contestable NNs in practice. Additionally, the contestable NNs framework could incorporate argumentative explanations, such as those proposed in (Rago et al., 2021, 2023), to aid deeper understanding of model relations during model refinement.

The scalability of contestable NNs also warrants attention. Extending the framework to accommodate high-dimensional datasets and more complex causal relationships will require optimisation techniques that maintain effective collaboration between models and SMEs. The uncertainty quantification capabilities mentioned earlier could guide SMEs towards the relationships that require the most attention, creating a dynamic and efficient feedback loop. We hypothesise that these enhancements would enable contestable NNs to accommodate the demands of large-scale, real-world applications.

In the empirical evaluation of Chapter 4, I played the part of the SME to illustrate how the contestation process could work. To assess the value added of Contestable NN, **user studies** could be conducted to evaluate the impact of our method on model understanding, error detection, debugging effectiveness, and usability. Participants—domain experts—would be divided into two groups: one using a standard NN with feature importance explanations and another using a contestable NN with causal graph injection and interactive debugging tools. Their performance would be measured through tasks such as explaining model decisions, identifying erroneous causal assumptions, and refining model outputs. Key metrics would include time and effectiveness in detecting errors, change in performance of refined models, and subjective ratings of trust and usability. By comparing both groups, this study would quantify how contestability improves interpretability, facilitates expert-driven refinements, and fosters trust in AI-driven decision-making.

Another promising direction is the development of NNs capable of not only predicting outcomes—corresponding to the bottom rung of Pearl and Mackenzie's *Ladder of Causation*—but also reasoning about the causes of outcomes, aligning with the top rung. Achieving this goal would require extending the Contestable NNs framework to incorporate **counterfactual reasoning**, enabling models to simulate interventions and provide causal explanations.

To this end, Zhang et al. (2024), inspired by our Contestable NNs, propose a framework for causality-informed neural networks (CINN) that integrates causal DAGs into NN architectures and loss functions. Their method incorporates domain knowledge to enable reasoning about interventions and facilitating "what-if" analyses for decision-making and risk assessment. Notably, their approach to embedding the "causal roles" of the variables into the loss function offers a pathway to equipping Contestable NNs with counterfactual reasoning capabilities.

Complementary approaches, such as those in (Germain et al., 2015), propagate structural assumptions through models to estimate intervention effects and could be adapted to refine this capability. Similarly, techniques proposed by (Pawlowski et al., 2020), which focus on counterfactual modelling but require input causal structures, offer additional avenues for enhancing counterfactual reasoning. Finally, methods like those in (Geiger et al., 2022; Xia et al., 2021; Zhang et al., 2020), which enforce interventional distributions in NNs through various mechanisms, address remaining challenges and could help fully realise the potential of counterfactual reasoning within Contestable NNs.

7.2.2 Shapley-based Causal Discovery

Our proposed Shapley-PC Algorithm offers several promising directions for future research. While its current implementation focuses on the orientation of v-structures, **extending Shapley-based rules** to the skeleton estimation phase could significantly improve robustness by reducing susceptibility to errors in adjacency determination. By incorporating Shapley values into edge inclusion tests, the algorithm could prioritise edges with higher explanatory power, thereby increasing the accuracy of causal skeletons. Furthermore, extending the use of SIVs to the third phase of the algorithm, where edge orientations are propagated through the graph (Meek, 1995), would provide a principled approach for assessing the conflicting orientations highlighted in (Tsagris, 2019).

Studying the value of Shapley-based rules using association measures **alternative to p-values** is another compelling opportunity to explore. Measures such as *mutual information* or other information-theoretic criteria (Glielmo et al., 2022; Shah and Peters, 2018) may offer a more robust assessment of association strength and improve causal discovery performance. Similarly, score-based metrics like the

BIC, often used in score-based causal discovery methods like GES (Chickering, 2002a; Ramsey, 2016), could replace or complement p-values in our Shapley-based decision rule. This integration could reduce the risk of spurious orientations and strengthen SIV-based structure identification.

Customising the Shapley-PC algorithm for specific applications represents another interesting avenue. Extensions to **handle latent variables, feedback loops, or additional inputs** from complementary algorithms such as GES (Chickering, 2002a) could expand its applicability to more complex causal structures. Such efforts have already shown promise in the literature, for instance, by extending the PC-max decision rule for GES (Ramsey, 2016) or adapting it to accommodate latent variables (Raghu et al., 2018). Incorporating these extensions into Shapley-PC would broaden its relevance for real-world datasets and complex scenarios.

The algorithm's flexibility also allows for **context-dependent adaptations**. For high-stakes domains like medicine, a more conservative approach could be developed, prioritising minimisation of false positives to ensure reliability and safety. This might involve considering only the lowest SIVs as colliders, and flagging other potential structures for review. Conversely, for exploratory scientific research, an alternative mode could emphasise hypothesis generation, facilitating the discovery of novel causal structures. These domain-specific adaptations would boost the algorithm's versatility.

7.2.3 Causal ABA

Causal ABA also presents numerous opportunities for future development. One critical challenge lies in its **scalability**, as the current implementation is limited to datasets with up to 10 variables. Overcoming this limitation will require optimisations in the processing of ABA extensions. Incremental solving techniques that avoid re-grounding rules during updates offer a promising solution. Additionally, approximate methods for handling large-scale ABA problems, such as those proposed by Cibier and Mailly (2024) using graph neural networks or graph attention networks, could be adapted to enable the analysis of more extensive datasets and complex causal graphs.

Extending Causal ABA to account for **latent confounders and feedback cycles** is another promising direction. Real-world systems can be influenced by hidden variables and exhibit cyclic dependencies, particularly when temporal factors are involved. Incorporating approaches like RFCI (Colombo et al., 2012) for latent confounders or methods specifically designed for cyclic graphs (Mooij and Claassen, 2020) would significantly expand its applicability to dynamic and intricate environments.

Exploring alternative **argumentation semantics** could also yield valuable insights into resolving conflicting evidence and reconstructing causal structures. Diverse semantics would enhance the

framework's flexibility to various problem domains, offering various degrees of conservatism. Native ABA solvers, such as the one implemented by Lehtonen et al. (2024), support alternative semantics and represent one promising avenue, though they may require separate grounding of the ABAF, posing scalability challenges. Gradual semantics, like those proposed by Rago et al. (2024) for structured argumentation, could be adapted to incorporate quantitative measures into the reasoning process. Furthermore, **integrating weighted arrows and edges**—generated by score-based and hybrid methods (Chickering, 2002a; Magliacane et al., 2016; Ramsey et al., 2017)—would enable more nuanced causal modelling, embedding quantitative confidence levels into the argumentation framework.

Another exciting area of exploration involves **interactive extensions** to Causal ABA. As envisaged in (Russo, 2023), incorporating real-time feedback from multiple experts would enable the framework to reconcile diverse perspectives into a consistent causal graph, resolving conflicts and fostering consensus-driven causal discovery. This human-in-the-loop approach bridges the gap between algorithmic insights and expert knowledge, enhancing the transparency and contestability of the resulting causal structures. Such capabilities would be particularly impactful in interdisciplinary research and policymaking, where synthesising diverse viewpoints into actionable insights is crucial.

7.2.4 Integrations and Deployment Challenges

Integrating the methodologies developed in this thesis presents an opportunity to create a unified framework for causal discovery and prediction. Contestable NNs, Shapley-PC, and Causal ABA share foundational principles, and their combination could significantly enhance the capabilities of AI systems. For instance, contestable NNs could incorporate causal graphs generated by Causal ABA or Shapley-PC to refine predictions while ensuring interpretability. Embedding SIVs within Causal ABA would enable the framework to prioritise arguments based on statistical significance, improving the precision of conflict resolution and bolstering the robustness of causal graphs. Together, these integrations could form a cohesive system that aligns with the principles of trustworthy AI, providing both predictive accuracy and interpretability.

However, implementing these techniques in real-world settings presents notable challenges. Issues such as managing high-dimensional data, ensuring scalability, and adapting to domain-specific workflows require careful study. Additionally, fostering trust and ensuring ease of use for diverse stakeholders are critical for successful adoption, particularly in high-stakes applications. By addressing these challenges and exploring the outlined directions, future research can build on the foundations of this thesis, driving advancements in trustworthy AI to meet the demands of complex, real-world scenarios.

7.3 Concluding Remarks

The research presented in this thesis advances the growing field of trustworthy AI by filling key gaps in robustness, transparency, and accountability. By integrating contestability (Leofante et al., 2024), XAI (Ali et al., 2023), robustness (Tocchetti et al., 2024), and fairness (Caton and Haas, 2024), this work contributes to a multidisciplinary effort aimed at aligning AI systems with societal values and ethical principles.

The methodologies introduced—including Contestable Neural Networks, the Shapley-PC algorithm, and the Causal ABA framework—demonstrate how causal discovery, expert feedback, game theory, and computational argumentation can be systematically incorporated into AI system design to improve reliability, interpretability, and stakeholder engagement. These contributions establish a foundation for deploying trustworthy AI across high-stakes domains such as healthcare, finance, and public policy, where the alignment of technical and ethical considerations is particularly paramount.

As AI continues to influence critical areas of decision-making, the importance of trustworthiness cannot be overstated (Bryant, 2023; Meissner and Narita, 2023). This thesis demonstrates that trustworthiness is not merely a technical property but a dynamic interplay of reliability, transparency, and stakeholder engagement. By equipping AI systems with tools to explain their decisions, integrate expert insights, and resolve uncertainties in their reasoning, this work envisions AI as a responsible and equitable partner in advancing societal goals. For example, the development of contestable neural networks enables stakeholders to interrogate and refine AI predictions, ensuring alignment with domain-specific knowledge and fostering trust through robustness, transparency and accountability.

Beyond methodological advancements, this thesis contributes to the broader discourse on trustworthy AI. Our findings demonstrate that transparency and accountability are not abstract ideals but practical imperatives that can be realised through innovative design and implementation. For example, the combination of causal graph injection with expert feedback ensures that neural networks respect domain knowledge while maintaining high predictive performance, challenging the traditional transparency-accuracy trade-off (Rudin, 2019).

Similarly, the Shapley-PC algorithm and the Causal ABA framework provide robust, interpretable tools for causal discovery, addressing critical challenges in data-driven decision-making (Matovski, 2024) by providing more reliable causal assumptions. Furthermore, Causal ABA's ability to integrate diverse forms of statistical evidence and expert feedback positions it as a transformative tool for causal discovery in high-stakes domains.

In conclusion, this thesis highlights the transformative potential of Causal Discovery for Trustworthy AI. The contributions made here represent significant strides toward realising this potential, while the proposed directions for future work point to a horizon rich with innovation and discovery. By establishing a foundation for robust, transparent, and accountable AI systems, this work aspires to shape a future where AI empowers society to make more informed and equitable decisions.

Bibliography

Joaquín Abellán, Manuel Gómez-Olmedo, and Serafín Moral. Some variations on the PC algorithm. In *Third European Workshop on Probabilistic Graphical Models, 12-15 September 2006, Prague, Czech Republic. Electronic Proceedings*, pages 1–8, 2006. URL http://www.utia.cas.cz/files/mtr/pgm06/41_paper.pdf.

Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. Contestable camera cars: A speculative design exploration of public AI that is open and responsive to dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 8:1–8:16. ACM, 2023a. URL https://doi.org/10.1145/3544548.3580984.

Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. Contestable AI by design: Towards a framework. *Minds Mach.*, 33(4):613–639, 2023b. URL https://doi.org/10.1007/s11023-022-09611-z.

Sajid Ali, Tamer Abuhmed, Shaker H. Ali El-Sappagh, Khan Muhammad, Jose Maria Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz Rodríguez, and Francisco Herrera. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, 99:101805, 2023. URL https://doi.org/10.1016/j.inffus.2023.101805.

Marco Almada. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial*

Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019, pages 2–11. ACM, 2019. URL https://doi.org/10.1145/3322640.3326699.

Dalal Alrajeh, Hana Chockler, and Joseph Y. Halpern. Combining experts' causal judgments. *Artif. Intell.*, 288:103355, 2020. URL https://doi.org/10.1016/j.artint.2020.103355.

Leila Amgoud and Claudette Cayrol. A reasoning model based on the production of acceptable arguments. *Ann. Math. Artif. Intell.*, 34(1-3):197–215, 2002. URL https://doi.org/10.1023/A:1014490210693.

Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artif. Intell.*, 173(3-4):413–436, 2009. URL https://doi.org/10.1016/j.artint.2008.11.006.

Bryan Andrews, Joseph D. Ramsey, Ruben Sanchez-Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow shrink trees. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c9cde817d04811ba28e44071bd9f76a5-Abstract-Conference.html.

André Assis, Jamilson Dantas, and Ermeson Andrade. The performance-interpretability trade-off: a comparative study of machine learning models. *Journal of Reliable Intelligent Environments*, 11(1):1, 2025. URL https://doi.org/10.1007/s40860-024-00240-0.

Bart Baesens, Daniel Roesch, and Harald Scheule. *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons, 2016.

Ricardo Baeza-Yates and Jeanna Matthews. Statement on principles for responsible algorithmic systems. Technical report, Association for Computing Machinery, 2022. URL https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf. Accessed: 2024-11-30.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999. URL http://www.doi.org/10.1126/science.286.5439.509.

Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4):365–410, 2011. URL https://doi.org/10.1017/S0269888911000166.

Barry G. Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20.

James O. Berger. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, 18 (1):1 – 32, 2003. URL https://doi.org/10.1214/ss/1056397485.

Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946. ISSN 00994987. URL http://www.doi.org/10.2307/3002000.

Amy Bernstein and Anand Raman. The great decoupling: An interview with erik brynjolfsson and andrew mcafee. output is on the rise, but workers aren't sharing in the bounty. *Harvard Business Review*, 2015. URL https://hbr.org/2015/06/the-great-decoupling.

Philippe Besnard and Anthony Hunter. A review of argumentation based on deductive arguments. *Handbook of formal argumentation*, 1(437-484):111, 2018. URL http://www0.cs.ucl.ac.uk/staff/a.hunter/papers/hofachapter.pdf.

Ansh Bhatnagar and Devyani Gajjar. How is artificial intelligence affecting society? Technical report, UK Parliament, May 2024. URL https://post.parliament.uk/how-is-artificial-intelligence-affecting-society/.

Andrei Bondarenko, Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997. URL https://doi.org/10.1016/S0004-3702(97)00015-5.

Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1435–1444. ACM, 2016. URL https://doi.org/10.1145/2939672.2939872.

Andrea Borghesi, Federico Baldo, and Michela Milano. Improving deep learning models via constraint-based domain knowledge: a brief survey. *CoRR*, abs/2005.10691, 2020. URL https://arxiv.org/abs/2005.10691.

Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. Elsevier, 2004. ISBN 978-1-55860-932-7. URL http://www.elsevier.com/wps/find/bookdescription.cws_home/702602/description.

Facundo Bromberg and Dimitris Margaritis. Improving the reliability of causal discovery from small data sets using argumentation. *J. Mach. Learn. Res.*, 10:301–340, 2009. URL https://dl.acm.org/doi/10.5555/1577069.1577081.

Kalina Bryant. How AI Is Impacting Society And Shaping The Future, 2023. URL https://www.forbes.com/sites/kalinabryant/2023/12/13/how-ai-is-impacting-society-and-shaping-the-future/. Accessed: 2024-11-25.

Erik Brynjolfsson and Andrew McAfee. The business of artificial intelligence. *Harvard Business Review*, 2017. URL https://hbr.org/2017/07/the-business-of-artificial-intelligence.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. URL https://doi.org/10.1214/14-AOS1260.

Martin Caminada, Samy Sá, João F. L. Alcântara, and Wolfgang Dvorák. On the difference between assumption-based argumentation and abstract argumentation. *FLAP*, 2(1):15–34, 2015. URL http://www.collegepublications.co.uk/downloads/ifcolog00003.pdf.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL https://books.google.de/books?id=0x_vAAAAMAAJ.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Joao Henrique Cavalcanti, Tibor Kovács, Andrea Ko, and Károly Pocsarovszky. Production system efficiency optimization through application of a hybrid artificial intelligence solution. *Int. J. Comput. Integr. Manuf.*, 37(6):790–807, 2024. URL https://doi.org/10.1080/0951192x.2023.2257661.

Zhengming Chen, Jie Qiao, Feng Xie, Ruichu Cai, Zhifeng Hao, and Keli Zhang. Testing conditional independence between latent variables by independence residuals. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2024. URL https://doi.org/10.1109/TNNLS.2024.3368561.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, 2002a. URL https://jmlr.org/papers/v2/chickering02a.html.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002b. URL https://jmlr.org/papers/v3/chickering02b.html.

Jawad Chowdhury, Rezaur Rashid, and Gabriel Terejanu. Evaluation of induced expert knowledge in causal structure learning by NOTEARS. In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2023, Lisbon, Portugal, February 22-24, 2023*, pages 136–146. SCITEPRESS, 2023. URL https://doi.org/10.5220/0011716000003411.

Paul Cibier and Jean-Guy Mailly. Graph convolutional networks and graph attention networks for approximating arguments acceptability. In *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18-20, 2014*, volume 388 of *Frontiers in Artificial Intelligence and Applications*, pages 25–36. IOS Press, 2024. URL https://doi.org/10.3233/FAIA240307.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 135–144. AUAI Press, 2011. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2231&proceeding_id=27.

Tom Claassen and Tom Heskes. A bayesian approach to constraint based causal inference. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 207–216. AUAI Press, 2012. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2283&proceeding_id=28.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014. URL https://dl.acm.org/doi/10.5555/2627435.2750365.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40 (1):294–321, 2012. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/41713636.

Anthony C. Constantinou, Zhigao Guo, and Neville Kenneth Kitson. The impact of prior knowledge on causal structure learning. *Knowl. Inf. Syst.*, 65(8):3385–3434, 2023. URL https://doi.org/10.1007/s10115-023-01858-x.

Kristijonas Cyras, Xiuyi Fan, Claudia Schulz, and Francesca Toni. Assumption-based argumentation: Disputes, explanations, preferences. *FLAP*, 4(8), 2017. URL http://www.collegepublications.co.uk/downloads/ifcolog00017.pdf.

Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4392–4399. ijcai.org, 2021. URL https://doi.org/10.24963/ijcai.2021/600.

Haoyue Dai, Rui Ding, Yuanyuan Jiang, Shi Han, and Dongmei Zhang. ML4C: seeing causality through latent vicinity. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pages 226–234. SIAM, 2023. URL https://doi.org/10.1137/1.9781611977653.ch26.

Alex Davies, Petar Velickovic, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomasev, Richard Tanburn, Peter W. Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with AI. *Nat.*, 600(7887):70–74, 2021. URL https://doi.org/10.1038/s41586-021-04086-x.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 12 1979. ISSN 0035-9246. URL https://doi.org/10.1111/j.2517-6161.1979.tb01052.x.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 518–528. PMLR, 2022. URL https://proceedings.mlr.press/v180/deleu22a.html.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995. URL https://doi.org/10.1016/0004-3702(94)00041-X.

Frederick Eberhardt. Introduction to the foundations of causal discovery. *Int. J. Data Sci. Anal.*, 3(2):81–91, 2017. URL https://doi.org/10.1007/s41060-016-0038-6.

Doris Entner, Patrik O. Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 256–264. JMLR.org, 2013. URL http://proceedings.mlr.press/v31/entner13a.html.

Paul Erdős and Alfréd Rényi. On random graphs I. *Publ. math. debrecen*, 6(290-297):18, 1959.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nat.*, 542(7639):115–118, 2017. URL https://doi.org/10.1038/nature21056.

European Commission. European approach to Artificial Intelligence, 2021a. URL https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence. Accessed: 2024-11-25.

European Commission. Coordinated Plan on Artificial Intelligence, 2021b. URL https://digital-strategy.ec.europa.eu/en/policies/plan-ai. Accessed: 2024-11-25.

European Parliament. 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, 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

European Union. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), May 2016. URL https://gdpr-text.com/read/article-22/.

Rosanna Fanni, Valerie Eveline Steinkogler, Giulia Zampedri, and Jo Pierson. Enhancing human agency through redress in artificial intelligence systems. *AI Soc.*, 38(2):537–547, 2023. URL https://doi.org/10.1007/s00146-022-01454-7.

FICO. Fico xml challenge found at community.fico.com/s/xml, 2017. URL https://community.fico.com/s/explainable-machine-learning-challenge.

Financial Conduct Authority. Mortgage conduct of business sourcebook: Mcob 11.6 - responsible lending. release 41., 2024. URL https://www.handbook.fca.org.uk/handbook/MCOB/11/6.html.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Martin Gebser, Roland Kaminski, and Torsten Schaub. Complex optimization in answer set programming. *Theory Pract. Log. Program.*, 11(4-5):821–839, 2011. URL https://doi.org/10.1017/S1471068411000329.

Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Multi-shot ASP solving with clingo. *Theory Pract. Log. Program.*, 19(1):27–82, 2019. URL https://doi.org/10.1017/S1471068418000054.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 881–889. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/germain15.html.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gábor Csányi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS Nexus*, 1(2):pgac039, 04 2022. ISSN 2752-6542. URL https://doi.org/10.1093/pnasnexus/pgac039.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Front. Genet., 04 June 2019 Sec. Statistical Genetics and Methodology*, 10:524, 2019. URL https://doi.org/10.3389/fgene.2019.00524.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022. URL https://doi.org/10.1098/rspa.2021.0068.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 585–592. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/d5cfead94f5350c12c322b5b664544c1-Abstract.html.

Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, 53(4):75:1–75:37, 2021. URL https://doi.org/10.1145/3397269.

Ezgi Gursel, Bhavya Reddy, Anahita Khojandi, Mahboubeh Madadi, Jamie Baalis Coble, Vivek Agarwal, Vaibhav Yadav, and Ronald L. Boring. Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance. *Nuclear Engineering and Technology*, 55(2):603–622, 2023. ISSN 1738-5733. URL https://doi.org/10.1016/j.net.2022.10.032.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14(1):3365–3383, 2013. URL https://dl.acm.org/doi/10.5555/2567709.2567770.

David Harrison and Daniel Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 03 1978. URL https://doi.org/10.1016/0095-0696(78)90006-2.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. ISBN 9780387848570. URL https://doi.org/10.1007/978-0-387-84858-7.

Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI Soc.*, 37(4):1397–1410, 2022. URL https://doi.org/10.1007/s00146-021-01251-8.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Alexander Hicks. Transparency, compliance, and contestability when code is(n't) law. In *Proceedings of the 2022 New Security Paradigms Workshop, NSPW 2022, North Conway, NH, USA, October 24-27, 2022*, pages 130–142. ACM, 2022. URL https://doi.org/10.1145/3584318.3584854.

Alexander Hicks. Transparency, compliance, and contestability when code is(n't) law. In *Proceedings of the 2022 New Security Paradigms Workshop*, NSPW '22, page 130–142, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398664. URL https://doi.org/10.1145/3584318.3584854.

High-Level Expert Group on AI. Ethics guidelines for trustworthy artificial intelligence. Technical report, European Commission, 2019. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Tad Hirsch, Kritzia Merced, Shrikanth S. Narayanan, Zac E. Imel, and David C. Atkins. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems, DIS '17, Edinburgh, United Kingdom, June 10-14, 2017*, pages 95–99. ACM, 2017. URL https://doi.org/10.1145/3064663.3064703.

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. URL https://doi.org/10.1016/0893-6080(89)90020-8.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 689–696. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1551–1560. ACM, 2018. URL https://doi.org/10.1145/3219819.3220104.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.*, 37:100270, 2020. URL https://doi.org/10.1016/j.cosrev.2020.100270.

H. M. James Hung, Robert T. O'Neill, Peter Bauer, and Karl Kohne. The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53(1):11–22, 1997. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533093.

Antti Hyttinen, Patrik O. Hoyer, Frederick Eberhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2391&proceeding_id=29.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pages 340–349. AUAI Press, 2014. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2469&proceeding_id=30.

Antti Hyttinen, Paul Saikko, and Matti Järvisalo. A core-guided approach to learning optimal causal graphs. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 645–651. ijcai.org, 2017. URL https://doi.org/10.24963/ijcai.2017/90.

T. Ichiishi. *Game Theory for Economic Analysis*. Economic Theory, Econometrics, and Mathematical Economics. Elsevier Science, 1983. ISBN 9780123701800. URL https://books.google.co.uk/books?id=zFm7AAAAIAAJ.

Christodoulos Ioannou and Loizos Michael. An automated arbitrator for contesting dialogues. *International Workshop on AI Value Engineering and AI Compliance Mechanisms (VECOMP 2024) associated with ECAI2024*, 10 2024. URL https://jurisinformaticscenter.github.io/VECOMP2024/AICOM/vecomp_aicomp_proceedings.pdf.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Fattaneh Jabbari, Joseph D. Ramsey, Peter Spirtes, and Gregory F. Cooper. Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part II*, volume 10535 of *Lecture Notes in Computer Science*, pages 142–157. Springer, 2017. URL https://doi.org/10.1007/978-3-319-71246-8_9.

Donald B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84, 1975. URL https://doi.org/10.1137/0204007.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger,

Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-021-03819-2.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, 2007. URL https://dl.acm.org/doi/10.5555/1314498.1314520.

Daniel N. Kluttz, Nitin Kohli, and Deirdre K. Mulligan. *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, page 137–152. Cambridge University Press, 2020. URL https://doi.org/10.1017/9781108610018.

Matthias König, Anna Rapberger, and Markus Ulbricht. Just a matter of perspective. In *Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14-16 September 2022*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, pages 212–223. IOS Press, 2022. URL https://doi.org/10.3233/FAIA220154.

Oleg Kovalevskiy, Juan Mateos-Garcia, and Kathryn Tunyasuvunakool. Alphafold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121(34):e2315002121, 2024. URL https://www.pnas.org/doi/abs/10.1073/pnas.2315002121.

Paul J. Krause, Simon Ambler, Morten Elvang-Gøransson, and John Fox. A logic of argumentation for reasoning under uncertainty. *Comput. Intell.*, 11:113–131, 1995. URL https://doi.org/10.1111/j.1467-8640.1995.tb00025.x.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. CASTLE: regularization via auxiliary causal graph discovery. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1068bceb19323fe72b2b344ccf85c254-Abstract.html.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based

neural DAG learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Wai-Yin Lam, Bryan Andrews, and Joseph D. Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 1052–1062. PMLR, 2022. URL https://proceedings.mlr.press/v180/lam22a.html.

Steffen L Lauritzen. *Graphical Models*. Oxford University Press, 05 1996. ISBN 9780198522195. URL https://doi.org/10.1093/oso/9780198522195.001.0001.

Ngoc-Bao-van Le, Yeong-Seok Seo, and Jun-Ho Huh. Artificial intelligence in finance: Coffee commodity trading big data for informed decision making. *IEEE Access*, 12:91780–91792, 2024. URL https://doi.org/10.1109/ACCESS.2024.3409762.

Tuomo Lehtonen, Anna Rapberger, Francesca Toni, Markus Ulbricht, and Johannes Peter Wallner. Instantiations and computational aspects of non-flat assumption-based argumentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 3457–3465. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/383.

Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. Contestable AI needs computational argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024. URL https://doi.org/10.24963/kr.2024/83.

Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. FIND: human-in-the-loop debugging deep text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 332–348. Association for Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.24.

Fraser I. Lewis and Benjamin J. J. McCormick. Revealing the complexity of health determinants

in resource-poor settings. *American Journal of Epidemiology*, 176(11):1051–1059, 11 2012. ISSN 0002-9262. URL https://doi.org/10.1093/aje/kws183.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/54f7125dee9b8b3dc798bb9a082b09e2-Abstract-Conference.html.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Henrietta Lyons, Eduardo Velloso, and Tim Miller. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. URL https://doi.org/10.1145/3449180.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4466–4474, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/f3d9de86462c28781cbe5c47ef22c3e5-Abstract.html.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10869–10879, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/39e98420b5e98bfbdc8a619bef7b8f61-Abstract.html.

Darko Matovski. Causal AI is transforming decision-making. Here's how, 2024. URL https://www.weforum.org/stories/2024/04/causal-ai-decision-making/. Accessed: 2024-11-25.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask

representation learning. *J. Mach. Learn. Res.*, 17:81:1–81:32, 2016. URL https://jmlr.org/papers/v17/15-242.html.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, pages 403–410. Morgan Kaufmann, 1995. URL https://doi.org/10.48550/arXiv.1302.4972.

Philip Meissner and Yusuke Narita. Artificial intelligence will transform decision-making. Here's how, 2023. URL https://www.weforum.org/stories/2023/09/how-artificial-intelligence-will-transform-decision-making/. Accessed: 2024-11-25.

Tim Miller. Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 333–342. ACM, 2023. URL https://doi.org/10.1145/3593013.3594001.

Joris M. Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 1159–1168. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/m-mooij20a.html.

Kasey Morris, Vincent S Huang, Mokshada Jain, BM Ramesh, Hannah Kemp, James Blanchard, Shajy Isac, Bidyut Sarkar, Vikas Gothalwal, Vasanthakumar Namasivayam, and Sema Sgaier. Ml and precision public health: Saving mothers and babies from dying in rural india. In *AI for Social Good Workshop at NeurIPS (2019)*, 2019. URL https://aiforsocialgood.github.io/neurips2019/accepted/track1/pdfs/82_aisg_neurips2019.pdf.

Joseph F Mudge, Leanne F Baker, Christopher B Edge, and Jeff E Houlahan. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PloS one*, 7(2):e32734, 2012. URL https://doi.org/10.1371/journal.pone.0032734.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI rmf 1.0). Technical report, U.S. Department of Commerce, 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

Ignavier Ng, Yujia Zheng, Jiji Zhang, and Kun Zhang. Reliable causal discovery with improved exact search and weaker assumptions. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20308–20320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a9b4ec2eb4ab7b1b9c3392bb5388119d-Abstract.html.

Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining, SDM 2022, Alexandria, VA, USA, April 28-30, 2022*, pages 424–432. SIAM, 2022. URL https://doi.org/10.1137/1.9781611977172.48.

Søren Holbech Nielsen and Simon Parsons. A generalization of dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In *Argumentation in Multi-Agent Systems, Third International Workshop, ArgMAS 2006, Hakodate, Japan, May 8, 2006, Revised Selected and Invited Papers*, volume 4766 of *Lecture Notes in Computer Science*, pages 54–73. Springer, 2006. URL https://doi.org/10.1007/978-3-540-75526-5_4.

OECD. Ai in society: Policy recommendations. *OECD Policy Papers*, 2021. URL https://www.oecd.org/en/topics/artificial-intelligence.html. Accessed: 2024-11-25.

OECD. Gpai and oecd unite to advance coordinated international efforts for trustworthy ai, 2024a. URL https://www.oecd.org/en/about/news/speech-statements/2024/07/GPAI-and-OECD-unite-to-advance-coordinated-international-efforts-for-trustworthy-AI.html. Accessed: 2024-11-25.

OECD. Oecd and un announce next steps in collaboration on artificial intelligence, 2024b. URL https://www.oecd.org/en/about/news/press-releases/2024/09/oecd-and-un-announce-next-steps-in-collaboration-on-artificial-intelligence.html. Accessed: 2024-11-25.

OECD. Oecd ai principles dashboard: Promoting trustworthy ai, 2024c. URL https://oecd.ai/en/dashboards/ai-principles/P7. Accessed: 2024-11-30.

I Ogunleye. Ai's redress problem: Recommendations to improve consumer protection from artificial intelligence. *Center for Long-Term Cyber Security*, 2022. URL https://cltc.berkeley.edu/wp-content/uploads/2022/08/AIs_Redress_Problem.pdf.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997. ISSN 0167-7152. URL https://doi.org/10.1016/S0167-7152(96)00140-X.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989. ISBN 978-0-08-051489-5. URL https://doi.org/10.1016/C2009-0-27609-4.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. URL https://doi.org/10.1017/CBO9780511803161.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Judea Pearl and Azaria Paz. Graphoids: graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *Proceedings of the 7th European Conference on Artificial Intelligence - Volume 2*, ECAI'86, page 357–363. North-Holland, 1986. ISBN 0444702792. URL https://ftp.cs.ucla.edu/pub/stat_ser/r53-L.pdf.

Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Comput.*, 27(3):771–799, 2015. URL https://doi.org/10.1162/NECO_a_00708.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15(1):2009–2053, 2014. URL https://dl.acm.org/doi/10.5555/2627435.2670315.

Sayeh Gholipour Picha, Dawood Al Chanti, and Alice Caplier. How far generated data can impact neural networks performance? In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2023, Volume 5: VISAPP, Lisbon, Portugal, February 19-21, 2023*, pages 472–479. SCITEPRESS, 2023. URL https://doi.org/10.5220/0011629000003417.

Thomas Ploug and Søren Holm. The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI. *Artif. Intell. Medicine*, 107:101901, 2020. URL https://doi.org/10.1016/j.artmed.2020.101901.

Vineet K. Raghu, Joseph D. Ramsey, Alison Morris, Dimitrios V. Manatakis, Peter Spirtes, Panos K. Chrysanthis, Clark Glymour, and Panayiotis V. Benos. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Int. J. Data Sci. Anal.*, 6(1):33–45, 2018. URL https://doi.org/10.1007/s41060-018-0104-3.

Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. Argumentation as a framework for interactive explanations for recommendations. In Diego Calvanese, Esra Erdem, and Michael Thielscher, editors, *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020*, pages 805–815, 2020. URL https://doi.org/10.24963/kr.2020/83.

Antonio Rago, Fabrizio Russo, Emanuele Albini, Pietro Baroni, and Francesca Toni. Forging argumentative explanations from causal models. In *Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021*, volume 3086 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-3086/paper3.pdf.

Antonio Rago, Fabrizio Russo, Emanuele Albini, Francesca Toni, and Pietro Baroni. Explaining classifiers' outputs with causal models and argumentation. *FLAP*, 10(3):421–509, 2023. URL https://www.collegepublications.co.uk/downloads/ifcolog00059.pdf.

Antonio Rago, Stylianos Loukas Vasileiou, Francesca Toni, Tran Cao Son, and William Yeoh. A methodology for gradual semantics for structured argumentation under incomplete information. *CoRR*, abs/2410.22209, 2024. URL https://doi.org/10.48550/arXiv.2410.22209.

Joseph Ramsey. Improving accuracy and scalability of the pc algorithm by maximizing p-value. *CoRR abs/1610.00378*, 2016. URL https://doi.org/10.48550/arXiv.1610.00378.

Joseph D. Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press, 2006. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1259&proceeding_id=22.

Joseph D. Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int. J. Data Sci. Anal.*, 3(2):121–129, 2017. URL https://doi.org/10.1007/s41060-016-0032-z.

Anna Rapberger, Markus Ulbricht, and Francesca Toni. On the correspondence of non-flat assumption-based argumentation and logic programming with negation as failure in the head. In *Proceedings of the 22nd International Workshop on Nonmonotonic Reasoning (NMR 2024) co-located with 21st International Conference on Principles of Knowledge Representation and Reasoning (KR 2024), Hanoi, Vietnam, November 2-4, 2024*, volume 3835 of *CEUR Workshop Proceedings*, pages 112–121. CEUR-WS.org, 2024. URL https://ceur-ws.org/Vol-3835/paper12.pdf.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27772–27784, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e987eff4a7c7b7e580d659feb6f60c1a-Abstract.html.

Thomas Richardson and Peter Spirtes. Automated discovery of linear feedback models. In *Computation, Causation, and Discovery*. AAAI Press, 05 1999. ISBN 9780262315821. URL https://doi.org/10.7551/mitpress/2006.003.0010.

Simon Rittel and Sebastian Tschiatschek. Specifying prior beliefs over dags in deep bayesian causal structure learning. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1962–1969. IOS Press, 2023. URL https://doi.org/10.3233/FAIA230487.

James Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90, 09 2003. URL https://doi.org/10.1093/biomet/90.3.491.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753. PMLR, 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. URL https://doi.org/10.1109/ACCESS.2020.2976199.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019. URL https://doi.org/10.1038/s42256-019-0048-x.

Fabrizio Russo. Argumentation for interactive causal discovery. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 7091–7092. ijcai.org, 2023. URL https://doi.org/10.24963/ijcai.2023/820.

Fabrizio Russo and Francesca Toni. Causal discovery and knowledge injection for contestable neural networks. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 2025–2032. IOS Press, 2023a. URL https://doi.org/10.3233/FAIA230495.

Fabrizio Russo and Francesca Toni. Shapley-PC: Constraint-based Causal Structure Learning with Shapley Values. *CoRR*, abs/2312.11582, 2023b. URL https://doi.org/10.48550/arXiv.2312.11582. (Under Review at CLeaR 2025).

Fabrizio Russo, Anna Rapberger, and Francesca Toni. Argumentative causal discovery. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024. URL https://doi.org/10.24963/kr.2024/88.

Andreas W. M. Sauter, Erman Acar, and Vincent François-Lavet. A meta-reinforcement learning algorithm for causal discovery. In *Conference on Causal Learning and Reasoning, CLeaR 2023,*

11-14 April 2023, Amazon Development Center, Tübingen, Germany, April 11-14, 2023, volume 213 of *Proceedings of Machine Learning Research*, pages 602–619. PMLR, 2023. URL https://proceedings.mlr.press/v213/sauter23a.html.

Claudia Schulz and Francesca Toni. Logic programming in assumption-based argumentation revisited - semantics and graphical representation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 1569–1575. AAAI Press, 2015. URL https://doi.org/10.1609/aaai.v29i1.9417.

Marco Scutari. Bayesian network repository. http://www.bnlearn.com/bnrepository, 2014.

Thomas Sellke, MJ Bayarri, and James O Berger. Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001. URL https://doi.org/10.1198/000313001300339950.

Sema K. Sgaier, Vincent Huang, and Grace Charles. The Case for Causal AI. *Stanford Social Innovation Review*, 18(3):50–55, 2020. URL https://doi.org/10.48558/KT81-SN73.

Azad Shademan, Ryan S Decker, Justin D Opfermann, Simon Leonard, Axel Krieger, and Peter CW Kim. Supervised autonomous robotic soft tissue surgery. *Science translational medicine*, 8(337): 337ra64, 2016. URL https://doi.org/10.1126/scitranslmed.aad9398.

Rajen Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48, 04 2018. URL https://doi.org/10.1214/19-AOS1857.

Lloyd S Shapley. A value for n-person games (1953). *Contribution to the Theory of Games*, 1953. URL https://www.rand.org/content/dam/rand/pubs/papers/2021/P295.pdf.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006. URL https://jmlr.org/papers/v7/shimizu06a.html.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer, 2016. URL https://doi.org/10.1186/s40535-016-0018-x.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000. ISBN 978-0-262-19440-

2. URL https://www.cs.cmu.edu/afs/andrew/scs/cs/15-381/archive/OldFiles/lib/cvsub/.g/group/sdss/.g/group2/g/scottd/fullbook.pdf.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, 2014. URL https://dl.acm.org/doi/10.5555/2627435.2670313.

Peter A Stott, Dáithí A Stone, and Myles R Allen. Human contribution to the european heatwave of 2003. *Nature*, 432(7017):610–614, 2004. URL https://doi.org/10.1038/nature03089.

Milan Studený. Attempts at axiomatic description of conditional independence. *Kybernetika*, 25(7): 72–79, 1989. URL http://www.kybernetika.cz/content/1989/7/72.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Jacopo Teneggi, Beepul Bharti, Yaniv Romano, and Jeremias Sulam. SHAP-XRT: the shapley value meets conditional independence testing. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=WFtTpQ47A7.

Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. A.i. robustness: a human-centered perspective on technological challenges and opportunities. *ACM Comput. Surv.*, 2024. ISSN 0360-0300. URL https://doi.org/10.1145/3665926.

Francesca Toni. A tutorial on assumption-based argumentation. *Argument Comput.*, 5(1):89–117, 2014. URL https://doi.org/10.1080/19462166.2013.869878.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16(66):2147–2205, 2015. URL http://jmlr.org/papers/v16/triantafillou15a.html.

Sofia Triantafilou, Ioannis Tsamardinos, and Ioannis G. Tollis. Learning causal structure from overlapping variable sets. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*,

volume 9 of *JMLR Proceedings*, pages 860–867. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/triantafillou10a.html.

Sofia Triantafilou, Ioannis Tsamardinos, and Anna Roumpelaki. Learning neighborhoods of high confidence in constraint-based causal discovery. In *Probabilistic Graphical Models - 7th European Workshop, PGM 2014, Utrecht, The Netherlands, September 17-19, 2014. Proceedings*, volume 8754 of *Lecture Notes in Computer Science*, pages 487–502. Springer, 2014. URL https://doi.org/10.1007/978-3-319-11433-0_32.

Michail Tsagris. Bayesian network learning with the PC algorithm: An improved and correct variation. *Appl. Artif. Intell.*, 33(2):101–123, 2019. URL https://doi.org/10.1080/08839514.2018.1526760.

Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Mach. Learn.*, 65(1):31–78, 2006. URL https://doi.org/10.1007/s10994-006-6889-7.

Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=qwBK94cP1y.

Andrea Aler Tubella, Andreas Theodorou, Virginia Dignum, and Loizos Michael. Contestable black boxes. In *Rules and Reasoning - 4th International Joint Conference, RuleML+RR 2020, Oslo, Norway, June 29 - July 1, 2020, Proceedings*, volume 12173 of *Lecture Notes in Computer Science*, pages 159–167. Springer, 2020. URL https://doi.org/10.1007/978-3-030-57977-7_12.

UK Government. Algorithmic transparency: records, 2024. URL https://www.gov.uk/algorithmic-transparency-records. Accessed: 2024-11-25.

Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. Contestability for content moderation. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):318:1–318:28, 2021. URL https://doi.org/10.1145/3476059.

J. H. van der Waerden. On the use of the likelihood ratio test. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, 43:165–176, 1940.

Joachim Vandekerckhove, Dora Matzke, and Eric-Jan Wagenmakers. Model Comparison and the Principle of Parsimony. *The Oxford handbook of computational and mathematical psychology*, 300, 2015. URL https://doi.org/10.1093/oxfordhb/9780199957996.013.14.

Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 284–293. ACM, 2020. URL https://doi.org/10.1145/3351095.3372876.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, page 255–270. Elsevier Science Inc., USA, 1990. ISBN 0444892648. URL https://ftp.cs.ucla.edu/tech-report/1991-reports/910020.pdf.

Laura von Rüden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 35(1): 614–633, 2023. URL https://doi.org/10.1109/TKDE.2021.3079836.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.*, 55(4), November 2022. ISSN 0360-0300. URL https://doi.org/10.1145/3527154.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer Texts in Statistics. Springer, 2004. ISBN 9780387402727. URL https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib,url,uid&db=nlebk&AN=2544135&site=ehost-live.

David S. Watson and Ricardo Silva. Causal discovery under a confounder blanket. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 2096–2106. PMLR, 2022. URL https://proceedings.mlr.press/v180/watson22a.html.

Darrell M West and John R Allen. How artificial intelligence is transforming the

world. *Brookings Institution.*, 2018. URL https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/.

Sewall Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934. URL https://www.jstor.org/stable/2957502.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.*, 135:364–381, 2022. URL https://doi.org/10.1016/j.future.2022.05.014.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10823–10836, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5989add1703e4b0480f75e2390739f34-Abstract.html.

Liqi Xia, Ruiyuan Cao, Jiang Du, and Jun Dai. Consistent complete independence test in high dimensions based on Chatterjee correlation coefficient. *Statistical Papers*, 66(1):3, December 2024. ISSN 1613-9798. doi: 10.1007/s00362-024-01618-1. URL https://doi.org/10.1007/s00362-024-01618-1.

Nadir Yalçın, Merve Kaşıkcı, Hasan Tolga Çelik, Karel Allegaert, Kutay Demirkan, Şule Yiğit, and Murat Yurdakök. Development and validation of a machine learning-based detection system to improve precision screening for medication errors in the neonatal intensive care unit. *Front. Pharmacol., 14 April 2023 Sec. Obstetric and Pediatric Pharmacology*, 14:1151560, 2023. URL https://doi.org/10.3389/fphar.2023.1151560.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3320–3328, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

H. Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985. URL https://doi.org/10.1007/BF01769885.

Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 134:1–134:21. ACM, 2023a. URL https://doi.org/10.1145/3544548.3581161.

Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9781450394215. URL https://doi.org/10.1145/3544548.3581161.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *Int. J. Approx. Reason.*, 151:101–129, 2022. URL https://doi.org/10.1016/j.ijar.2022.09.004.

Zhalama, Jiji Zhang, Frederick Eberhardt, and Wolfgang Mayer. Sat-based causal discovery under weaker assumptions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/234.pdf.

Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/02ed812220b0705fabb868ddbf17ea20-Abstract.html.

Hao Zhang, Yewei Xia, Kun Zhang, Shuigeng Zhou, and Jihong Guan. Conditional independence test based on residual similarity. *ACM Trans. Knowl. Discov. Data*, 17(8):117:1–117:18, 2023. URL https://doi.org/10.1145/3593810.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, August 7-10 2003*, pages 632–639. Morgan Kaufmann, 2003. URL https://dl.acm.org/doi/pdf/10.5555/2100584.2100661.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan.

gcastle: A python toolbox for causal discovery, 2021. URL https://doi.org/10.48550/arXiv.2111.15155.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 647–655. AUAI Press, 2009. URL https://www.auai.org/uai2009/papers/UAI2009_0064_3ba42a41c0d20bb784c0d8a0d7b71ded.pdf.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 804–813. AUAI Press, 2011. URL https://doi.org/10.48550/arXiv.1202.3775.

Xiaoge Zhang, Xiangyun Long, Yu Liu, Kai Zhou, and Jinwu Li. A generic causality-informed neural network (cinn) methodology for quantitative risk analytics and decision support. *Risk Analysis*, 44(11):2677–2695, 2024. doi: https://doi.org/10.1111/risa.14347. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14347.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with NO TEARS: continuous imization for structure learning. In *Prooceedings of the 31st Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9492–9503, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric dags. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR, 2020. URL http://proceedings.mlr.press/v108/zheng20a.html.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph D. Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *J. Mach. Learn. Res.*, 25: 60:1–60:8, 2024. URL https://jmlr.org/papers/v25/23-0970.html.

Appendix A

Appendix to Contestable Neural Networks

A.1 Further Details for Experiments

For all experiments[1] we choose 3 hidden layers ($M = 3$) of sizes $2 * |V|, 2/3 * |V|, 2 * |V|$ respectively, with ReLU activations. Experiments with smaller networks are provided for synthetic data in §A.1.2. All networks are initialized and seeded identically and use the Adam optimizer with a learning rate of 0.001 for a maximum of 1000 steps (T). A patience (T_s) of 50 steps on the loss on the validation set is used to stop training. Results for the synthetic dataset are then reported for 10 randomly generated DAGs. For the real data experiments, given the hyper-parameters optimisation of the threshold τ, we have used 5-fold nested cross validation and we report the average over the resulting 25 runs.

A.1.1 Experiments with Real Data

Here we provide additional details for the experiments in §4.6: we start from the datasets to then cover the optimisation of the threshold τ, part of experiment (i) in §4.6.1.

We report details for the datasets in Table A.1. For the regression tasks, with Boston (Harrison and Rubinfeld, 1978) and California Housing (Pace and Barry, 1997), we used the data out of the box from scikit-learn[2]. For the classification tasks, with HELOC[3] (FICO, 2017) and Adult Income[4] (Becker and

[1] Our code is available at: https://github.com/briziorusso/causal-injection-FFNN/tree/main

[2] https://scikit-learn.org/stable/datasets/toy_dataset.html

[3] Data and pre-processing mapping (keys.csv) is downloadable from our repo at https://github.com/briziorusso/causal-injection-FFNN/tree/main/data/fico/;

[4] Data and pre-processing mapping are from Penn Machine Learning Datasets repository at https://github.com/EpistasisLab/pmlb/blob/master/datasets/adult/

Table A.1: Real-world dataset details. Type is Regression (R) for real-valued target or Classification (C) for binary target.

Dataset	**Sample Size**	**Features**	**Type**
Boston Housing (BH)	506	14	R
California Housing (CH)	16512	8	R
Home Equity Line Of Credit (HELOC)	7844	23	C
Adult Income (IN)	48842	14	C

Kohavi, 1996), we used pre-processed data. Additionally, for Adult, we sampled 20000 observations and we balanced the proportion of positive and negative examples in target feature to 50-50 (from 25-75).

Details on Statistical Tests

This section provides the details necessary to reproduce the observed significance levels reported in Table 4.1 of the main text. Table A.2 summarises the means and standard deviations for all experiments, complementing the results in the main text.

As described in the main text, a two-tailed, two-sample t-test was used to assess differences in means. Each sample consisted of 25 runs within a 5-fold nested cross-validation setup, yielding 48 degrees of freedom for the Student's t-distribution. The calculated t-statistics and corresponding p-values for these tests are presented in Table A.3, providing a detailed account of the statistical validation of our results.

Table A.2: MSE or AUC (std) for regression and classification, respectively, across different sample sizes of the training data (N). We report the mean ± std across different training data sample sizes (N). Best results are in bold.

		N**=100**	N**=500**	N**=1K**	N**=2K**	N**=5K**	N**=10K**	N**=20K**
AUC ↑	**Adult**							
	CASTLE	0.67±0.03	0.72±0.04	0.75±0.03	0.74±0.03	0.75±0.03	0.75±0.02	0.76±0.02
	Injected	**0.69±0.04**	**0.74±0.02**	**0.76±0.03**	**0.77±0.01**	**0.79±0.03**	**0.85±0.01**	**0.86±0.01**
	Partial	0.66±0.02	0.71±0.02	0.74±0.03	0.76±0.03	0.76±0.02	0.76±0.02	0.77±0.02
	Refined	**0.69±0.04**	**0.74±0.02**	**0.76±0.02**	**0.77±0.02**	**0.79±0.03**	**0.85±0.01**	**0.86±0.01**
	HELOC							
	CASTLE	**0.75±0.02**	**0.79±0.01**	0.78±0.01	**0.79±0.01**	0.79±0.01	0.80±0.01	
	Injected	0.74±0.04	0.78±0.01	0.78±0.01	0.78±0.01	0.79±0.01	0.79±0.01	
MSE ↓	**California**							
	CASTLE	7.05±12.81	2.33±1.39	2.96±4.12	3.86±3.68	4.91±7.41	1.74±1.70	**0.66±0.08**
	Injected	**2.94±2.63**	**2.25±1.07**	**1.68±1.14**	**1.71±0.57**	**1.51±0.62**	**1.16±0.31**	1.02±0.35
	Boston							
	CASTLE	112.04±91.06	21.95±6.84					
	Injected	**86.17±13.75**	**20.45±5.12**					

Table A.3: t-statistic (p-value) comparing each column of Table A.2 against the respective CASTLE baseline for each dataset and sample size.

	Adult			HELOC	California	Boston
Data	***Injected***	***Partial***	***Refined***	***Injected***	***Injected***	***Injected***
100	2.000 (0.051)	1.387 (0.172)	2.000 (0.051)	1.118 (0.269)	1.571 (0.123)	1.405 (0.167)
500	2.236 (0.030)	1.118 (0.269)	2.236 (0.03)	3.536 (0.001)	0.228 (0.821)	0.878 (0.384)
1000	1.179 (0.224)	1.179 (0.244)	1.387 (0.172)	0.000 (1.000)	1.497 (0.141)	NA
2000	4.743 (0.000)	2.357 (0.023)	4.160 (0.000)	3.536 (0.001)	2.887 (0.006)	NA
5000	4.714 (0.000)	1.387 (0.172)	4.714 (0.000)	0.000 (1.000)	2.286 (0.027)	NA
10000	22.361 (0.000)	1.768 (0.083)	22.361 (0.000)	3.536 (0.001)	1.678 (0.100)	NA
20000	22.361 (0.000)	1.768 (0.083)	22.361 (0.000)	NA	5.014 (0.000)	NA

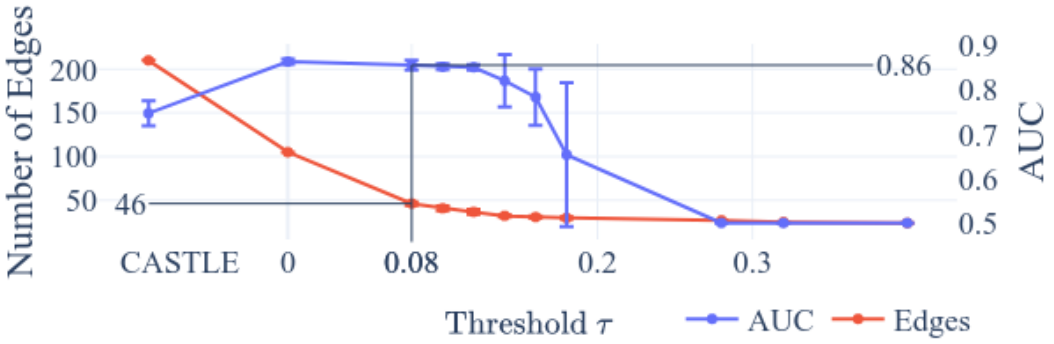
Threshold Optimisation

This section details the optimisation of the threshold τ, used to select a DAG without making qualitative causal judgments, as part of the experiment in §4.6.1. Results for all datasets are shown in Figure A.1. The optimisation evaluates the impact of τ on the number of edges (red) and the predictive performance metric (blue), with τ increasing along the x-axis, progressively masking more edges. As expected, the number of edges decreases monotonically with higher τ, while MSE/AUC trends highlight the trade-off between parsimony and predictive performance.

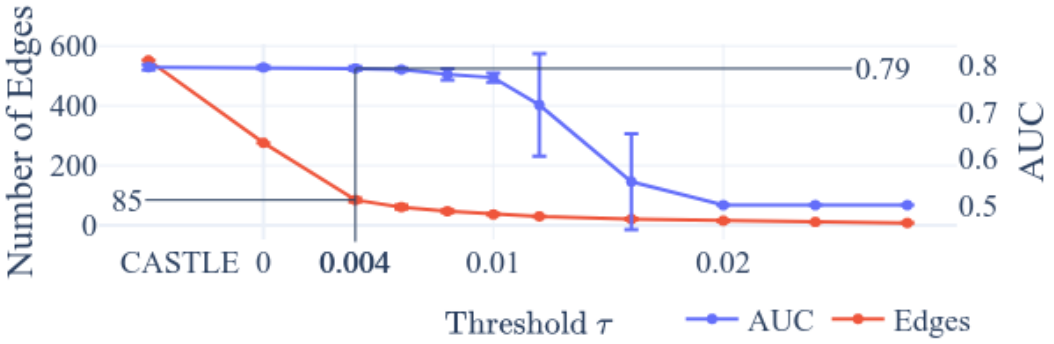
For the Adult dataset (Figure A.1a), $\tau = 0.08$ is selected as it retains the highest AUC while reducing the number of edges by over 50% compared to $\tau = 0$. Increasing τ further leads to diminishing returns in parsimony and a decline in performance. Similarly, for the FICO HELOC dataset (Figure A.1b), $\tau = 0.004$ achieves the best balance, minimising edge count without sacrificing AUC. Although $\tau = 0.1$ and 0.01 are viable alternatives, these thresholds trade slightly better parsimony for a more significant reduction in predictive performance.

The Boston dataset (Figure A.1c) follows a similar pattern, with $\tau = 0.13$ chosen to minimise MSE. Beyond this threshold, MSE increases significantly, indicating reduced performance. The California dataset (Figure A.1d) exhibits a different trend: injecting edges consistently worsens predictive performance. Here, $\tau = 0.05$ is selected as it achieves the lowest MSE. Notably, Table 4.1 in the main text shows that smaller sample sizes mitigate this negative effect for injected NNs.

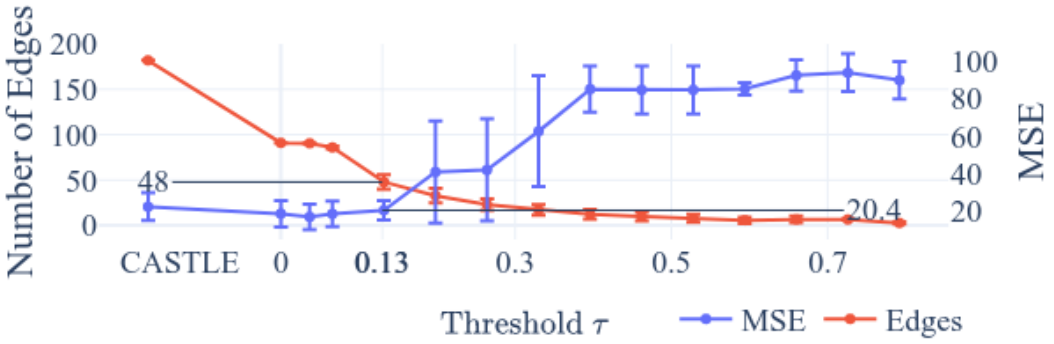
These findings illustrate the importance of carefully tuning τ to balance parsimony and performance, with optimal values varying by dataset and experimental context.



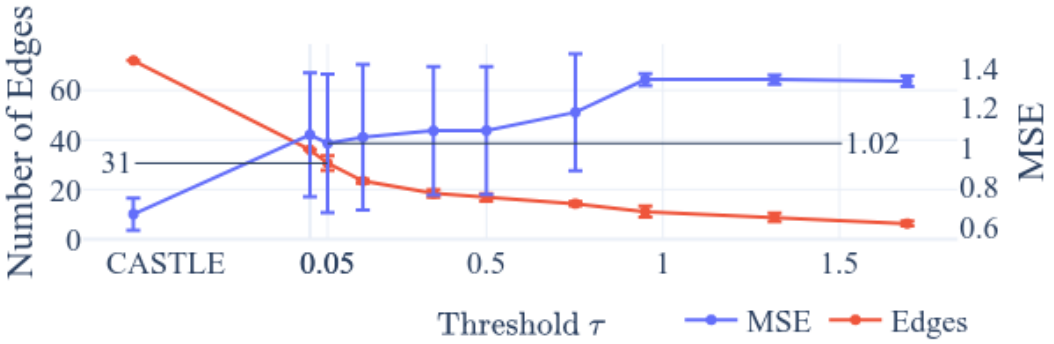
(a) Adult Income



(b) FICO HELOC



(c) Boston Housing



(d) California Housing

Figure A.1: Threshold τ optimisation for experiment in §4.6.1. Here $\tau < 0$ corresponds to the application of "unconstrained" CASTLE Kyono et al. (2020). Chosen thresholds are in bold on the x-axis. The number of edges and the predictive performance of the network injected with the DAG derived with the chosen τ are reported on the y-axes.

A.1.2 Parameter Study for Synthetic Data

This section expands on the results reported in §4.6.2 of the main text, which focused on a fixed proportion of edges injected (20%), presenting additional analyses to explore the impact of varying:

- The percentage of edges injected.
- The proportion of edges to nodes ($e = |E|/|V|$) in the DAG.
- The number of nodes in the DAG.
- The size of the NN.

The results of these analyses corroborate the findings in §4.6, providing further insights into the behaviour of our proposed method under different scenarios.

Percentage of Known Edges

Figure A.2 compares the performance of our algorithm and CASTLE+ when injecting different proportions of edges (10%, 20%, and 50%). Injecting 50% of edges results in diminishing returns, as the average reconstruction accuracy does not increase proportionally compared to the injected amount. By contrast, injecting 10% of edges often results in disproportionately higher gains, particularly for denser DAGs ($e \in \{2, 5\}$), where reconstruction is more challenging. These findings highlight that the optimal amount of injected edges depends on the density of the underlying DAG.

Proportion of Edges to Nodes in the DAG

The effect of varying the proportion of edges per node ($e = |E|/|V|$) is shown in Figure A.2. Sparse DAGs ($e = 1$) yield the highest reconstruction accuracy, averaging around 75%, while denser DAGs ($e = 5$) show significantly reduced performance, with accuracy dropping below 40%. These results mirror the trends observed for small sample sizes ($s = 50$), as reported in Figure 4.6, emphasising the challenges posed by denser graphs for both CASTLE+ and our method.

Number of Nodes in the DAG

Figure A.3 presents results for different numbers of nodes ($|V|$) in the DAG. For small DAGs ($|V| = 10$), both CASTLE+ and our method perform above the average reconstruction accuracy observed in Figure 4.6. However, as the DAG size increases, CASTLE+ experiences a sharp decline in performance, whereas our method demonstrates greater resilience to the increasing complexity of the graph.

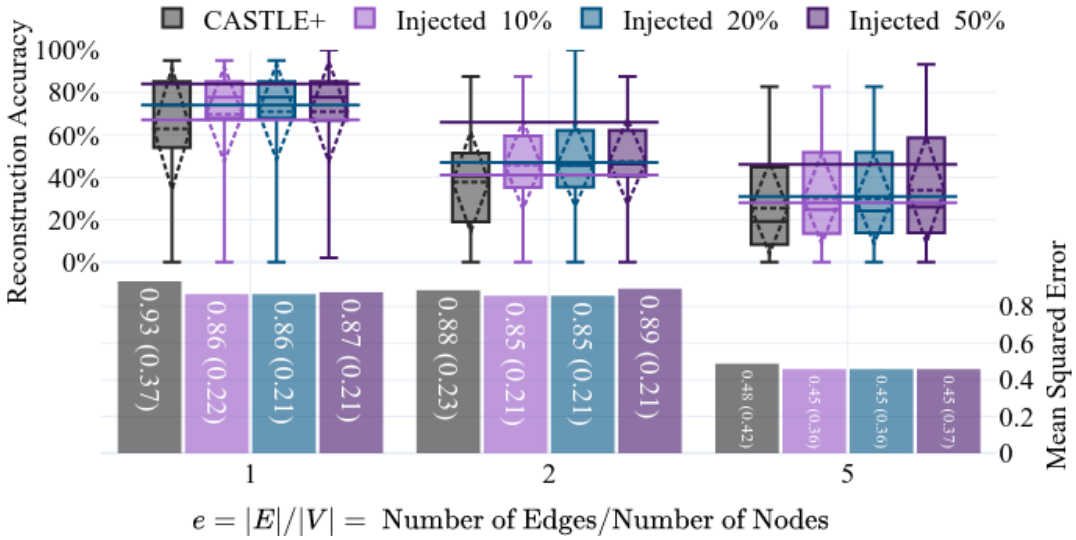


Figure A.2: Reconstruction Accuracy and MSE by the proportion of edges over nodes (see §A.1.2). The amount of known edges injected is 10%, 20% (as in Figure 4.6) and 50% (see §A.1.2).

Network Size

In addition to analysing the number of nodes, Figure A.3 compares the performance of NNs with three hidden layers (used in the experiments of §4.6.2) to smaller networks with a single hidden layer containing neurons equal to 3.2 times the input features ($M = 1, h = (d + 1) * 3.2$). Interestingly, while smaller networks improve the MSE for the target variable Y, they result in lower reconstruction accuracy for the causal graph. These findings suggest that better predictive performance does not always align with better causal discovery, underscoring the need for careful consideration of network size in causal discovery tasks.

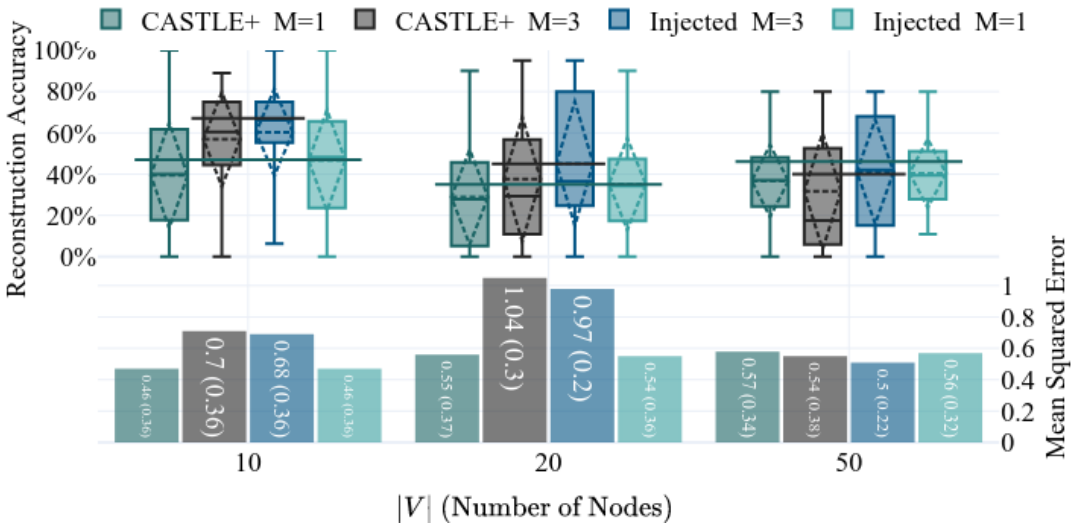


Figure A.3: Reconstruction Accuracy and MSE by the number of nodes in the DAG underpinning the data (see §A.1.2) and the numbers of layers (M=3 in Figure 4.6, see §A.1.2).

Appendix B

Appendix to Shapley-PC

B.1 Further Details for Experiments

In this section we provide additional details for the experiments in §5.4 of the main text. We provide an implementation of Shapley-PC based on the `causal-learn` python package (Zheng et al., 2024). Within `causal-learn`, we define a new PC function that accommodates our decision rule. The code is available at the following repository: https://github.com/briziorusso/ShapleyPC. In the repository, we also made available the code to reproduce all experiments and we saved all the plots, presented herein and in the main text, in HTML format. Downloading and opening them in a browser allows the inspection of all the numbers behind the plots in an interactive way.

Hyperparameters For all the methods we used Fisher's Z test (Fisher, 1970), as implemented in `causal-learn`, with significance threshold $\alpha = 0.01$ for the `bnlearn` dataset and with decreasing α for increasing number of nodes in the fully synthetic simulations: we used $\alpha = 0.1, 0.05, 0.01$ for number of nodes $|\mathbf{V}| = 10, 20, 50$, respectively.

B.1.1 Evaluation Metrics

In line with (Ramsey, 2016; Ramsey et al., 2006), we analyse the ability to identify v-structures (colliders), which is the focus of our proposed algorithm, alongside the overall performance in recovering the causal arrows of the true graph. All the metrics are calculated on the output CPDAGs hence the (binary) adjacency matrices can have entries for both (X_i, X_j) and (X_j, X_i), in which case the edge is undirected.

For the accuracy in classifying v-structures, we use precision and recall in classifying correctly

identified UTs and summarise it with the F1 Score. Specifically:

- V-Precision = V-TP/(V-TP + V-FP)
- V-Recall = V-TP/(V-TP + V-FN)
- V-F1 Score = $2 \times$ (V-P $\times$ V-R)/(V-P + V-R)

where V-True Positive (V-TP) is the number of correctly estimated v-structures; V-False Positive (V-FP) is the number of UTs wrongly deemed as v-structures; V-False Negative (V-FN) is the number of v-structures not deemed as such.

Also to evaluate the accuracy in identifying causal directions in the true graph we use F1, but calculated on the number of (in)correct arrowheads (AH-F1) as follows:

- AH-Precision = AH-TP/(AH-TP + AH-FP)
- AH-Recall = AH-TP/(AH-TP + AH-FN)
- AH-F1 Score = $2 \times$ (AH-P $\times$ AH-R)/(AH-P + AH-R)

where AH-True Positive (AH-TP) is the number of estimated edges with correct direction; False Positive (AH-FP) is the number of extra arrowheads; False Negative (AH-FN) is the number of missing arrowheads.

In addition to the metrics focusing on orientations and v-structures, we report two other commonly used metrics in CSL (see e.g. Constantinou et al. (2023)): Structural Hamming Distance (SHD) (Tsamardinos et al., 2006) and Structural Intervention Distance (SID) (Peters and Bühlmann, 2015).

SHD = E + M + R, where Extra (E) is the set of extra edges, Missing (M) are the ones missing from the skeleton of the estimated graph and Reversed (R) have incorrect direction.

SID quantifies the agreement to a causal graph in terms of interventional distributions. It aims at quantifying the incorrect causal inference estimations stemming out of a mistake in the causal graph estimation, akin to a downstream task error on a pre-processing step where the task is causal inference and the pre-processing step is finding the right graph to inform it. Both missing/extra edges and incorrect orientation will play a role in the incorrect causal inferences. See Definition C.1.1 for the formal definition of SID and a discussion on its properties.

B.1.2 Synthetic Data

Here we provide detail for the simulation study presented in §5.4, in particular Table 5.1 and 5.2.

DGP Details

In each experiment, we first generate 10 random graphs with maximum degree of 10 for each combination of three parameters:

- graph type: Erdös-Rényi (ER (Erdős and Rényi, 1959)) and Scale Free (SF (Barabási and Albert, 1999));
- number of nodes: $|\mathbf{V}| \in \{10, 20, 50\}$;
- density: $d = \{1, 2, 4\}$, with $|\mathbf{E}| = |\mathbf{V}| \times d$.

Given the ground truth DAGs DAG, we simulate Structural Equation Models (SEMs) belonging to the Additive Noise Model, formally:

$$X_j = f_j(\text{pa}(DAG, X_j)) + u_j \quad \forall\, j \in [1, \ldots, |\mathbf{V}|] \tag{B.1}$$

where f_j is linear function with coefficients $\mathbf{W}$ and u_j are samples from a noise distribution.

The coefficients $\mathbf{W}$ are sampled from a uniform distribution with parameters $[-1.5, -0.5] \cup [0.5, 1.5]$ for 95% of the effects, and $[-0.001, 0.001]$ for the remaining 5%. Sampling effect magnitudes from the range $[-0.001, 0.001]$ simulates the presence of weak edges to induce violations of orientation faithfulness (see §5.4 and (Ramsey et al., 2006)).

Finally, we sample $\mathbf{X} = \mathbf{W}^T\mathbf{X} + u$ where the noise u is generated from the following four distributions:

- Gaussian: $u \sim \mathcal{N}(0, 1)$
- Gumbel: $u \sim G(0, 1)$
- Exponential: $u \sim E(1)$
- Uniform: $u \sim U(-1, 1)$

We vary the number of drawn samples (N) in function of the number of nodes $N = s \times |\mathbf{V}|$, $s \in \{100, 500, 1000\}$ and refer to s as the proportional sample size. After sampling from the described DGPs we standardise the data using the standard scaler from `sklearn`. Code to reproduce the simulated data is provided in our repository.

Statistical Tests

Here we provide details for the statistical tests used to measure the significance of the difference in the results presented in Table 5.1 in the main text. In Tables B.4, B.6, B.5 and B.7 we provide t-statistics and p-values for graphs ER and SF graphs of 10 and 50 nodes, respectively.

Table B.1: AH-F1 and V-F1 Scores ± std for ER1 and SF1 graphs of nodes $|\mathbf{V}| \in \{10, 50\}$. No significant differences according to a t-test, $\alpha = 0.05$.

	Method	**ER1**		**SF1**	
		$\|\mathbf{V}\| = 10$	$\|\mathbf{V}\| = 50$	$\|\mathbf{V}\| = 10$	$\|\mathbf{V}\| = 50$
AH-F1	PC-Stable	0.91±0.14	0.82±0.14	0.9±0.12	0.91±0.07
	CPC	0.97±0.08	0.87±0.11	0.99±0.03	1.0±0.01
	MPC	0.96±0.09	0.87±0.11	0.98±0.05	0.99±0.02
	PC-Max	0.96±0.09	0.87±0.11	0.95±0.08	0.96±0.04
	Shapley-PC	0.93±0.12	0.9±0.07	0.95±0.08	0.96±0.04
V-F1	PC-Stable	0.97±0.09	0.92±0.16	0.99±0.03	0.98±0.04
	CPC	1.0±0.0	0.93±0.13	1.0±0.0	1.0±0.0
	MPC	1.0±0.0	0.93±0.13	1.0±0.01	1.0±0.0
	PC-Max	1.0±0.0	0.97±0.08	1.0±0.0	1.0±0.0
	Shapley-PC	0.99±0.03	1.0±0.02	1.0±0.0	1.0±0.0

In each table we present pairwise comparisons of means, for V-F1 and AH-F1 scores presented in Table 5.1 of the main text. We use two-sample, unequal variance t-tests, with degrees of freedom of 39 (10 seeds and 4 noise distributions, minus 1).

Additional Results

Sparsest Graphs (d=1) The results presented in Table 5.1 in the main text show AH-F1 and V-F1 for ER and SF graphs of density $d = \{2, 4\}$. Here we complete the picture and provide results for the sparsest graphs analysed: $d = 1$. We can see from Table B.1, that all methods perform quite well of very sparse graphs. This result is in line with (Kalisch and Bühlmann, 2007). Given the limited opportunity for improvement, no significant differences between the various methods is observed.

Graphs of 20 Nodes The results presented in Table 5.1 in the main text show AH-F1 and V-F1 for ER and SF graphs of 10 anf 50 nodes. Here we complete the picture and provide results for graphs of 20 nodes. The results corroborate the ones presented in the main paper. From Table B.2, we can see that Shapley-PC outperforms all other methods on ER4, SF2 and SF4. On ER2 it is not significantly different from PC-Max (according to a t-test, $\alpha = 0.05$), but better than all other methods.

Proportional Sample Size The results presented in Table 5.1 in the main text show AH-F1 and V-F1 for proportional sample size $s = 1000$. The proportional sample size is the number of samples per node in the dataset, with total number of samples $N = s * |V|$. Here we show the trends for $s \in \{100, 500, 1000\}$, in Fig. B.1 and Fig. B.2 for AH-F1 and V-F1, respectively. From the plots, we notice that the trends are mostly flat, demonstrating that none of the methods is very "data-hungry."

Table B.2: AH-F1 and V-F1 Scores ± std for ER2, ER4, SF2, and SF4 graphs of 20 nodes. Bold if significantly different from the runner-up (according to a t-test, $\alpha = 0.05$).

	Method	ER2	ER4	SF2	SF4
AH-F1	PC-Stable	0.28±0.26	0.08±0.10	0.59±0.30	0.22±0.25
	CPC	0.37±0.25	0.11±0.11	0.52±0.35	0.25±0.27
	MPC	0.31±0.26	0.11±0.11	0.50±0.36	0.25±0.26
	PC-Max	**0.51±0.24**	0.08±0.11	0.63±0.32	0.37±0.28
	Shapley-PC	**0.56±0.19**	**0.17±0.12**	**0.81±0.04**	**0.61±0.09**
V-F1	PC-Stable	0.35±0.35	0.14±0.21	0.71±0.37	0.31±0.38
	CPC	0.52±0.36	0.23±0.26	0.63±0.43	0.36±0.40
	MPC	0.42±0.37	0.19±0.24	0.60±0.43	0.36±0.38
	PC-Max	**0.71±0.34**	0.22±0.32	0.77±0.39	0.54±0.41
	Shapley-PC	**0.82±0.27**	**0.42±0.33**	**0.99±0.02**	**0.97±0.05**

Noise Distributions Plots by noise distribution are provided as interactive plots in our repository, as they would not be easily displayed on A4 paper. No major differences are observed across noise distributions.

Additional Metrics

Precision and Recall The results presented in Table 5.1 in the main text show F1 scores for arrowheads and v-structures' classification. Here we provide a breakdown of the F1 scores into their components: Precision and Recall. In Table B.8 we can see that Shapley-PC is significantly better (according to a t-test, $\alpha = 0.05$) than all other methods, for both precision and recall wrt ArrowHead classifications. Results are analogous for v-structure classifications, shown in Table B.9.

SHD and SID In addition to the results that focus on the v-structures and arrowhead orientations, as presented in main text and the additional results in this section, we also present results using SID (Peters and Bühlmann, 2015) in Table B.10 and SHD (Tsamardinos et al., 2006) in Table B.11. Both metrics measure error, hence the lower the better. Since we calculate all metrics on CPDAGs, SID estimates a best and worst scenario (SID-Low and High, respectively) depending on the orientation of the undirected edges in the output CPDAG

From Table B.10, we can see that Shapley-PC is significantly better than all other methods for the best case scenario (SID-Low) on 10 nodes graphs. For the worst case scenario (SID-High) all methods are on par for ER4, and Shapley-PC and PC-Max are on par, and better than all others for SF4. Shapley-PC is significantly better than all other methods for the remaining types of 10 nodes graphs. For 50 nodes graphs, which are sparser (see results discussion in the main text §5.4), Shapley-PC is

better than PC, CPC and MPC, but not significantly better than PC-Max.

Comparison of Shapley-PC with our baselines based SHD are shown in Table B.11. We can see that PC-Stable is significantly worse than all other methods for ER1, ER2 and SF2, while no significant differences are observed for the remaining three graph types.

Overall, Shapley-PC is never worse than any other baseline, based on both SID and SHD.

We remark that SHD and SID are more general graphical metrics, that do not take into account that there can be errors in skeleton and orientations, and that these can be isolated one from the other. With ArrowHead F1, we measure the orientation capabilities of the different methods, that with these metrics are confounded by errors in the skeleton.

B.1.3 Pseudo-Real Data

Table B.3: Details of the BNs from `bnlearn` used to generate the pseudo-real datasets in §5.4

Dataset Name	Type	Cat	Cont	$\|N\|$	$\|E\|$	d
ALARM	Discrete	37	0	37	46	1.24
CHILD	Discrete	20	0	20	25	1.25
HEPAR2	Discrete	70	0	70	123	1.76
INSURANCE	Discrete	27	0	27	52	1.93
ARTH150	Gaussian	0	107	107	150	1.40
ECOLI70	Gaussian	0	46	46	70	1.52
MAGIC-IRRI	Gaussian	0	64	64	102	1.59
MAGIC-NIAB	Gaussian	0	44	44	66	1.50
MEHRA	Linear Gaussian	8	16	24	71	2.96
SANGIOVESE	Linear Gaussian	1	14	15	55	3.67

`bnlearn` datasets

For the experiments on pseudo-real data, we used ten datasets from the `bnlearn` repository. This repository is widely used for research in Causal Discovery and hosts a number of commonly used benchmarks, some of which result from real published experiments or from the collection of expert opinions on the causal graph and the conditional probability tables necessary to create a BN. The BNs used in our experiments are from all three categories in the repository: Discrete, Gaussian and Conditional Linear Gaussian. The number of nodes varies from 15 to 107 and the number of edges from 25 to 150. Details on the number of nodes, edges and density of the DAGs underlying these data are reported in Table B.3, together with links to a more detailed description from the `bnlearn` repository.

Having downloaded all the .bif or .rda files from the repository, we load the Bayesian network and the associated conditional probability tables and sample 50000 observations, with 10 different seeds.

We encoded the labels using the label encoder from `sklearn` for categorical variables and applied standard scaling from `sklearn` for the continuous ones. The BNs, together with the code to reproduce the dataset, is provided in our repository.

Statistical Tests

Here we present details of the statistical tests used to measure the significance of the difference in the results presented in Fig. 5.1 in the main text. In Tables B.12 and B.13 we provide t-statistics and p-values for the Alarm, Insurance, Ecoli70 and Mehra datasets. In each table we present pairwise comparisons of means (shown in brackets together with standard deviations), for the AH-F1 and V-F1 presented in Fig. 5.1 of the main text.

Additional Metrics

In Fig. 5.1 in the main text, we show the results on four of the ten dataset detailed in Table B.3, according to ArrowHead F1 Score. In this section we report additional metrics, in line with the experiments on synthetic data. In particular, we visualise V-F1 (Fig. B.3), SHD (Fig. B.4) and SID (Fig. B.5). Precision and Recall are left out because they show very similar trends to AH-F1 and V-F1 presented herein, but are provided in our repository as interactive plots.

In Fig. B.3, we report V-F1 scores for the same set of datasets as in the main text. For AH-F1 (Fig. 5.1 in the main text), Shapley-PC is significantly better than all other methods on Alarm and Insurance. For V-F1, Shapley-PC is better than all other methods on Alarm, Insurance and Mehra. For Ecoli70, we are on par with PC-Max, and better than all others. According to SHD (Fig. B.4) and SID (Fig. B.5), no significant differences are observed across datasets and methods.

Additional Datasets

The results presented in Fig. 5.1 show AH-F1 for four datasets out of the ten analysed. We show results for the remaining six datasets (Arth150, Child, Hepar2, Magic-irri, Magic-niab and Sangiovese) in Fig. B.6 (AH-F1) and Fig. B.7 (V-F1). Out of these six datasets, Shapley-PC results to be significantly better than all other methods according to AH-F1 on Arth150 and Sangiovese. According to V-F1, Shapley-PC is better than all others on Sangiovese, and on par with CPC, improving on all other methods, on Arth150. No significant differences are observed on the remaining four datasets, apart from PC-Stable being worse than all other methods on the Magic-irri and Magic-niab datasets.

Table B.4: Two-sample, unequal variance t-tests for difference in means for ER graphs with $|\mathbf{V}| = 10$. Significance levels: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1. DoF: $n_a = n_b = 39$.

Type	Metric	Methods	Means±Std	t	p-value
ER2	V-F1	PC-S vs CPC	0.46 ± 0.4 vs 0.59 ± 0.3	-1.64	0.104
		PC-S vs MPC	0.46 ± 0.4 vs 0.58 ± 0.3	-1.54	0.128
		PC-S vs PC-M	0.46 ± 0.4 vs 0.67 ± 0.3	-2.80	0.006**
		PC-S vs SPC	0.46 ± 0.4 vs 0.86 ± 0.2	-6.53	0.000***
		CPC vs MPC	0.59 ± 0.3 vs 0.58 ± 0.3	0.11	0.916
		CPC vs PC-M	0.59 ± 0.3 vs 0.67 ± 0.3	-1.12	0.265
		CPC vs SPC	0.59 ± 0.3 vs 0.86 ± 0.2	-4.73	0.000***
		MPC vs PC-M	0.58 ± 0.3 vs 0.67 ± 0.3	-1.23	0.222
		MPC vs SPC	0.58 ± 0.3 vs 0.86 ± 0.2	-4.85	0.000***
		PC-M vs SPC	0.67 ± 0.3 vs 0.86 ± 0.2	-3.66	0.001***
	AH-F1	PC-S vs CPC	0.36 ± 0.3 vs 0.42 ± 0.3	-0.98	0.331
		PC-S vs MPC	0.36 ± 0.3 vs 0.42 ± 0.3	-0.89	0.377
		PC-S vs PC-M	0.36 ± 0.3 vs 0.50 ± 0.2	-2.52	0.014*
		PC-S vs SPC	0.36 ± 0.3 vs 0.63 ± 0.2	-5.39	0.000***
		CPC vs MPC	0.42 ± 0.3 vs 0.42 ± 0.3	0.09	0.926
		CPC vs PC-M	0.42 ± 0.3 vs 0.50 ± 0.2	-1.52	0.133
		CPC vs SPC	0.42 ± 0.3 vs 0.63 ± 0.2	-4.42	0.000***
		MPC vs PC-M	0.42 ± 0.3 vs 0.50 ± 0.2	-1.62	0.109
		MPC vs SPC	0.42 ± 0.3 vs 0.63 ± 0.2	-4.54	0.000***
		PC-M vs SPC	0.50 ± 0.2 vs 0.63 ± 0.2	-3.08	0.003**
ER4	V-F1	PC-S vs CPC	0.23 ± 0.3 vs 0.23 ± 0.3	-0.08	0.940
		PC-S vs MPC	0.23 ± 0.3 vs 0.22 ± 0.3	0.03	0.973
		PC-S vs PC-M	0.23 ± 0.3 vs 0.09 ± 0.2	2.15	0.035*
		PC-S vs SPC	0.23 ± 0.3 vs 0.36 ± 0.4	-1.64	0.106
		CPC vs MPC	0.23 ± 0.3 vs 0.22 ± 0.3	0.11	0.915
		CPC vs PC-M	0.23 ± 0.3 vs 0.09 ± 0.2	2.19	0.032*
		CPC vs SPC	0.23 ± 0.3 vs 0.36 ± 0.4	-1.55	0.125
		MPC vs PC-M	0.22 ± 0.3 vs 0.09 ± 0.2	2.05	0.044*
		MPC vs SPC	0.22 ± 0.3 vs 0.36 ± 0.4	-1.64	0.106
		PC-M vs SPC	0.09 ± 0.2 vs 0.36 ± 0.4	-3.54	0.001***
	AH-F1	PC-S vs CPC	0.15 ± 0.2 vs 0.15 ± 0.2	0.01	0.989
		PC-S vs MPC	0.15 ± 0.2 vs 0.15 ± 0.2	0.09	0.930
		PC-S vs PC-M	0.15 ± 0.2 vs 0.07 ± 0.1	2.45	0.017*
		PC-S vs SPC	0.15 ± 0.2 vs 0.25 ± 0.2	-2.73	0.008**
		CPC vs MPC	0.15 ± 0.2 vs 0.15 ± 0.2	0.07	0.943
		CPC vs PC-M	0.15 ± 0.2 vs 0.07 ± 0.1	2.38	0.020*
		CPC vs SPC	0.15 ± 0.2 vs 0.25 ± 0.2	-2.70	0.009**
		MPC vs PC-M	0.15 ± 0.2 vs 0.07 ± 0.1	2.26	0.027*
		MPC vs SPC	0.15 ± 0.2 vs 0.25 ± 0.2	-2.74	0.008**
		PC-M vs SPC	0.07 ± 0.1 vs 0.25 ± 0.2	-5.59	0.000***

Table B.5: Two-sample, unequal variance t-tests for difference in means for SF graphs with $|\mathbf{V}| = 10$. Significance levels: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1. DoF: $n_a = n_b = 39$.

Type	Metric	Methods	Means±Std	t	p-value
SF2	V-F1	PC-S vs CPC	0.81 ± 0.3 vs 0.88 ± 0.2	-1.36	0.179
		PC-S vs MPC	0.81 ± 0.3 vs 0.88 ± 0.2	-1.35	0.183
		PC-S vs PC-M	0.81 ± 0.3 vs 0.90 ± 0.2	-1.55	0.125
		PC-S vs SPC	0.81 ± 0.3 vs 0.99 ± 0.0	-3.93	0.000***
		CPC vs MPC	0.88 ± 0.2 vs 0.88 ± 0.2	0.01	0.989
		CPC vs PC-M	0.88 ± 0.2 vs 0.90 ± 0.2	-0.36	0.721
		CPC vs SPC	0.88 ± 0.2 vs 0.99 ± 0.0	-3.73	0.001***
		MPC vs PC-M	0.88 ± 0.2 vs 0.90 ± 0.2	-0.37	0.712
		MPC vs SPC	0.88 ± 0.2 vs 0.99 ± 0.0	-3.72	0.001***
		PC-M vs SPC	0.90 ± 0.2 vs 0.99 ± 0.0	-2.56	0.014*
	AH-F1	PC-S vs CPC	0.67 ± 0.2 vs 0.74 ± 0.1	-1.72	0.091
		PC-S vs MPC	0.67 ± 0.2 vs 0.74 ± 0.1	-1.73	0.090
		PC-S vs PC-M	0.67 ± 0.2 vs 0.73 ± 0.2	-1.32	0.191
		PC-S vs SPC	0.67 ± 0.2 vs 0.82 ± 0.1	-3.95	0.000***
		CPC vs MPC	0.74 ± 0.1 vs 0.74 ± 0.1	-0.01	0.992
		CPC vs PC-M	0.74 ± 0.1 vs 0.73 ± 0.2	0.20	0.840
		CPC vs SPC	0.74 ± 0.1 vs 0.82 ± 0.1	-3.36	0.001**
		MPC vs PC-M	0.74 ± 0.1 vs 0.73 ± 0.2	0.21	0.834
		MPC vs SPC	0.74 ± 0.1 vs 0.82 ± 0.1	-3.34	0.001**
		PC-M vs SPC	0.73 ± 0.2 vs 0.82 ± 0.1	-2.57	0.013*
SF4	V-F1	PC-S vs CPC	0.49 ± 0.4 vs 0.54 ± 0.4	-0.52	0.606
		PC-S vs MPC	0.49 ± 0.4 vs 0.59 ± 0.4	-1.09	0.280
		PC-S vs PC-M	0.49 ± 0.4 vs 0.59 ± 0.4	-1.12	0.266
		PC-S vs SPC	0.49 ± 0.4 vs 0.84 ± 0.2	-4.67	0.000***
		CPC vs MPC	0.54 ± 0.4 vs 0.59 ± 0.4	-0.61	0.546
		CPC vs PC-M	0.54 ± 0.4 vs 0.59 ± 0.4	-0.66	0.511
		CPC vs SPC	0.54 ± 0.4 vs 0.84 ± 0.2	-4.43	0.000***
		MPC vs PC-M	0.59 ± 0.4 vs 0.59 ± 0.4	-0.08	0.936
		MPC vs SPC	0.59 ± 0.4 vs 0.84 ± 0.2	-3.70	0.000***
		PC-M vs SPC	0.59 ± 0.4 vs 0.84 ± 0.2	-3.38	0.001**
	AH-F1	PC-S vs CPC	0.32 ± 0.3 vs 0.34 ± 0.2	-0.29	0.774
		PC-S vs MPC	0.32 ± 0.3 vs 0.36 ± 0.2	-0.69	0.493
		PC-S vs PC-M	0.32 ± 0.3 vs 0.39 ± 0.3	-1.16	0.252
		PC-S vs SPC	0.32 ± 0.3 vs 0.54 ± 0.2	-4.52	0.000***
		CPC vs MPC	0.34 ± 0.2 vs 0.36 ± 0.2	-0.44	0.662
		CPC vs PC-M	0.34 ± 0.2 vs 0.39 ± 0.3	-0.96	0.341
		CPC vs SPC	0.34 ± 0.2 vs 0.54 ± 0.2	-4.74	0.000***
		MPC vs PC-M	0.36 ± 0.2 vs 0.39 ± 0.3	-0.55	0.583
		MPC vs SPC	0.36 ± 0.2 vs 0.54 ± 0.2	-4.25	0.000***
		PC-M vs SPC	0.39 ± 0.3 vs 0.54 ± 0.2	-3.23	0.002**

Table B.6: Two-sample, unequal variance t-tests for difference in means for ER graphs with $|\mathbf{V}| = 50$. Significance levels: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1. DoF: $n_a = n_b = 39$.

Type	Metric	Methods	Means±Std	t	p-value
ER2	V-F1	PC-S vs CPC	0.31 ± 0.4 vs 0.50 ± 0.4	-2.05	0.043*
		PC-S vs MPC	0.31 ± 0.4 vs 0.42 ± 0.4	-1.26	0.210
		PC-S vs PC-M	0.31 ± 0.4 vs 0.81 ± 0.3	-6.24	0.000***
		PC-S vs SPC	0.31 ± 0.4 vs 0.98 ± 0.0	-11.45	0.000***
		CPC vs MPC	0.50 ± 0.4 vs 0.42 ± 0.4	0.79	0.429
		CPC vs PC-M	0.50 ± 0.4 vs 0.81 ± 0.3	-3.53	0.001***
		CPC vs SPC	0.50 ± 0.4 vs 0.98 ± 0.0	-6.95	0.000***
		MPC vs PC-M	0.42 ± 0.4 vs 0.81 ± 0.3	-4.56	0.000***
		MPC vs SPC	0.42 ± 0.4 vs 0.98 ± 0.0	-8.56	0.000***
		PC-M vs SPC	0.81 ± 0.3 vs 0.98 ± 0.0	-3.13	0.003**
	AH-F1	PC-S vs CPC	0.25 ± 0.3 vs 0.40 ± 0.4	-2.02	0.047*
		PC-S vs MPC	0.25 ± 0.3 vs 0.35 ± 0.3	-1.32	0.191
		PC-S vs PC-M	0.25 ± 0.3 vs 0.63 ± 0.3	-5.86	0.000***
		PC-S vs SPC	0.25 ± 0.3 vs 0.75 ± 0.1	-10.22	0.000***
		CPC vs MPC	0.40 ± 0.4 vs 0.35 ± 0.3	0.70	0.485
		CPC vs PC-M	0.40 ± 0.4 vs 0.63 ± 0.3	-3.21	0.002**
		CPC vs SPC	0.40 ± 0.4 vs 0.75 ± 0.1	-6.12	0.000***
		MPC vs PC-M	0.35 ± 0.3 vs 0.63 ± 0.3	-4.11	0.000***
		MPC vs SPC	0.35 ± 0.3 vs 0.75 ± 0.1	-7.42	0.000***
		PC-M vs SPC	0.63 ± 0.3 vs 0.75 ± 0.1	-2.75	0.009**
ER4	V-F1	PC-S vs CPC	0.08 ± 0.2 vs 0.14 ± 0.2	-1.38	0.172
		PC-S vs MPC	0.08 ± 0.2 vs 0.08 ± 0.2	-0.06	0.954
		PC-S vs PC-M	0.08 ± 0.2 vs 0.12 ± 0.3	-0.88	0.380
		PC-S vs SPC	0.08 ± 0.2 vs 0.48 ± 0.4	-6.32	0.000***
		CPC vs MPC	0.14 ± 0.2 vs 0.08 ± 0.2	1.30	0.198
		CPC vs PC-M	0.14 ± 0.2 vs 0.12 ± 0.3	0.35	0.731
		CPC vs SPC	0.14 ± 0.2 vs 0.48 ± 0.4	-4.86	0.000***
		MPC vs PC-M	0.08 ± 0.2 vs 0.12 ± 0.3	-0.82	0.414
		MPC vs SPC	0.08 ± 0.2 vs 0.48 ± 0.4	-6.21	0.000***
		PC-M vs SPC	0.12 ± 0.3 vs 0.48 ± 0.4	-4.96	0.000***
	AH-F1	PC-S vs CPC	0.04 ± 0.1 vs 0.06 ± 0.1	-1.10	0.275
		PC-S vs MPC	0.04 ± 0.1 vs 0.04 ± 0.1	-0.08	0.933
		PC-S vs PC-M	0.04 ± 0.1 vs 0.05 ± 0.1	-0.48	0.635
		PC-S vs SPC	0.04 ± 0.1 vs 0.19 ± 0.2	-5.63	0.000***
		CPC vs MPC	0.06 ± 0.1 vs 0.04 ± 0.1	1.00	0.320
		CPC vs PC-M	0.06 ± 0.1 vs 0.05 ± 0.1	0.55	0.584
		CPC vs SPC	0.06 ± 0.1 vs 0.19 ± 0.2	-4.52	0.000***
		MPC vs PC-M	0.04 ± 0.1 vs 0.05 ± 0.1	-0.39	0.695
		MPC vs SPC	0.04 ± 0.1 vs 0.19 ± 0.2	-5.50	0.000***
		PC-M vs SPC	0.05 ± 0.1 vs 0.19 ± 0.2	-4.91	0.000***

Table B.7: Two-sample, unequal variance t-tests for difference in means for SF graphs with $|\mathbf{V}| = 50$. Significance levels: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1. DoF: $n_a = n_b = 39$.

Type	Metric	Methods	Means±Std	t	p-value
SF2	V-F1	PC-S vs CPC	0.67 ± 0.4 vs 0.58 ± 0.5	0.85	0.397
		PC-S vs MPC	0.67 ± 0.4 vs 0.57 ± 0.5	1.03	0.307
		PC-S vs PC-M	0.67 ± 0.4 vs 0.61 ± 0.5	0.56	0.579
		PC-S vs SPC	0.67 ± 0.4 vs 1.00 ± 0.0	−5.30	0.000***
		CPC vs MPC	0.58 ± 0.5 vs 0.57 ± 0.5	0.17	0.864
		CPC vs PC-M	0.58 ± 0.5 vs 0.61 ± 0.5	−0.27	0.787
		CPC vs SPC	0.58 ± 0.5 vs 1.00 ± 0.0	−5.44	0.000***
		MPC vs PC-M	0.57 ± 0.5 vs 0.61 ± 0.5	−0.44	0.661
		MPC vs SPC	0.57 ± 0.5 vs 1.00 ± 0.0	−5.58	0.000***
		PC-M vs SPC	0.61 ± 0.5 vs 1.00 ± 0.0	−5.08	0.000***
	AH-F1	PC-S vs CPC	0.63 ± 0.4 vs 0.53 ± 0.4	1.07	0.290
		PC-S vs MPC	0.63 ± 0.4 vs 0.51 ± 0.4	1.29	0.201
		PC-S vs PC-M	0.63 ± 0.4 vs 0.56 ± 0.4	0.75	0.454
		PC-S vs SPC	0.63 ± 0.4 vs 0.90 ± 0.0	−4.58	0.000***
		CPC vs MPC	0.53 ± 0.4 vs 0.51 ± 0.4	0.21	0.832
		CPC vs PC-M	0.53 ± 0.4 vs 0.56 ± 0.4	−0.29	0.773
		CPC vs SPC	0.53 ± 0.4 vs 0.90 ± 0.0	−5.26	0.000***
		MPC vs PC-M	0.51 ± 0.4 vs 0.56 ± 0.4	−0.50	0.617
		MPC vs SPC	0.51 ± 0.4 vs 0.90 ± 0.0	−5.50	0.000***
		PC-M vs SPC	0.56 ± 0.4 vs 0.90 ± 0.0	−4.85	0.000***
SF4	V-F1	PC-S vs CPC	0.28 ± 0.4 vs 0.46 ± 0.4	−1.87	0.065.
		PC-S vs MPC	0.28 ± 0.4 vs 0.42 ± 0.4	−1.39	0.167
		PC-S vs PC-M	0.28 ± 0.4 vs 0.69 ± 0.4	−4.34	0.000***
		PC-S vs SPC	0.28 ± 0.4 vs 0.99 ± 0.0	−11.16	0.000***
		CPC vs MPC	0.46 ± 0.4 vs 0.42 ± 0.4	0.46	0.649
		CPC vs PC-M	0.46 ± 0.4 vs 0.69 ± 0.4	−2.28	0.025*
		CPC vs SPC	0.46 ± 0.4 vs 0.99 ± 0.0	−7.40	0.000***
		MPC vs PC-M	0.42 ± 0.4 vs 0.69 ± 0.4	−2.75	0.007**
		MPC vs SPC	0.42 ± 0.4 vs 0.99 ± 0.0	−8.07	0.000***
		PC-M vs SPC	0.69 ± 0.4 vs 0.99 ± 0.0	−4.42	0.000***
	AH-F1	PC-S vs CPC	0.25 ± 0.4 vs 0.40 ± 0.4	−1.80	0.076.
		PC-S vs MPC	0.25 ± 0.4 vs 0.36 ± 0.4	−1.29	0.201
		PC-S vs PC-M	0.25 ± 0.4 vs 0.59 ± 0.4	−4.12	0.000***
		PC-S vs SPC	0.25 ± 0.4 vs 0.83 ± 0.1	−10.12	0.000***
		CPC vs MPC	0.40 ± 0.4 vs 0.36 ± 0.4	0.50	0.619
		CPC vs PC-M	0.40 ± 0.4 vs 0.59 ± 0.4	−2.15	0.034*
		CPC vs SPC	0.40 ± 0.4 vs 0.83 ± 0.1	−6.77	0.000***
		MPC vs PC-M	0.36 ± 0.4 vs 0.59 ± 0.4	−2.68	0.009**
		MPC vs SPC	0.36 ± 0.4 vs 0.83 ± 0.1	−7.57	0.000***
		PC-M vs SPC	0.59 ± 0.4 vs 0.83 ± 0.1	−4.04	0.000***

Table B.8: ArrowHead Precision and Recall for ERd and SFd graphs of 10 and 50 nodes. d is the number of edges per node in the true DAG. Bold if significantly different from the runner-up (according to a t-test, $\alpha = 0.05$).

		Method	ER2	ER4	SF2	SF4
$\|\mathbf{V}\| = 10$	**Precision**	PC-Stable	0.4±0.3	0.16±0.18	0.73±0.25	0.42±0.35
		CPC	0.46±0.29	0.16±0.19	0.82±0.14	0.44±0.29
		MPC	0.45±0.28	0.16±0.19	0.82±0.14	0.47±0.29
		PC-Max	0.55±0.24	0.08±0.13	0.80±0.23	0.50±0.33
		Shapley-PC	**0.70±0.18**	**0.29±0.20**	**0.91±0.11**	**0.71±0.20**
	Recall	PC-Stable	0.34±0.26	0.14±0.15	0.62±0.22	0.26±0.22
		CPC	0.39±0.25	0.14±0.16	0.68±0.13	0.28±0.19
		MPC	0.39±0.25	0.14±0.16	0.68±0.13	0.30±0.19
		PC-Max	0.47±0.21	0.07±0.11	0.68±0.19	0.32±0.22
		Shapley-PC	**0.59±0.16**	**0.22±0.14**	**0.76±0.08**	**0.44±0.14**
$\|\mathbf{V}\| = 50$	**Precision**	PC-Stable	0.29±0.35	0.06±0.13	0.65±0.38	0.27±0.38
		CPC	0.46±0.41	0.11±0.18	0.56±0.46	0.44±0.42
		MPC	0.40±0.39	0.06±0.14	0.53±0.46	0.39±0.42
		PC-Max	0.72±0.31	0.08±0.18	0.58±0.46	0.63±0.40
		Shapley-PC	**0.86±0.05**	**0.33±0.25**	**0.93±0.04**	**0.89±0.07**
	Recall	PC-Stable	0.22±0.27	0.03±0.06	0.61±0.36	0.24±0.33
		CPC	0.36±0.32	0.04±0.07	0.50±0.42	0.38±0.37
		MPC	0.31±0.30	0.03±0.07	0.49±0.43	0.34±0.36
		PC-Max	0.56±0.25	0.03±0.08	0.54±0.42	0.56±0.36
		Shapley-PC	**0.67±0.07**	**0.14±0.11**	**0.86±0.05**	**0.79±0.08**

Table B.9: V-structure Precision and Recall for ERd and SFd graphs of 10 and 50 nodes. d is the number of edges per node in the true DAG. Bold if significantly different from the runner-up (according to a t-test, $\alpha = 0.05$).

		Method	ER2	ER4	SF2	SF4
$\|\mathbf{V}\| = 10$	**Precision**	PC-Stable	0.46±0.37	0.35±0.41	0.88±0.27	0.56±0.45
		CPC	0.60±0.36	0.32±0.41	0.96±0.09	0.65±0.42
		MPC	0.60±0.37	0.31±0.40	0.95±0.09	0.68±0.40
		PC-Max	**0.72±0.33**	0.19±0.39	0.92±0.22	0.65±0.42
		Shapley-PC	**0.85±0.21**	0.55±0.47	**1.0±0.01**	**0.86±0.25**
	Recall	PC-Stable	0.48±0.37	0.27±0.39	0.77±0.30	0.47±0.40
		CPC	0.60±0.35	0.25±0.36	0.84±0.22	0.50±0.36
		MPC	0.60±0.35	0.24±0.37	0.85±0.22	0.56±0.37
		PC-Max	0.67±0.32	0.09±0.24	0.88±0.23	0.56±0.39
		Shapley-PC	**0.91±0.14**	**0.35±0.42**	**0.99±0.04**	**0.84±0.24**
$\|\mathbf{V}\| = 50$	**Precision**	PC-Stable	0.32±0.38	0.07±0.16	0.70±0.41	0.31±0.42
		CPC	0.50±0.44	0.12±0.21	0.59±0.49	0.47±0.45
		MPC	0.42±0.41	0.07±0.16	0.57±0.50	0.42±0.45
		PC-Max	0.80±0.34	0.12±0.26	0.62±0.49	0.69±0.43
		Shapley-PC	**0.97±0.04**	**0.44±0.34**	**1.0±0.0**	**0.98±0.03**
	Recall	PC-Stable	0.30±0.36	0.09±0.20	0.64±0.39	0.27±0.38
		CPC	0.50±0.44	0.18±0.30	0.58±0.48	0.46±0.44
		MPC	0.42±0.42	0.10±0.23	0.56±0.49	0.41±0.44
		PC-Max	0.82±0.35	0.13±0.29	0.61±0.48	0.68±0.43
		Shapley-PC	**0.99±0.02**	**0.53±0.39**	**1.0±0.0**	**0.99±0.02**

Table B.10: Structural Interventional Distance (SID, the lower the better) for ERd and SFd graphs of 10 and 50 nodes. d is the number of edges per node in the true DAG. Bold if significantly different from the runner-up (according to a t-test, $\alpha = 0.05$).

		Method	**ER2**	**ER4**	**SF2**	**SF4**
$\|\mathbf{V}\| = 10$	**SID-Low**	PC-Stable	42±17	72±7	14±10	37±14
		CPC	39±14	70±6	13±10	42±12
		MPC	40±13	70±6	13±10	42±12
		PC-Max	35±11	71±9	11±7	36±14
		Shapley-PC	**29±13**	**65±10**	**7±4**	**28±12**
	SID-High	PC-Stable	59±14	78±6	27±12	57±14
		CPC	57±11	76±5	26±12	60±14
		MPC	57±10	76±5	26±11	57±15
		PC-Max	56±8	79±6	22±8	**50±13**
		Shapley-PC	**51±11**	77±6	**20±7**	**48±13**
$\|\mathbf{V}\| = 50$	**SID-Low**	PC-Stable	792±196	1906±155	189±84	394±172
		CPC	559±168	1904±189	128±59	308±116
		MPC	644±156	1880±157	**119±57**	315±107
		PC-Max	**494±127**	**1682±197**	**113±57**	**216±109**
		Shapley-PC	**471±140**	**1569±172**	**93±43**	**173±85**
	SID-High	PC-Stable	1073±254	2110±139	263±96	484±189
		CPC	900±244	2188±110	221±74	426±155
		MPC	953±229	2136±120	**199±64**	419±143
		PC-Max	**803±251**	**2012±191**	**186±75**	**311±147**
		Shapley-PC	**792±268**	**1962±172**	**170±54**	**272±125**

Table B.11: Structural Hamming Distance (SHD, the lower the better) for ERd and SFd graphs of 10 and 50 nodes. d is the number of edges per node in the true DAG. Bold if significantly different from the runner-up (according to a t-test, $\alpha = 0.05$).

Nodes	ER1	ER2	ER4	SF1	SF2	SF4
10	SHD	SHD	SHD	SHD	SHD	SHD
PC-Stable	1.3±1.8	13.0±4.2	35.4±3.6	2.0±1.9	6.5±2.4	18.5±3.4
CPC	0.9±1.3	12.8±4.1	34.9±3.1	1.6±1.5	6.2±2.4	20.4±3.4
MPC	0.9±1.3	13.1±4.2	34.8±3.0	1.6±1.6	6.2±2.3	20.1±3.4
PC-Max	0.9±1.3	12.0±3.1	36.7±2.5	1.7±1.6	5.9±2.3	**18.3±3.6**
Shapley-PC	1.1±1.6	10.7±4.1	**34.4±3.5**	1.6±1.6	**4.9±1.8**	**17.1±3.2**
50						
PC-Stable	6.9±4.1	53.3±8.6	188.0±10.8	5.5±2.5	17.5±5.9	34.6±13.0
CPC	**4.8±2.3**	42.5±9.9	180.5±8.7	4.5±1.8	**14.6±4.0**	31.7±11.4
MPC	**4.9±2.5**	45.4±8.4	183.3±12.6	4.6±1.8	**14.6±4.5**	32.6±11.2
PC-Max	**4.8±2.4**	**39.5±7.6**	178.1±8.5	4.7±1.9	**13.5±4.0**	28.7±13.1
Shapley-PC	**4.9±2.4**	**38.5±8.1**	171.7±11.0	4.7±2.0	**13.8±4.7**	26.7±11.9

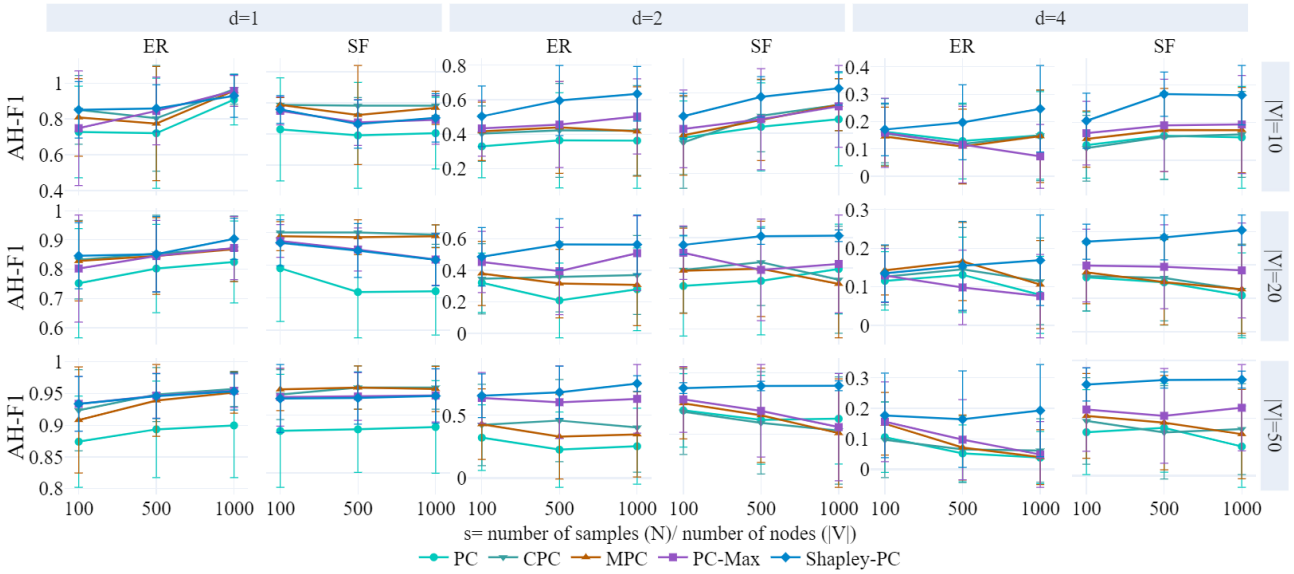


Figure B.1: ArrowHead F1 scores by proportional sample size ($s \in \{100, 500, 1000\}$) for the fully synthetic data in §5.4.

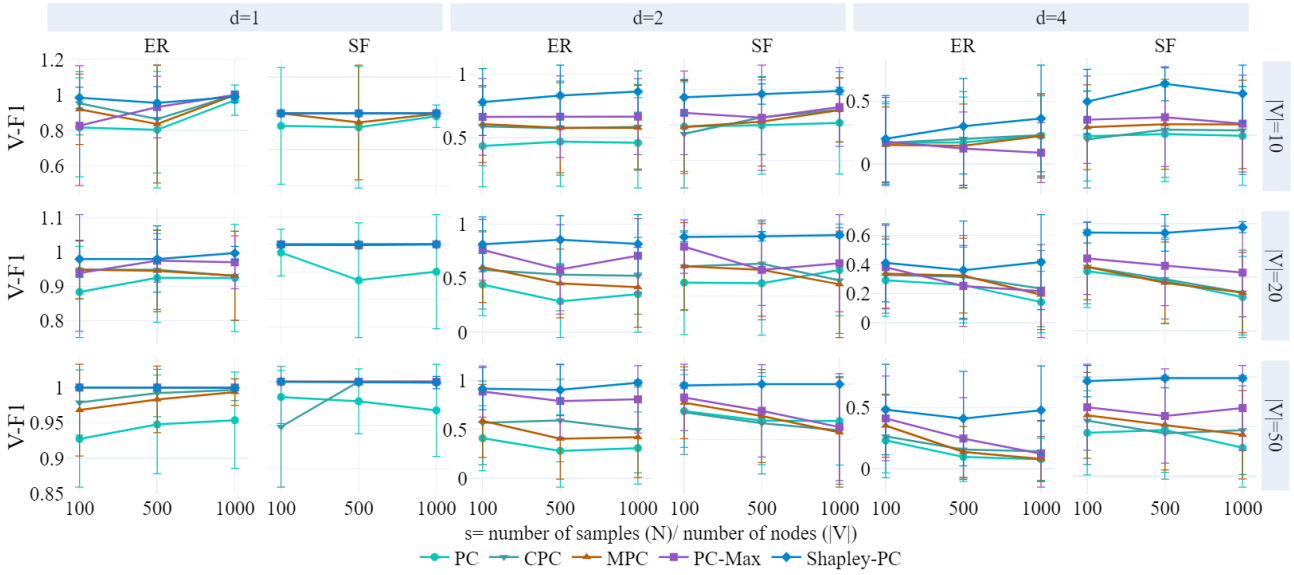


Figure B.2: V-structure F1 scores by proportional sample size ($s \in \{100, 500, 1000\}$) for the fully synthetic data in §5.4.

Table B.12: Two-sample, unequal variance t-tests for difference in means for Alarm and Insurance Data. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1. DoF: $n_a = n_b = 9$.

Dataset	Metric	Methods	Means±Std	t	p-value
Alarm	V-F1	CPC vs MPC	0.53 ± 0.4 vs 0.24 ± 0.4	1.75	0.098.
		CPC vs PC-S	0.53 ± 0.4 vs 0.18 ± 0.3	2.37	0.030*
		CPC vs PC-M	0.53 ± 0.4 vs 0.24 ± 0.4	1.73	0.101
		CPC vs SPC	0.53 ± 0.4 vs 0.85 ± 0.1	−2.67	0.024*
		MPC vs PC-S	0.24 ± 0.4 vs 0.18 ± 0.3	0.39	0.700
		MPC vs PC-M	0.24 ± 0.4 vs 0.24 ± 0.4	−0.01	0.991
		MPC vs SPC	0.24 ± 0.4 vs 0.85 ± 0.1	−5.04	0.001***
		PC-S vs PC-M	0.18 ± 0.3 vs 0.24 ± 0.4	−0.40	0.692
		PC-S vs SPC	0.18 ± 0.3 vs 0.85 ± 0.1	−7.19	0.000***
		PC-M vs SPC	0.24 ± 0.4 vs 0.85 ± 0.1	−4.97	0.001***
	AH-F1	CPC vs MPC	0.37 ± 0.3 vs 0.16 ± 0.3	1.78	0.091.
		CPC vs PC-S	0.37 ± 0.3 vs 0.13 ± 0.2	2.33	0.032*
		CPC vs PC-M	0.37 ± 0.3 vs 0.16 ± 0.3	1.80	0.089.
		CPC vs SPC	0.37 ± 0.3 vs 0.57 ± 0.0	−2.38	0.041*
		MPC vs PC-S	0.16 ± 0.3 vs 0.13 ± 0.2	0.33	0.743
		MPC vs PC-M	0.16 ± 0.3 vs 0.16 ± 0.3	0.01	0.994
		MPC vs SPC	0.16 ± 0.3 vs 0.57 ± 0.0	−4.81	0.001***
		PC-S vs PC-M	0.13 ± 0.2 vs 0.16 ± 0.3	−0.33	0.749
		PC-S vs SPC	0.13 ± 0.2 vs 0.57 ± 0.0	−6.64	0.000***
		PC-M vs SPC	0.16 ± 0.3 vs 0.57 ± 0.0	−4.84	0.001***
Insurance	V-F1	CPC vs MPC	0.05 ± 0.1 vs 0.02 ± 0.1	0.73	0.474
		CPC vs PC-S	0.05 ± 0.1 vs 0.03 ± 0.1	0.31	0.757
		CPC vs PC-M	0.05 ± 0.1 vs 0.07 ± 0.2	−0.35	0.732
		CPC vs SPC	0.05 ± 0.1 vs 0.42 ± 0.2	−4.54	0.001***
		MPC vs PC-S	0.02 ± 0.1 vs 0.03 ± 0.1	−0.34	0.736
		MPC vs PC-M	0.02 ± 0.1 vs 0.07 ± 0.2	−0.89	0.391
		MPC vs SPC	0.02 ± 0.1 vs 0.42 ± 0.2	−5.17	0.000***
		PC-S vs PC-M	0.03 ± 0.1 vs 0.07 ± 0.2	−0.59	0.567
		PC-S vs SPC	0.03 ± 0.1 vs 0.42 ± 0.2	−4.71	0.000***
		PC-M vs SPC	0.07 ± 0.2 vs 0.42 ± 0.2	−3.83	0.001**
	AH-F1	CPC vs MPC	0.04 ± 0.1 vs 0.02 ± 0.1	0.72	0.484
		CPC vs PC-S	0.04 ± 0.1 vs 0.02 ± 0.1	0.59	0.563
		CPC vs PC-M	0.04 ± 0.1 vs 0.04 ± 0.1	−0.10	0.920
		CPC vs SPC	0.04 ± 0.1 vs 0.21 ± 0.1	−3.78	0.002**
		MPC vs PC-S	0.02 ± 0.1 vs 0.02 ± 0.1	−0.11	0.911
		MPC vs PC-M	0.02 ± 0.1 vs 0.04 ± 0.1	−0.75	0.465
		MPC vs SPC	0.02 ± 0.1 vs 0.21 ± 0.1	−4.76	0.000***
		PC-S vs PC-M	0.02 ± 0.1 vs 0.04 ± 0.1	−0.64	0.532
		PC-S vs SPC	0.02 ± 0.1 vs 0.21 ± 0.1	−4.54	0.000***
		PC-M vs SPC	0.04 ± 0.1 vs 0.21 ± 0.1	−3.47	0.003**

Table B.13: Two-sample, unequal variance t-tests for difference in means for Ecoli70 and Mehra Data. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1. DoF: $n_a = n_b = 9$.

Dataset	Metric	Methods	Means±Std	t	p-value
Ecoli70	V-F1	CPC vs MPC	0.01 ± 0.0 vs 0.60 ± 0.1	-19.57	0.000***
		CPC vs PC-S	0.01 ± 0.0 vs 0.29 ± 0.2	-3.93	0.003**
		CPC vs PC-M	0.01 ± 0.0 vs 0.72 ± 0.3	-7.88	0.000***
		CPC vs SPC	0.01 ± 0.0 vs 0.89 ± 0.1	-23.53	0.000***
		MPC vs PC-S	0.60 ± 0.1 vs 0.29 ± 0.2	3.88	0.002**
		MPC vs PC-M	0.60 ± 0.1 vs 0.72 ± 0.3	-1.33	0.210
		MPC vs SPC	0.60 ± 0.1 vs 0.89 ± 0.1	-6.19	0.000***
		PC-S vs PC-M	0.29 ± 0.2 vs 0.72 ± 0.3	-3.72	0.002**
		PC-S vs SPC	0.29 ± 0.2 vs 0.89 ± 0.1	-7.39	0.000***
		PC-M vs SPC	0.72 ± 0.3 vs 0.89 ± 0.1	-1.74	0.108
	AH-F1	CPC vs MPC	0.01 ± 0.0 vs 0.55 ± 0.1	-27.65	0.000***
		CPC vs PC-S	0.01 ± 0.0 vs 0.32 ± 0.2	-4.27	0.002**
		CPC vs PC-M	0.01 ± 0.0 vs 0.60 ± 0.2	-8.52	0.000***
		CPC vs SPC	0.01 ± 0.0 vs 0.73 ± 0.1	-28.17	0.000***
		MPC vs PC-S	0.55 ± 0.1 vs 0.32 ± 0.2	3.12	0.011*
		MPC vs PC-M	0.55 ± 0.1 vs 0.60 ± 0.2	-0.68	0.510
		MPC vs SPC	0.55 ± 0.1 vs 0.73 ± 0.1	-5.64	0.000***
		PC-S vs PC-M	0.32 ± 0.2 vs 0.60 ± 0.2	-2.82	0.011*
		PC-S vs SPC	0.32 ± 0.2 vs 0.73 ± 0.1	-5.42	0.000***
		PC-M vs SPC	0.60 ± 0.2 vs 0.73 ± 0.1	-1.80	0.099.
Mehra	V-F1	CPC vs MPC	0.45 ± 0.4 vs 0.26 ± 0.4	1.04	0.311
		CPC vs PC-S	0.45 ± 0.4 vs 0.01 ± 0.0	3.49	0.007**
		CPC vs PC-M	0.45 ± 0.4 vs 0.28 ± 0.5	0.91	0.377
		CPC vs SPC	0.45 ± 0.4 vs 0.76 ± 0.3	-1.98	0.065.
		MPC vs PC-S	0.26 ± 0.4 vs 0.01 ± 0.0	1.88	0.092.
		MPC vs PC-M	0.26 ± 0.4 vs 0.28 ± 0.5	-0.10	0.925
		MPC vs SPC	0.26 ± 0.4 vs 0.76 ± 0.3	-3.10	0.007**
		PC-S vs PC-M	0.01 ± 0.0 vs 0.28 ± 0.5	-1.89	0.091.
		PC-S vs SPC	0.01 ± 0.0 vs 0.76 ± 0.3	-8.30	0.000***
		PC-M vs SPC	0.28 ± 0.5 vs 0.76 ± 0.3	-2.85	0.012*
	AH-F1	CPC vs MPC	0.27 ± 0.2 vs 0.14 ± 0.2	1.24	0.230
		CPC vs PC-S	0.27 ± 0.2 vs 0.01 ± 0.0	3.49	0.007**
		CPC vs PC-M	0.27 ± 0.2 vs 0.15 ± 0.2	1.18	0.253
		CPC vs SPC	0.27 ± 0.2 vs 0.41 ± 0.2	-1.45	0.166
		MPC vs PC-S	0.14 ± 0.2 vs 0.01 ± 0.0	1.82	0.101.
		MPC vs PC-M	0.14 ± 0.2 vs 0.15 ± 0.2	-0.04	0.965
		MPC vs SPC	0.14 ± 0.2 vs 0.41 ± 0.2	-2.99	0.009**
		PC-S vs PC-M	0.01 ± 0.0 vs 0.15 ± 0.2	-1.83	0.100.
		PC-S vs SPC	0.01 ± 0.0 vs 0.41 ± 0.2	-8.24	0.000***
		PC-M vs SPC	0.15 ± 0.2 vs 0.41 ± 0.2	-2.87	0.011*

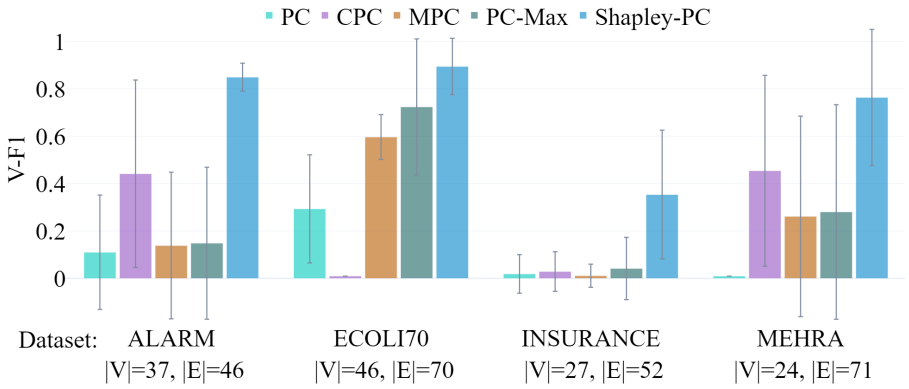


Figure B.3: V-structure F1 for the datasets in Fig. 5.1 in the main text.

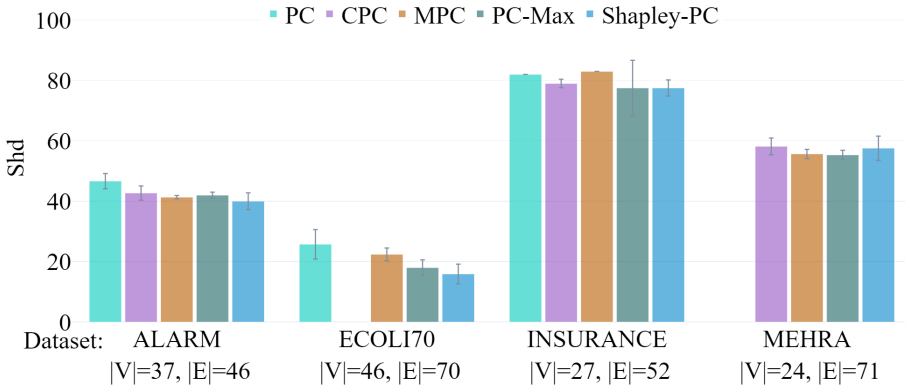


Figure B.4: SHD, the lower the better, for the datasets in Fig. 5.1 in the main text.

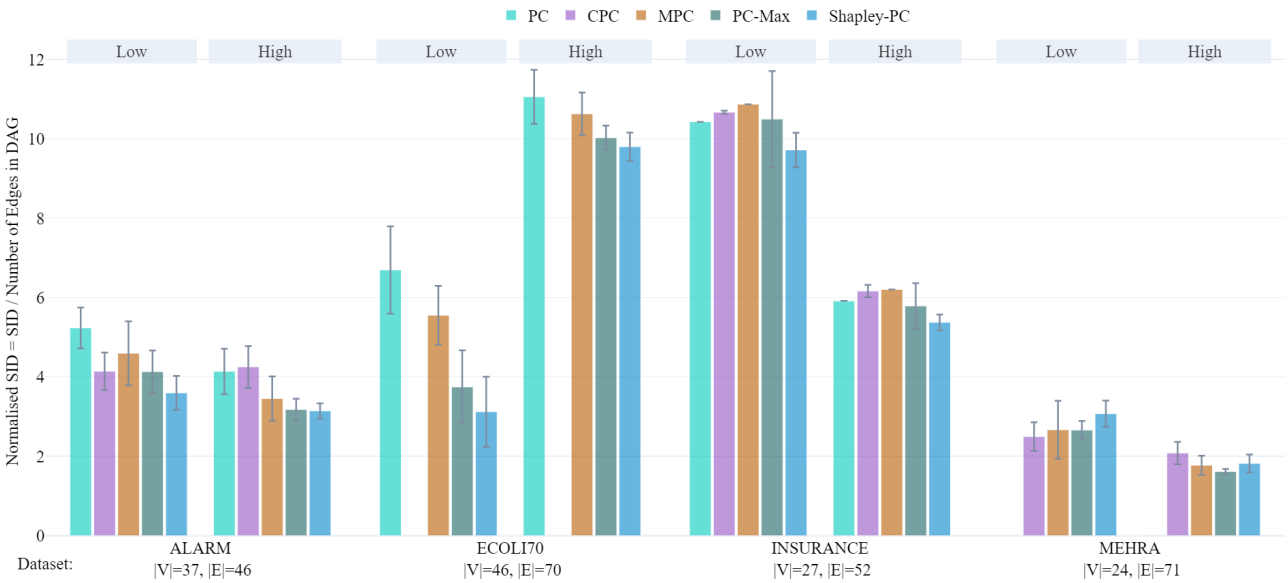


Figure B.5: Normalised SID, the lower the better, for the datasets in Fig. 5.1 in the main text.

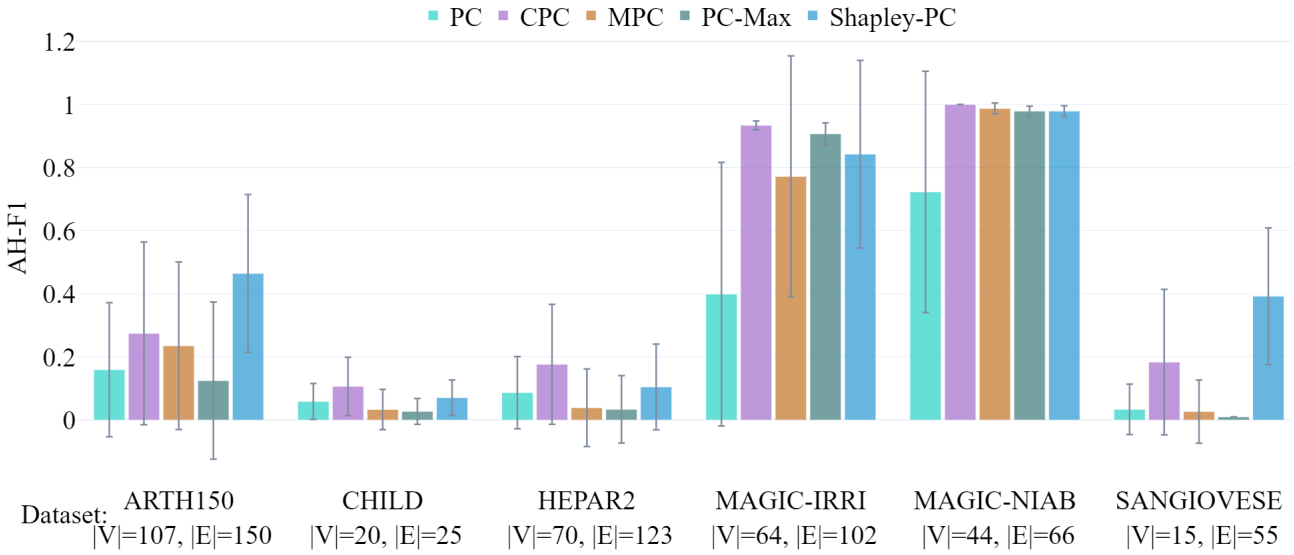


Figure B.6: ArrowHead F1 for additional datasets in the `bnlearn` repository.

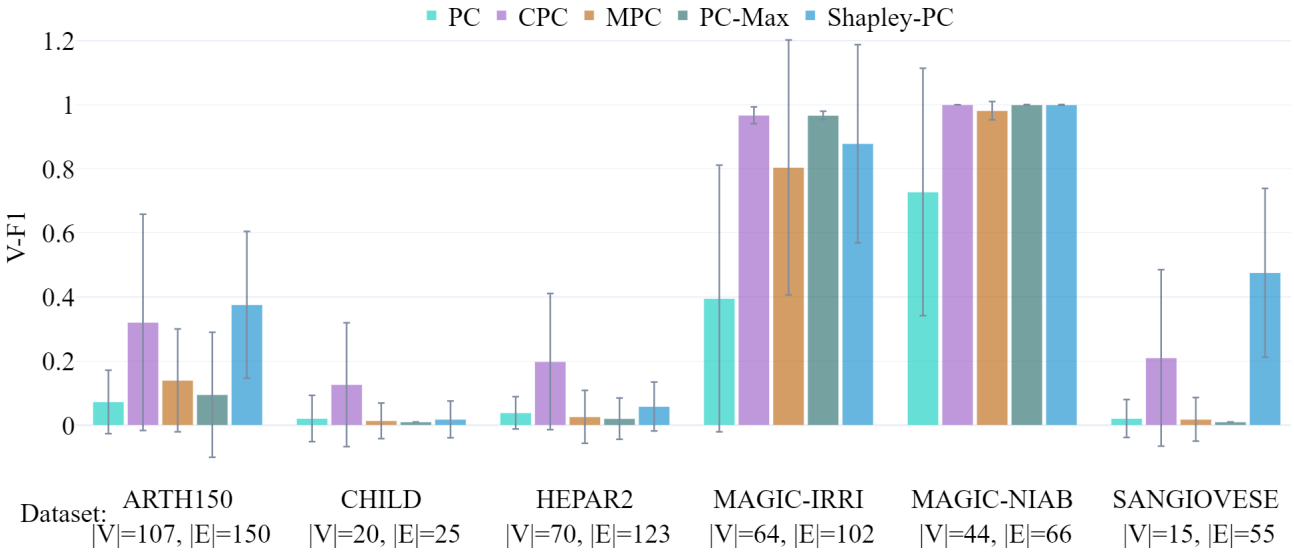


Figure B.7: V-structure F1 for additional datasets in the `bnlearn` repository.

Appendix C

Appendix to Causal ABA

C.1 Further Details for Experiments

We provide an implementation of ABA-PC using clingo (Gebser et al., 2019) version 5.6.2 and python 3.10. The code is available at the following repository: `https://github.com/briziorusso/ArgCausalDisco`. In the repository, we also made available the code to reproduce all experiments and we saved all the plots, presented herein and in the main text, in HTML format. Downloading and opening them in a browser allows the inspection of all the numbers behind the plots interactively.

Hyperparameters We used default parameters for all the methods. For MPC, Shapley-PC and ABAPC (ours) we used Fisher Z test (Fisher, 1970), as implemented in `causal-learn`, with significance threshold $\alpha = 0.05$.

C.1.1 Pseudo-real Data

For our empirical evaluation (see §6.4), we used four datasets from the `bnlearn` repository. Specifically, we use the Asia, Cancer, Earthquake and Survey datasets, as reported in Table C.1, which represent problems of decision-making in the medical, law and policy domains. The datasets are from the small categories with number of nodes varying from 5 to 8. Details on the number of nodes, edges and density of the DAGs underlying the BNs, together with links to a more detailed description on the `bnlearn` repository can be found in Table C.1.

Having downloaded all the .bif files from the repository, we load the BNand the associated conditional probability tables and sample 5000 observations with 50 different seeds to measure variance and confidence intervals. We also run all the experiments shown with 2000 samples and resulted in analogous results, hence omitted.

Table C.1: Details of dataset from `bnlearn`.

Dataset Name	$\|N\|$	$\|E\|$	$\|E\|/\|N\|$
CANCER	5	4	0.8
EARTHQUAKE	5	4	0.8
SURVEY	6	6	1
ASIA	8	8	1

C.1.2 Evaluation Metrics

We evaluated the estimated graphs with five commonly used metrics in causal discovery (as in e.g. (Colombo and Maathuis, 2014; Lachapelle et al., 2020; Ramsey et al., 2017; Zheng et al., 2020)):

- Structural Intervention Distance (SID)
- Structural Hamming Distance (SHD) = E + M + R
- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F1 = 2×Precision*Recall/(Precision+Recall)

The Structural Intervention Distance (SID), introduced by (Peters and Bühlmann, 2015), evaluates how well an estimated causal graph aligns with a ground-truth graph in terms of interventional distributions. SID measures the potential errors in causal inference arising from inaccuracies in the estimated graph, effectively treating these errors as downstream task failures caused by mistakes in the pre-processing step.

Definition C.1.1 (SID (Peters and Bühlmann, 2015))**.** *Let $\mathbb{G}$ be the space of DAGs over d variables. Structural Intervention Distance is defined as*

$$\text{SID} : \mathbb{G} \times \mathbb{G} \to \mathbb{N}, \quad (\mathcal{G}, \mathcal{H}) \mapsto \#\{(i,j), i \neq j \mid \textit{the intervention distribution from } i \textit{ to } j \textit{ is falsely inferred by } \mathcal{H} \textit{ with respect to } \mathcal{G}\}$$

SID counts discrepancies at the level of intervention effects rather than edges. A single edge misrepresentation in the graph can result in multiple incorrect interventional implications due to the transitive nature of causal relationships. For example, an erroneous edge may propagate its effects across downstream nodes, leading to several falsely inferred intervention distributions. As defined,

SID does not satisfy all properties of a metric, particularly it is not symmetric. However, Peters and Bühlmann (2015) show that it is a (pre-)metric, and can be simmetrised.

In contrast, the Structural Hamming Distance (SHD) is a graphical metric that quantifies discrepancies between an estimated directed graph and the true graph by counting the number of edge insertions, deletions, and reversals needed to transform one into the other. Specifically, SHD consists of three components: Extra (E), representing superfluous edges; Missing (M), denoting edges absent in the estimated graph's skeleton; and Reversed (R), capturing edges with incorrect orientations. A lower SHD indicates a more accurate causal graph.

While SHD focuses on the correctness of individual edges, SID assesses the ability of a graph to predict the effects of interventions. Since interventional distributions can often be estimated using only the causal ordering (Bühlmann et al., 2014), SID is less sensitive to arbitrary choices in edge-pruning procedures or thresholds. Intuitively, SID prioritises the causal order, whereas SHD prioritises the accuracy of individual edges.

Precision and Recall measure the proportion of correct orientations based on the estimated graph and the true one, respectively. In the formulae, True Positive (TP) is the number of estimated edges with correct direction; False Positive (FP) is an edge which is not in the skeleton of the true graph. True Negative (TN) and False Negative (FN) are edges that are not in the true graph and correctly (resp, incorrectly) removed from the estimated graph. Finally, F1 score is the harmonic mean of precision and recall.

We carried out the evaluation in the main text on CPDAGs. As discussed in §2.1, CPDAGs represent Markov Equivalence Classes. MECs are all that can be inferred from a given set of independence relations, since multiple DAGs can entail the same set of independencies. Both our method and NOTEARS-MLP output DAGs rather than CPDAGs (which are the output of the other two baselines used, FGS and Shapley-PC). For the methods returning DAGs, we first transform the DAG to a CPDAG (isolating skeleton and v-structures) and then calculate the metrics presented above. Evaluation on DAGs is provided in §C.1.3 for completeness.

C.1.3 Additional Results

Here we provide results that complement the ones in §6.4 in the main text. Specifically, we evaluate our method based on SHD, F1 score, Precision and Recall (see §C.1.2 for details).

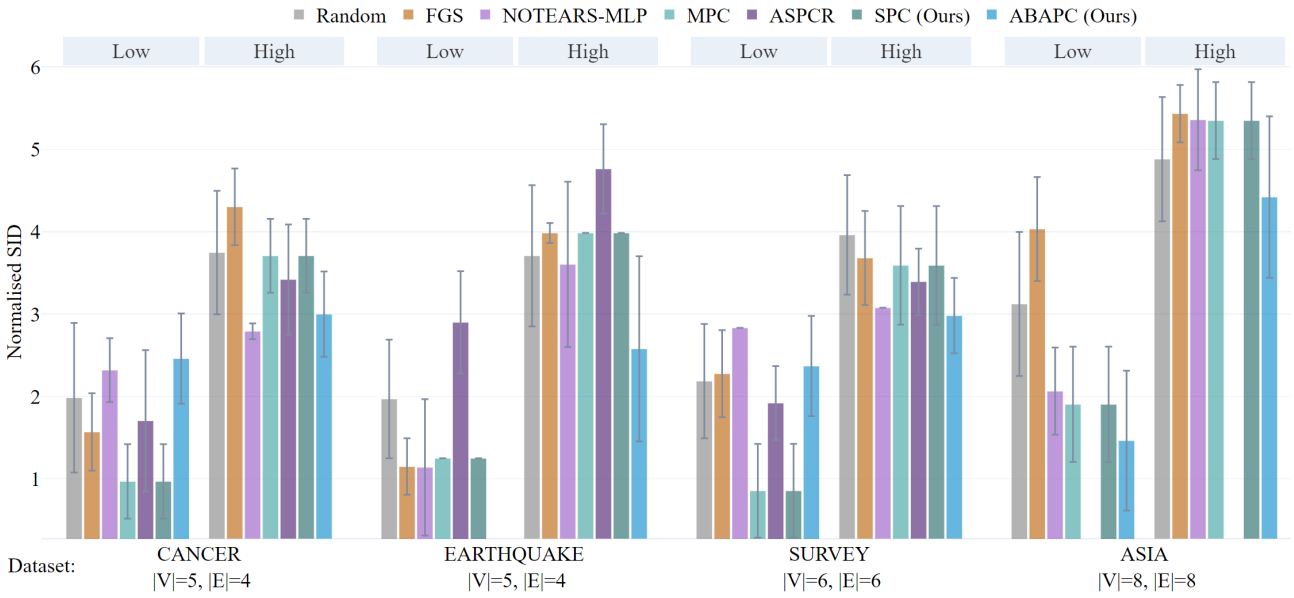


Figure C.1: Normalised Structural Interventional Distance (NSID) for four datasets from the `bnlearn` repository. Lower values are better. Low (resp. High) represents the SID for the best (resp. worst) DAG within the estimated CPDAGs.

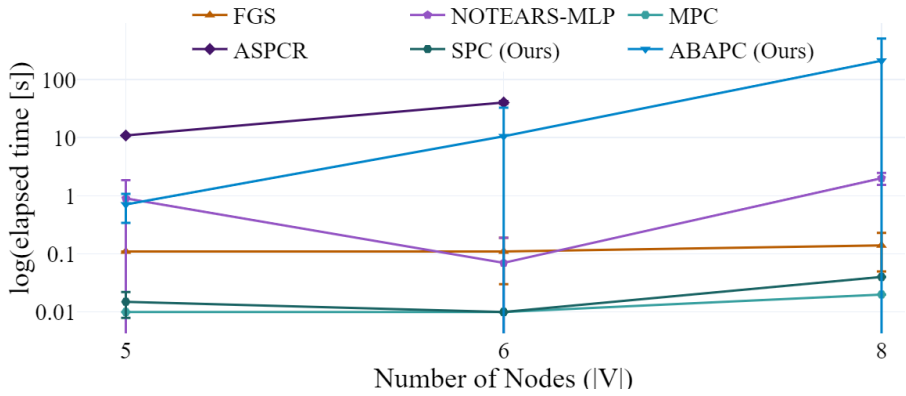


Figure C.2: Mean and standard deviation of elapsed time (log scale) by the number of nodes, averaged over 50 runs.

Additional Baselines

In Figure C.1 we show the results of the SID metric for the additional baselines. Specifically, we add MPC and ASPCR to the comparison. These were excluded from the main text because the results are not significantly different from the other baselines. It is still interesting to compare the performance of our method with these two additional baselines as MPC is the closest method to Shapley-PC, and ASPCR uses as well ASP. Note that the results for ASPCR are missing since the method is not able to handle the datasets with more than 6 nodes. In Figure C.2 we show the runtime comparison of the additional baselines. We can see that MPC has runtime comparable to Shapley-PC, while ASPCR is significantly slower and can only process datasets with up to 6 nodes.

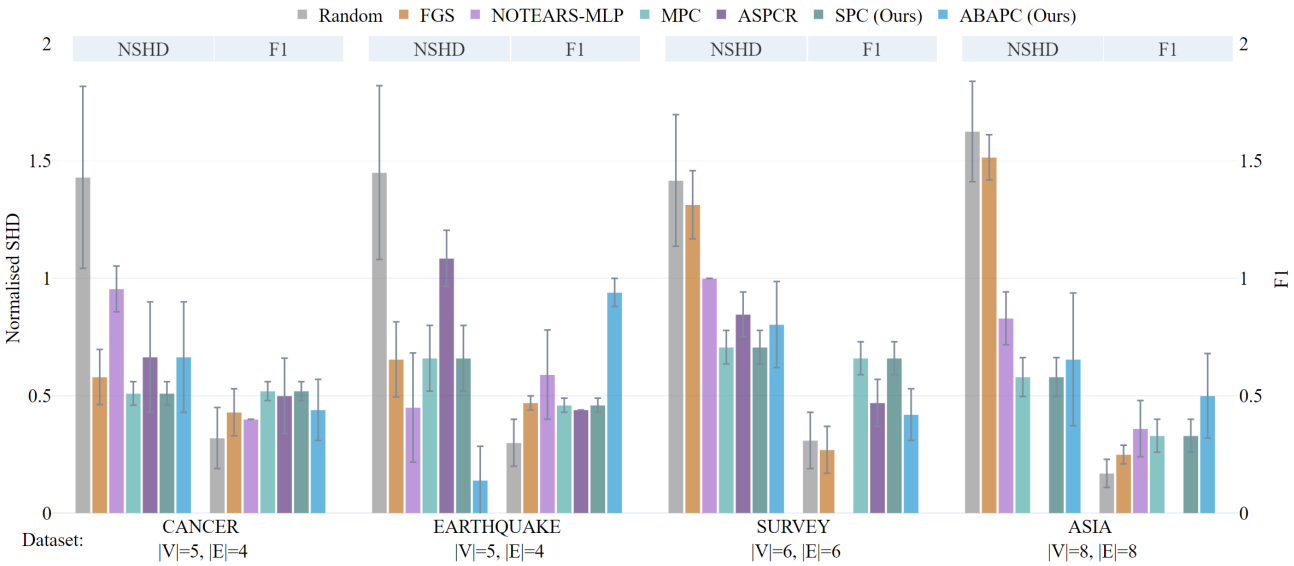


Figure C.3: Normalised Structural Hamming Distance (SHD, left y-axis) and F1 score (right y-axis) for the estimated CPDAGs for four datasets in the `bnlearn` repository. Lower is better for NSHD and higher is better for F1.

Additional Metrics

Additionally, to the results shown in Figure 6.3 in the main text, we evaluate our proposed method with four other metrics commonly used in Causal Discovery: SHD, F1 (Figure C.3), Precision and Recall (Figure C.4). Furthermore, we show the average size of the estimated CPDAGs in (Figure C.5). From Figure C.3, we can see that, according to Normalised SHD, ABA-PC is the best method for the Earthquake dataset and not significantly different from MPC (ranking first) for the other datasets. In terms of F1 score, ABA-PC is significantly better than all baselines for the Earthquake and Asia datasets, on par with all the others for Cancer and ranking second, after MPC, for the Survey dataset. Precision and Recall provide a breakdown of the F1 score (which is their harmonic mean), hence follows similar patterns. From the estimated graph size (the number of edges in the estimated graph) in Figure C.5 we can observe that ABA-PC is generally in line with the true graph size, apart from the Survey dataset for which some of the edges are missed.

DAGs Evaluation

In the main text, we evaluated our method based on the estimated CPDAG compared to the ground truth one. Here, for completeness, we report results based on DAGs. Indeed, both our method and NOTEARS-MLP have DAGs in output. Note that this evaluation penalises the methods outputting CPDAGs. In transforming CPDAGs to DAGs, we had to only select the directed arrows in order to

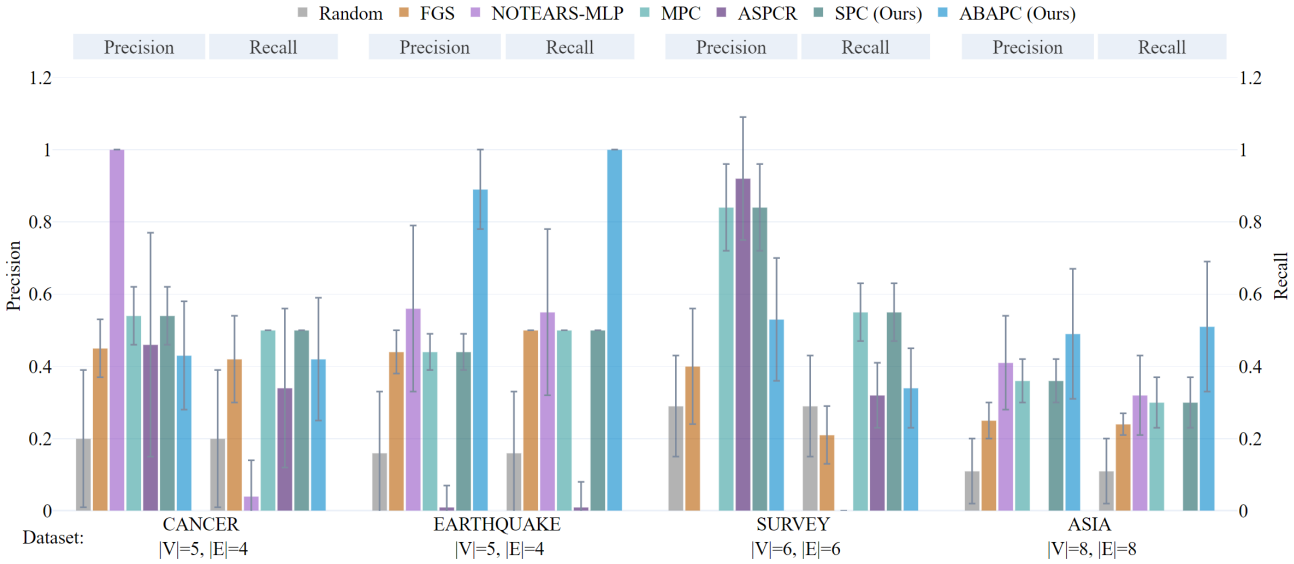


Figure C.4: Precision (left y-axis) and Recall (right y-axis) for the estimated CPDAGs for four datasets in the `bnlearn` repository. Higher is better for both metrics.

extract DAGs from the output CPDAGs (see Figure C.5 and C.6 for the average size of the estimated CPDAGs and DAGs, respectively).

We report the evaluation of DAGs in Figure C.7 (NSID and NSHD) and Figure C.8 (Precision and Recall). The plot of the F1 score is provided in our repository both as image and interactive files. According to NSHD, we can see that ABA-PC performs significantly better than all baselines on two out of the four datasets (Survey and Survey). For the Earthquake and Asia datasets ABA-PC is on par with MPC and significantly better than the other baselines. According to NSID, ABA-PC is not significantly different from MPC, ranking 1^{st} for the Earthquake and Asia datasets, and in line with all other baselines for the other two datasets. According to precision and recall, ABAPC is better than MPC for three out of the four datasets when evaluating on recall, while maintaining the same precision, again, three out of four times. For the Asia dataset ABAPC trades some precision for a significantly higher recall.

C.1.4 Statistical Tests

Here we present details of the statistical tests used to measure the significance of the difference in the results presented in Figure 6.3 in the main text. In tables C.2, C.3, C.4 and C.5 we provide t-statistics and p-values for the Cancer, Earthquake, Survey and Asia datasets, respectively. In each table we present pairwise comparisons of means (shown in brakets together with standard deviations), for the best and worst case SID (High and Low, resp.) presented in Figure 6.3 of the main text.

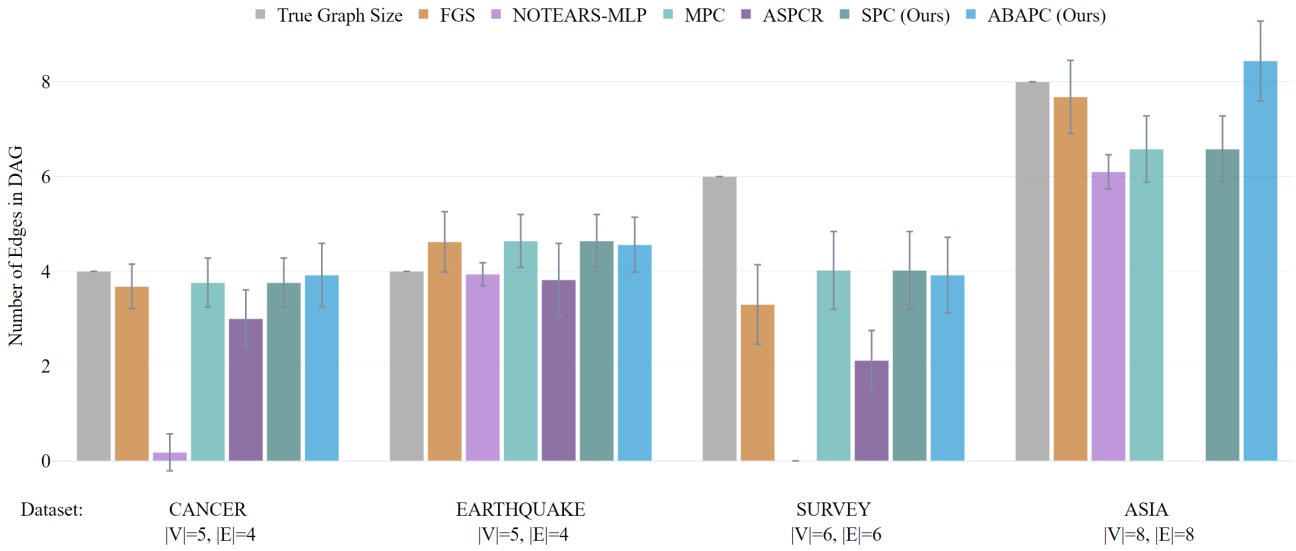


Figure C.5: Number of edges in the estimated CPDAGs compared to ground truth.

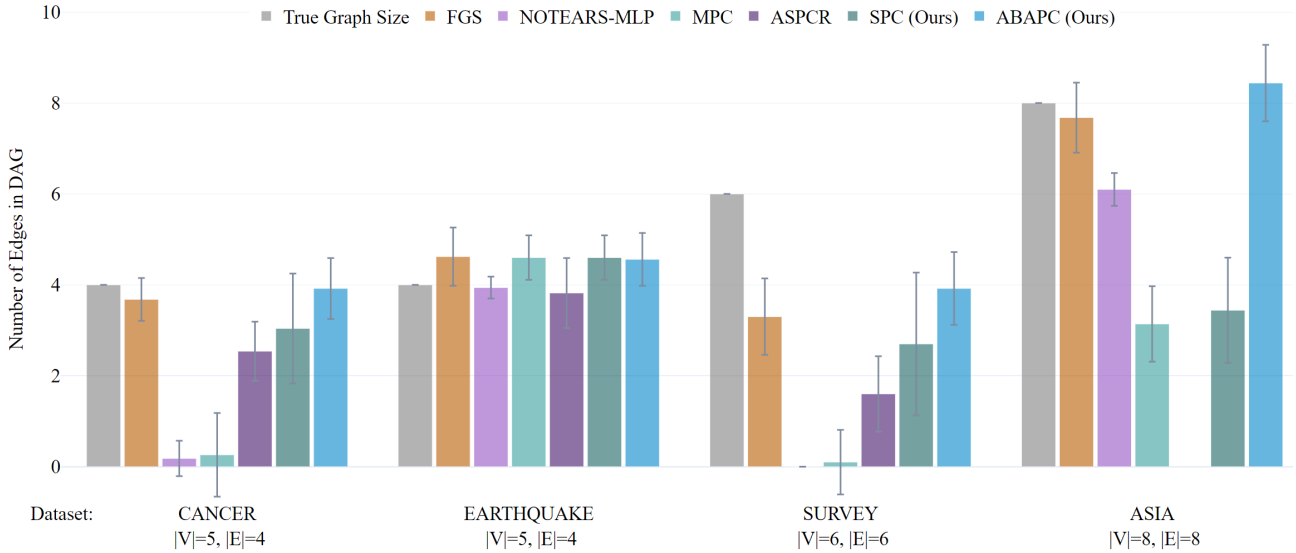


Figure C.6: Number of edges in the estimated DAGs compared to ground truth.

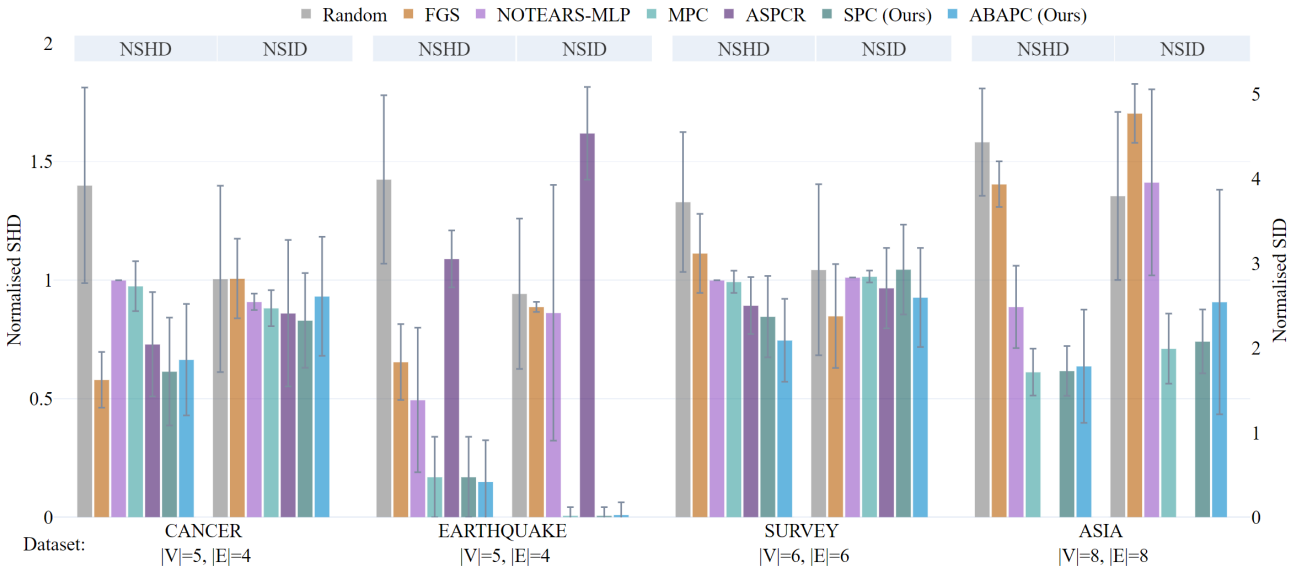


Figure C.7: Normalised Structural Hamming Distance (SHD, left y-axis) and Structural Intervention Distance (SID, right y-axis) for the estimated DAGs for four datasets in the `bnlearn` repository. Lower is better for both metrics.

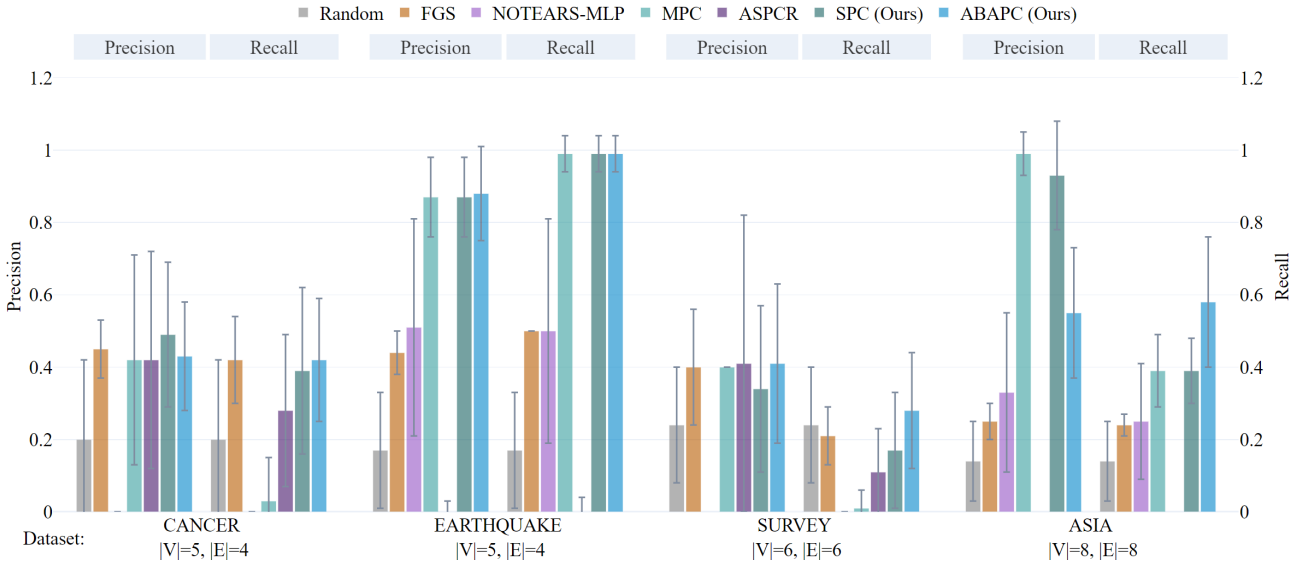


Figure C.8: Precision (left y-axis) and Recall (right y-axis) for the estimated DAGs for four datasets in the `bnlearn` repository. Higher is better for both metrics.

Table C.2: Two-sample, unequal variance t-tests for difference in means for Cancer dataset. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Metric	Methods	Means±Std	t	p-value
NSID (low)	APC vs ASPCR	9.8 ± 2.2 vs 6.8 ± 3.4	5.247	0.000***
	APC vs FGS	9.8 ± 2.2 vs 6.3 ± 1.9	8.722	0.000***
	APC vs MPC	9.8 ± 2.2 vs 3.9 ± 1.8	14.833	0.000***
	APC vs NT	9.8 ± 2.2 vs 9.3 ± 1.6	1.476	0.144
	APC vs RND	9.8 ± 2.2 vs 7.9 ± 3.6	3.169	0.002**
	APC vs SPC	9.8 ± 2.2 vs 3.9 ± 1.8	14.833	0.000***
	ASPCR vs FGS	6.8 ± 3.4 vs 6.3 ± 1.9	0.976	0.332
	ASPCR vs MPC	6.8 ± 3.4 vs 3.9 ± 1.8	5.360	0.000***
	ASPCR vs NT	6.8 ± 3.4 vs 9.3 ± 1.6	-4.621	0.000***
	ASPCR vs RND	6.8 ± 3.4 vs 7.9 ± 3.6	-1.586	0.116
	ASPCR vs SPC	6.8 ± 3.4 vs 3.9 ± 1.8	5.360	0.000***
	FGS vs MPC	6.3 ± 1.9 vs 3.9 ± 1.8	6.503	0.000***
	FGS vs NT	6.3 ± 1.9 vs 9.3 ± 1.6	-8.706	0.000***
	FGS vs RND	6.3 ± 1.9 vs 7.9 ± 3.6	-2.871	0.005**
	FGS vs SPC	6.3 ± 1.9 vs 3.9 ± 1.8	6.503	0.000***
	MPC vs NT	3.9 ± 1.8 vs 9.3 ± 1.6	-16.024	0.000***
	MPC vs RND	3.9 ± 1.8 vs 7.9 ± 3.6	-7.078	0.000***
	MPC vs SPC	3.9 ± 1.8 vs 3.9 ± 1.8	0.000	1.000
	NT vs RND	9.3 ± 1.6 vs 7.9 ± 3.6	2.401	0.019*
	NT vs SPC	9.3 ± 1.6 vs 3.9 ± 1.8	16.024	0.000***
	RND vs SPC	7.9 ± 3.6 vs 3.9 ± 1.8	7.078	0.000***
NSID (high)	APC vs ASPCR	11.0 ± 2.1 vs 12.7 ± 2.7	-3.521	0.001***
	APC vs FGS	11.0 ± 2.1 vs 16.3 ± 1.9	-13.231	0.000***
	APC vs MPC	11.0 ± 2.1 vs 13.9 ± 1.8	-7.315	0.000***
	APC vs NT	11.0 ± 2.1 vs 10.2 ± 0.4	2.794	0.007**
	APC vs RND	11.0 ± 2.1 vs 14.0 ± 3.0	-5.801	0.000***
	APC vs SPC	11.0 ± 2.1 vs 13.9 ± 1.8	-7.315	0.000***
	ASPCR vs FGS	12.7 ± 2.7 vs 16.3 ± 1.9	-7.651	0.000***
	ASPCR vs MPC	12.7 ± 2.7 vs 13.9 ± 1.8	-2.523	0.013*
	ASPCR vs NT	12.7 ± 2.7 vs 10.2 ± 0.4	6.584	0.000***
	ASPCR vs RND	12.7 ± 2.7 vs 14.0 ± 3.0	-2.300	0.024*
	ASPCR vs SPC	12.7 ± 2.7 vs 13.9 ± 1.8	-2.523	0.013*
	FGS vs MPC	16.3 ± 1.9 vs 13.9 ± 1.8	6.503	0.000***
	FGS vs NT	16.3 ± 1.9 vs 10.2 ± 0.4	22.465	0.000***
	FGS vs RND	16.3 ± 1.9 vs 14.0 ± 3.0	4.442	0.000***
	FGS vs SPC	16.3 ± 1.9 vs 13.9 ± 1.8	6.503	0.000***
	MPC vs NT	13.9 ± 1.8 vs 10.2 ± 0.4	14.130	0.000***
	MPC vs RND	13.9 ± 1.8 vs 14.0 ± 3.0	-0.321	0.749
	MPC vs SPC	13.9 ± 1.8 vs 13.9 ± 1.8	0.000	1.000
	NT vs RND	10.2 ± 0.4 vs 14.0 ± 3.0	-8.934	0.000***
	NT vs SPC	10.2 ± 0.4 vs 13.9 ± 1.8	-14.130	0.000***
	RND vs SPC	14.0 ± 3.0 vs 13.9 ± 1.8	0.321	0.749

Table C.3: Two-sample, unequal variance t-tests for difference in means for Earthquake dataset. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Metric	Methods	Means±Std	t	p-value
NSID (low)	APC vs ASPCR	0.0 ± 0.0 vs 11.6 ± 2.5	-33.074	0.000***
	APC vs FGS	0.0 ± 0.0 vs 4.6 ± 1.4	-23.742	0.000***
	APC vs MPC	0.0 ± 0.0 vs 5.0 ± 0.0	$-\infty$	0.000***
	APC vs NT	0.0 ± 0.0 vs 4.6 ± 3.3	-9.741	0.000***
	APC vs RND	0.0 ± 0.0 vs 7.9 ± 2.9	-19.347	0.000***
	APC vs SPC	0.0 ± 0.0 vs 5.0 ± 0.0	$-\infty$	0.000***
	ASPCR vs FGS	11.6 ± 2.5 vs 4.6 ± 1.4	17.470	0.000***
	ASPCR vs MPC	11.6 ± 2.5 vs 5.0 ± 0.0	18.818	0.000***
	ASPCR vs NT	11.6 ± 2.5 vs 4.6 ± 3.3	12.036	0.000***
	ASPCR vs RND	11.6 ± 2.5 vs 7.9 ± 2.9	6.921	0.000***
	ASPCR vs SPC	11.6 ± 2.5 vs 5.0 ± 0.0	18.818	0.000***
	FGS vs MPC	4.6 ± 1.4 vs 5.0 ± 0.0	-2.065	0.044*
	FGS vs NT	4.6 ± 1.4 vs 4.6 ± 3.3	0.079	0.937
	FGS vs RND	4.6 ± 1.4 vs 7.9 ± 2.9	-7.272	0.000***
	FGS vs SPC	4.6 ± 1.4 vs 5.0 ± 0.0	-2.065	0.044*
	MPC vs NT	5.0 ± 0.0 vs 4.6 ± 3.3	0.940	0.352
	MPC vs RND	5.0 ± 0.0 vs 7.9 ± 2.9	-7.071	0.000***
	MPC vs SPC	5.0 ± 0.0 vs 5.0 ± 0.0	NaN	NaN
	NT vs RND	4.6 ± 3.3 vs 7.9 ± 2.9	-5.351	0.000***
	NT vs SPC	4.6 ± 3.3 vs 5.0 ± 0.0	-0.940	0.352
	RND vs SPC	7.9 ± 2.9 vs 5.0 ± 0.0	7.071	0.000***
NSID (high)	APC vs ASPCR	9.3 ± 4.5 vs 18.1 ± 2.2	-12.373	0.000***
	APC vs FGS	9.3 ± 4.5 vs 15.0 ± 0.5	-8.796	0.000***
	APC vs MPC	9.3 ± 4.5 vs 15.0 ± 0.0	-8.847	0.000***
	APC vs NT	9.3 ± 4.5 vs 13.5 ± 4.0	-4.812	0.000***
	APC vs RND	9.3 ± 4.5 vs 13.9 ± 3.5	-5.649	0.000***
	APC vs SPC	9.3 ± 4.5 vs 15.0 ± 0.0	-8.847	0.000***
	ASPCR vs FGS	18.1 ± 2.2 vs 15.0 ± 0.5	9.894	0.000***
	ASPCR vs MPC	18.1 ± 2.2 vs 15.0 ± 0.0	10.138	0.000***
	ASPCR vs NT	18.1 ± 2.2 vs 13.5 ± 4.0	7.187	0.000***
	ASPCR vs RND	18.1 ± 2.2 vs 13.9 ± 3.5	7.356	0.000***
	ASPCR vs SPC	18.1 ± 2.2 vs 15.0 ± 0.0	10.138	0.000***
	FGS vs MPC	15.0 ± 0.5 vs 15.0 ± 0.0	0.000	1.000
	FGS vs NT	15.0 ± 0.5 vs 13.5 ± 4.0	2.669	0.010*
	FGS vs RND	15.0 ± 0.5 vs 13.9 ± 3.5	2.266	0.028*
	FGS vs SPC	15.0 ± 0.5 vs 15.0 ± 0.0	0.000	1.000
	MPC vs NT	15.0 ± 0.0 vs 13.5 ± 4.0	2.689	0.010**
	MPC vs RND	15.0 ± 0.0 vs 13.9 ± 3.5	2.289	0.026*
	MPC vs SPC	15.0 ± 0.0 vs 15.0 ± 0.0	NaN	NaN
	NT vs RND	13.5 ± 4.0 vs 13.9 ± 3.5	-0.558	0.578
	NT vs SPC	13.5 ± 4.0 vs 15.0 ± 0.0	-2.689	0.010**
	RND vs SPC	13.9 ± 3.5 vs 15.0 ± 0.0	-2.289	0.026*

Table C.4: Two-sample, unequal variance t-tests for difference in means for Survey dataset. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Metric	Methods	Means±Std	t	p-value
NSID (low)	APC vs ASPCR	14.2 ± 3.6 vs 11.5 ± 2.7	4.211	0.000***
	APC vs FGS	14.2 ± 3.6 vs 13.7 ± 3.2	0.819	0.415
	APC vs MPC	14.2 ± 3.6 vs 5.1 ± 3.4	12.854	0.000***
	APC vs NT	14.2 ± 3.6 vs 17.0 ± 0.0	−5.386	0.000***
	APC vs RND	14.2 ± 3.6 vs 13.1 ± 4.2	1.405	0.163
	APC vs SPC	14.2 ± 3.6 vs 5.1 ± 3.4	12.854	0.000***
	ASPCR vs FGS	11.5 ± 2.7 vs 13.7 ± 3.2	−3.640	0.000***
	ASPCR vs MPC	11.5 ± 2.7 vs 5.1 ± 3.4	10.387	0.000***
	ASPCR vs NT	11.5 ± 2.7 vs 17.0 ± 0.0	−14.405	0.000***
	ASPCR vs RND	11.5 ± 2.7 vs 13.1 ± 4.2	−2.284	0.025*
	ASPCR vs SPC	11.5 ± 2.7 vs 5.1 ± 3.4	10.387	0.000***
	FGS vs MPC	13.7 ± 3.2 vs 5.1 ± 3.4	12.940	0.000***
	FGS vs NT	13.7 ± 3.2 vs 17.0 ± 0.0	−7.450	0.000***
	FGS vs RND	13.7 ± 3.2 vs 13.1 ± 4.2	0.730	0.467
	FGS vs SPC	13.7 ± 3.2 vs 5.1 ± 3.4	12.940	0.000***
	MPC vs NT	5.1 ± 3.4 vs 17.0 ± 0.0	−24.593	0.000***
	MPC vs RND	5.1 ± 3.4 vs 13.1 ± 4.2	−10.490	0.000***
	MPC vs SPC	5.1 ± 3.4 vs 5.1 ± 3.4	0.000	1.000
	NT vs RND	17.0 ± 0.0 vs 13.1 ± 4.2	6.595	0.000***
	NT vs SPC	17.0 ± 0.0 vs 5.1 ± 3.4	24.593	0.000***
	RND vs SPC	13.1 ± 4.2 vs 5.1 ± 3.4	10.490	0.000***
NSID (high)	APC vs ASPCR	16.4 ± 2.8 vs 18.9 ± 2.4	−4.797	0.000***
	APC vs FGS	16.4 ± 2.8 vs 20.7 ± 3.5	−6.764	0.000***
	APC vs MPC	16.4 ± 2.8 vs 20.1 ± 4.4	−5.065	0.000***
	APC vs NT	16.4 ± 2.8 vs 17.0 ± 0.0	−1.481	0.145
	APC vs RND	16.4 ± 2.8 vs 22.4 ± 4.4	−8.092	0.000***
	APC vs SPC	16.4 ± 2.8 vs 20.1 ± 4.4	−5.065	0.000***
	ASPCR vs FGS	18.9 ± 2.4 vs 20.7 ± 3.5	−2.910	0.005**
	ASPCR vs MPC	18.9 ± 2.4 vs 20.1 ± 4.4	−1.700	0.093.
	ASPCR vs NT	18.9 ± 2.4 vs 17.0 ± 0.0	5.587	0.000***
	ASPCR vs RND	18.9 ± 2.4 vs 22.4 ± 4.4	−4.848	0.000***
	ASPCR vs SPC	18.9 ± 2.4 vs 20.1 ± 4.4	−1.700	0.093.
	FGS vs MPC	20.7 ± 3.5 vs 20.1 ± 4.4	0.686	0.494
	FGS vs NT	20.7 ± 3.5 vs 17.0 ± 0.0	7.480	0.000***
	FGS vs RND	20.7 ± 3.5 vs 22.4 ± 4.4	−2.151	0.034*
	FGS vs SPC	20.7 ± 3.5 vs 20.1 ± 4.4	0.686	0.494
	MPC vs NT	20.1 ± 4.4 vs 17.0 ± 0.0	5.060	0.000***
	MPC vs RND	20.1 ± 4.4 vs 22.4 ± 4.4	−2.560	0.012*
	MPC vs SPC	20.1 ± 4.4 vs 20.1 ± 4.4	0.000	1.000
	NT vs RND	17.0 ± 0.0 vs 22.4 ± 4.4	−8.633	0.000***
	NT vs SPC	17.0 ± 0.0 vs 20.1 ± 4.4	−5.060	0.000***
	RND vs SPC	22.4 ± 4.4 vs 20.1 ± 4.4	2.560	0.012*

Table C.5: Two-sample, unequal variance t-tests for difference in means for Asia dataset. Significance levels: 0 '***', 0.001 '**', 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Metric	Methods	Means±Std	t	p-value
SID (low)	APC vs FGS	11.7 ± 6.8 vs 32.3 ± 5.0	-17.164	0.000***
	APC vs MPC	11.7 ± 6.8 vs 15.2 ± 5.6	-2.828	0.006**
	APC vs NT	11.7 ± 6.8 vs 16.5 ± 4.2	-4.243	0.000***
	APC vs RND	11.7 ± 6.8 vs 25.0 ± 7.0	-9.622	0.000***
	APC vs SPC	11.7 ± 6.8 vs 15.2 ± 5.6	-2.828	0.006**
	FGS vs MPC	32.3 ± 5.0 vs 15.2 ± 5.6	15.960	0.000***
	FGS vs NT	32.3 ± 5.0 vs 16.5 ± 4.2	16.895	0.000***
	FGS vs RND	32.3 ± 5.0 vs 25.0 ± 7.0	5.970	0.000***
	FGS vs SPC	32.3 ± 5.0 vs 15.2 ± 5.6	15.960	0.000***
	MPC vs NT	15.2 ± 5.6 vs 16.5 ± 4.2	-1.290	0.200
	MPC vs RND	15.2 ± 5.6 vs 25.0 ± 7.0	-7.690	0.000***
	MPC vs SPC	15.2 ± 5.6 vs 15.2 ± 5.6	0.000	1.000
	NT vs RND	16.5 ± 4.2 vs 25.0 ± 7.0	-7.322	0.000***
	NT vs SPC	16.5 ± 4.2 vs 15.2 ± 5.6	1.290	0.200
	RND vs SPC	25.0 ± 7.0 vs 15.2 ± 5.6	7.690	0.000***
SID (high)	APC vs FGS	33.5 ± 7.9 vs 41.7 ± 2.8	-6.883	0.000***
	APC vs MPC	33.5 ± 7.9 vs 41.0 ± 3.8	-6.046	0.000***
	APC vs NT	33.5 ± 7.9 vs 41.1 ± 5.0	-5.739	0.000***
	APC vs RND	33.5 ± 7.9 vs 37.2 ± 6.1	-2.633	0.010**
	APC vs SPC	33.5 ± 7.9 vs 41.0 ± 3.8	-6.046	0.000***
	FGS vs MPC	41.7 ± 2.8 vs 41.0 ± 3.8	1.023	0.309
	FGS vs NT	41.7 ± 2.8 vs 41.1 ± 5.0	0.745	0.458
	FGS vs RND	41.7 ± 2.8 vs 37.2 ± 6.1	4.702	0.000***
	FGS vs SPC	41.7 ± 2.8 vs 41.0 ± 3.8	1.023	0.309
	MPC vs NT	41.0 ± 3.8 vs 41.1 ± 5.0	-0.091	0.928
	MPC vs RND	41.0 ± 3.8 vs 37.2 ± 6.1	3.732	0.000***
	MPC vs SPC	41.0 ± 3.8 vs 41.0 ± 3.8	0.000	1.000
	NT vs RND	41.1 ± 5.0 vs 37.2 ± 6.1	3.478	0.001***
	NT vs SPC	41.1 ± 5.0 vs 41.0 ± 3.8	0.091	0.928
	RND vs SPC	37.2 ± 6.1 vs 41.0 ± 3.8	-3.732	0.000***