Causal modelling and validation based on observational data and domain knowledge

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Physik der Universität Regensburg

vorgelegt von

Daniel Grünbaum

aus Kelheim

im Jahr 2023

Promotionsgesuch eingereicht am: 21.04.2023
Die Arbeit wurde angeleitet von: Prof. Dr. Elmar Lang

Prüfungsausschuss:
Prof. Dr. Dieter Weiss (Vorsitzender)
Prof. Dr. Elmar Lang (Erstgutachter)
Prof. Dr. Rainer Spang (Zweitgutachter)
Prof. Dr. Tilo Wettig (weiterer Hochschullehrer)

ABSTRACT

By assessing the effect of hypothetical actions without the need to directly interact with the real world, causal inference offers valuable tools for data science and artificial intelligence. However, a consensus on how to combine different causal algorithms into a holistic analysis workflow, as well as a universally agreed-upon validation strategy for causal models are yet to be established. In this thesis, a causal end-to-end analysis is proposed as a combination of multiple methods of graph-based causal inference [1] from observational data and domain knowledge. Quantitative probing [2] is introduced as a model-agnostic causal validation strategy in accordance with Popper's falsificationist view on scientific discovery [3]. The effectiveness of the strategy is evidenced by a thorough simulation study that includes a discussion of its current limits at the example of malfunctioning validation runs. In order to provide application scenarios for the methodological contributions, selected use cases from the domain of manufacturing light-emitting diodes are presented. Open-source Python packages for executing the causal end-to-end analysis and benchmarking the quantitative probing validation strategy are provided [4, 5].

CONTENTS

Part I

INTRODUCTION

Denken ist Handeln im vorgestellten Raum. / Thinking is acting in an imagined space.
– Konrad Lorenz, *Die Rückseite des Spiegels* [6]

Il faut cultiver notre jardin. / We must cultivate our garden.
– Voltaire, *Candide, ou l'optimisme* [7]

When we make use of data science techniques, there is often a mismatch between the employed methods and the ultimate goal: The methods are mostly descriptive or predictive in nature, passively describing what is or what will be, whereas the intent usually is to find an active manipulation of the reality that helps us achieve certain goals. A company faced with horrible sales forecasts will not stoically accept that customers do not want their product, but ask how they can prevent the predicted scenario. The management could decide that a new sales strategy should be implemented and ask the data scientist to predict the company's success in the actively manipulated scenario. It is tempting to query the same model that has produced the bad sales forecast and ask for an updated prediction factoring in the introduction of the new sales strategy. However, due to spurious correlations and hidden biases that might be present in the data, it is hard to distill genuinely causal mechanisms from observed data that can be exploited for intelligent decision making. Gaining knowledge about the effect of hypothetical changes to a system is precisely the task of *causal inference*. The concerns about purely data-driven causal inference have lead to slogans such as "No causation without manipulation" [8], suggesting that experiments, as opposed to only passively observed data, are necessary to uncover and exploit the underlying causal structure. Nevertheless, causal inference researchers lead by Judea Pearl and Donald Rubin have succeeded in developing techniques that provably solve the above task by supplementing the observational data with additional assumptions [1, 9]. These efforts have been recognized by the scientific community and their influence is evidenced by Pearl winning the ACM Turing Award [10] and Rubin's collaborator Guido Imbens winning the Nobel Memorial Prize in Economics [11]. Considering the goal of artificial intelligence (AI) research, namely creating machines or algorithms that can truly think, causal inference provides an exciting new perspective: By acting in an imagined space, which is made possible by the different classes of models developed by the causal inference community, it is feasible to think ahead and assess the effect of various actions. On this basis, intelligent strategies can be devised without directly interacting with the real

world. The techniques for manual construction or even data-driven learning of causal models that satisfy the demands of downstream AI tasks are already available [1, 12]. However, so far there is no consensus on how to combine the different causal methods that solve subtasks of the causal inference spectrum, and no common interface for the integration of application-specific domain knowledge. An even greater concern lies in the absence of agreed-upon tools for validating the causal models once they have been constructed. As we will see, the validation task is considerably harder for causal models than for purely predictive ones, and still an active area of research. Without reassurance that our causal model actually does what it promises, it can be misleading and dangerous to base any downstream decisions on its output. The main contributions of this thesis therefore focus on these two questions:

1. How can we create a causal end-to-end analysis that combines appropriate causal model types and algorithms into a holistic strategy for causal modelling based on observational data and domain knowledge?

2. How can we validate the resulting causal models?

The remainder of the thesis is structured as follows: Part II starts by highlighting the differences between causal and purely associational queries, in order to demonstrate the need for causal inference methods. Techniques for answering causal queries are presented and graphical causal models are introduced as an appropriately complex model type for answering interventional questions. Consequently, causal discovery techniques are discussed as a method of recovering the causal graph from observational data and domain knowledge in situations where it cannot be manually constructed by domain experts. The discussed tools are subsequently combined into the causal end-to-end analysis, which is a holistic strategy for causal analysis based on observational data and domain knowledge.

Given that the causal modelling process is a complex workflow, in which errors can have drastic consequences for the intended downstream task, Part III focusses on the validation of causal models using quantitative domain knowledge. Therefore, the state of the art for the validation of both correlation-based and causal models is briefly reviewed together with methods of exploiting domain knowledge for causal modelling. Based on the gaps in the literature, quantitative probing is developed as a largely model-agnostic causal validation strategy that integrates quantitative domain knowledge and follows the logic of scientific discovery. The proposed validation strategy is evaluated in a

thorough simulation study, which serves to highlight both strengths and weaknesses of the concept.

Although the main focus of this thesis is on the development of causal analysis methods, Part IV briefly illustrates the presented techniques by discussing application cases from the domain of LED manufacturing. The usefulness of quantitative probing together with the causal end-to-end analysis is shown using the example of color point shifts during the manufacturing process. A real-world instance of Simpson's paradox is resolved for evaluating the benefit of an additional phosphor conversion process step. Finally, an ongoing project about the holistic optimization of the production processes serves to highlight the potential of combining causal inference methods with the techniques of reinforcement learning, before Part V concludes the thesis by shortly reviewing its contents and main findings.

Part II

THEORY AND PRACTICE OF CAUSAL END-TO-END ANALYSIS

In the first part of this thesis, we want to introduce the central elements of causal inference from observational data and domain knowledge. We follow the graph-based framework that was pioneered by Judea Pearl [1], as opposed to Rubin's potential outcomes framework [9], due to its clearer interface for integrating domain knowledge. The first chapter highlights the difference between associational queries, which can be answered using correlation-based statistical tools, and genuinely causal queries, in order to demonstrate the need for methods of causal inference. Subsequently, different methods for answering interventional queries are explained and graphical models are presented as a viable solution that strikes the balance between overly complex structural causal models and insufficiently powerful purely statistical models. Therefore, the methods of causal discovery from observational data and domain knowledge are briefly introduced, in order to tackle situations where the graphical structure of the data generating process is not known to the analyst. Finally, the presented methods for learning the causal graph, identifying an unbiased estimand for the causal effect of interest, and estimating the latter are combined into a causal end-to-end analysis. The cause2e package is provided as an open-source Python implementation of this analysis strategy [4].

1. CAUSALITY VS. CORRELATION

Causality $\neq$ *correlation* is a common warning issued at the beginning of statistics courses. However, lectures tend to proceed by ignoring all causal questions and focussing purely on analyzing correlations, which is clearly not an option for this thesis about causal inference. In order to acquire sufficient working knowledge about the difference between the two terms, we will first examine a number of counterintuitive situations where correlation-based statistical tools are not enough to resolve the problem. Subsequently, we briefly introduce a popular exemplary problem that helps illustrate all the methodology presented in the remainder of the thesis. Finally, a terminology of queries is presented, which allows us to select the right tools for tackling different types of statistical problems.

1.1 Shortcomings of correlation-based analysis

Before we dive into the fascinating, but admittedly sometimes confusing and counterintuitive world of causal inference, we want to see whether the additional methodology is really necessary. Why should we not simply use traditional correlation-based methods to determine causal effects from observational data? Discussing serious problems arising from missing causal tools in statistical analysis, we aim to eliminate all such doubts about the necessity of proper causal inference methodology.

1.1.1 Hidden variables

The first common pitfall is the omission of important variables from an analysis. Ice cream sales and shark attacks are likely to be positively correlated, but the two direct causal conclusions - sharks preferring humans with ice cream filling or humans resorting to stress eating from watching their fellow citizens being devoured - both seem unlikely. Indeed, this is an example of a *spurious correlation* that does not imply direct causation. The explanation is given

by including a third important variable that has previously been hidden, namely the temperature: On hot days, people are more likely both to buy ice cream and to go swimming in the shark-infested sea. Such a variable that influences treatment and outcome in a causal study is called a *confounder* and we need to account for it explicitly in the analysis to avoid spurious correlations. While the above example admittedly sounds constructed and easy to defuse for experienced statisticians, not all confounding is equally harmless: In a heatedly debated study [13], researchers empirically confirmed a correlation between soda consumption and violent tendencies among teenagers. Although the authors refrained from drawing causal conclusions and others cited socioeconomic status as an important neglected confounder, media coverage partly portrayed soda consumption as the direct cause of the violent tendencies. The above examples suggest that we should condition on all available covariates, in order to avoid confounding bias due to hidden variables. However, this strategy has major drawbacks: Firstly, subdividing the available data into a potentially enormous number of subpopulations and calculating statistical measures on each of them separately leads to unnecessarily small effective sample sizes and imprecise estimates. Secondly, blindly searching for correlations in large amounts of data leads to false positives that arise purely for statistical reasons and cannot be explained by confounding, as is illustrated in Figure 1.1. The figure is reprinted with permission from [14], where circa 30000 such correlations have been collected by automated data mining methods.

As we will see later on, causal inference allows us to resolve both problems by explicitly using qualitative knowledge about the data generating process as an additional input.

1.1.2 Berkson's paradox

In addition to the above mentioned issue with sample size, there are cases where conditioning on additional variables does not reduce but increase bias. Recent studies investigated the peculiar negative correlation between smoking and COVID-19 severity [15] among different parts of the population. As an explanation, [16] suggests that the counterintuitive observation is not causal, but a case of *collider bias* or *Berkson's paradox*. By selecting a subpopulation, such as patients admitted to a hospital, which is equivalent to conditioning on a covariate, we introduce a bias: Persons that have been admitted, but do not suffer from severe COVID-19, must suffer from some other disease to

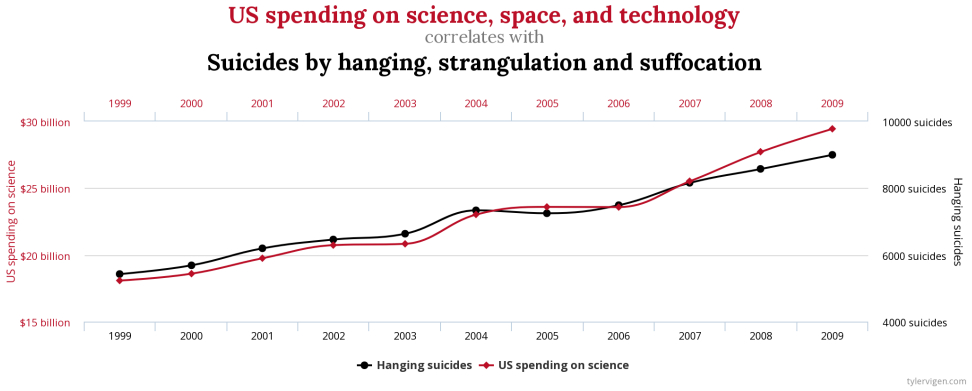

Fig. 1.1: A Pearson correlation of 0.9979 seems to suggest a clear causal relation between government spending on science and certain types of suicide. Reprinted with permission from [14].

justify their admission. This disease might in turn be positively correlated to (and even caused by) smoking. Therefore, severe symptoms of COVID-19 can be negatively correlated to being a smoker under the right numerical circumstances. The paradox is named after Joseph Berkson, who published a seminal article about the still relevant problems of using fourfold tables in observational clinical studies in 1946 [17]. As an entertaining example, the Wikipedia entry on Berkson's paradox shows how talent and attractiveness are negatively correlated among celebrities, even if the traits are independent among the general population: By considering only celebrities, we introduce a collider bias, since people that are ugly should at least be talented to become famous [18]. In summary, we see that conditioning on the wrong covariates can be just as dangerous as forgetting to condition on a confounder.

1.1.3 Simpson's paradox

Simpson's paradox is named after Edward Simpson who described the phenomenon of opposite causal effects in subpopulations and the overall population in 1951 [19]. In illustrating perhaps the most surprising paradox on our list, we follow Pearl's introduction [20] that clearly demonstrates how data alone is not enough to draw causal conclusions. Table 1.1 shows data from an observational study about the effectiveness of a drug for treating a certain illness. A decision maker who is presented with only the first two rows

would certainly see the data as evidence that the drug increases the chances of recovery: Both the male (93% vs. 87%) and the female subpopulation (73% vs. 69%) show better recovery rates when using the drug. However, considering the coarser aggregate data, the effect is reversed: Patients that have taken the drug are less likely to recover (78% vs. 83 %). Therefore, a decision maker who is presented with only the data for the overall population would most likely decide against recommending the drug to future patients.

	Recovery rates		
	Drug	No drug	Overall
Men	81/87 (93%)	234/270 (87%)	315/357 (88%)
Women	192/263 (73%)	55/80 (69%)	247/343 (72%)
Combined data	273/350 (78%)	289/350 (83%)	562/700 (80%)

Tab. 1.1: Simpson's paradox: Although both sex-specific subpopulations seem to indicate a beneficial effect of the drug, the overall population seems to show the opposite effect. Adapted from [20].

The confusing behavior can be explained by examining the number of patients in each of the finer subpopulations: Both men and women have contributed to the study in roughly equal proportions. However, it appears that women are considerably more likely than men to take the drug (263 vs. 87 cases) and at the same time have lower overall chances of recovery (72% vs. 88%). These observations hint at the solution for the seemingly paradox behavior of the numbers: Since women are generally more likely to die from the illness, which might be explained by biological factors, and at the same time more likely to take the drug, a spurious correlation is established between taking the drug and not recovering in the non-adjusted overall population data. As with sharks and ice cream, the solution is given by conditioning on the confounder (sex), such that the subpopulation trends lead to the correct conclusion: Taking the drug is advantageous for both men and women.

Note how it is impossible to draw this conclusion from the data alone: Table 1.2 holds the exact same data for the study of a different illness, but the labelling has changed. This time, we record the patients' blood pressure after the treatment instead of their sex, and the column labels have been switched. Is it again correct to condition on all covariates, leading to the conclusion that the drug is harmful? The answer is no: The illness is caused by high blood pressure, which is counteracted by the drug. Consistent with this observation,

the combined data shows that the recovery rate is better for treated patients. By erroneously conditioning on the post-treatment blood pressure, we would condition away the effect of the drug. The only effect remaining to be seen in the subpopulations would be harmful side effects of the drug, therefore suggesting a negative treatment effect on recovery rates. Falsely conditioning on all available covariates is again an instance of Berkson's paradox.

	Recovery rates		
	No drug	Drug	Overall
Low BP	81/87 (93%)	234/270 (87%)	315/357 (88%)
High BP	192/263 (73%)	55/80 (69%)	247/343 (72%)
Combined data	273/350 (78%)	289/350 (83%)	562/700 (80%)

Tab. 1.2: Simpson's paradox: Although both blood-pressure-specific subpopulations seem to indicate a harmful effect of the drug, the overall population seems to show the opposite effect. Adapted from [20].

Even the statistically literate reader is probably confused by the above scenarios and their different resolutions that are far from obvious from inspecting the raw data. We highlight that these problems are hard to solve for cases with only three variables and the situation will not improve for considerably more complex real-world applications. Unfortunately, Simpson's paradox does indeed not only appear in carefully constructed numerical examples, but also in studies whose conclusions are likely to influence important political decisions [21, 22, 23]. An often cited study by Bickel et al. from 1975 focussed on seemingly obvious bias against women in the process of graduate school admissions at Berkeley [21]. Although the different ratios of admission between men and women displayed a massive discrepancy in favor of male applicants, closer examination of the underlying mechanisms revealed a small, but significant bias against men. Women were simply applying more often to departments with a higher rejection rate, which lead to their overall worse chances of success. In summary, we see that it is hard and even dangerous to answer causal questions on the basis of non-causal, purely data-driven methods. Therefore, we will introduce techniques of causal inference and combine them into a holistic strategy for causal analysis based on data and domain knowledge.

1.2 Sprinkler example

In order to make all the theoretical and practical concepts in this thesis more tangible, we introduce an exemplary scenario that is widely used within the causal inference community: Pearl's sprinkler example [1]. As we will see later on, it has been carefully crafted to illustrate many aspects of causality while still being accessible to researchers from all application domains due to its conceptual simplicity. This makes it ideal for our purposes and we will therefore prefer this example over more domain specific applications scenarios from the optical semiconductor industry. We take the liberty to add, emphasize and omit aspects of Pearl's original formulation whenever it is beneficial for this thesis. In short, the sprinkler example is about inferring causal effects from observational data and possibly domain knowledge. Suppose that a gardener has logged the same five variables daily over the course of a year, in order to learn more about his lawn watering strategy:

- What is the current season?
- Is it raining?
- Is the lawn sprinkler turned on?
- Is the lawn wet?
- Is the lawn slippery?

Additionally, he might know of some direct causal influences, e.g. the fact that rain makes the lawn wet. A graphical representation of the data generating process, unknown to the gardener, is shown in Figure 1.2. The gardener is interested in finding out how these variables influence each other. In particular, he is concerned whether sprinkler activation is responsible for an increased slipperiness of the lawn. The easiest way to answer the question would be to conduct a direct experiment: Each morning, the gardener would throw a coin and activate the sprinkler depending on the outcome. Assuming a constant number of visitors per day, he counts how many of them slip and compares the numbers between days with and without sprinkler activation. The major drawback of this simple strategy is the need to put people's health in jeopardy for the sake of answering the causal question. Furthermore, the experiment answers only one question, whereas other causal effects between the variables remain unknown to the gardener. In light of the already logged

and readily available observational data, it seems preferable to run a statistical analysis that avoids any collateral damages caused by experimentation. The naive approach is the calculation of correlation coefficients or the examination of linear regression coefficients. However, Section 1.1 has already highlighted that correlation-based techniques are not an adequate match for causal queries. The correlation coefficients are by definition symmetrical, contrary to the directed nature of causation, and linear regression coefficients depend on the inclusion of the correct covariates to be effective. At least for the challenge of covariate selection, it should be possible to incorporate the gardener's aforementioned domain knowledge, but the strategy for doing so is not clear. In the remainder of this thesis, we will, among other things, introduce methodology for solving the sprinkler example and more generally the inference of causal effects from observational data and domain knowledge.

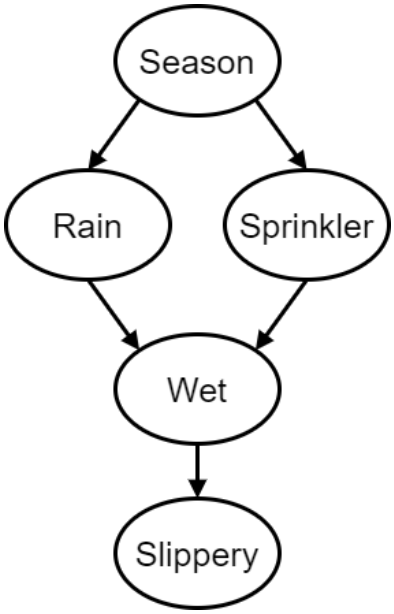

Fig. 1.2: Visualization of Pearl's sprinkler example: The season directly influences the probability of rain and the activation of the sprinkler. The latter two variables are responsible for the wetness of the lawn, which in turn determines its slipperiness. All arrows represent direct causal influences, as opposed to indirect effects, which correspond to directed paths.

1.3 Terminology of queries

In order to understand what queries cannot be answered by correlation-based or purely probabilistic analysis, it is helpful to have a clear terminology of the queries themselves.

1.3.1 Associational queries

The queries that we aim to answer with traditional probabilistic or statistical methods are *associational* in nature. "What is the probability that the lawn is wet, given that it rains? What is the probability that it rains, given that the lawn is wet?" Answering these questions requires the computation of *conditional probabilities* or *conditional probability distributions (CPDs)* of the form

$$p(Y|X = x) \tag{1.1}$$

where Y is the prediction target and X a variable whose value is observed to be x. Note that the value of X is not actively manipulated in any way: We are merely observing its value in a passive role, in an attempt to gather additional information for the prediction of Y. Apart from such predictions, another type of associational query aims at relating variables to each other: "Does the observation that it is raining have any benefit for predicting the state of the sprinkler or vice versa? What if we already know the current season?" As the questions already suggest, they can be answered using CPDs. Two variables X and Y are called *(statistically) independent*, written as

$$X \perp\!\!\!\perp Y, \tag{1.2}$$

if and only if

$$p(Y, X) = p(Y) \cdot p(X) \tag{1.3}$$

holds, which is equivalent to having

$$p(Y|X) = p(Y) \tag{1.4}$$

and

$$p(X|Y) = p(X). \tag{1.5}$$

Otherwise, they are called *(statistically) dependent* or *associated*. Notably, whenever two variables X and Y are independent, the symmetry of condition (1.3) with respect to X and Y shows that the same holds true with

the roles of X and Y exchanged. For the above question, we can intuitively say that rain and sprinkler are dependent in Europe: The sprinkler is more likely to be activated in the hot months of the year, which in turn show a lower probability of rain compared to the colder months in autumn. In order to answer the question whether the rain variable gives any additional information about the status of the sprinkler, given that we already know the season, we need to introduce the concept of *conditional independence*: Given a variable Z, two variables X and Y are called *conditionally independent given* Z, written as

$$(X \perp\!\!\!\perp Y \mid Z), \tag{1.6}$$

if and only if

$$p(X, Y|Z) = p(Y|Z) \cdot p(X|Z) \tag{1.7}$$

holds, with the same implications concerning symmetry as above. Looking at Figure 1.2, it seems plausible that the rain and the sprinkler are conditionally independent given the season. We will present tools for reading of the full set of conditional independencies from graphical representations in Section 2.2.2. Alternatively, (conditional) independence conditions can be tested by comparing the probability distributions that appear in the definitions. The area of conditional independence testing is still an active field of research due to its importance for causal inference and statistics in general [24, 25, 26], but we will steer clear of this challenge and assume that we have access to an oracle answering conditional independence questions for the remainder of this thesis. Perhaps the most frequent associational query is about the *correlation* between two variables, which is encoded in the *Pearson correlation coefficient*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \tag{1.8}$$

for the *covariance*

$$\mathrm{Cov}(X, Y) = \mathrm{E}[X \cdot Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y] \tag{1.9}$$

and the *standard deviations*

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\mathrm{E}[(X - \mathrm{E}(X))^2]}. \tag{1.10}$$

and σ_Y. The correlation coefficient $\rho_{X,Y}$ can be interpreted as a measure of linear association between the two random variables X and Y. In practical

applications, it can be estimated from a dataset $(x_i, y_i)_{i=1,\dots,n}$ as

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\hat{\sigma}_X \cdot \hat{\sigma}_Y} \tag{1.11}$$

with the usual maximum likelihood estimators

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.12}$$

and $\bar{y}$ for the means as well as

$$\hat{\sigma}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.13}$$

and $\hat{\sigma}_Y$ for the standard deviations. It is easy to see that both $\rho_{X,Y}$ and its empirical counterpart $\hat{\rho}_{X,Y}$ take values in $[-1, 1]$ with the extremes only realized in case of a linear relation between X and Y. The empirical correlation tells us how well the data points fit on a line, i.e. it answers the question: "How strongly do the data suggest that an increase in Y is linearly associated with an increase in X?" It is most important to note that this question is distinct from the causal question about the effect that an increase in X has on Y. The *Reichenbach common cause principle*, formulated by physicist Hans Reichenbach [27], connects the notions of association and causation: Two variables A and B are associated if and only if there is a third variable C that causally affects both A and B. As a special case, C can coincide with either A or B.

1.3.2 Interventional queries

Going beyond association, the first type of genuinely causal queries that we want to state are *interventional queries*. As the name already indicates, these queries ask about the behavior of one or more random variables in response to an active intervention in the data generating process. As opposed to a relation on an associational level, an interventional cause/effect relation is not symmetric: Even if an intervention on a variable X leads to a change in the distribution of another variable Y, the converse does not need to hold. We can change the measurement on a thermometer by manipulating the temperature of the environment, but writing a different number on the display of the

thermometer will have no effect on the temperature. The most basic type of an intervention is given by the *atomic intervention*, which intervenes on a single *treatment variable* by forcing it to take on a single value, and thereby entails a distribution over an *outcome variable* of interest. Our sprinkler example is focussed on estimating the effect of such an atomic intervention: We want to predict what happens to the slipperiness of the lawn Y if we force the sprinkler X to be turned on. Considering the Reichenbach common cause principle, we cannot describe the resulting distribution over Y by estimating $p(Y|X = x)$ from observational data because of possible confounding. Even if the sprinkler had no effect on the slipperiness, there could be a third variable that influences both the sprinkler and the slipperiness such that we would measure a spurious correlation in the data. Therefore, we use a different notation for distinguishing active interventions from passive observations. Variables that are actively intervened upon are accompanied by the word *do*:

$$p(Y|do(X = x)) \tag{1.14}$$

The *covariate-specific intervention* is a slight variation of the atomic intervention. There is again a single intervention variable that is assigned a single value, but this time we have additionally observed the value of other variables, leading to a query of the form

$$p(Y|do(X = x), Z = z) \tag{1.15}$$

where Z is the set of observed covariates and z the set of observed values. Another extension is given by a *stochastic* or *soft intervention*, where we assign the value of the treatment variable not in a deterministic, but in a stochastic manner. The assigned values of X can be seen as realizations of another stochastic function σ, leading to a query of the form

$$p(Y|do(X = \sigma)). \tag{1.16}$$

Instead of varying the number of available observed covariates or the assignment mechanism of the treatment variable, we can as well choose multiple treatment or outcome variables. It is furthermore possible to combine all these query types to form more complex interventions. In order to fill the mysterious do-expressions with life, we can interpret them as ordinary conditionals of an *entailed* or *interventional distribution*: Our intervention on X changes the DGP such that the remaining variables behave according to

a new probability distribution $p_{do(X=x)}$ instead of the original *observational distribution* p. The interventional distribution is a probability distribution in its own right and we can use it to formulate quantities of interest for causal analysis. For ease of presentation, we assume both treatment and outcome to be binary random variables, but extensions to more general settings are well-established in the causal literature [1]. The *average treatment effect (ATE)* τ quantifies the change in the outcome Y if we compare interventional distributions for both assignments of the treatment X:

$$\tau = p_{do(X=1)}(Y = 1) - p_{do(X=0)}(Y = 1) \tag{1.17}$$

If we want to evaluate causal effects in more specific scenarios, e.g. the effect of the sprinkler on the slipperiness specifically during winter, we can compute the more refined *conditional average treatment effect (CATE)* ρ.

$$\rho_{Z=z} = p_{do(X=1)}(Y = 1|Z = z) - p_{do(X=0)}(Y = 1|Z = z) \tag{1.18}$$

In other cases, we are not interested in overall effects of an intervention, but only in specific mechanisms: Does the sprinkler make the lawn more slippery only indirectly by making it more wet, or is there a direct causal effect that is not mediated by the wetness? Such a question can be answered by the *controlled direct effect (CDE)* λ that measures changes to the outcome under a given treatment while keeping all other variables at fixed values.

$$\lambda_{W=w} = p_{do(X=1,W=w)}(Y = 1) - p_{do(X=0,W=w)}(Y = 1) \tag{1.19}$$

The set W is given by all considered variables except for X and Y, and w indicates a special configuration of W. Note how the do-notation is not only employed to convey the active notion of treatment, but also for the fixing of the rest of the variables to $W = w$, which illustrates its usefulness for mediation analysis in a more general context. We are still lacking the tools to determine the interventional distribution or any quantities derived from it, but this gap will be filled in Chapter 2.

1.3.3 Counterfactual queries

Ascending on the ladder of query complexity, we advance to *counterfactual queries*. Whereas interventional queries ask about the expected outcome of a target variable Y, given an intervention $do(X = x)$ on the treatment variable

and possibly observations $Z = z$ of other variables, counterfactual queries add another type of observation. We assume that we have observed what happened for a different choice of intervention $X = \tilde{x}$. Intuitively, humans ask such questions to think about hypothetical outcomes of situations, had they chosen to act differently. Having observed that the lawn became wet after the sprinkler was turned on in summer, what would have been the effect of not turning it on in the same situation? Although these types of queries are of minor importance in the remainder of the thesis, it is helpful to mention them explicitly, in order to establish the boundaries of interventional queries $p(Y|do(X = x))$ not only from one side by associational queries $p(Y|X = x)$, but also from the other side by counterfactual queries. These take on forms such as

$$p(Y_{X=x} = y|X = \tilde{x}, Y = \tilde{y}) \tag{1.20}$$

to describe the probability of observing $Y = y$ after the hypothetical active manipulation $X = x$, given that we have actually observed $Y = \tilde{y}$ and $X = \tilde{x}$.

Contrary to interventional queries, not even experiments can answer such a query because there is no way of having both $X = \tilde{x}$ and $X = x$ occurring for the same subject at the same time. Nevertheless, counterfactuals have real implications for performing mediation analysis. The *natural direct effect (NDE)* μ, a special instance of the CDE in Eq. (1.19), counts only causal contributions that are not mediated by other variables. However, instead of allowing arbitrary values w for the configuration of the remaining variables W, we prescribe $W = w_0$, meaning that the remaining variables take on exactly the value that they would have taken on for an intervention $do(X = 0)$. The *natural indirect effect (NIE)* ν on the other hand excludes direct influences and counts only indirect contributions to the causal effect. Perhaps surprisingly, we cannot just define $\nu = \tau - \mu$ except in the case of linear DGPs. Furthermore, it is possible to use counterfactual language to define a number of intricate causal effects that can serve as metrics for otherwise inaccessible attribution problems. Given that such questions are not the main focus of this thesis, we refer the reader to Pearl's short primer for an intuitive introduction with examples from various application domains [20].

1.3.4 The ladder of causation

Assuming that we can answer interventional queries, it is clear that we can also answer corresponding associational queries by intervening on none of the variables. In the same way, we can rephrase interventional queries as counterfactual queries by not changing the circumstances encountered by the sample under consideration. The converse is not true: Interventional queries can only restated as associational queries if we provide additional ingredients such as a causal graph. Therefore, it is not possible to answer interventional queries using only observational data. A similar border appears between interventional and counterfactual queries. The grouping of queries into observational, interventional and counterfactual queries that we have illustrated in the previous sections can be justified rigorously in the the *ladder of causation* or *Pearl causal hierarchy (PCH)* [28]. For us, the important takeaway is that we are operating on the interventional level, such that our models are necessarily more complex than observational ones, but still coarser than counterfactual ones.

2. ANSWERING INTERVENTIONAL QUERIES

As we have seen in Section 1.3.2, several causal effects of interest can be formulated in terms of do-probabilities of the form

$$p(y|do(x), z) \tag{2.1}$$

where x, y, z are realizations of random variables X, Y and Z. The expression (2.1) denotes the probability of observing an outcome $Y = y$, given that we have also passively observed a covariate $Z = z$ and actively enforced the treatment $X = x$. Consequently, we need to be able to determine these quantities, which leads us to consider different options:

1. Active intervention: Enforce $X = x$ in the physical world or a simulation and observe what happens.

2. Algebraic solution: Reduce all interventional probabilities to ordinary non-interventional probabilities and estimate these from observational data.

Both options are used in practical settings and we want to briefly explain the methodology behind them, as well as their respective strengths and weaknesses.

2.1 Active methods

It is no surprise that the effect of an active intervention in a data generating process can be assessed by actively intervening and observing the result. However, direct interaction with the environment is not always feasible or desirable, such that different strategies have been developed. Even though we will focus on methods using observational - as opposed to interventional - data, it is helpful to briefly introduce active methods as the gold standard that we are trying to mimic with observational methods of causal inference.

2.1.1 Experiments and simulations

The easiest method of answering queries on the higher levels of the Pearl causal hierarchy is to execute reproducible experiments under controlled conditions. If we are interested in the effect of a binary intervention on an outcome variable, we simply carry out two iterations of the experiment: We set the treatment to 0 in the first one and observe the outcome, and then repeat the experiment under the same conditions with the treatment set to 1. If the setup includes a probabilistic component, multiple experiment runs can be executed to gather the necessary statistics. By the reproducible nature of the experiment, we can even answer counterfactual queries simply by having access to multiple "copies of the same world" such that we can answer hypothetical questions. In practical applications, the necessary controlled conditions might be hard to obtain: There is no way of assessing the effect of the sprinkler in summer if it is currently winter. If we want to know whether the sprinkler makes the lawn more slippery even if it is already raining, we cannot test this assumption if the sky is currently unclouded. However, this obstacle can be overcome if the dynamics of the problem are sufficiently well-understood. The dynamics, e.g. known system equations in a physical context, can serve as the basis for a numerical simulation. Such a simulation gives us full control of all tunable parameters and we can again have many attempts of manipulating our environment. Under these conditions, causal inference is not really needed because we can simply answer all our questions by direct interaction with the simulated environment. The focus then typically shifts more towards finding the optimal manipulations for achieving a predefined goal, which is precisely the domain of reinforcement learning (RL) [29]. For example, Wankerl, Luce et al. carried out this strategy and optimized components of light-emitting diodes using RL and a simulation based on the transfer matrix method [30, 31]. In this thesis, we want to examine the methods of causal inference as a valuable alternative for situations where the dynamics are not known such that no simulations can be set up.

2.1.2 Randomized controlled trials

For many causal inference tasks, detailed simulations or reproducible experiments are out of reach: The environment is simply too complex to understand, let alone replicate all the underlying physics. If we want to answer a query that is concerned with some causal effect of a medical treatment on a pa-

tient, we cannot simulate all the biochemical processes in the patient's body. Similarly, it is not possible to put the patient into the same state twice to compare the effect of subsequently administering the treatment or not. In the medical domain, *randomized controlled trials* serve to answer at least interventional questions about the effect of a treatment on the patients' well-being [32]. The procedure is simple: In order to avoid spurious correlations by differing propensity of treatment between subpopulations with different chances of recovery, the doctors randomly decide for each patient whether they should be treated or receive a placebo. Afterwards, the interventional distributions for treated and untreated patients can be compared, in order to answer interventional queries as in Section 1.3.2. For example, the ATE of the treatment on the recovery can be calculated by simply comparing recovery rates in both patient groups. In our sprinkler example, an RCT can be performed by activating the sprinkler on a randomly selected set of days. The main drawback of RCTs lies in their side effects: Whereas the slipping of some people on the lawn might be considered acceptable collateral damage, the deliberate non-treatment of ill patients or the deliberate treatment of patients with a potentially harmful drug poses serious ethical concerns. However, RCTs are still the default method in medical settings due to a lack of alternatives, although there is active research on adaptive versions that optimize the outcome for the study population by changing treatment assignment during the study [33]. The success of RCTs reaches so far that they are also used outside of the medical field for similarly sensitive questions, such as development economics, which has sparked debate due to the unavoidable side effects [34]. In the remainder of this thesis, we will show causal inference methods for avoiding RCTs in the simple sprinkler example, but research on using the same methods for more critical settings has already started [35].

2.2 Causal models

If we want to be able to perform causal inference without intervening in the physical world, we need a model from which we can derive causal effects. Figure 2.1 connects causal modelling to the well-known *probabilistic/statistical modelling*: Given a probabilistic model, such as a joint distribution over all variables of interest, we can predict future outcomes Y from a set of observations X, by evaluating $p(Y|X)$. This process is called *probabilistic*

reasoning. Conversely, if we are given a set of observations X and a set of corresponding outcomes Y, we can try to recover the underlying probabilistic model using methods of *statistical learning*. The obtained model can then again be used to perform probabilistic reasoning, in order to predict outcomes for hypothetical observations that have not been part of the dataset. As we have seen above, predicting future outcomes by probabilistic reasoning is subject to an important constraint. A probabilistic model only describes how samples behave if we passively observe them being drawn from an unchanged data generating process. If we want to understand the behavior of samples that are drawn from a data generating process that is subject to being changed by active interventions, we need to turn to a new class of *causal models*. Given that we can choose to "intervene" by doing nothing, every causal model must already be a probabilistic model in itself. This means that we do not lose any flexibility by choosing causal models over purely probabilistic ones. However, it might still be preferable to use a probabilistic model for strictly probabilistic queries, as the additional flexibility of causal models necessarily comes at a cost. In this section, we introduce different types of causal models and explain how they can be exploited to answer different types of causal and probabilistic queries.

2.2.1 Structural causal models

A *structural causal model (SCM)* codifies the intuition of causal mechanisms as a set of functions that tell us how a set of variables causally depends on each other and possibly random influences. Formally, an SCM S is a quadruple

$$S = (\mathcal{X}, U, F, p_U), \tag{2.2}$$

consisting of

- a set of *endogenous random variables* $\mathcal{X}$,
- a family of *exogenous random variables* or *error terms* $U = (U_X)_{X \in \mathcal{X}}$,
- a family of functions $F = (f_X)_{X \in \mathcal{X}}$ and
- a probability distribution p_U over the exogenous variables.

The family F contains one function

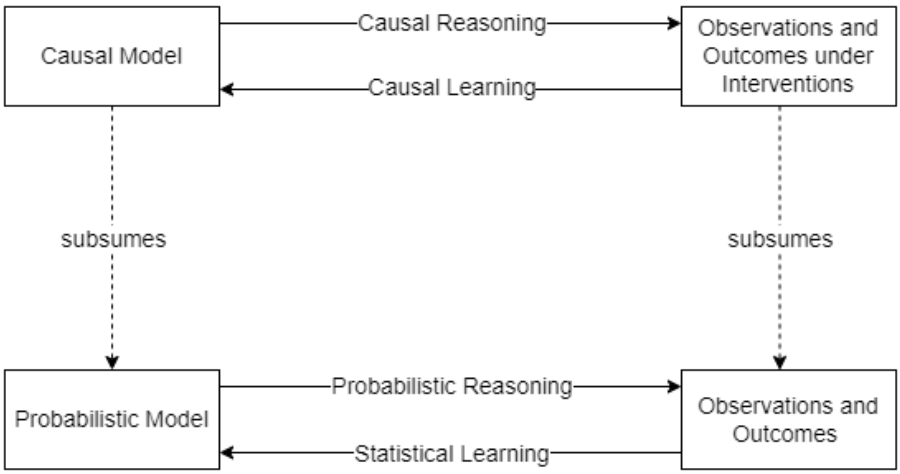

Fig. 2.1: Probabilistic reasoning and statistical learning both connect probabilistic models and observational data. Analogously, causal learning and causal reasoning connect causal models to both observational and interventional data. While observational and interventional data can obviously be seen as a generalization of purely observational data, it is worth noting that causal models can be seen as a generalization of probabilistic models. Redrawn based on a figure in [12].

$$f_X \colon \prod_{Y \in \mathcal{Y}_X} R_Y \times R_{U_X} \to R_X, (y, u) \mapsto f_X(y, u) \tag{2.3}$$

for each of the endogenous variables $X \in \mathcal{X}$, where $\mathcal{Y}_X \subset \mathcal{X} \setminus \{X\}$ is a subset of the endogenous variables and R_Z denotes the support of a random variable Z. This function deterministically computes the value of X, given all the relevant values of other variables. We call the endogenous variables $\mathcal{Y}_X$ that appear as arguments in f_X the *(causal) parents* or *direct causes* of X. In order to incorporate randomness into the model, the exogenous variables are governed by the probability distribution p_U, whereas each endogenous variable deterministically depends on a subset of the endogenous and exogenous variables.

In Pearl's sprinkler example, the corresponding SCM could look like this:

$$S = (\mathcal{X}, U, F, p_U) \tag{2.4}$$

with

- $\mathcal{X} = \{Season, Rain, Sprinkler, Wet, Slippery\}$,
- $U = \{U_X \mid X \in \mathcal{X}\}$ with each U_X following a continuous probability distribution over $[0, 1]$,
- $F = (f_X)_{X \in \mathcal{X}}$,
- $p_U = \prod_{X \in \mathcal{X}} p_{U_X}$.

The most interesting part of the SCM, namely the functions f_X could take on the following form: The season is a random variable that does not depend on any of the other endogenous variables.

$$f_{Season} \colon R_{U_{Season}} \to \{Spring,\ Summer,\ Fall,\ Winter\}, \tag{2.5}$$

$$u \mapsto \begin{cases} Spring & \text{if } u \in [0, 0.2), \\ Summer & \text{if } u \in [0.2, 0.5), \\ Fall & \text{if } u \in [0.5, 0.75), \\ Winter & \text{if } u \in [0.75, 1] \end{cases} \tag{2.6}$$

The rain depends on the season and on its associated exogenous variable:

$$f_{Rain} \colon R_{Season} \times R_{U_{Rain}} \to \{0, 1\}, \tag{2.7}$$

$$(y, u) \mapsto \begin{cases} 1 & \text{if } y = Spring \text{ and } u > 0.5, \text{or} \\ & \text{if } y = Summer \text{ and } u > 0.8, \text{or} \\ & \text{if } y = Fall \text{ and } u > 0.4, \text{or} \\ & \text{if } y = Winter \text{ and } u > 0.7, \\ 0 & \text{else} \end{cases} \tag{2.8}$$

The same holds true for the sprinkler, which is more likely to be activated in a hot summer. This is why we have explicitly included the season as an endogenous variable, even though its value could have been fully encoded in U_{Season} by choosing an appropriate discrete probability distribution. None of the exogenous variables are allowed to appear in multiple functions. The wetness of the lawn depends on both the sprinkler and the rain variable, as well as on another exogenous error term, whereas the slipperiness of the lawn is modelled to depend only on its wetness and another error term. For brevity, we refrain from explicitly spelling out the latter three functional mechanisms.

Note that we are modelling the situation such that the sprinkler does not depend on the rain variable. This means that the sprinkler will be activated regardless of the outside weather, as one could imagine in a scenario where the sprinkler is governed by a microcontroller that has no rain sensor. Further, it may be criticized that the sprinkler activation does not directly depend on the slipperiness of the lawn: If we assume that the sprinkler is manually activated by someone, this person could injure themselves on the way to their duty on the slippery lawn, rendering them unable to activate the sprinkler. These two examples show that an SCM, albeit very concise in its nature of specifying the causal mechanisms, is still subject to possible criticism. The main goal of the precision in the above formulation is not to make the model more correct, but to make the assumptions of the modeler more transparent. Not only the signatures of the causal mechanisms, qualitative in nature, can be scrutinized, but also their functional form or even specific parameter values of the function. Why should it rain more often in spring than in Winter? The parameters in f_{Rain} could be criticized depending on the climate zone.

Keeping the example in mind, the path to calculating interventional distributions of the form $p(Y|do(X = x)$ based on an SCM is layed out clearly:

1. Firstly, we need to understand how to compute the probability distribution $p_S(\mathcal{X})$ that is associated with a given SCM S.

2. In a second step, we can transform the SCM S into a new SCM $S_{do(X=x)}$ that reflects the intervention $do(X = x)$ and calculate the associated probability distribution $p_{S_{do(X=x)}}(\mathcal{X})$.

3. Finally, we can answer the interventional query via the equation
$$p(Y|do(X = x)) = p_{S_{do(X=x)}}(Y) \tag{2.9}$$
based on the observation that $p_{S_{do(X=x)}}(\mathcal{X})$ is precisely the interventional distribution associated with the manipulation $do(X = x)$.

The first point can be achieved simply by propagating the uncertainty over the exogenous variables through the structural mechanisms F. We start by topologically sorting $\mathcal{X}$ with respect to F, i.e. we sort the endogenous variables as $(X_1, ..., X_n)$ such that no X_i appears in f_{X_j} if we have $i > j$ [36]. As a next step, we write down the associated *trivial factorization*

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(X_i = x_i|X_1 = x_1, ...X_{i-1} = x_{i-1}) \tag{2.10}$$

where we omit the index S for convenience. Each factor on the right hand side is called a *conditional probability distribution (CPD) over* X_i and it can be separately calculated using F:

$$p(X_i = x_i | X_1 = x_1, ..., X_{i-1} = x_{i-1}) = \int_{R_{U_i}} \mathbb{1}_{g^{-1}_{x_1,...,x_{i-1}}(x_i)}(u) dp_U \tag{2.11}$$

In this expression that looks considerably more challenging than it actually is, we use the notation $\mathbb{1}$ for the indicator function of a set, namely the preimage of x_i under the map

$$g_{x_1,...,x_{i-1}} : R_{U_i} \to R_{X_i} \tag{2.12}$$

that inputs the necessary values into the function f_{X_i} to determine x_i and discards the rest of the x_j. Broadly speaking, for the configuration $(X_1 = x_1, ..., X_{i-1} = x_{i-1})$, we check which values of U_i are mapped to $X_i = x_i$ via f_{X_i} and weight their contributions according to the distribution p_U, in order to arrive at the probability of interest. As a byproduct of the explicit calculation, we notice that we do not need all of the preceding endogenous variables as input for the CPD over X_i, but only the causal parents Π_i that appear in the structural mechanism f_{X_i}. Therefore, we can replace the trivial factorization in Equation (2.10) with the more compact *causal Bayesian factorization associated with* S

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(X_i = x_i | \Pi_i = \pi_i) \tag{2.13}$$

which requires fewer parameters and is independent of the particular ordering of the variables.

Since we are interested in the probability distribution that occurs after the active intervention $do(X = x)$, we need to use a slightly altered version of S as the basis for the above calculations. Luckily, the SCM formalism naturally lends itself to incorporating atomic interventions: We arrive at the desired SCM $S_{do(X=x)}$ by replacing f_X with the constant function that has x as its only value, whereas the rest of the SCM stays the same. By calculating the probability distribution associated with $S_{do(X=x)}$ and using Equation (2.9), we can finally answer interventional queries. Note that the process of modifying the original SCM and drawing samples from the entailed distribution can also be used to construct a simulated environment as described in Section 2.1.1. Furthermore, we can avoid recalculating the interventional distribution for

each new intervention by directly modifying the causal Bayesian factorization associated with the original SCM S according to

$$p(x_1, ..., x_k, ..., x_n | do(X_k = v)) = \begin{cases} \prod_{i \neq k} p_S(X_i = x_i | \Pi_i = \pi_i) & \text{if } x_k = v, \\ 0 & \text{else} \end{cases} \tag{2.14}$$

which is called the *truncated product rule*. In summary, SCMs are a powerful tool to compute interventional probabilities, but correctly specifying them is obviously not an easy task. However, a closer look at the above calculations reveals that we never explicitly use the structural mechanisms F, but only their signatures: In order to evaluate the truncated product rule, we only need to know which variables appear as inputs for each of the functions in F. If we can provide the associated CPDs $p_S(X_i = x_i | \Pi_i = \pi_i)$, we are able to answer the interventional queries of interest. Therefore, specifying the full SCM might not even be necessary for our purposes, as long as we know that it exists in the background.

2.2.2 Causal Bayesian networks and causal graphs

Inspired by the preceding observations, the second class of causal models that we want to consider are *causal Bayesian networks (CBNs)*. In general, a *Bayesian network (BN)* $B = (G, C)$ is characterized by

- a *directed acyclic graph (DAG)* $G = (\mathcal{X}, E)$ with vertices X and edges $E \subset \mathcal{X} \times \mathcal{X}$ and
- a set of CPDs $C = (p(X = x | \Pi_X = \pi_X))_{X \in \mathcal{X}}$

where $\Pi_X = \{Y \in \mathcal{X} \mid (Y, X) \in E\}$ denotes the *parents* of X in G.

Starting from an SCM S, there is an *associated CBN* $B(S)$ that can be seen as its logical graphical representation: The vertex set of the graph are all endogenous variables in S, and the edges are given by

$$E = \{(X, Y) \in \mathcal{X} \times \mathcal{X} \mid f_Y \textit{ explicitly depends on } X\}. \tag{2.15}$$

Put simply, we look at each of the functions given by S and draw an arrow from each endogenous input variable to the output variable. Note that for each variable, the causal parents in S are precisely the parents in $B(S)$. The

CPDs of $B(S)$ are obtained from S by propagating the uncertainty of the exogenous variables through the functions F, as we have already seen in Equation (2.11).

In summary, we have created a coarser description of the SCM S: The endogenous variables directly correspond to the nodes $\mathcal{X}$ in G and the functional signatures in F are preserved as the edges E, but the exact form of the functions is no longer accessible. For each $X \in \mathcal{X}$, the CPD $p(X|\Pi_X)$ combines the probabilistic information in p_{U_X} and the deterministic assignment f_X into another probabilistic object. It is clear that the map $(p_{U_X}, f_X) \mapsto p(X|\Pi_X)$ is not injective, meaning that we lose information. As discussed at the end of the previous Section, this loss of information is irrelevant for the computation of interventional probabilities, since we still have enough knowledge to evaluate the truncated product rule in Equation (2.14). Furthermore, ordinary statistical methods can be used to learn the CPDs from data and the CBN can again be used as a simulator for interventional data. The limited ability of the CBN to model certain aspects of causality, as opposed to the fully specified SCM, would only matter if our queries were of counterfactual nature (cf. Section 1.3.3) because the exogenous variables and functional mechanisms need to be separately available for these purposes [1]. The resulting CBN for the sprinkler example can be seen in Figure 2.2.

Going one step further in the quest of reducing model complexity, we can even strip the CBN of its CPDs and consider only the graph G, which we call the *causal graph*. Just as the CBN, the causal graph justifies its existence as a causal model by being a derived entity from the fully specified SCM. Therefore, we can talk about *causal parents* and *direct causes* based on the graphical representation, without knowing p_U or the exact functional forms of F beyond the signatures.

2.2.3 Noncausal Bayesian networks

While the above view on CBNs and causal graphs sees them only as a weaker representation of the SCM, it is clear that directed graphs are an interesting mathematical object in their own right. We could draw many different directed graphs over the endogenous variables and talk about parents of nodes with respect to each of these other graphs, but we would lose any justification for calling them *causal parents* or *direct causes*.

The same holds true for CBNs: (Possibly non-causal) Bayesian networks are again interesting mathematical objects in their own right, and they can

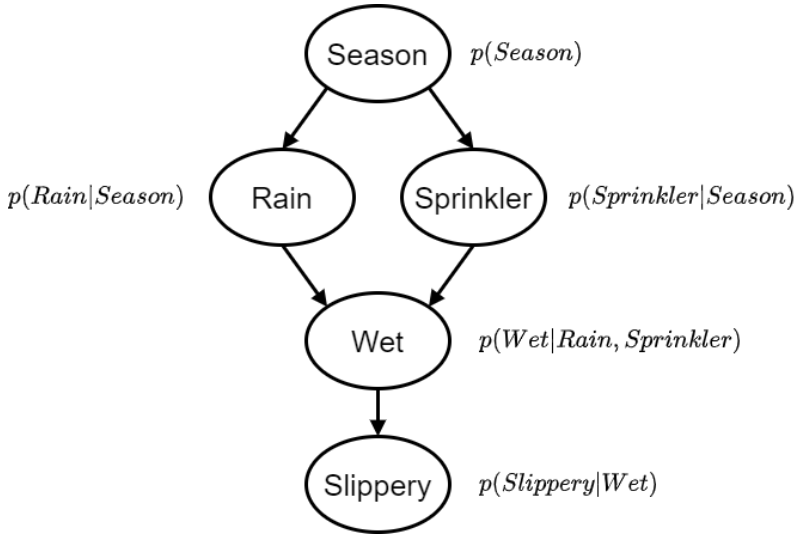

Fig. 2.2: Bayesian networks incorporate both graphical and probabilistic information. The displayed BN is the causal one since its DAG and CPDs are derived from the SCM that fully describes the underlying causal mechanisms of the DGP.

be used for efficiently modelling the joint distribution over the endogenous random variables. Non-causal BNs are not useful for answering genuinely causal queries, but it is helpful to understand their strengths when compared to non-graphical probabilistic models such as a joint distribution. A graphical criterion called *d-separation* has been developed as an analogue of the probabilistic notion of independence between random variables [37]. A given (possibly noncausal) DAG G over the endogenous variables is eligible as basis for a BN whenever each graphical d-separation between the variables implies the corresponding independence statement in the joint distribution. This criterion can be written as

$$(X \perp_d Y \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z) \tag{2.16}$$

where the left-hand side denotes that the nodes X and Y are d-separated in the DAG given another set of nodes Z. Since a definition or even a deeper discussion of the d-separation formalism is unfortunately out of scope for this work, we refer the interested reader to the seminal work by Koller and Friedman [37]. For us, it suffices to know that reading off d-separation statements

from the graph is a trivial task and Condition (2.16), which is called the *global Markov condition*, allows us to easily read off independence statements from a DAG. The global Markov condition is equivalent to the *local Markov condition*, which states that each endogenous variable needs to be independent of its nondescendants given its parents in G. The importance of this condition becomes apparent in light of our above derivation of the causal Bayesian factorization (2.13) from the trivial factorization (2.10). This derivation was enabled precisely by the local Markov condition that is justified for the causal DAG by the signatures of the structural mechanisms in F. Consequently, whenever another DAG satisfies the local Markov condition, we can use it as the basis for an associated Bayesian factorization. Although such a noncausal Bayesian factorization cannot be used to answer interventional queries, the decomposition of the joint distribution gives us a compact description of the DGP on an associational level. Therefore, research on Bayesian networks has already been conducted before their capabilities in causal settings had been discovered [38, 39]. Two DAGs in which the same set of d-separation statements hold are called *Markov equivalent* and this equivalence relation separates the set of DAGs over the endogenous variables into disjoint *Markov equivalence classes*. We know that a DAG whose d-separations exactly correspond to the conditional independencies of a joint distribution lies in the same Markov equivalence class as the causal graph that corresponds to the CBN. However, we cannot answer interventional queries using its associated truncated factorization if the BN is not the causal one, as the justification for the truncated product rule in Section 2.2.1 depended on the underlying SCMs, which makes it invalid for noncausal BNs. Nevertheless, the d-separation formalism evolved into a causal calculus that lets us treat the raw causal graph as a non-generative causal model, as we will see in the next section.

2.2.4 Pearl's do-calculus

An alternative to the potentially harmful RCTs and the complex creation of a surrogate model is given by Pearl's do-calculus [40, 41]. The central idea is to restate the original interventional query in terms of do-free observational expressions. Once this is achieved, the well-established tools of statistics can be employed to estimate the resulting expression, which is called an *estimand*. As we have seen in Section 1.1, it is not always possible to simply replace $do(x)$ by x in the conditional probability. The causal structure of the problem, encoded in the causal graph G and its d-separations, determines

which substitutions are allowed for expressions involving groups of variables X, Y, Z, W and their respective values x, y, z, w:

1. Observation insertion/deletion

$$p(y|do(x), z, w) = p(y|do(x), w) \quad \text{if } (Y \perp_d Z \mid X, W)_{G^X} \tag{2.17}$$

2. Action/observation exchange

$$p(y|do(x), do(z), w) = p(y|do(x), z, w) \quad \text{if } (Y \perp_d Z \mid X, W)_{G^X_Z} \tag{2.18}$$

3. Action insertion/deletion

$$p(y|do(x), do(z), w) = p(y|do(x), w) \quad \text{if } (Y \perp_d Z \mid X, W)_{G^{X,Z(W)}} \tag{2.19}$$

The graph G^X denotes the graph that is obtained by deleting all edges from G that point into variables in X. Analogously, G_X denotes the graph that is obtained by deleting all edges from G that originate in variables in X. Both notations can be combined. The set $Z(W)$ consists of all nodes in Z that are not ancestors of any node in W based on G^X.

From a high-level perspective, the do-calculus is a machinery that automatically converts interventional expressions into do-free estimands. It is both *sound* and *complete*, in the sense that this conversion can be obtained by applying the three rules 2.17 to 2.19 sequentially [42, 43], whenever the effect is *identifiable*, i.e. expressible in terms of the observed distribution. There are *non-identifiable* cases where the existence and position of an unmeasured confounder in the causal graph are known, but there is no way of expressing the causal effect in terms of only the observed variables. Note that the functional form of the DGP does not matter except for the part encoded in the causal graph such that a table of identifiable effects together with the suitable estimand can be compiled for common graph structures. The do-calculus even allows us to infer causal effects in many scenarios with unmeasured variables, as long as we know the location of these variables in the causal graph and they do not appear in the estimand provided by the algebraic machinery. Furthermore, uncertainty about specific edges in

the causal graph can be ignored if all possible versions of the graph lead to the same estimand. However, since the studies presented in this thesis were concerned with situations where all variables of interest have been observed, we will sidestep discussions of identifiability and point to [1, 44, 42] for further reading.

In the sprinkler example, we need to estimate the expression

$$p(Slippery|do(Sprinkler)) \tag{2.20}$$

in order to calculate the average treatment effect of interest. This can be achieved by applying the law of total probability to introduce the season as a condition, before using an action/observation exchange on $Z = Sprinkler$ in the first factor and an action deletion on $Z = Sprinkler$ in the second factor of each summand:

$$\begin{aligned} & p(Slippery|do(Sprinkler)) \\ &= \sum_s p(Slippery|do(Sprinkler), Season = s) \cdot p(Season = s|do(Sprinkler)) \\ &= \sum_s p(Slippery|Sprinkler, Season = s) \cdot p(Season = s) \end{aligned} \tag{2.21}$$

To justify the resulting expression, we can observe that the season acts as a confounder in the DGP: The sprinkler is more likely to be turned on in some seasons, and similarly the lawn is more likely to be slippery in some seasons due to the rain variable. By keeping the season fixed in each summand, we eliminate spurious correlations. This can be seen as an instance of the *backdoor criterion* or *adjustment rule* that can be derived from the do-calculus. It indicates that the interventional query $p(y|do(x))$ can be answered by a reweighted sum of conditional probabilities including the causal parents $\Pi(X)$ of the treatment variable, rather than just the unweighted conditional probability $p(y|x)$. There is some discussion whether the heavy do-calculus machinery is necessary to arrive at such transformations [45]. However, this approach has the advantage that more intricate criteria such as the *frontdoor criterion* for partially unobserved scenarios can be derived by the same calculus when human intuition reaches its limits [1].

Coming back to the sprinkler example, we can rewrite the ATE of the sprinkler on the slipperiness on the lawn as

$$\begin{aligned} & p(Slippery = 1|do(Sprinkler = 1)) - p(Slippery = 1|do(Sprinkler = 0)) \\ = & \sum_s p(Slippery = 1|Sprinkler = 1, Season = s) \cdot p(Season = s) \\ - & \sum_s p(Slippery = 1|Sprinkler = 0, Season = s) \cdot p(Season = s) \end{aligned} \tag{2.22}$$

such that all involved quantities are computable from the observed distribution. If we further assume a *homogeneous effect*, i.e. if there is a $c \in \mathbb{R}$ such that for all seasons s, we have

$$p_{1,s} - p_{0,s} = c \tag{2.23}$$

using the notation $p_{i,s} = p(Slippery = 1|Sprinkler = i, Season = s)$, the above computation of the ATE simplifies and yields precisely c. A special case for the homogeneous situation is given by a linear SCM of the form

$$\begin{aligned} p(Slippery = 1) &= Wet \\ p(Wet = 1) &= c \cdot Sprinkler + d \cdot Rain \end{aligned} \tag{2.24}$$

where the exogenous variables (omitted for ease of notation) act by flipping the outcome of each variable with a small probability.

In summary, we observe that we can estimate ATEs in fully observed scenarios from the observational data. In linear SCMs, we can employ linear regressions by applying the backdoor criterion, regressing the outcome on the treatment and the parents of the treatment, and finally reading of the coefficient associated with the treatment. Under the same strict assumptions, similar procedures serve to determine NDEs and NIEs as regression coefficients. They are presented in detail in [20], accompanied by a discussion of the additional obstacles in the general nonlinear case.

3. CAUSAL DISCOVERY: LEARNING CAUSAL GRAPHS FROM DATA AND DOMAIN KNOWLEDGE

As anyone who tries to draw the causal graph even for a moderately complex problem will immediately notice, it is not an easy task to decide which edges should or should not be in the graph. This is alarming, since the validity of graph-based causal inference is first and foremost based on the validity of the graphical assumptions. However, we are rarely in the situation of having to draw a graph out of thin air, but in many cases we will have access to observational data that stems from the underlying SCM. Given that this SCM links the causal graph and the generated data, there are several methods of exploiting this connection to learn the causal graph from data. On the other hand, it is clear that data alone cannot be sufficient to infer the causal graph, as this would again imply that causal inference is possible from data alone. Therefore, we will also take the time to consider how domain knowledge can be used to augment the presented learning algorithms.

3.1 Different algorithmic approaches

The main classes of causal discovery algorithms are given by constraint-based and score-based algorithms. Whereas the constraint-based algorithms exploit the correspondence of graphical and statistical independence, the score-based algorithms view the graph as part of a generative model and reduce the problem to the well-known maximum likelihood formalism from statistical learning [46]. There are other interesting approaches to causal discovery, such as the LiNGAM method that exploits independence constraints on the error terms, but we will focus on the two main classes. The goal is not a textbook treatment of all available algorithms, but only reaching the level of understanding that is necessary for the remainder of this thesis. For the same reason, we will not discuss concrete implementations of the ideas, as these can be used interchangeably in all of the subsequently discussed applications.

Comprehensive surveys on the subject offer further detail for the interested reader [47, 48, 49].

3.1.1 Constraint-based algorithms

As we have seen in Section 2.2.2, the causal graph fulfills Markov properties, which link graphical d-separation criteria to probabilistic conditional independence statements. Constraint-based discovery algorithms exploit this knowledge in order to check candidate graphs with respect to their ability to represent the statistical independencies that the data suggest. For each d-separation $(X \perp_d Y \mid Z)$ that we can read from the graph, we inspect the data and verify if the corresponding conditional independence statement $(X \perp\!\!\!\perp Y \mid Z)$ holds. If this not not the case, we know that the data cannot be generated from an SCM whose structure corresponds to the graph in question. In principle, we can now list all possible candidate graphs and apply this procedure to filter out one ill-suited graph after the other. Such a naive algorithm, however, has two important technical weaknesses. First, it is in many cases not practically possible to enumerate all DAGs with the required number of nodes, as their number grows superexponentially in the number of nodes: For $n = 0, 1, 2, 3, ...$ nodes, the number of possible DAGs with n vertices is given by the rapidly growing sequence $1, 1, 3, 25, 543, 29281, 3781503, ...$ [50]. This obstacle can be overcome by reversing the above procedure. Instead of blindly trying out all possible DAGs, it is much more efficient to look at the independencies in the data first and reverse-engineer a suitable graph from this information. The most prominent example of this strategy is the famous PC-algorithm, which is named after its creators Peter Spirtes and Clark Glymour [51]. Second, in the above formulation, we have assumed that we can flawlessly read off conditional independencies from observational data. This is an unrealistic assumption, as the data is likely noisy or simply to small to allow such inferences, which poses a considerably challenge to current conditional independence tests [25]. Apart from these technical difficulties, there is also a severe conceptual shortcoming of constraint-based causal discovery algorithms: We are filtering out all graphs whose graphical independence structure does not match the statistical independencies in the data, but are we left with only one, namely the causal, graph in the end? As we have already discussed in 2.2.2, all other graphs that are Markov-equivalent to the causal graph will generate the same independence constraints, meaning that constraint-based causal discovery can only find the correct Markov-equivalence

class of graphs instead of the correct graph. Given that Markov classes can consist of many different graphs and only the causal representative can be used in downstream causal inference tasks, this is a major problem. However, we will shortly see how domain knowledge can be used to mitigate the issue.

3.1.2 Score-based algorithms

Score-based algorithms reduce the discovery problem to a maximum likelihood problem by building a generative model based on the candidate graph and checking how likely such a model is to have generated the observed data. Given the candidate graph, we have to prescribe a parameterized form

$$p(X_i|\Pi_i, \theta_i) \tag{3.1}$$

of the associated CPDs for each of the nodes, such that we can use the resulting Bayesian network as generative model. We can then build up the joint distribution over all nodes by multiplying the CPDs and calculate the likelihood of the data as

$$L(G) = p(D|G, \theta_G) \tag{3.2}$$

with

$$\theta_G = \underset{\theta}{\operatorname{argmax}}\, p(D|G, \theta) \tag{3.3}$$

and

$$\theta = (\theta_i)_{i=1,\dots,n} \tag{3.4}$$

Given a candidate graph G, finding the associated optimal parameterization θ_G is a standard optimization problem and the same algorithms as for any other optimization problem can be used. Provided that this optimization can be solved and that the computation of the likelihood has a tractable form, we are again at the same dilemma as in the constraint-based case. The number of DAGs is too high to compute the respective likelihood for each of them. As in the PC algorithm, the problem can be mitigated by building up suitable candidate graphs in a strategic way. On the other hand, this time we have an explicit figure of merit, namely the likelihood function L that we want to maximize, so we can see this as an instance of an optimization problem over a discrete search space. There are different ways of dealing with the discreteness that prevents us from using generic optimization techniques, such as the well-known gradient descent. Using a simple greedy approach makes

the discrete search a tractable problem by transforming it into a sequence of problems with a restricted search space: We start out with a graph that contains no edges, and compute the likelihood for all graphs that we can reach by adding a single edge. We greedily add the edge with the highest score and repeat the procedure, until no further improvement is possible. Subsequently, a pruning phase can be employed to remove superfluous edges and thereby reduce the complexity of the resulting DAG [52].

A second approach transforms the discrete search space into a continuous one by a slight reformulation, using the concept of an *adjacency matrix*: For a graph $G = (X, E)$ with vertices X and edges E, the associated adjacency matrix is given by $A = (a_i j)_{i,j=1,..n}$ with

$$a_i j = \begin{cases} 1 & \text{if } (X_i, X_j) \in E \\ 0 & \text{else} \end{cases} \tag{3.5}$$

If we allow arbitrary values in $[0, 1]$ instead of only binary ones, and make the generative model depend smoothly on these values, we can perform gradient based optimization on A. By thresholding, the result can be retransformed into a binary matrix, which can then be interpreted as a graph again. In order to restrict the search space to contain only acyclic graphs, a penalty term that enforces acyclicity in terms of the adjacency matrix must be introduced, before Lagrange optimization is possible [53, 54]. As in the constraint-based case, several problems remain. First, the score-based approach requires the specification of additional information (e.g. the functional form of CPDs) to turn the graph into a generative model that can be used to calculate a likelihood. If this information is misspecified, the quality of the causal discovery procedure will deteriorate accordingly. Second, the purely likelihood-based approach leads to overfitting: An overparameterized model with many unnecessary edges can always perform at least as well as a model that contains only a subset of the edges, since unnecessary parameters can simply be set to zero. It is therefore essential to add a penalty term that punishes model complexity [55]. Third, most scores are borrowed from structure learning for general, possibly non-causal, Bayesian networks, where it is seen as advantageous to have equal scores for graphs in the same Markov-equivalence class, as they can represent the same probability distributions. As discussed above, this is a severe disadvantage if we are interested only in the causal graph as a special representative of the optimal equivalence class.

3.2 Domain knowledge in causal discovery

As we have seen above, there are several obstacles in purely data-based causal discovery, such as the enormous search space and the indistinguishability of Markov-equivalent graphs. Luckily, we are often in a position to support the algorithms with domain knowledge. In most applications, it is not hard for a human to specify some edges which are obviously required if the graph is to be the causal one, as well as some other edges which can never occur in the causal graph. Often, this knowledge stems from a temporal order: In industrial manufacturing, the results of the first processing step could influence variables in the second processing step, whereas edges in the antitemporal direction can be excluded. Another source of knowledge are basic physical facts: The temperature will determine the value that is displayed on a thermometer, the reverse direction is out of question. A third category of domain knowledge is expert knowledge: A production expert often knows a set of causal implications that hold true in the manufacturing process, e.g. the logistics of how intermediate products are distributed across the different machines for the next process step.

Given some edges that must or must not be present in the causal graph, we need a way to communicate this information to the aforementioned causal discovery algorithms. Unfortunately, most of the algorithms were not designed with this requirement in mind, so there is no unified interface for passing domain knowledge to an arbitrary algorithm. However, a closer manual inspection often leads to at least one feasible option per algorithm: In the case of an algorithm that relies on naive enumeration of all DAGs, in order to subsequently check independency constraints or evaluate a likelihood score, we can just leave out all of the candidate graphs that do not meet our requirements, which saves us from unnecessary computational efforts. In the PC algorithm, we can initialize the graph scaffold with all our required edges and refrain from adding any forbidden edges. The score-based algorithms even allow multiple possibilities: For a greedy algorithm that adds one edge at a time, we can again start by adding all required edges and refrain from adding any forbidden edges afterwards. Another possibility is to add explicit penalty terms to the score, which is a softer way of preventing the algorithm from selecting unwanted edges. Conversely, there can be reward terms for selecting required edges. These soft constraints have the advantage of being fault tolerant with respect to misspecified domain knowledge: Even an expert can make mistakes in enumerating many required or forbidden edges, and

in such a case, the optimization can still overrule this knowledge if the data demands it, given that the hyperparameters are chosen suitably. If the expert wants to explicitly specify a degree of certainty about each domain knowledge item, we can transform the above likelihood optimization into a *maximum a posteriori (MAP)* optimization by introducing a prior $p(G)$ over all graphs:

$$\tilde{L}(G) = p(G|D) \propto p(D|G) \cdot p(G) = L(G) \cdot p(G) \tag{3.6}$$

Such a prior can even be further decomposed into a prior over certain edge probabilities. On the other hand, if we set $p(G) = 1$ for all graphs that are compatible with the domain knowledge and $p(G) = 0$ for the rest, we recover the above method of simply excluding all non-compatible graphs from the search space.

It is worth noting that the relation between domain knowledge specification and causal discovery is not unidirectional: Running causal discovery and inspecting the result can in turn serve as a basis for discussing specific edges with domain experts. For instance, the initial discovery result could show edges that the expert considers unrealistic, and these can subsequently be added to the list of forbidden edges. Afterwards, another run of the discovery algorithm with the improved domain knowledge produces a second proposed causal graph. Iterating this process ideally leads to an increase in both the quality of the resulting graphs and the amount of codified domain knowledge.

4. CAUSAL END-TO-END ANALYSIS

After the above discussion of causal inference techniques, it is time to distill a holistic strategy for conducting causal analyses. The main challenge that we want to address is answering interventional queries based on available domain knowledge and observational data. For this purpose, the *(causal) end-to-end analysis* workflow was developed and implemented by the author during his PhD studies. An open-source implementation as a Python package is available on GitHub [4] and maintained using continuous integration techniques to ensure a working state for other researchers and practitioners.

4.1 Components of a holistic causal analysis

In order to select suitable components for the end-to-end analysis from the available modelling strategies discussed in the previous sections, we work our way back from the goal. If we want to answer interventional queries, we should only consider models on the second layer of the Pearl Causal Hierarchy. More powerful models, such as a full SCM, are more complex to build and we do not need their additional third layer capabilities to answer our queries of interest. This observation already suggests that we should focus on graphical models, which leads to a necessary decision between two alternatives:

1. A causal graph, together with Pearl's do-calculus and an estimation technique, can answer all identifiable interventional queries: Our query is stated in terms of do-expressions, which the do-calculus can often convert into do-free probabilities. The problem is then reduced to estimating the latter probabilities from the observational data, which is the central task of statistics and therefore a well-studied area.

2. A Causal Bayesian Network offers even stronger functionality, considering that we can directly use it as a generative model to sample from the interventional distribution of interest. The price we pay for this

functionality is the need to know the functional form of each involved CPD as well as the distribution over the exogenous variables.

Since the first approach is sufficient to answer our queries and does not require the explicit modelling of the exogenous variables, we select the bare causal graph as our model of choice. As for the estimation strategy, we can use linear regression as a well-established, hyperparameter-free default setting and switch it out for more sophisticated techniques if the DGP is suspected to be nonlinear. With these pillars in place, our strategy for answering causal queries based on a causal graph and observational data is mapped out. The central remaining obstacle is the distillation of the causal graph from domain knowledge and observational data, but we have already introduced causal discovery as a solution for exactly this problem in Section 3. In order to keep track of the domain knowledge, which turned out to be an obstacle in discussions with domain experts, the author developed the concept of the *knowledge graph*. It features the same nodes as the causal graph, but visually indicates all edges whose presence is required or forbidden in the causal discovery result (cf. Figure 4.2). As mentioned in Section 3, many discovery algorithms only provide the Markov-equivalence class of the causal graph. Therefore, edges under whose orientation the equivalence class is invariant are left unoriented in the causal discovery result. In order to advance to a fully oriented DAG, which is in general necessary to proceed to the identification step, we suggest that the orientation of these edges is manually decided in cooperation with a domain expert. An alternative route for dealing with partially oriented graphs would be a sensitivity analysis, i.e. the remainder of the end-to-end analysis is carried out for all possible orientation choices (cf. Section 5.1.1). However, the effort required for this strategy grows exponentially with the number of unoriented edges, such that it is not always feasible in practice. The three main pillars of causal discovery, estimand identification and estimation are supplemented by data preprocessing and reporting steps.

4.2 Step-by-step description

In summary, by *causal end-to-end analysis*, we mean the following procedure for a given dataset and given causal effects of interest.

1. We preprocess the data by deleting, adding, rescaling or combining variables.

2. We pass domain knowledge by specifying which edges must or must not be part of the causal graph.
3. We run a causal discovery algorithm that respects the domain knowledge.
4. We postprocess the proposed causal graph by deleting, adding, reversing or orienting a subset of edges in the causal discovery result.
5. We identify an unbiased statistical estimand for each effect of interest by applying the do-calculus to the causal graph.
6. We estimate the estimands by a method of our choice.
7. We report the results of the analysis.

Figure 4.1 illustrates how the steps of the analysis depend on various inputs. We can ignore the depicted validation step and the distinction between qualitative and quantitative domain knowledge for now, as these aspects will be developed and discussed in great detail in Part III of this thesis.

4.3 Software contributions

In order to perform the above steps of a causal end-to-end analysis, the author developed the open-source cause2e Python package [4]. Internally, the package relies on pycausal [56], which is a Python wrapper around the popular TETRAD Java application [57], to perform the causal discovery step, whereas identification and estimation are delegated to DoWhy [58]. The cause2e package takes care of the interface between pycausal and DoWhy, including functionality for postprocessing partially oriented causal discovery results. Furthermore, methods for reading and preprocessing data from different formats are provided. Domain knowledge can be efficiently passed, managed and visualized using custom classes and the knowledge graph. Methods for validating causal models, including quantitative probing and domain-agnostic refutation checks (cf. Part III for a discussion of these methods) are provided to ensure reliable results. Each analysis is summarized automatically in a detailed pdf report that can be used to communicate application-specific assumptions and the results of the analysis. Tutorial notebooks can be found on the cause2e GitHub page [4]. It has to be mentioned that the DoWhy package has recently sparked the creation of the PyWhy organization with

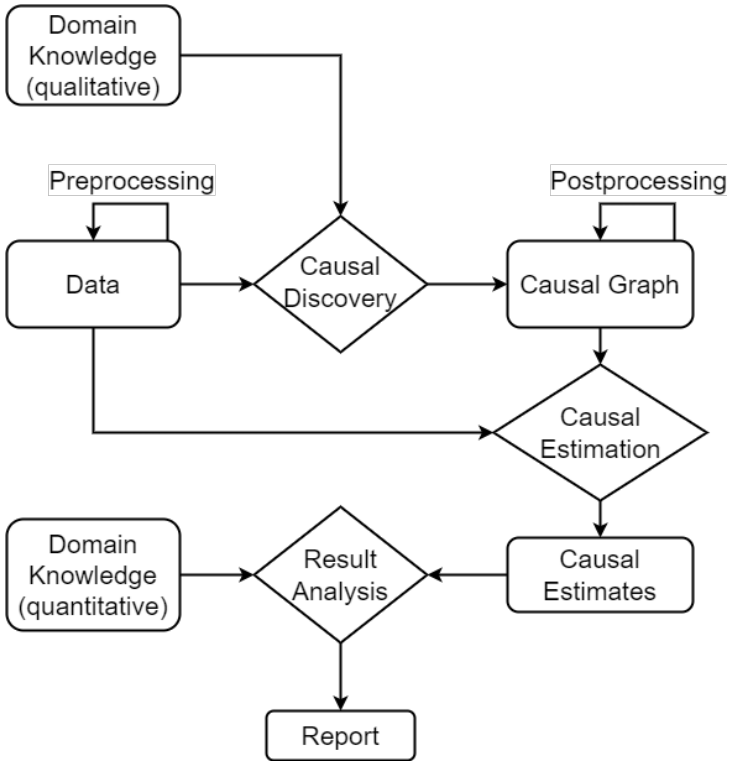

Fig. 4.1: The causal end-to-end analysis includes data preprocessing, domain knowledge management, causal discovery, identification, estimation and validation, before an automated report is generated.

the goal of creating an open-source ecosystem for causal machine learning [59]. PyWhy's commitment is to "build and host interoperable libraries, tools, and other resources spanning a variety of causal tasks and applications, connected through a common API on foundational causal operations and a focus on the end-to-end analysis process". With a growing open source community behind it (instead of a single PhD student) that extends the core DoWhy functionality [60], PyWhy will hopefully make cause2e obsolete soon and provide a durable and continuously supported implementation of the above causal end-to-end analysis.

4.4 Sprinkler example

In order to illustrate the concept of the causal end-to-end analysis depicted in Figure 4.1, we apply it to our sprinkler example and finally solve the mystery whether sprinkler activation leads to a slippery lawn.

4.4.1 Setup and preprocessing

As a ground truth, we prescribe a data generating process following Figure 1.2 and generate 10000 samples from it that are provided as observational data for our exemplary analysis. Preprocessing is not necessary, since all the variables except for the season are chosen to be binary with zero corresponding to False and one corresponding to True. A use case requiring real preprocessing is described in Section 8.2.

4.4.2 Causal discovery

In addition to the observational data, we provide some domain knowledge items for recovering the causal graph:

1. We forbid all edges that originate from "Slippery".
2. We forbid all edges that go into "Season".
3. We forbid the edges "Sprinkler" $\rightarrow$ "Rain" and "Season" $\rightarrow$ "Wet".
4. We require the edges "Sprinkler" $\rightarrow$ "Wet" and "Rain" $\rightarrow$ "Wet"

In this scenario, we could naturally prescribe most of the graph from domain knowledge, but we leave multiple edges unspecified for illustrative purposes. The resulting knowledge graph in Figure 4.2 (left) gives a comprehensive graphical overview of the communicated domain knowledge. Subsequently, the causal graph is recovered by running fast greedy equivalence search, a score-based causal discovery algorithm [61], and the result in Figure 4.2 (right) indeed corresponds to the known DGP. Postprocessing of the graph is not necessary, since our domain knowledge was sufficient to both find the correct Markov equivalence class and exclude all non-causal representatives, such that no edge has remained unoriented.

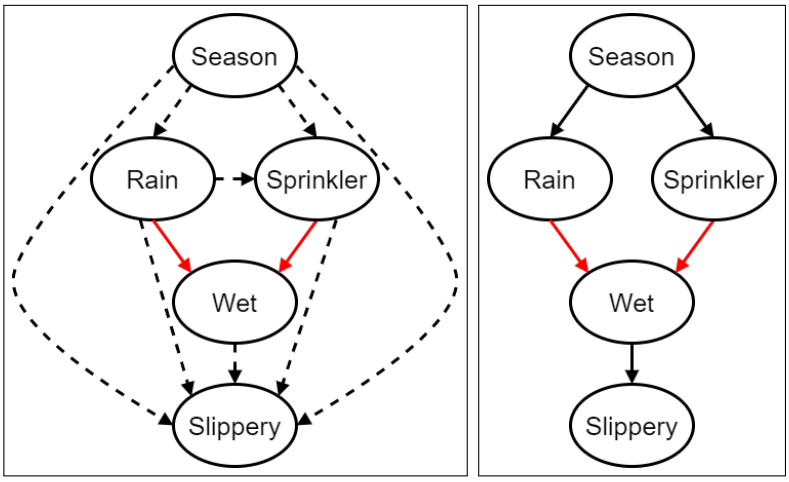

Fig. 4.2: Causal discovery using observational data and domain knowledge: The knowledge graph (left) indicates edges that are required by domain knowledge in red. Forbidden edges are omitted from the graph whereas the remaining possible edges are drawn dotted. The causal discovery result (right) contains all required edges (red), but only a subset of the possible edges (black) has been selected by the discovery algorithm based on the observational data.

4.4.3 Effect identification and estimation

The graph is then used to identify a suitable linear regression for estimating the quantitative probes. Even without rigorously applying the do-calculus, it should be clear that it is sufficient to include the season as a confounder, since it influences both the sprinkler and the slippery variable. Conversely, the wetness must not be adjusted for in the linear regression, since this would incorrectly block the causal pathway from the sprinkler to the slipperiness. The remaining rain variable can optionally be included as a covariate, but it is not necessary. Controlling for the season and fitting a linear regression on the available data, the causal model yields an estimate of 0.52 for the ATE of "Sprinkler" on "Slippery". The result makes sense: Turning on the sprinkler makes the lawn more slippery because of the increased wetness. We do not need to content ourselves with this particular ATE between our target variables: Using the do-calculus and our causal graph, we can formulate and estimate unbiased estimands for a multitude of other causal effects. Given the

observational data and the knowledge about the necessary linear regression covariates, the estimation consists only of fitting the linear regression models on the observational data and reading off the causal effects from the resulting regression coefficients. As an example, we estimate all possible ATEs, NDEs and NIEs between the variables in the causal graph. Unfortunately, it is not clear what a causal effect with a categorical treatment or outcome variable is supposed to describe: Does "treatment" mean changing the season from winter to spring, or from fall to summer, or from summer to winter? In order to avoid any ambiguity, we simply filter the dataset to include only samples recorded in winter or spring and binarize the resulting feature. Winter is encoded as zero whereas spring is encoded as one. Alternatively, we could use the whole dataset and cycle over all possible binary treatment/outcome encodings for the season variable, in order to calculate an aggregation of the causal effects. Cause2e provides different aggregation strategies, but we skip this exercise while keeping in mind that categorical variables need to be treated with additional caution compared to binary ones.

4.4.4 Result analysis

At the end of the causal analysis, the results of the various effect estimations can be evaluated using different tables and visualizations. Figures 4.3 and 4.4 provide a quick overview of the causal mechanisms that govern the behavior of the involved variables: Figure 4.3 shows heatmaps of the overall ATEs, the direct NDEs and the indirect NIEs. Especially the black areas in the heatmaps are helpful for discerning variable pairs that do not interact in the sense of the effect type under consideration. Figure 4.4 helps to quantify the most important influences by listing the largest effects for each effect type. If we already know a specific effect of interest, such as the ATE of the sprinkler on the slipperiness of the lawn, we can look it up in the full effect tables provided in Figure 4.5. We can see that sprinkler activation makes the lawn considerably more slippery: the ATE of 0.52 is halfway between the minimal effect of zero and the maximal effect of one. Therefore our analysis finally tells the gardener that sprinkler activation is only possible at the cost of endangering people walking on the lawn. Additionally, we can read off that the effect is fully mediated by the wetness of the lawn, meaning that the danger is only caused indirectly by the increased wetness.

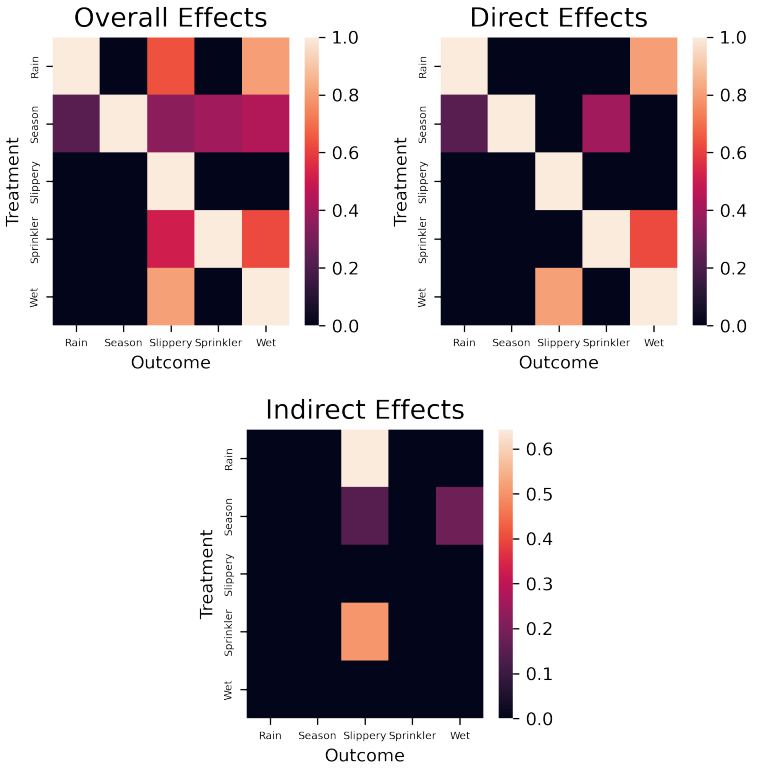

Fig. 4.3: Heatmaps serve as a visualization of the causal effects between various treatment-output-combinations. Different effect types, such as the ATE (upper left), NDE (upper right) and NIE (bottom) can be illustrated in different heatmaps.

10 Largest Overall Effects:

Treatment	Outcome	Estimated_effect
Wet	Slippery	0.81
Rain	Wet	0.80
Rain	Slippery	0.64
Sprinkler	Wet	0.62
Sprinkler	Slippery	0.52
Season	Wet	0.44
Season	Sprinkler	0.40
Season	Slippery	0.35
Season	Rain	0.23
Slippery	Rain	0.00

10 Largest Direct Effects:

Treatment	Outcome	Estimated_effect
Wet	Slippery	0.81
Rain	Wet	0.80
Sprinkler	Wet	0.62
Season	Sprinkler	0.40
Season	Rain	0.23
Slippery	Rain	0.00
Slippery	Season	0.00
Slippery	Wet	0.00
Slippery	Sprinkler	0.00
Rain	Slippery	0.00

10 Largest Indirect Effects:

Treatment	Outcome	Estimated_effect
Rain	Slippery	0.64
Sprinkler	Slippery	0.50
Season	Wet	0.18
Season	Slippery	0.14
Slippery	Rain	0.00
Slippery	Season	0.00
Slippery	Wet	0.00
Slippery	Sprinkler	0.00
Rain	Season	0.00
Rain	Wet	0.00

Fig. 4.4: In an exploratory causal analysis, the largest causal effects can be examined to gain an overview over the most important causal relationships governing the DGP.

Overall Effects

	Rain	Season	Slippery	Sprinkler	Wet
Rain	1.00	0.00	0.64	0.00	0.80
Season	0.23	1.00	0.35	0.40	0.44
Slippery	0.00	0.00	1.00	0.00	0.00
Sprinkler	0.00	0.00	0.52	1.00	0.62
Wet	0.00	0.00	0.81	0.00	1.00

Direct Effects

	Rain	Season	Slippery	Sprinkler	Wet
Rain	1.00	0.00	0.00	0.00	0.80
Season	0.23	1.00	0.00	0.40	0.00
Slippery	0.00	0.00	1.00	0.00	0.00
Sprinkler	0.00	0.00	0.00	1.00	0.62
Wet	0.00	0.00	0.81	0.00	1.00

Indirect Effects

	Rain	Season	Slippery	Sprinkler	Wet
Rain	0.00	0.00	0.64	0.00	0.00
Season	0.00	0.00	0.14	0.00	0.18
Slippery	0.00	0.00	0.00	0.00	0.00
Sprinkler	0.00	0.00	0.50	0.00	0.00
Wet	0.00	0.00	0.00	0.00	0.00

Fig. 4.5: For a detailed analysis of all causal effects, the full causal effect tables are provided.

Part III

CAUSAL MODEL VALIDATION

After having introduced the methods of graph-based causal inference from observational data, we can now apply them to a wide range of problems. However, an important part of the analysis is still missing: The presented methodology allows us to predict the effect of interventions in settings where direct experimentation would be too costly or risky such that inference from purely observational data is preferable. By the same argumentation, it is of paramount importance to validate the results of the causal analysis before decision makers use the findings to manipulate the sensitive target domains. In the remainder of this thesis, we will therefore discuss the problem of causal model validation, which will complete the notion of a causal end-to-end analysis. To reach this goal, the concept of quantitative probing will be introduced as a largely model-agnostic causal validation strategy that exploits quantitative domain knowledge. Following the motivation based on Popper's falsificationist validation ideas, a thorough simulation study is presented as evidence for the effectiveness of quantitative probing. Limitations of the concept are discussed and a practitioner's guide aims at facilitating the incorporation of the method in causal data science applications. Remaining open questions are finally revisited, in order to provide a starting point for future research.

The following content has in parts already been included in a research article (soon to appear in the Journal of Causal Inference) by the author and his advisors [2] or in a preprint version of the same article [62]. In order to avoid self-plagiarism, direct quotes are listed explicitly here:

- Except for the newly added description of the train/validation/test split, the review of correlation-based model validation in Section 5.1.1 is a verbatim quote of the corresponding section in [62] that has been omitted from [2] in response to the reviewers' feedback.

- The reviews of refutation-based causal model validation in Section 5.1.2 and exploiting causal domain knowledge in Section 5.2 are taken verbatim from the "Related work" section in [2].

- The stepwise description of the simulation setup in Section 7.1.2 is a slightly reworked version of the description in [2] and contains direct quotes.

- Figures 6.1, 7.1 to 7.3 and 7.5 to 7.10, as well as their captions are taken from [2] without any modifications.

The overall structure and content naturally also overlaps with [62] and [2], given that the same subject matter is presented as part of this thesis. However, the thesis explains many points in greater detail and discusses ideas that were not included in the above articles.

5. STATE OF THE ART

In order to motivate quantitative probing as a largely model-agnostic causal validation strategy that exploits quantitative domain knowledge, we review the existing literature and highlight the gaps that quantitative probing is aiming to fill.

5.1 Model validation

As quantitative probing is a strategy for validating causal models, we will first review the literature on validation procedures for both correlation-based and causal models. Unfortunately, we will see that the currently available methods either cannot be applied to causal modelling at all, are restricted to specific types of causal models, or are unable to incorporate various forms of domain knowledge.

5.1.1 Correlation-based models

In order to gauge the difficulty of the challenge that validation poses for causal inference procedures, it is worth taking a step back and recapitulating the reasoning behind the predominant strategy for validating traditional correlation-based statistical learning methods. These methods, which encompass both classification and regression algorithms, such as linear regression, support vector machines or numerous variations of deep learning with neural networks, have one crucial assumption in common [63]: Every sample that we have observed in fitting the model as well as every sample that we will need to feed into the final model for classification or regression is drawn from the same distribution, and they all are drawn independently of each other. This assumption is commonly referred to as the *i.i.d. assumption*, which stands for *independent and identically distributed*. However, it is seldomly explicitly mentioned because of how deeply ingrained it is in all correlation-based thinking about machine learning. The two parts of the term i.i.d. have important

consequences for how we train or validate machine learning models:

The independence assumption is the implicit foundation of the current practices for model training: If we did not assume that all samples are drawn independently from each other, the likelihood

$$p(X, Y|\phi) = \prod_{i=1}^{n} p(x_i, y_i|\theta) \tag{5.1}$$

of observing samples with features X and labels Y for a given parameterization ϕ would not factorize over all the samples (x_i, y_i) for a lower-dimensional parameterization θ. The consequence would be that the commonly used error metrics, e.g. the mean squared error or mean absolute error, would lose their theoretical backing: All of them are based on summing up independently computed prediction errors of the model for each sample, which is justified precisely by the factorization property of the likelihood (or equivalently the summation property of the log-likelihood).

The assumption of identical distribution enables the use of the train/test split for model validation: If all the samples that we will ever need to classify stem from the same distribution as the observed labelled data, we can fit our model on the observed data and be confident that the obtained model will also be suitable for classifying the new incoming data. Even the thereby caused risk of overfitting, i.e. learning overly specific characteristics of the training data that fail to generalize and lead to a worse than expected performance on new data, can be mitigated using the same distributional assumption: If we do not train the model on all the labelled data, but only on a subset of it (the *training set*), we can evaluate its performance on the rest of it (the *test set*), given that the correct labels for the test set are available to us. Additionally, we have not used the test set for model training and it is statistically identical to the new data that we will have to classify in the actual task, because it is drawn from the same distribution. Therefore, the expected value of the prediction error for any unseen sample is equal to the mean error on the test set. In summary, we can base our confidence in the predictions of the model on its performance on the test data, which is enabled precisely by the assumption of identical distribution.

If we want to compare multiple candidate models and select the best among them, we cannot base the comparison on the error on the training set, given that these measures are possibly tainted by overfitting. On the other hand, we also cannot use the test set for model selection because the test

set would no longer constitute new, unseen data after the selection process. This issue can be overcome by the *train/validation/test split*, which divides the non-test samples into a training and a validation set: After all models have been trained on the training set, they can be compared based on their performance on the validation set. The best model is selected accordingly and the test set now can serve to judge the expected fitness of the model with respect to new data.

If we now try to transfer these techniques to the training and validation of *causal* models, we will inevitably face severe problems. The observational data that we use to train our causal models can very well follow the i.i.d. assumption. The notion of using test samples from the same distribution to evaluate how well the model performs on hypothetical queries, however, is diametrically opposed to the task of causal inference: We want to predict what happens under certain interventions, and an intervention is precisely the act of changing the data generating process. A change in the data generating process, of course, generally entails a change in the distribution from which the samples are drawn. Therefore, the train/test split cannot be used to evaluate the performance of causal models.

5.1.2 Causal models

The unavailability of the train/test split has already lead to different streams of validation research in the causal inference community. We will briefly review the main directions and point out why there is still no universally agreed upon solution that would constitute a causal analogue to the model-agnostic train/test split.

Sensitivity analysis

Sensitivity analysis is a well-established tool in statistical modelling with a simple premise: If we suspect that the model M in question might be misspecified in one or multiple ways, but we are not able to correct all the errors, we should accept them and change the form of output instead. Suppose that we are able to identify all of the possible weaknesses as the degree to which the true parameters might deviate from the model under scrutiny. Using this knowledge, we can turn the initial point-estimate of the prediction $y = M(x)$ into a range of predictions that reflect our beliefs about the remaining uncertainty. For a set $\mathcal{M}$ of models that we consider reasonable,

we can then still report $f(M)$ as our estimate, but with the addition that the true value might as well lie anywhere within the set

$$Y = \{\tilde{M}(x) | \tilde{M} \in \mathcal{M}\}. \tag{5.2}$$

Although we give up on the goal of a single precise estimate, the resulting set Y of possible estimates can in practice be sufficient for subsequent decision-making, since all estimates outside of Y are ruled out. The evaluation of the additional models often comes at a negligible cost, when $\mathcal{M}$ is given as a parameterized family of models and the same estimation procedure can be reused. In a causal scenario, the uncertainty could arise from multiple sources and different model-specific strategies for addressing it are proposed in the literature [64, 65, 66, 67]. An exemplary source of uncertainty is given by doubts about the presence or absence of an edge in the causal graph. To avoid an overconfident but incorrect estimate, we use both variants of the causal graph to identify and estimate the causal effect of interest and report the set of plausible values. Downstream decisions can then be based on this set to account for the underlying uncertainty, instead of being based only on an unstable point estimate. While sensitivity analyses are easy to use and allow for an honest communication of remaining doubts about the examined model, they also suffer from a crucial drawback: In order to specify the set $\mathcal{M}$ of possible models, it is necessary to identify not only the dimensions in which the model might be incorrect, but also to what extent a deviation in these dimensions is possible. Both questions are in general hard to answer and any mistake threatens to void the guarantees that the sensitivity analysis is supposed to give. Another pitfall depends on the intentions with which the model was built: If its only goal is to predict a single causal effect, before being abandoned, the set Y can be a sufficient outcome of the validation procedure. If, however, we are planning to reuse the model or even just components of it for future tasks, or if we are interested in the model itself for understanding the underlying process, the sensitivity analysis does not indicate how certain we can be that our initial model is indeed correct.

Benchmarking on simulated data

Another stream of research does not try to directly validate the causal model, but the algorithm that was used to produce it [68]. As we do not have a ground truth for our actual modelling task, we make up for it by applying one or multiple simulations in the following procedure:

1. We prescribe the DGP, e.g. a Causal Bayesian Network, ourselves and draw a number of samples from it.

2. We specify a task to be solved for the simulated scenario, e.g. estimating an average treatment effect between two of the involved variables.

3. We use the same algorithm as in the actual modelling task, e.g. causal discovery via greedy equivalence search, identification via the do-calculus and estimation via linear regression, to build and evaluate a causal model for the simulated data.

4. We check whether the analysis was successful and use this information to decide whether the same algorithm should be applied to the actual task of interest.

Note that the last step is only possible because we have a ground truth in the simulated scenario, given that we have full knowledge about the DGP. The main idea is that an algorithm that will perform well on the simulated tasks will also perform well on the real task. In principle, it is not necessary to resort to simulated data if we can provide benchmark scenarios from the real world where the underlying DGP is sufficiently well-understood. Instances of such benchmarking approaches are given by various papers in the causal discovery community that evaluate algorithms on the genomic Sachs dataset [69, 48, 70]. Although circumventing the issue of missing causal ground truth by the introduction of benchmark datasets is elegant and, especially in the case of simulated ones, allows for precise customization, there is again a major pitfall: It is simply not known how well the performance on the benchmarking task transfers to the real task. Certainly, not every task is sufficient for ensuring the quality of the candidate algorithm: If the real task is a complex one with hundreds of variables but we only evaluate the algorithm on a simulated variant with three variables, the benchmark will be of limited use. Generalizing this observation, the missing piece can be phrased as a statement about Lipschitz continuity of maps between the space of DGPs and the performance of causal models. In [71], a Taylor expansion on the space of DGPs is used as a first-order approximation: The error of the candidate model for the real DGP is estimated by its error on a simulated DGP, corrected by the product of an influence function and the distance between the two DGPs. It remains problematic to obtain this distance without detailed knowledge about the real DGP, which would make any causal analysis obsolete.

Refutation tests

Given the shortcomings of simulation-based benchmarks, much of the current research on causal model validation is focused on providing refutation checks that can be directly applied to the causal model under scrutiny without the need to go back and forth between simulated and real DGPs:

In the potential outcomes community, model criticism for Bayesian causal inference [72] has been developed based on posterior predictive checks [73, 74]. The causal model is separated into a treatment model and an outcome model, which are criticized independently. Both are generative parameterized models and, for a given candidate model, discrepancy functions are evaluated to summarize properties of the data generated from it, using a suitable prior. The model is then evaluated by a comparison of these results and the discrepancy that has been realized by the actually observed data. Drawbacks of the procedure lie in its restriction to a special case of a potential outcomes model where the posterior factorizes across outcome and assignment parameters, the need to choose suitable discrepancy functions and the missing interface for incorporating domain knowledge.

In out-of-sample causal tuning [75], a graphical causal model induces a set of predictive models, namely one for each of the nodes. If the underlying graph is misspecified, some of the predictive models will not rely on the correct inputs (the Markov blanket) to predict the node. Each predictive model can then be evaluated against domain knowledge about the actual distribution over the respective nodes, such that wrong models can be detected. As presented in [75], the method is restricted to probabilistic graphical models and formulated for checks of non-interventional distributions that will be passed not only by the causal model, but any model in the same Markov equivalence class. We will build on the point of view that all variables in a graphical model can be used for refutation checks and extend it to the interventional setting.

In [76], domain- and model-agnostic refutation tests are employed to probe candidate models. An example would be to replace the data for the treatment or outcome variable by random data, which is independent of all other variables. If the model predicts a non-zero causal effect, it should clearly be refuted. Other tests include the synthetic addition of random and unobserved common causes, as well as replacing the original dataset by a subset or a bootstrapped version of itself. These tests serve as a filter to refute implausible models. Although such checks are well in line with the scientific

method [3], these generic tests might be too weak a filter for distinguishing the correct model from plausible, but incorrect models. The authors explicitly call for the extension of their methods by more domain-expert guided validation tests to improve their practical relevance.

In summary, the existing validation methods are either tightly coupled to a type of causal model or unable to incorporate problem-specific domain knowledge that could be provided by a domain expert without deep knowledge in causal inference.

5.2 Exploiting causal domain knowledge

Incorporating domain knowledge in the causal discovery step is an established concept in graph-based causal inference, as we have seen in Section 3. The presented approaches have in common that only knowledge about the presence or absence of certain edges in the causal graph can be directly exploited. Due to its relation with the discrete properties of the graph, we refer to this type of knowledge as *qualitative domain knowledge*. However, this represents only a fraction of the causal knowledge that could be available to domain experts, which becomes clear by a slight rephrasing: The presence or absence of an edge from node A to node B in the causal graph is equivalent to a set of controlled direct effects of variable A on variable B being nonzero or zero, respectively. Knowledge about other types of causal effects, such as the average treatment effect (ATE), the natural indirect effect (NIE) or the conditional average treatment effect (CATE) cannot be used in the above procedures. In order to separate this wider notion of knowledge from the restricted qualitative domain knowledge, we refer to it as *quantitative domain knowledge*. It furthermore describes not only constraints demanding that certain effects be either nonzero or zero, but also includes knowledge about the effect strength on the full spectrum of real numbers. As its name suggests, quantitative probing will make all quantitative domain knowledge available for causal model validation.

6. APPLYING THE LOGIC OF SCIENTIFIC DISCOVERY TO CAUSAL INFERENCE

In the following, we want to explore a line of thought that we consider a solid foundation for the validation of causal models and scientific theories in general. Naturally, most of the following reasoning belongs to the field of philosophy, which is not the domain of expertise of the author. Given that the same statement will also hold true for most of the readership, we will refrain from a lengthy and analytical discourse, which would require a precise definition of terms that philosophers have struggled to define for many centuries. The arguments will be presented in a necessarily imprecise, but tangible and hopefully instructive manner, such that the validation of causal models can be connected to a much broader and better investigated spectrum of scientific questions. For a more mature treatment, we refer the reader to Popper's seminal work on the subject [3].

6.1 A brief excursion into the natural sciences

In order to understand the challenge that the validation of causal models poses, it is beneficial to briefly step away from the comparatively young field of causal inference and to inspect a similar challenge in the more mature natural sciences. In these sciences, which include physics, chemistry and biology, the central goal is the formulation of theories that explain the behavior of our surrounding physical reality. Such a theory could in principle take many forms, such as a verbal description, a drawing or a mathematical formula. If we allow for such freedom in the syntax of possible theories, the thereby entailed freedom in the semantics is enormous. Any children's drawing could be considered a scientific theory, which is clearly not in line with our understanding of research in the natural sciences. If we demand that a theory must answer a prespecified question (whatever that means), most of these nonsensical theories are immediately excluded. For the remaining

theories, it is of course desirable to verify whether they indeed give the correct answer to the question. A collection of beautiful mathematical formulas might look convincing and elegant, but it is only useful as a scientific theory if its predictions actually correspond to the reality that it is intended to model. Deciding strictly whether a theory gives correct answers to our scientific questions is necessarily impossible here, given that we have not defined strictly what we mean by *question* and *answer*. However, it is clear that different theories will often make conflicting statements and in most cases, we want to resolve this conflict by declaring most of the candidate theories as invalid ones, until no further inconsistencies are entailed by the remaining set of theories. Well-known principles, such as Occam's razor, advise us to discard theories that are unnecessarily complex in favor of simpler ones that seem equally suitable for explaining our reality, and have already found their way into probabilistic modelling [77]. Nevertheless, we are still required to judge whether a theory, however complex or simple it may be, does indeed possess the ability to explain what is happening in the real world.

The challenge here does not only reside in the already mentioned difficulty of precisely stating what this is supposed to mean, but there is an inherent problem with proving theories: In order to prove any statement, it is necessary to start with a collection of statements that are known to be true, and then to logically derive the statement in question from these assumptions. This method is called *deductive reasoning*. Such a collection of true statements about the physical world, alas, is impossible to obtain, as philosophers have long established by the use of simple thought experiments [78]. Contrary to the paradise of pure mathematics, it is therefore impossible to employ only deductive logic in the natural sciences, simply because we cannot provide a starting point for the deductive machinery. This does by no means exclude the use of deductive logic after having accepted some initial theories to be true: Using Newton's general laws of motion, many laws for special cases, such as the different pendulum setups that are taught in every mechanics class, can be derived in the true sense of the word. For the initial theories, it is by definition impossible to use deductive reasoning, such that *inductive reasoning* has become the method of choice in the natural sciences. This line of thought can be seen as an inverse of the deductive method: Instead of deriving many special cases from one established general law, we observe many special cases and try to find a more general law that is consistent with each of the special cases. In other words, we do not believe in the theory of gravity and therefore expect an apple to fall on our head, but we observe the

fall of the apple and speculate that there is an underlying theory of gravity that describes the general mechanisms behind this particular observation. When dealing with the actual physical world, inductive reasoning has one decisive advantage over deductive reasoning: Although we cannot prove that a theory is true in the sense that it correctly explains all data in the entirety of our spatio-temporal universe, we can easily prove the same theory wrong by providing one data point that is in conflict with the theory. A theory that predicts all sheep to be white can easily be refuted by showing a single black sheep. This refutation makes for an efficient tool in theory selection or restriction, since the large majority of the above-mentioned nonsensical theories are immediately and effortlessly disproven by the presentation of a single counterexample.

Deductive and inductive logic can even be combined in a simple procedure. Observations are primarily used to inspire a theory that explains them (inductive), before special cases of the theory are derived by sound logic (deductive). If we now find any datum, be it in the original or in an additional set of observations, that is inconsistent with any of the derived versions of the theory, we can directly conclude that the original theory must be wrong. It is important to realize that the absence of conflicting data does not warrant the conclusion that the theory must be correct, since we cannot exclude the possibility of conflicting, but yet unobserved data. Therefore, the effectiveness of such a refutation process depends on the availability of possibly conflicting data. The sheep-theory cannot be refuted if we do not observe any sheep (no data). The same holds true if we restrict our data collection to white sheep only: A large amount of data is still not useful, unless we ensure an process of data collection that is not biased in favor of the theory. Unfortunately, the problems can even arise before the data collection by formulating theories that purposefully evade the possibility of being refuted: A theory suggesting that every man who is born in the year 3023 will die in 3054 cannot be refuted by any data available to researchers in the foreseeable future. Although the theory will never be refuted before the year 3055, we are not inclined to believe it, simply because it has not earned our trust by withstanding any refutation attempts despite being in conflict with our prior beliefs. In order to be considered a worthy candidate, a theory must be *falsifiable*, otherwise we can consider it useless. A falsifiable theory, on the other hand, can gain our trust by not only conforming to the original data that lead to its formulation, but also to subsequent data that is collected with the aim of disproving it. As Popper writes: "But I shall certainly admit

a system as empirical or scientific only if it is capable of being tested by experience. These considerations suggest that not the verifiability but the falsifiability of a system is to be taken as a criterion of demarcation." [3] To summarize this philosophical excursion, the current practice of establishing theories in the natural sciences can be described as follows:

1) Initial data is collected.

2) A falsifiable theory that explains the data is formulated.

3) Additional data is collected with the goal of possibly falsifying the theory.

Step 3 can lead to different outcomes: If the additional data contradicts the theory and any errors in measurement can be excluded, the theory is clearly not correct and needs to be adapted (step 2), before more data is collected. Contrarily, if the additional data does not contradict the theory, its conformance with the theory supports it and increases the trust in its correctness.

It is worth noting that this so-called *scientific method* has proven effective to such a degree that it is now even used in disciplines where formal proof would be an option: In mathematics, the Riemann hypothesis is believed to be true by most mathematicians because more than 10^{13} falsification attempts have failed [79, 80]. Although this would be considered compelling evidence in most fields of science, it is seen only as a crutch by most mathematicians who would prefer a proper proof or a counterexample to eliminate the remaining doubts and settle the matter for all eternity [81]. This fear is nourished by known examples, such as Skewes's numbers [82, 83], where all numerical evidence supports a conjecture, until a large number in the order of 10^{316} finally disproves it.

In software engineering, it is common practice to verify all parts of a program, the *units*, by *unit tests*. These tests ensure the correct functionality of each unit by verifying that it performs as expected for various given inputs. If enough unit tests fail to expose a unit as flawed, it is trusted to work as expected. Otherwise, the unit has to be rewritten until it passes all the tests without error. This approach to software validation is in sharp contrast to the ideas of pioneers like Dijkstra, who advocated for the use of formal verification: "Today a usual technique is to make a program and then to test it. But: program testing can be a very effective way to show the presence

of bugs, but is hopelessly inadequate for showing their absence. The only effective way to raise the confidence level of a program significantly is to give a convincing proof of its correctness." [84] Although formal verification had been abandoned as a validation method by most software developers, the more recent advent of automated tools has lead to an increase in its popularity, especially for safety critical applications [85, 86].

6.2 Challenges in the falsification of causal models

Coming back to the realm of causal inference, it seems straightforward to apply the logic of causal discovery to the challenge of validating causal models. Analogously to physical theories, such as the law of gravity, causal models are also theories that aim at explaining our surrounding reality. Therefore, we can apply the logic of scientific discovery and collect data in an attempt to falsify the causal model after its original formulation. At this point, however, we should realize a crucial difference between causal models and other theories in the natural sciences. In physics, we can usually observe the quantities that we are reasoning about by setting up experiments in a safe environment that we consider similar enough to the target domain. If we want to model the behavior of a giant spring on Mars using Hooke's law, we can build a small spring on Earth and observe the data that it generates. After gathering sufficient data in these inexpensive settings, we can then decide that we trust the model enough to predict the behavior of the giant spring on Mars, which saves us the inconvenient and costly journey that would be necessary to perform the experiment in the target domain. In the causal setting, an analogous scenario could be the testing of a new drug for medical patients. Just as the journey to Mars, it is desirable to avoid direct testing on the patients with a randomized controlled trial (RCT): If the drug works, we have possibly sacrificed the lives of the patients in the control group for the study. If the drug has no effect at all, both the treated and the control groups have foregone the chance of being treated with another possibly helpful therapy. If the drug has a negative effect, again both groups are harmed. Therefore, we want to exploit a theory that gives us the desired information about the effectiveness of the drug without performing the otherwise necessary experiment, so we build a causal model that can answer interventional queries, such as the effect of the drug on the patients' recovery. When it comes to validating our theory (the causal model), we note that it is hard to find an analogous experiment that

we can perform in a safe environment. We could of course perform an RCT on a "validation group" of patients different from the original one, just as Hooke's law could be validated with data from a different spring. However, this immediately poses two new problems: Firstly, it is hard to find such a validation group of patients that would constitute a "safe environment" for experimentation without admitting the questionable hypothesis that some human lives are more valuable than others. If the lives in the validation group deserve the same protection as those in the original group, we gain nothing by performing an experiment on the validation group: Validating the causal model with the resulting data and thereby avoiding experimentation on the original group causes the same harm as directly performing the experiment on the original group. Secondly, even if we manage to find a validation group that we would be willing to experiment on, for example a species of mice or even apes, it is questionable whether the data is useful to validate a causal model that should work for humans. The underlying problem seems to be that laws in the natural sciences are far more universal than most causal models that are intended to be used only for a very narrow domain of application. In physics, it could even be worth doing the actual experiment because, despite its cost, we might gain a greater understanding of the world by the formula that offers additional insights other than the direct result of the experiment. For causal models, the aspect of understanding the underlying mechanisms is often secondary, as we are primarily interested in the answer to one specific causal query, such as a single average treatment effect between two variables.

6.3 Quantitative probing

In order to overcome the above challenges in applying the logic of scientific discovery to the validation of causal models, we propose the method of *quantitative probing*, which has been co-developed by the author during his PhD studies [2]. As explained above, it is possible and helpful to first derive one or more falsifiable statements from the original causal model by the use of deductive logic. These can subsequently serve as targets of our falsification attempts. Such a derived statement could be a specific causal effect between two variables in the causal model, and the deductive logic that is employed to calculate it could be Pearl's do-calculus in the case of an unparameterized graphical causal model, or a simple probabilistic calculation in the case of a Causal Bayesian Network. Other derived statements are in principle possible,

but the investigated concept of quantitative probing focusses exclusively on quantitative causal effects predicted by the model. Since these effects will be used to probe the quality of the model, they are referred to as *quantitative probes* or simply *probes*. It is important to note here that the selected quantitative probes do not have to correspond to the actual causal effect of interest of our investigation, which is referred to as the *target effect*. In the drug example in Section 6.2, we could select the natural direct effect of gender on recovery as a quantitative probe, even if we are ultimately interested in predicting the average treatment effect of drug use on recovery. If the causal model is trustworthy, it should correctly predict the natural direct effect of gender on recovery. Otherwise, the model is falsified and should not be trusted to give the right answer to our original query about the average treatment effect of drug use on recovery. In fact, it is even advantageous to use *non-target effects* as quantitative probes: Our knowledge about the target effect is necessarily restricted, which is why we are performing the causal analysis in the first place. Non-target effects, on the other hand, can be well-understood and serve as falsification tools in a validation strategy based on domain knowledge, without introducing any circular reasoning.

The general procedure for quantitative probing as a validation method can be summarized as follows (see Figure 6.1):

1. Probe selection and specification: Select one or multiple non-target effects and specify their true values.

2. Modelling: Construct a causal model.

3. Probe prediction: Use the causal model to predict values for the quantitative probes.

4. Probe evaluation: Evaluate the prediction of each probe by comparing the predicted values to the true values.

5. Model evaluation: Based on the probe evaluation, accept or refute the causal model.

6. Target prediction: If the causal model has been accepted, use it to predict the target effect.

The proposed procedure can be applied to any type of causal model, as long as it is able to predict both the target effect and the quantitative

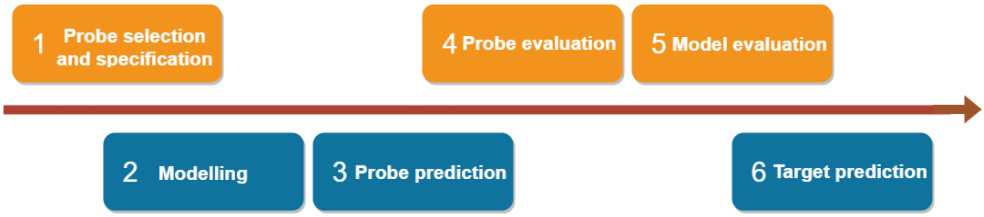

Fig. 6.1: The six steps of a causal modelling workflow that uses quantitative probing as a validation strategy. Steps that are part of the validation itself are colored in orange, whereas steps that are performed by the modelling algorithm are marked in blue.

probes. Even though there is still no direct usage of the quantitative domain knowledge for building the causal model, the formulation of the quantitative probes provides a way of indirectly incorporating the knowledge into the falsification-based validation of the model. Combining these two advantages, quantitative probing represents the first causal validation strategy that is both model-agnostic and able to incorporate quantitative domain knowledge.

From a high-level perspective, we have reduced the problem of determining the target causal effect to the problem of determining one or even multiple other causal effects between the variables in the causal model. This might at first seem disillusioning, but we have made progress: We are unable to directly determine the target effect, which is why we are conducting the causal analysis in the first place. For the quantitative probes, however, it is entirely possible that they are either easy to determine or even already known. We have therefore reduced an intractable original problem to one or more tractable new problems, which is a classic way of problem solving. Unfortunately, unlike in mathematics, knowing the answer to all the new problems does not yield a definitive answer to the original problem: It is possible that our model passes all tests by correctly recovering the quantitative probes, but this only means that the model is not falsified. Due to the remaining uncertainty that is inherent to the scientific method, the model could still be wrong and therefore yield an incorrect estimate of the target effect, as we will see in Section 7.4. This is lamentable, but it is an inherent and unavoidable flaw of falsificationist model validation that we have to accept, as opposed to a weakness specific to the method of quantitative probing.

6.4 The sprinkler example revisited

Before we step into the statistical analysis of the effectiveness of the proposed validation method, we illustrate the full procedure using Pearl's sprinkler example. The goal is again to evaluate the average treatment effect of the sprinkler on the slipperiness of the lawn. Is the risk of slipping increased if we turn on the sprinkler? In order to have a ground truth, we prescribe the usual binary data generating process (cf. Figure 1.2) and generate 10000 samples from it that are provided as observational data for our exemplary analysis.

6.4.1 Probe selection and specification

In this example, we could arguably write down a multitude of quantitative probes without consulting a domain expert. For a clearer presentation, however, we only choose two of them:

1. Our knowledge about sprinklers tells us that activating the sprinkler should have a positive overall effect on the chance of the lawn being wet, so we expect the ATE to be greater than zero.

2. The same holds true for the effect of the wetness on the slipperiness of the lawn, which yields another quantitative probe in the form of a positive expected ATE.

Note that quantitative probing can be applied to any type of causal effect, as well as to any type of probe specification: We could have chosen to use probes involving the natural direct effect, or probes where we do not only expect a positive value, but more precisely a value in a given compact interval.

6.4.2 Modelling

For the modelling part, we use the previously introduced strategy of causal end-to-end analysis, since the tools for it are readily available in the open-source cause2e package [4], which was developed by the author during his PhD studies. In order to show how both correctly and incorrectly specified causal models are received by quantitative probing, we build two independent models in parallel that only differ in the specified qualitative domain knowledge. The first one, in addition to the observational data, exploits the following domain knowledge items for recovering the causal graph:

1. We forbid all edges that originate from "Slippery".
2. We forbid all edges that go into "Season".
3. We forbid the edges "Sprinkler" → "Rain" and "Season" → "Wet".
4. We require the edges "Sprinkler" → "Wet" and "Rain" → "Wet"

The second model receives qualitative domain knowledge for the same edges, but the two required edges are specified in the wrong direction, whereas the forbidden edges are left unchanged. The corresponding knowledge graphs in Figure 6.2 indicate that seven edges are still unconstrained if we exclude cycles. Subsequently, two causal graphs (cf. Figure 6.3) are obtained by running fast greedy equivalence search, a score-based causal discovery algorithm [61], on the respective sets of qualitative domain knowledge. The graphs are then used to identify suitable linear regressions for estimating the quantitative probes.

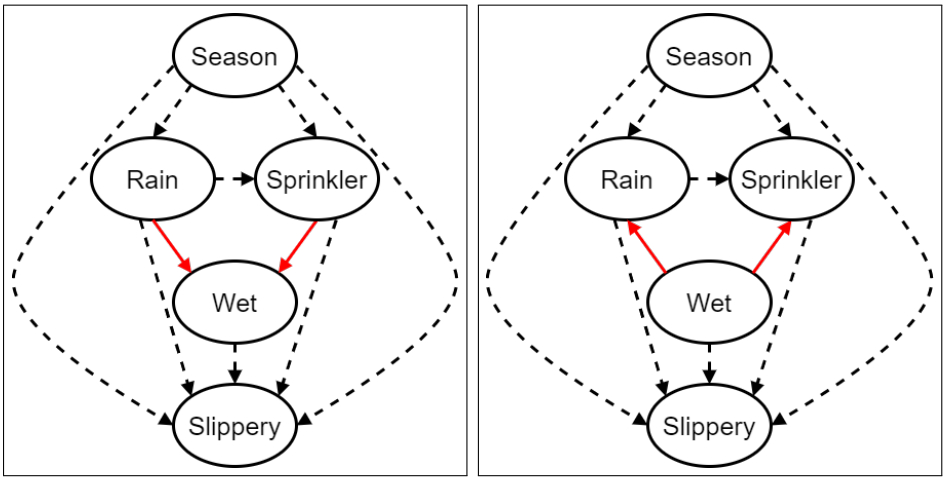

Fig. 6.2: The knowledge graphs visualize the correct (left) and incorrect (right) sets of qualitative domain knowledge. Edges prescribed by the respective sets of domain knowledge are marked in red, forbidden edges are not drawn at all, and unconstrained edges are drawn dotted.

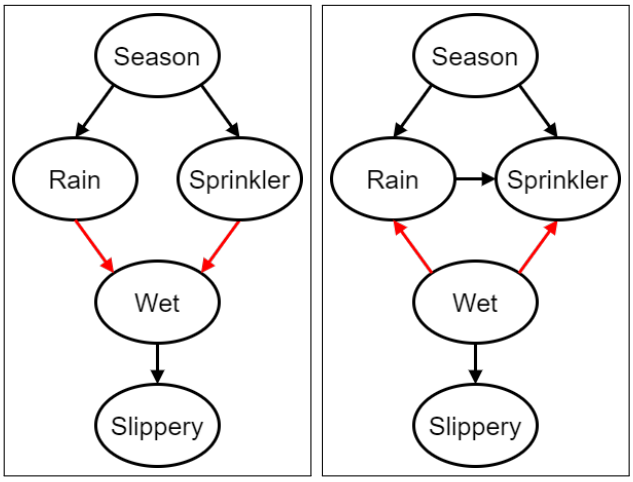

Fig. 6.3: The recovered graphs correctly (left) or incorrectly (right) represent the data generating process. Edges prescribed by the respective sets of domain knowledge are marked in red, the rest of them has been added by the causal discovery algorithm.

6.4.3 Probe prediction

Given the observational data and the knowledge about the necessary linear regression covariates, probe prediction consists only of fitting the linear regression models on the observational data and reading off the causal effects from the resulting regression coefficients. The correct causal model yields an estimate of 0.62 for the ATE of "Sprinkler" on "Wet" and 0.81 for the ATE of "Wet" on "Slippery". The incorrect causal model yields estimates of 0 and 0.81, respectively.

6.4.4 Probe evaluation

In both cases, the probe evaluation is straightforward: For the correct causal model, both probe estimates are evidently greater than zero, which is in accordance with the previously specified expectations. Note that the hypothetical extreme values of the causal effects are -1 and 1, respectively, meaning that there is even a comfortable margin of error around both estimates. For the incorrect causal model, the ATE of "Wet" on "Slippery" is still greater than 0. This illustrates that incorrect causal models can still succeed at recovering

quantitative probes. However, the second probe is estimated to be 0. In this case, we can even exclude a statistical error, since there is no directed path from treatment to outcome in the corresponding causal graph. Therefore, the estimate for the effect will always be 0, regardless of the supplied observational data, and we do not have to worry about tolerances.

6.4.5 Model evaluation

The correct causal model is evaluated to be trustworthy, since it has successfully recovered all quantitative probes. Contrarily, the incorrect causal model has been detected by one of the probes, leading to its rejection, possibly followed by further investigation and subsequent adaptation. It is probable that the erroneous reversal of the two required edges, which might have been due to a typing error, will be detected, such that the correction of this error will lead to the true causal model. Note, however, that the above two probes cannot be used to evaluate any reworked versions of the model: The rework process constitutes an instance of overfitting to the probes, such that an independent evaluation of the result is no longer possible. This issue is discussed in more detail in Section 7.5.5.

6.4.6 Target prediction

The correct causal model yields an estimate of 0.52 for the ATE of "Sprinkler" on "Slippery", therefore successfully recovering the fact that turning on the sprinkler makes the lawn more slippery because of the increased wetness. The incorrect (not reworked) causal model would yield an estimate of 0, since there is no direct path from "Sprinkler" to "Slippery" in its causal graph. To summarize, the quantitative probing method has declared the correct model to be safe for use, leading to an accurate answer to the original causal query, whereas the incorrect model has been prevented from giving a potentially harmful inaccurate answer by being marked as unsafe.

7. SIMULATION STUDY

In the preceding sections, we have motivated the quantitative probing strategy and illustrated the procedure in an exemplary analysis. In order to provide empirical evidence for the effectiveness of the concept, we supplement the above arguments for the strategy with an extensive simulation study. The main hypothesis that has been investigated is the following:

Hypothesis 1 (H1): *Although even a perfect prediction of the quantitative probes does not guarantee a correct estimation of the target effect, the success of the target effect estimation should on average increase with the number of correctly predicted probes.*

7.1 Setup

The simulation-based evaluation of a validation strategy necessarily involves a complex chain of data generation, analysis and assessment steps. Therefore, we attempt to logically derive each of the steps from our goal of corroborating hypothesis H1, before separately summarizing the concrete choice of setup and the selected parameters.

7.1.1 Deriving the components of the setup

The above hypothesis H1 generates some constraints for the simulation setup:

1. Since the success of each analysis needs to be measured, we need to use scenarios with a known ground truth.

2. Since the statement is about the behavior of an ensemble of analyses, we need to evaluate many independent analysis runs.

3. The previous two points leave simulation as the only option, as opposed to real-world data where often either the ground truth is not known, or not enough data can be provided.

4. Since the overall relation between success at probe recovery and success at target effect recovery needs to be measured, either a quantitative or a visual evaluation of the aggregated results is necessary. Given the unknown nature of the relation, we refrain from calculating correlation coefficients and choose a visual approach based on suitable plots.

These points already prescribe the main pillars of the study. Regarding the model-agnostic formulation of the quantitative probing strategy, it is still unclear what kind of analysis should be performed in each of the runs. Another open choice for the design of the study is the type of DGP that should be used as ground truth. For this work, the author decided to focus on graph-based causal inference. The DGP is therefore chosen as a randomly generated Causal Bayesian Network, which is a fully customizable type of DGP and at the same time a classic causal model in itself, allowing for precise comparison of ground truth and causal model. In accordance with this choice, a configuration of the earlier presented causal end-to-end analysis (cf. Chapter 4) is selected as a graph-based modelling strategy. The graph-based nature of both DGP and modelling strategy enable the evaluation of a second hypothesis with only minor additional effort:

Hypothesis 2 (H2): *Although even a perfect prediction of the quantitative probes does not guarantee a correct recovery of the causal graph, the success of the causal discovery should on average increase with the number of correctly predicted probes.*

In order to avoid needlessly complicated DGPs and analysis strategies, only ATEs in a binary data setting are evaluated. However, the concept naturally translates to other causal effects and data types: These choices can be seen as implementation details of the modelling steps in Figure 6.1, which leave the necessary validation steps unaltered. The remaining free choices are now only related to parametrization details of the DGP and modelling strategy, as well as to the measures that are used to evaluate the success of each analysis. Concerning the latter point, we have chosen two different measures for H1 and H2 to account for the different target entities.

1. The target effect estimation is evaluated by comparing the estimate $\hat{\tau}$ and the true value τ of the target causal effect on an absolute and on a relative scale, leading to the measures

$$|\tau - \hat{\tau}| \text{ and } \frac{|\tau - \hat{\tau}|}{|\tau|}.$$

There is no risk of dividing by zero, since we are only using DGPs with *nontrivial target effects*: We require that there exists a directed path from the treatment to the outcome in the causal graph. Otherwise, any ATE is trivially zero or one, depending on whether treatment and outcome coincide.

2. Similarly, the estimation of each probe is evaluated using the absolute error. In order to account for numerical errors and statistical fluctuations, we allow an absolute deviation of ϵ_{probe} from the true value for a probe estimate to be considered successful. The *hit rate* is defined as the proportion of probes that have been correctly recovered by the analysis.

3. The graph recovery is evaluated using the *Structural Hamming Distance (SHD)* $\Delta(G, \tilde{G})$ [87], which is defined for two graphs $G = (V, E)$ and $\tilde{(}G) = (V, \tilde{E})$ that share a vertex set V, but whose edges E and $\tilde{E}$ possibly differ This measure checks for each node pair $(A, B) \in V$ whether G and $\tilde{G}$ agree on the type of edge between A and B, and then returns the number of disagreements:

$$\Delta(G, \tilde{G}) = |E \triangle \tilde{E}| - \left|\{e \in E \mid r(e) \in \tilde{E}\}\right| \tag{7.1}$$

where $\triangle$ denotes the symmetric difference between two sets, and $r(e)$ is the edge that results from reversing edge e. The second term avoids double-counting reversed edges. If one of the graphs is obvious from the context, we shorten the notation to ΔG.

7.1.2 Step-by-step procedure

The above considerations can be put into practice by executing the following steps multiple times:

1. Parameterization: Choose n (number of nodes), p_{edge} (edge probability), m (number of samples), p_{hint} (hint probability), p_{probe} (probe probability) and ϵ_{probe} (probe tolerance).

2. Ground truth generation:

   (a) Draw a random DAG with n nodes $x_1, ..., x_n$. Random means that for each of the n^2 possible directed edges, we include the edge with

a probability p_{edge}. After all the edges have been selected, check whether the result is a DAG with no isolated nodes. If not, repeat the procedure.

(b) Draw a random binary CPD for each node x_i, given its causal parents Π_i. The entries $p(x_i = 1|\Pi_i = \pi_i)$, which fully determine the CPD, are sampled from a uniform distribution on $[0, 1]$.

(c) Draw m samples from the resulting joint distribution over $(x_1, ..., x_n)$.

(d) Select a proportion p_{hint} of all the edges (rounded down) in the causal graph and add their presence to the qualitative domain knowledge.

(e) Randomly choose a nontrivial target effect and $p_{probe} \cdot n^2$ (rounded down) other treatment-outcome pairs that will serve as quantitative probes.

(f) Calculate the corresponding ATEs for the target effect and the probes from the fully specified Causal Bayesian Network.

3. Causal analysis: Run a causal end-to-end analysis (cf. Chapter 4), using the m observational samples, the qualitative domain knowledge, fast greedy equivalence search, the do-calculus and linear regression. If edges remain unoriented after the causal discovery step, their orientation is chosen randomly with equal probability.

4. Evaluation:

(a) Report the discovered causal graph, the estimate of the target effect and the hit rate for the quantitative probes.

(b) Report the number of edges that differ between the true and the discovered graph, as well as the absolute and relative error of the target effect estimate.

We refer to the execution of these steps as a *run*. By aggregating the results of many runs that all share the same parameterization, we accumulate the necessary statistics to judge the effectiveness of the quantitative probing strategy.

7.1.3 Parameter choices

In the next sections, we report the results for runs with the following parameter choices:

1. $n = 7$ nodes are used in the DGP to ensure that the scenario is sufficiently challenging for the causal analysis, but not needlessly complicated. The graphs provide sufficient options for backdoor paths, while still being easy to visualize and investigate.

2. An edge probability of $p_{edge} = 0.1$ produces graphs that feature on average $0.1 \cdot 7^2 = 4.9$ (allowing for loops) or $0.1 \cdot 7 \cdot 6 = 4.2$ (excluding loops) edges. Note that selection bias towards fewer edges is introduced by our approach to resample when a cycle appears. On the other hand, the constraint that there are no isolated nodes produces a bias towards more edges, since at least 4 edges are needed to avoid isolated nodes.

3. For each DGP, $m = 1000$ samples are generated, such that sufficient data is available to the causal discovery and estimation algorithms.

4. $p_{hint} = 0.3$ ensures that we pass 30% (rounded down) of the correct causal edges as domain knowledge to the causal discovery procedure. This choice is supposed to mimic real applications where domain experts will be able to determine the presence and orientation of some, but not all possible edges.

5. Analogously, $p_{probe} = 0.5$ ensures that we use half of the possible causal effects as quantitative probes because not all of the effect strengths will be known to domain experts in real applications.

6. By choosing $\epsilon_{probe} = 0.1$, we consider probe estimates successful if they deviate no more than 0.1 from the true value on an additive scale. To put the magnitude of the number into context, we note that all ATEs in the binary setting lie between -1 and 1 such that a maximal discrepancy of 2 between true and estimated ATE is theoretically possible. The additive scale, as opposed to the relative one, is chosen to avoid numerical instabilities for ATEs close to 0.

7.2 Software contributions

The implementation of the setup mainly rests on two open-source Python packages that were developed by the author during his PhD studies:

1. Each causal analysis is performed via the aforementioned cause2e package [4], enabling the sequential execution of causal discovery, identification and estimation. Even the quantitative probing functionality is already built into cause2e, as the idea was created during the early development phase of the package. Internally, the package relies on pycausal [56], which is a Python wrapper around the popular TETRAD Java application [57], to perform the causal discovery step, whereas identification and estimation are delegated to DoWhy [58]. Cause2e, however, is only designed for carrying out independent causal analyses and does not provide any functionality for benchmarking the validation strategy itself by aggregating the results.

2. In order to evaluate the success of the overall strategy, the qprobing package was created [5]. It internally uses cause2e to implement the above described simulation setup and furthermore provides methods for filtering and visualizing the results by incorporating additional third-party open-source libraries. For the ground truth generation, random DAGs are created using networkx [88] and extended to fully parameterized Causal Bayesian Networks using pgmpy [89]. After all the runs have been performed, the results are displayed in plots based on matplotlib [90].

Both packages are freely available, open-sourced, thoroughly tested by continuous integration techniques on a regular basis, and documented, in order to ensure their usability by other researchers.

7.3 Results

In this section, we discuss the results of a simulation study that consists of 1378 runs following the setup in Section 7.1. Initially, 2200 runs were performed, but 793 of them were aborted because of problems in the modelling process that would have required manual intervention. Of the remaining 1407 runs, 29 were excluded because the true causal effect was so close to zero that numerical instabilities occurred during the calculation of the relative estimation error.

As predicted above, our selection for acyclic DGPs introduced a bias on the edge structure in the underlying causal graph that brought the mean number of edges to 6.0, as opposed to the theoretical estimates of 4.9 (including loops) or 4.2 (excluding loops).

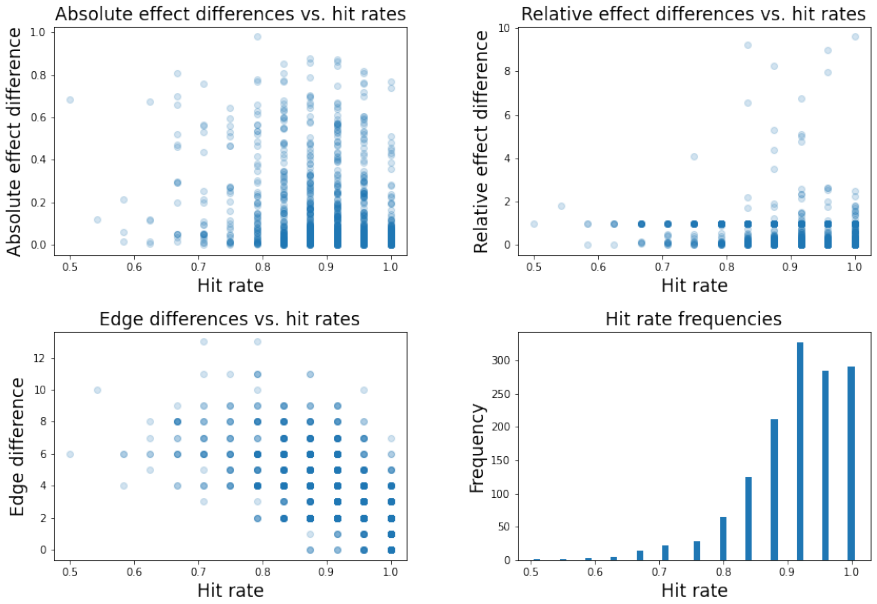

Fig. 7.1: The top row shows plots of the absolute (left) and relative (right) differences between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the structural hamming distance between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right). In the three scatterplots, the number of overlapping points is indicated by the level of opacity.

Figure 7.1 visualizes the relations between the entities in the hypotheses H1 and H2 in three separate plots. Contrary to the hypotheses, the expected downward trend in estimation error or SHD with increasing hit rate is not visible. In the two plots relating estimation error and hit rate (top row), it even seems that increasing hit rates lead to a higher estimation error. At least this concern can be explained by the bottom right plot in Figure 7.1: The vast majority of the runs achieved a hit rate of at least 0.8, such that the apparent increase in runs with a higher estimation error is simply due

to the overall increase in runs. An explanation for the high hit rate lies in the selection of the probes: In contrast to the target effect, which was constrained to be nontrivial, the probes were allowed to be trivial. By our parameter choice $p_{edge} = 0.1$, the causal graphs are only sparsely connected and most effects are indeed trivially zero due to the absence of directed paths between treatment and outcome. For such a probe, the estimation is successful for any analysis that does not mistakenly introduce a directed path during causal discovery. Given our parameter choice $p_{hint} = 0.3$, the causal discovery algorithm received sufficient qualitative domain knowledge to avoid such errors in most cases.

In order to address the main problem of the missing downward trend, we reduce the complexity of the plots by only showing the mean of each hit rate column in Figure 7.2. After this modification, the error plots for both target effect estimation and causal graph recovery show a clear downward trend, at least in the sufficiently populated hit rate regions. The behavior in these regions even looks linear, although there is no obvious theoretical justification for this observation (cf. Section 7.6.2).

While Figure 7.2, contrary to Figure 7.1, decidedly supports our hypotheses, we need to refer to a third visualization to account for the statistical variability of the simulation and analysis process. As Figure 7.2 only shows the means of the quantities, we enrich the contained information by the use of boxplots in Figure 7.3. Only hit rate columns with at least 20 data points are included in the boxplots, in order to avoid generating the impression of reliable error bounds for columns with insufficient data. The empirically determined quartile bounds, whiskers, medians and means still exhibit the downward trend, with the notable exception of the quantities in the relative estimation error plot (top right): The third quartiles and part of the medians are precisely 1, which represents the many cases where an estimation error was caused by the erroneous elimination of all directed paths between the target variables in the causal discovery step.

In summary, the simulation study provides convincing empirical evidence for both of our hypotheses: On average, a high hit rate in probe recovery serves as a reliable indicator for the success of an analysis in both target effect estimation and causal discovery. However, the reported success is only based on aggregate measures. For an individual analysis, there is still a risk that target effect estimation or causal discovery fail despite a high or even perfect hit rate, as is evidenced by Figure 7.1. In order to carefully assess the thereby evidenced shortcomings of the strategy, we will investigate exemplary failing

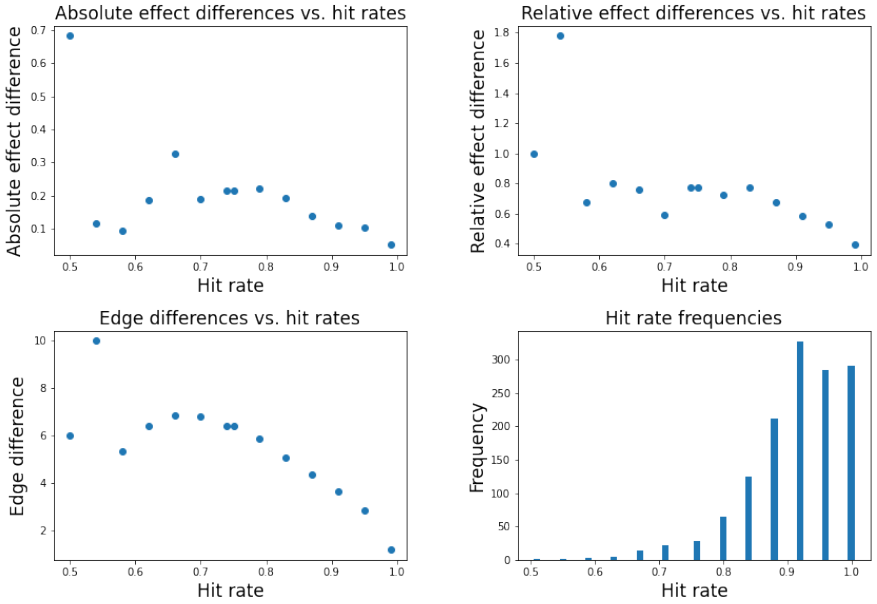

Fig. 7.2: Aggregated results (means only): The top row shows plots of the mean absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the mean structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right).

runs more closely in Section 7.4 and explore the reasons for the unexpected behavior. To complement this evaluation, Section 7.5 provides a guide for practitioners that serves as a reference for applying the strategy safely to real-world causal inference tasks.

7.4 Outlier analysis

In this section, we want to look more closely at runs that simultaneously display a high hit rate and bad performance with respect to the target effect estimation. To make this more precise, we define an *outlier* to be a run with a perfect hit rate of 100 % and an absolute estimation error of at least 0.2. The strict hit rate bound reflects our belief that models cannot be trusted if they fail any falsification test. The error bound is admittedly a rather

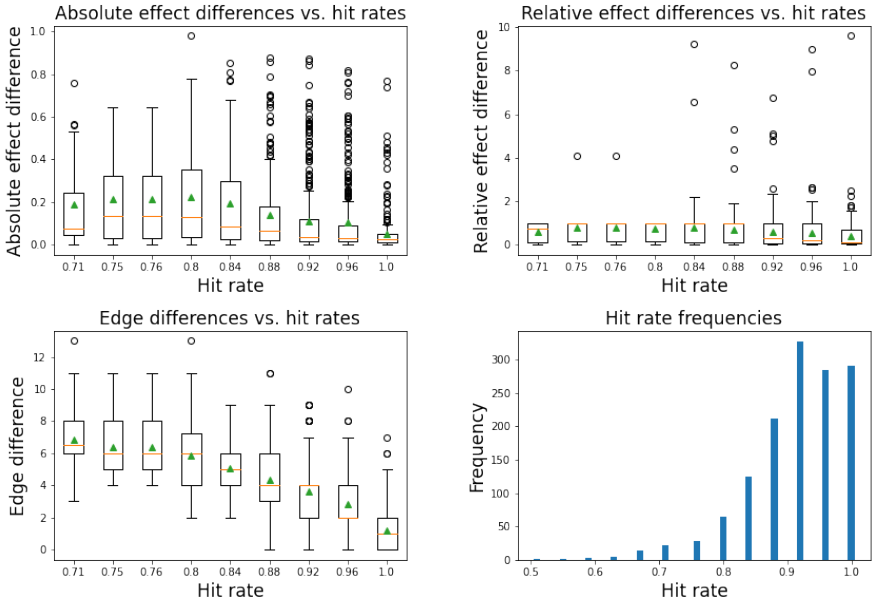

Fig. 7.3: Aggregated results: The top row shows boxplots of the absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a boxplot of the structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right). For each box in the boxplots, the green triangle indicates the mean whereas the orange line indicates the median. The lower and upper bounds of the boxes indicate the first and third quartiles, respectively, and the whiskers around the boxes use the standard interquartile range scaling factor of 1.5. Points that lie outside of this range are plotted as singular outliers. Only hit rate columns with at least 20 data points have been included in the boxplots.

arbitrary value, given that there is no hypothetical domain background for our simulated DGP that could provide a task-specific threshold for considering an estimation attempt a failure. To put the magnitude of the number into context, we note that all ATEs in the binary setting lie between -1 and 1, such that a maximal discrepancy of 2 between true and estimated ATE is theoretically possible. If we filter the dataset in the simulation study accordingly, 15 of the 1378 runs (highlighted in Figure 7.4) remain to be further investigated.

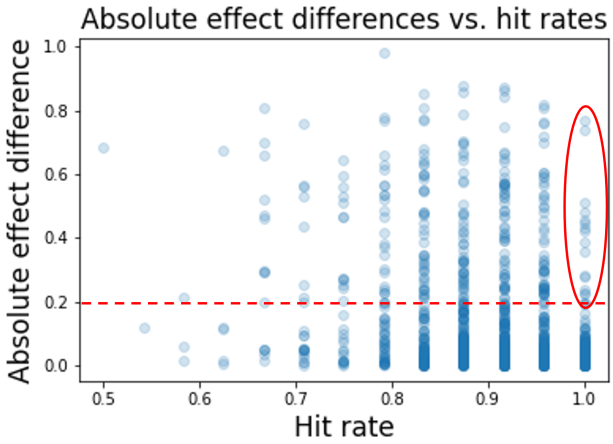

Fig. 7.4: Using an absolute estimation error threshold of 0.2 (dashed line) and a hit rate threshold of 1.0 (rightmost column), 15 outliers (circled) are selected for further investigation.

7.4.1 Connectivity

The true and discovered causal graphs of the first outlier are shown in Figure 7.5. Although the true target ATE of x_3 on x_5 is 0.51, the causal analysis has estimated it to be precisely 0. Looking at the graphs, this can already be explained by the absence of directed paths from x_3 to x_5 in the discovery result. Although the discrepancy between estimated and true value is large, all of the probes have been correctly recovered: This is possible because the causal graph has two components, whose respective skeletons are identical between ground truth and discovery result. For the larger component, not only the skeleton, but also the orientation of all edges has been correctly recovered. Therefore, all probes that include at least one variable from the larger component seem to corroborate the model. As the smaller component contains only the two variables that appear in the target effect, there are no probes left to select that would lie entirely in the smaller component. This explains why perfect probe recovery has been achieved. In summary, the strategy has failed due to the probes and the target effect being entirely unrelated.

In order to verify that the empirical findings in the previous section are still valid for connected causal graphs, we filter out all runs with DGPs

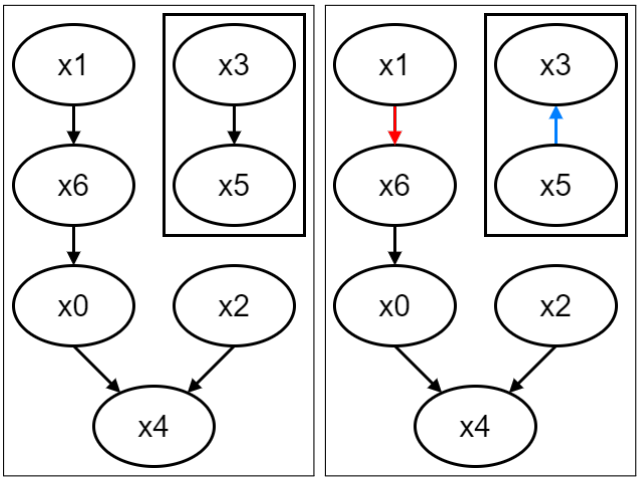

Fig. 7.5: Connectivity: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_3 and outcome x_5 (surrounded by small box). In the discovered graph, the red edge $x_1 \to x_6$ has been required by domain knowledge. The edge $x_3 \to x_5$ (blue) has been oriented incorrectly.

based on disconnected graphs and visualize the results for the remaining 653 runs in Figures 7.6, 7.7 and 7.8. The plots look largely similar to those in the corresponding Figures 7.1, 7.2 and 7.3. Above, our selection for acyclic DGPs without isolated nodes introduced a bias on the edge structure in the underlying causal graph that brought the mean number of edges to 6.0, as opposed to the theoretical estimates of 4.9 (including loops) or 4.2 (excluding loops). By selecting only connected graphs, we expect denser structures, which is confirmed by recalculating the mean number of edges to be 7.0. For the sake of completeness, we report the mean number of edges for graphs with more than one connected component to be 5.1. Since 11 of the 15 initially identified outliers have also been filtered out, we need to find additional explanations only for the 4 remaining ones.

7.4.2 Probe coverage

One of the remaining outliers illustrates an issue that is bound to appear in practice: We do not have unlimited quantitative domain knowledge, such that not every cause-effect pair can be used to formulate a quantitative probe. Looking at Figure 7.9, the target effect of x_5 on x_6 has been erroneously

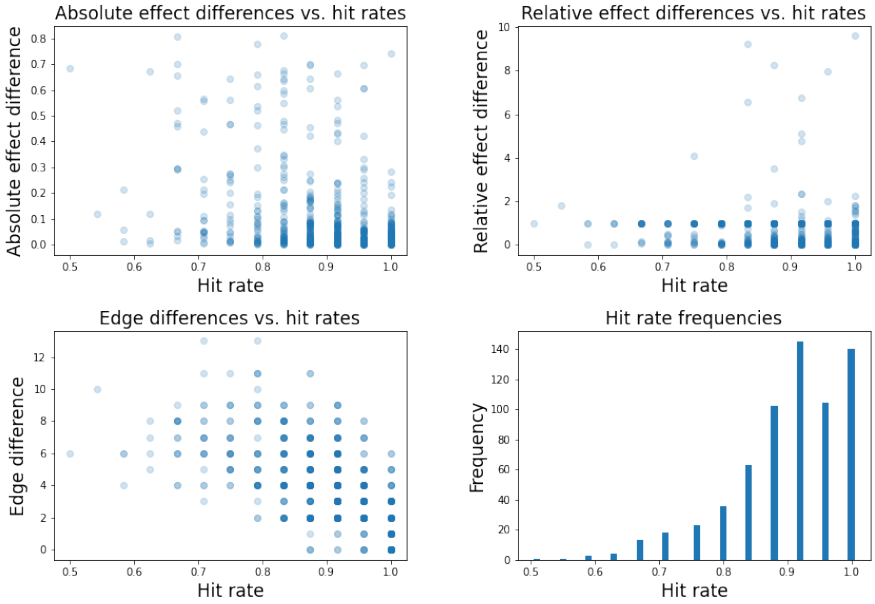

Fig. 7.6: Results for all runs with connected causal graphs: The top row shows plots of the absolute (left) and relative (right) differences between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the number of structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right). In the three scatterplots, the number of overlapping points is indicated by the level of opacity.

estimated to be 0 instead of 0.28. The discrepancy is again due to an edge reversal that leads to a trivially vanishing estimate. Contrary to Figure 7.5, the target variables are connected to the rest of the causal graph, such that we can hope to falsify the candidate model by a suitable probe. A manual inspection shows that a probe that would detect the error is given by the ATE of x_3 on x_6, which is trivially zero in the candidate model and non-zero (in the non-degenerate case) in the true model. However, our simulation setup was designed to include only half of the non-target effects as probes, in order to simulate more realistic conditions, and this probe was not selected. The same holds true for the other relevant probes given by the ATEs of x_6 on x_5, x_6 on x_1, x_5 on x_2, x_5 on x_4, x_3 on x_2 and x_3 on x_4, respectively, such

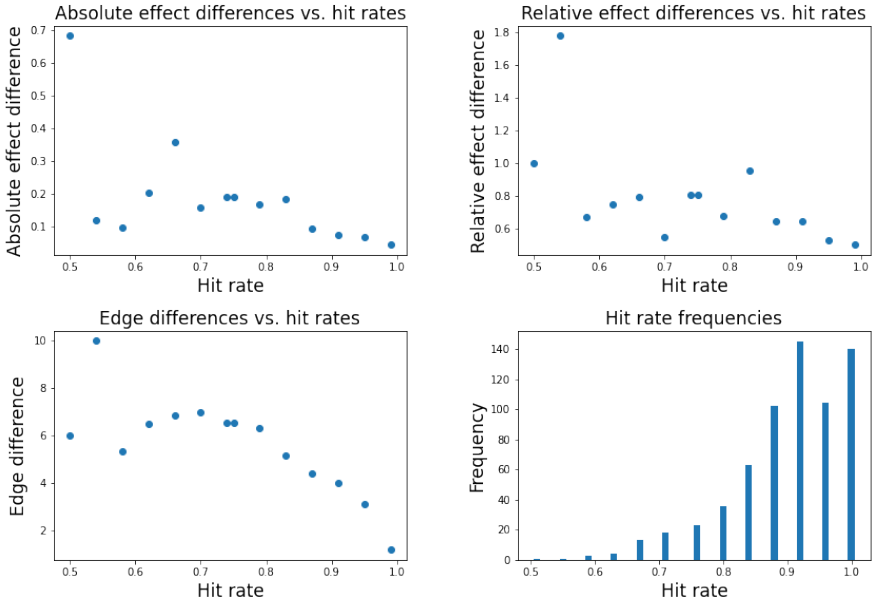

Fig. 7.7: Aggregated results for all runs with connected causal graphs (means only): The top row shows plots of the mean absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the mean structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right).

that the selected probes were not suitable for detecting the error. It seems unlucky that all of these probes have not been selected, but with the high number of runs some of these cases are bound to appear: The probability of not selecting all 7 relevant probes is $p_{probe}^{7} = 0.5^{7} \approx 0.78\%$, such that our 653 runs with connected causal graphs should produce roughly 5 such cases if we assume a constant number of relevant probes. Indeed, one of the remaining three outliers can also be fully explained by not having any of its relevant probes selected. We can therefore deduce that it is important to provide sufficient coverage of the model by quantitative probes.

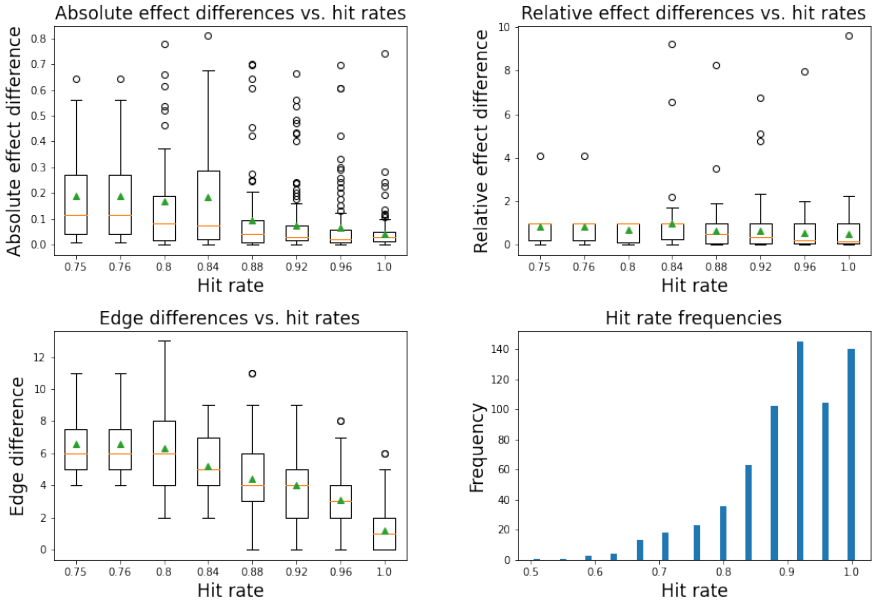

Fig. 7.8: Aggregated results for all runs with connected causal graphs: The top row shows boxplots of the absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a boxplot of the structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right). For each box in the boxplots, the green triangle indicates the mean whereas the orange line indicates the median. The lower and upper bounds of the boxes indicate the first and third quartiles, respectively, and the whiskers around the boxes use the standard interquartile range scaling factor of 1.5. Points that lie outside of this range are plotted as singular outliers. Only hit rate columns with at least 20 data points have been included in the boxplots.

7.4.3 *Probe tolerance*

Even in cases where we have enough probes and they are related to all important parts of the DGP, the success of the procedure is not guaranteed. For the model in Figure 7.10, the target ATE of x_1 on x_2 was estimated to be 0 instead of -0.24. Similarly to the last outlier, manual inspection reveals that the ATE of x_4 on x_2 must differ between the two models. Although the probe

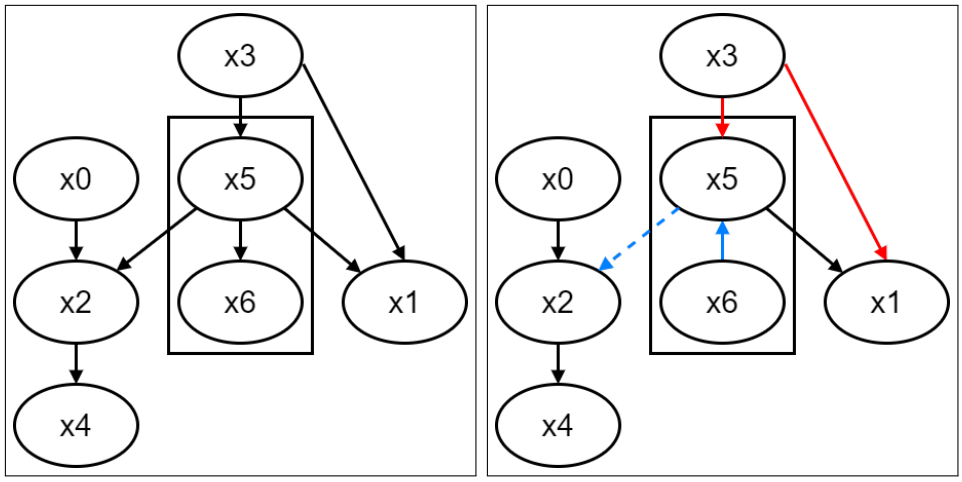

Fig. 7.9: Probe coverage: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_5 and outcome x_6 (surrounded by small box). In the discovered graph, the red edges $x_3 \to x_5$ and $x_3 \to x_1$ have been required by domain knowledge. The edge $x_5 \to x_6$ (blue) has been oriented incorrectly and the edge $x_5 \to x_2$ (blue dotted) is missing.

has been selected in this case, the incorrect causal model still has not been detected. A closer examination of the probe reveals the problem: The true ATE of x_4 on x_2 is indeed nonzero, but its absolute value of 0.07 is smaller than our selected tolerance parameter ϵ_{probe}. Therefore, the estimated value of 0 lies within the acceptance interval of $(-0.03, 0.17)$. This example illustrates the importance of the tolerance parameter for the overall effectiveness of the strategy. Similarly, the relevant probes given by the ATEs of x_4 on x_5, x_1 on x_0 and x_1 on x_4, respectively, have suffered from too lenient probe tolerances. On the other hand, the previously discovered probe coverage issue also comes into play, as the ATEs of x_1 on x_5 and x_1 on x_5 have not been selected in the first place and have therefore not detected the misspecified model. For the last remaining outlier run, a similar mixture of probe coverage and probe tolerance issues has lead to the failure of the validation strategy.

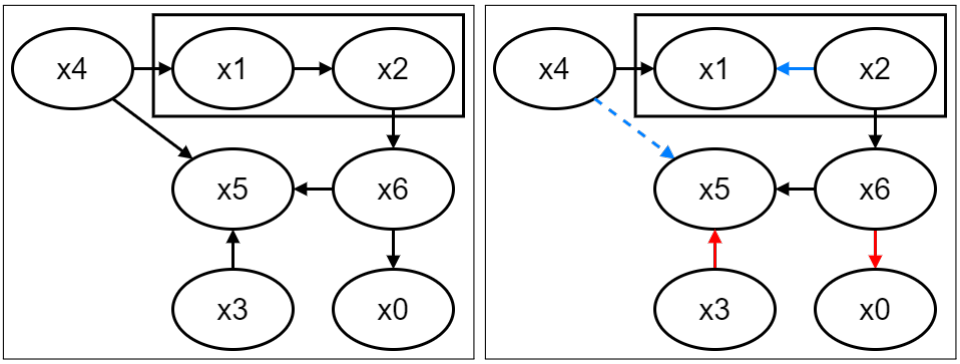

Fig. 7.10: Probe tolerance: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_1 and outcome x_2 (surrounded by small box). In the discovered graph, the red edges $x_6 \to x_0$ and $x_3 \to x_5$ have been required by domain knowledge. The edge $x_1 \to x_2$ (blue) has been oriented incorrectly and the edge $x_4 \to x_5$ (blue dotted) is missing.

7.5 Practical considerations for quantitative probing

Before we summarize the findings of the above sections and formulate questions for further research, we want to briefly revisit the individual steps of the quantitative probing strategy from a practitioner's viewpoint. Figure 6.1 and the accompanying stepwise description in Section 6.3 can serve as a rough guideline for causal modellers, but in order to enable a fruitful realization of all the proposed steps, it is necessary to provide additional advice on how to apply each step in practice.

7.5.1 Probe selection and specification

The quantitative probes are the central element of the validation strategy. As we have seen in Section 7.4.2, it is important to provide a sufficient number of them. This is also evident from the falsificationist point of view that motivates the quantitative probing strategy: Each probe provides another falsification test that has the potential to detect a misspecified causal model before we use it for the target estimation and risk downstream damages. However, the exemplary outlier in Section 7.4.1 shows that it makes no sense to blindly maximize the number of probes: In cases where the probes are not sufficiently

connected to the variables in the target effect, they fail to uncover problems in the decisive region of the causal graph. It is probable that the usefulness of a probe could be judged by different metrics, such as its distance to the target variables, but there are no studies to the knowledge of the author. Although we might be tempted to say that additional probes, even if ineffective, can never lead to a worse performance of the overall validation procedure, there are substantial arguments in favor of a careful probe selection: Firstly, the hit rate is no longer a reliable figure of merit if the probes are too weak. In Section 7.4.1, we have witnessed an analysis with perfect hit rate and substantial target estimation error. If practitioners still prefer to use all their available domain knowledge, they can overcome this obstacle by weighting the contribution of each probe to the hit rate individually. Secondly, it is evident that probe selection and specification cannot be performed independently of each other: There is no use in specifying probes if we cannot obtain their true values, and each additionally selected probe requires work to find its true value. The amount of work depends on the procedure that is used to find this true value, and there are two main possibilities:

Experimental data

After the distillation of the falsification targets, we can collect additional data, in order to compare the empirically measured causal effect to the one that is entailed by the causal model in question. Although it seems counterintuitive to avoid a direct experiment on the target effect, only to then validate the surrogate causal model using even more experiments, there are scenarios where this approach is justified: It is conceivable that the data collection for testing the derived statements is possible at a considerably reduced cost, when compared to performing the original experiment that we are determined to avoid.

Quantitative domain knowledge

In other cases where experiments for obtaining the true values of the derived statements are costly or even impossible, we can frequently resort to the preferred route of using expert knowledge: Although it might be impossible to evaluate the natural direct effect of gender on recovery by the use of a randomized controlled trial, medical experts could provide us with the answer in some cases. These cases are desirable, since we do not have to perform

any experiments anymore, with the minor disadvantage of having to find an expert and occupying their time. In order to distinguish this form of expert knowledge from other forms, such as the knowledge about the presence or absence of specific edges in the causal graph, we refer to any knowledge about the numeric causal effects as *quantitative domain knowledge*, whereas any knowledge about the discrete nature of the causal graph will be summarized under the term *qualitative domain knowledge*.

Despite the apparent difference in form and intended use of the knowledge items (cf. Section 5.2), it is plausible that the same persons will be experts for both qualitative and quantitative domain knowledge. This leads to synergy effects in the process of building and validating the causal model if a graphical modelling approach is used: The expert for qualitative domain knowledge is already required for partially or fully specifying the relevant edges in the causal graph. If the same person is able to answer our questions about quantitative causal effects, the additional effort for quantitative probing is minimized. It is particularly worth noting that the fear of having to keep the domain expert around while the model is fitted is baseless: Even though we have to provide the fully specified causal model for the comparison of the expected and predicted values of the quantitative probes, their initial formulation does not require any model. We only need to know which variables will be included in the final model, but this is already a prerequisite for eliciting the qualitative domain knowledge, such that no additional constraints arise by applying quantitative probing. However, it is of course beneficial to reconvene with the domain expert at the different stages illustrated in Figure 6.1, e.g. when a refutation test fails and we suspect the error to be on the probe side instead of the modelling side.

7.5.2 Modelling

Building a causal model refers not only to constructing the correct causal graph, but to all the steps that we need to perform in order to arrive at a causal model that can predict both the target effect and the probes. It is worth noting that quantitative probing can be seen as a model-agnostic validation method: There is no need for us to know the internals of the model. For example, it is irrelevant whether the model is a causal graph with supplementary observational data and an estimation strategy, a causal Bayesian network, a fully specified structural causal model or even a non-graphical potential outcomes model. Neither does it matter how we arrived

at the final model. In the case of a graphical model, we do not need to know whether the graph has been prescribed by domain knowledge, derived purely from the data via causal discovery, or resulted from a combination of both approaches. All that we expect from a model that is suitable for quantitative probing is the ability to predict both the target effect and the probes. This is of great practical advantage, since it allows researchers to independently modify the internals of the models and the validation procedure, as long as the interface between the two entities remains unchanged. Therefore, we refrain from recommending particular modelling strategies, as this step is not a part of the validation itself.

7.5.3 Probe prediction

After a model has been provided, we query it about the values of certain causal effects, which are given by the quantitative probes. Again, the model can be treated as a blackbox for this step. In order to adapt this point of view, it is necessary to consider the concrete prediction strategy an internal detail of the model itself. A causal graph together with observational data could otherwise give different predictions for the causal effects, depending on whether we use linear regressions or more sophisticated estimators. This does not diminish the previously mentioned benefits of causal models with respect to explainability and interpretability, and neither does it free the modeller from choosing an appropriate estimation technique. It only signals that the validator does not need to know the details of the model, which leads to a desirable separation of concerns, comparable to the clean division between causal discovery, identification and estimation steps within graph-based modelling itself.

7.5.4 Probe evaluation

This step seems to be the simplest one. In order to evaluate whether a causal effect has been correctly predicted, we only need to check whether the predicted value of a quantitative probe is in accordance with the one that we specified at the beginning of the validation. However, it is not trivial to define what we mean by *accordance*. In most cases, it is unrealistic to expect a perfect match between true and predicted value. The discrepancy might arise due to random statistical fluctuations, slight misspecifications of the true value or even imprecisions in floating point manipulations. Acknowledging

these hindrances as unavoidable, it is necessary to specify a tolerance within which a quantitative probe is judged to be correctly recovered by the model. The tolerance specification should ideally be a part of the probe specification at the beginning of the process, in order to ascertain an unbiased evaluation of the probes (cf. the setup in Section 7.1.2). Nevertheless, we have chosen to discuss this hyperparameter here where we face the consequences of the decision. In Section 7.4.3, we have seen that a careless tolerance choice can render decisive probes and thereby the whole validation strategy ineffective. Therefore, we recommend selecting the tolerance individually for each of the probes. The exact value of each tolerance should be guided by the domain knowledge expert and can include specifications on both an absolute and a relative scale.

7.5.5 Model evaluation

The model evaluation inherits its difficulty from the underlying probe evaluations. Both the selection of an acceptance threshold for the hit rate and the considerate usage of quantitative probing for adapting or selecting causal models pose considerable challenges to practitioners.

Setting the acceptance threshold

On the one hand, the scientific method requires us to refute a model as falsified whenever it fails to correctly recover any of the quantitative probes. On the other hand, knowing about the possibility of ill-advised tolerance specifications, it seems pedantic to reject a model that is in accordance with our expectations for one thousand probes, while only violating one. In such a case, it might be preferable to attribute the discrepancy to an error on the side of the validating entity instead of the model. As a consequence, the acceptance threshold could be lowered, such that models that pass not all, but most of the falsification attempts, say 95 %, are considered trustworthy. Note that each softening of the standard is linked to a suspected weakness on the side of the domain expert. Therefore, it seems preferable to address these weaknesses directly by searching the domain knowledge for misspecified probes together with an expert and subsequently removing these, in order to avoid the need to lower the acceptance threshold. If we do not want to remove an uncertain probe completely, it is again preferable to specifically reduce the impact of this probe on the overall hit rate by employing weighted

hit rates. This strategy accounts for the uncertainty at the right place, as opposed to introducing an overall lenience via a reduced acceptance threshold on the hit rate.

Adapting rejected models

This is an optional step, depending on the goal of the analysis. If the model evaluation has judged the model to be suitable for predicting the target effect, we can directly skip it. Otherwise, we have two possibilities. First, we could accept that our candidate model has not passed the tests and therefore abandon the analysis, in order to avoid producing an erroneous estimate of the target effect. A second possibility, which is often preferable, both in academic and industrial applications, is a rework of the candidate model. This gives us another chance to pass the validation with the modified model. Machine learning practitioners might rightfully be wary of adapting a model after it has seen its benchmarking routines, since this allows for accidental overfitting or even malevolent data snooping: It is trivial to build a model that correctly predicts all the probes if the ideal values are already known in advance.

While it is true that the validation procedure becomes easier to pass with each iteration, we should keep in mind that we are not interested in the hardest possible validation procedure, but in the best possible model. Incorporating previous failures into building a better model is an integral part of all of science, just as classical Newton mechanics was replaced by the more powerful theory of quantum mechanics after newly observed "probes" were found to be in conflict with the old theory [91, 92]. The unsatisfactory alternative would be to give up after the first initial failure, until another researcher wants to probe a completely independently developed model. This means that we are forced to allow a model rework in practice, even at the necessary price of a diminishing effectiveness of the provided quantitative probes.

In order to mitigate the risk of gradually wearing out our probes, we can apply a procedure similar to the above discussed train/validation/test split (cf. Section 5.1.1), which we call *causal validation/test split*: After probe selection and specification, part of the probes, say 20%, could be hidden from the modeller until a candidate model passes all of the other probes and therefore no longer needs to be reworked, whereas the rest of them is used for model validation. The suitability of the final model is then judged only

by checking its performance at recovering the previously unknown probes. It would bring the same benefits as dividing the non-training samples into a validation set, which is used for iterative model building and selection, and a test set, which is used for evaluating the performance of the final model. Given that the probes are never directly used for model fitting, the *train* part is omitted. The causal validation/test split solves the problem of degrading probes, but it cannot answer what happens if the final model evaluation on the test set is unsatisfactory. In practice, we would still be forced to return to a rework stage. Another drawback is the necessity to temporarily ignore all domain knowledge in the test set, which happens at the cost of thinning out our valuable domain knowledge. Note that our simulation study has in no way addressed the possibility of model rework, since no suitable way to automatically adapt rejected models could be determined. Therefore, it is not possible at this point to recommend a safe procedure for dealing with rejected candidate models.

Model selection

A related idea to using quantitative probing for an iterative modelling process is using it for model selection: If we are presented with a set of candidate models, it is tempting to probe all of them and select the one with the highest hit rate. Since there is no possibility of actively overfitting to the probes, we can then use the best model for target effect estimation. However, it is important to keep in mind that the best candidate model could still be severely misspecified if the initial set of models was inappropriate to begin with: Among the blind, the one-eyed is king. In order to avoid a one-eyed causal model, it is necessary to validate the chosen model independently after it has been selected. In correlation-based machine learning, this is again achieved by using the validation set for model selection and the disjoint test set for estimating the unbiased performance of the chosen model. To the best of our knowledge, there exists no other model-agnostic and domain-sensitive validation strategy that could be recommended for this final validation step. Therefore, we advise against using quantitative probing for model selection. If it is unavoidable, the causal validation/test split could again be used as outlined above with the same drawbacks.

7.5.6 Target prediction

After having successfully performed all of the above steps, we have reached a point where we decide that our causal model can be trusted to correctly predict the actual target effect. Note that probe prediction and target prediction can be carried out simultaneously if this yields computational benefits, as long as we are disciplined enough to not use the target estimate prematurely. Similarly to the step of probe prediction, the concrete mechanism that is used for target effect prediction is part of the causal model, as opposed to being a part of the validation procedure, such that no further discussion is needed here.

7.6 Discussion and open questions

In summary, the simulation study has provided compelling evidence for our hypotheses H1 and H2: High hit rates at probe recovery, on average, indicate good performance of the candidate model at target effect estimation and causal discovery. However, the outlier analysis in Section 7.4 has illustrated that passing all falsification attempts is no guarantee for having found an adequate causal model. Following the structure of the outlier analysis, Figure 7.11 recapitulates how a probe needs to be relevant (e.g. connected to the target effect), selected and equipped with the appropriate tolerance to detect a modelling error.

Furthermore, the study has evaluated only a specific setup with binary DGPs, ATEs as causal effects, causal end-to-end analysis with fast greedy equivalence search and linear regression as modelling strategy (cf. Section 7.1.2), as well as specific parameter choices (cf. Section 7.1.3). Various extensions of the discussed study are possible and we want to briefly explore different questions of interest for further research.

7.6.1 Parameter studies

A particular configuration of the parameters n, p_{edge}, m, p_{hint}, p_{probe}, ϵ_{probe} has been investigated (cf. Section 7.1.3). We believe that their values have been chosen appropriately with respect to probable conditions of practical application scenarios. However, similar studies could be carried out while varying one or multiple of these parameters to further investigate their respective influences on the overall success of the quantitative probing approach.

Fig. 7.11: A quantitative probe can only help detect a misspecified candidate model if it is relevant, selected and equipped with the appropriate tolerance. A successful validation refutes the incorrect model, whereas the model is failed to be exposed as untrustworthy if no error is detected by the probe. If multiple probes are employed, it suffices for any of the probes to detect an error, even if the rest of them raises no suspicions.

1. The parameters n and p_{edge} determine the nature of the DGP by prescribing the number of variables and the density of the causal graph. By varying them, an evaluation of the effectiveness of quantitative probing in DGPs of different complexity can be carried out.

2. The parameters m and p_{hint} seem to be of minor interest: The number

of samples is more likely to influence the performance of the modelling, as opposed to the validation strategy. Likewise, p_{hint} is a parameter specific to modelling strategies based on causal discovery and decoupled from the actual validation part.

3. The parameter p_{probe} plays an important role by determining the coverage of the DGP by quantitative probes. As discussed in Sections 7.9 and 7.5.1, it is vital to include both a sufficient number of probes and probes related to the target variables. By varying p_{probe} and replacing the currently uniform sampling distribution over the probes with a reweighted alternative, our qualitative observations can be quantified. The results will help practitioners in judging how many probes they need to gather and which probes to select if they want to maximize the effectiveness of the strategy subject to time and budget constraints.

4. The parameter ϵ_{probe} has great influence on the overall success of the validation procedure, given that it determines the acceptance bounds of each probe estimation. By varying it, the dynamics of balancing over- and underreject can be studied. Going one step further, tolerance thresholds based on relative error scales can be investigated, in order to avoid outliers as in Section 7.10. Simulation-based research following our recommendation of individually selected acceptance bounds (cf. Section 7.5.4) seems out of reach due to the required amount of manual bound specification.

7.6.2 Theoretical analysis

The above simulation study only provides empirical evidence for the effectiveness of quantitative probing. Strictly speaking, it is even limited to proving its effectiveness for modelling based on the concept of causal end-to-end analysis. Although the latter concern can be fixed simply by exploiting the model-agnostic nature of quantitative probing and repeating the analysis for different modelling approaches, settling the debate by a theoretical analysis would be optimal. Prompted by the approximately linear relation between hit rate and estimation/discovery performance in Figure 7.2, the author has attempted to follow this route but failed even for the comparably simple binary DGP setup. To put the goal into equations, the aim for hypothesis H1

is to prove a relationship of the form

$$\mathbb{E}(\Delta\tau|\eta) \approx \beta_\tau \cdot \eta \tag{7.2}$$

where $\Delta\tau = |\tau - \hat{\tau}|$ denotes the absolute estimation error, η denotes the hit rate and $\beta_\tau < 0$ describes the slope of the observed linear relation, or the analogous statement for a relative error measurement. Similarly, the aim for hypothesis H2 is to prove a relationship of the form

$$\mathbb{E}(\Delta G|\eta) \approx \beta_G \cdot \eta \tag{7.3}$$

where ΔG denotes the Structural Hamming Distance (SHD) between the true and the discovered causal graph. Using the tools of graph-based causal modelling, we describe the situation in Figure 7.12:

Even though some components, such as the discovery, identification and estimation strategies are kept constant in our simulation setup, the modelling of the remaining parts seems out of reach. In order to simplify the task and identify a crucial problem, we replace Figure 7.12 by Figure 7.13.

The main message of this illustration is that our three quantities of interest are causally linked by a fork structure: Both the target estimation error $\Delta\tau$ and the hit rate η depend on the quality of the causal graph, which is measured by ΔG. In our simple DGP, a misspecified causal graph is indeed the main source of estimation errors, aside from statistical fluctuations, although it is debatable whether this information is sufficiently represented by ΔG alone. However, even if we agree to all these simplifications, we still need to provide the two conditional probability distributions $p(\Delta\tau|\Delta G)$ and $p(\eta|\Delta G)$. A theoretical derivation of these expressions requires a complete understanding of how the misspecification of single edges changes the probability of correctly estimating some target effect. This is an unrealistic goal, since this heavily depends on the individual graph structure under consideration: The incorrect edge could introduce a new backdoor path, change a fork into a collider, or have no relevance at all for some causal effects. Therefore, the author could not think of a strategy to analytically derive the two CPDs. A different route is given by reexamining Figure 7.12: It might be impossible to model all the involved CPDs but we note that both $\Delta\tau$ and η depend on the same set of causal parents: The identified estimands prescribe the do-free expression to be estimated, and the estimation strategy carries out this task using the generated observational data. If we replace the continuous error term $\Delta\tau$ by the binary *target hit*

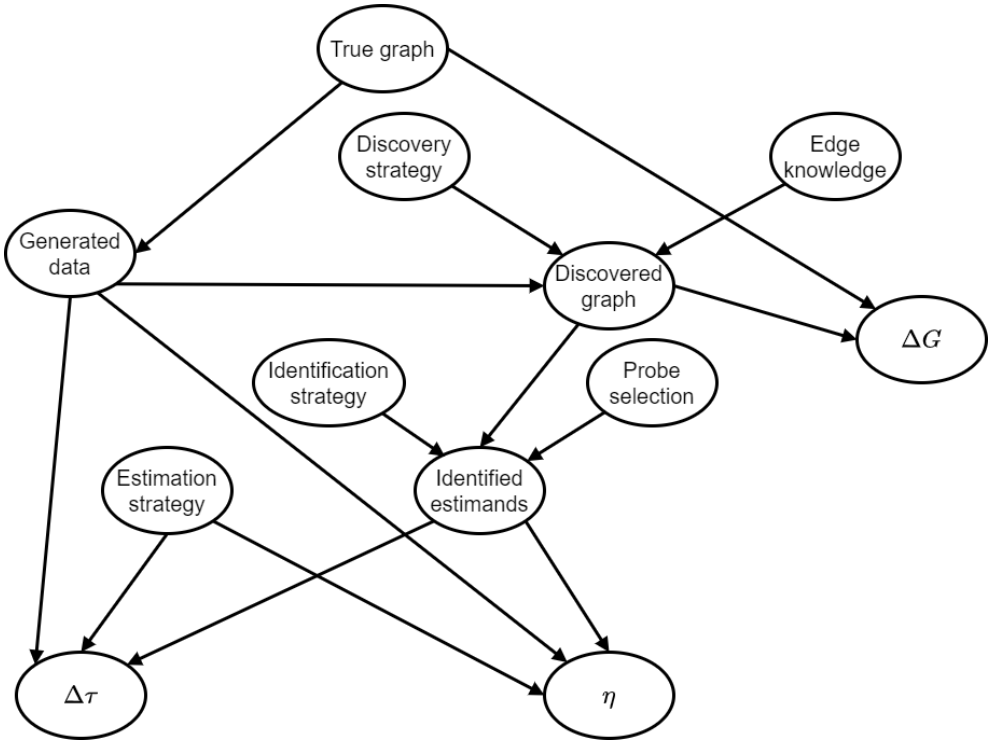

Fig. 7.12: Many variables contribute to the success of causal discovery, probe estimation and target effect estimation. Note that the estimation error $\Delta\tau$ and the hit rate η depend on the same set of causal parents: The identified estimands prescribe the do-free expression to be estimated, and the estimation strategy carries out this task using the generated observational data.

$$y_\tau = \begin{cases} 1 & \text{if } \Delta\tau \geq 0.1 \\ 0 & \text{else} \end{cases} \tag{7.4}$$

both the hit rate and the target hit are measures of success based on estimating one or multiple causal effects with an absolute error of less than 0.1. Similar to Figure 7.2, we can plot the mean target hits for each hit rate in Figure 7.14. Given the binary nature of y_τ, we refrain from including the non-aggregated plot and the boxplot.

The data in the most populated hit rate columns again seems to suggest an equation of the form

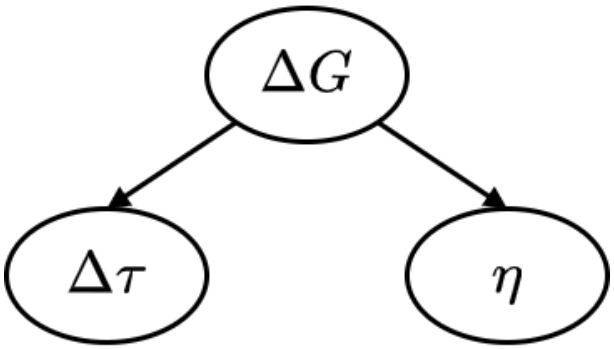

Fig. 7.13: The reduced version of Figure 7.12 symbolizes that the estimation of both the target effect and the quantitative probes depends on the quality of the discovered causal graph. As a measure of the latter, we use the SHD ΔG.

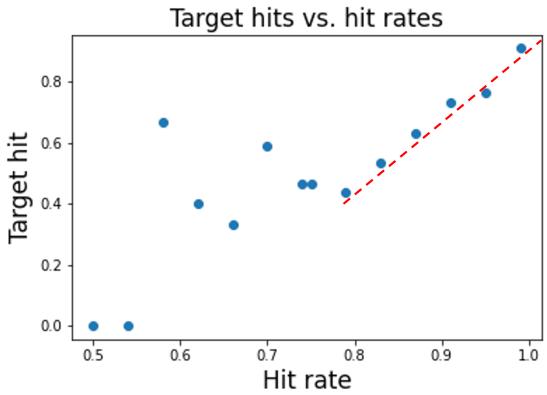

Fig. 7.14: Target hits (means only): As the hit rate increases, it is more likely to estimate the target effect with an absolute error of less than 0.1. As in Figure 7.2, the data for the most populated hit rate columns suggest a linear relationship between the two quantities (indicated by red line).

$$\mathbb{E}(y_\tau|\eta) \approx \beta_y \cdot \eta \tag{7.5}$$

and the above reasoning might even lead us to believe that $\beta_y = 1$ holds: Both the probes in the hit rate calculation, as well as the target effect are randomly chosen ATEs that need to be estimated with the same precision. However, a closer look at Section 7.1.2 reminds us that the target effect is

constrained to be nontrivial. Therefore, the probes that make up the hit rate are on average easier to recover than the nontrivial target effect, which explains why the fraction of target hits for each mean in Figure 7.14 tends to be smaller than the corresponding hit rate. It does not seem worthwhile pursuing closed form relations between the hit rate in the presented simulation study and other quantities: In our data, all probes share a tolerance of 0.1, whereas in practical applications all probes should be given specific tolerance bounds, as discussed in Section 7.5.4. Together with various biases that influence the probe or target effect selection, such as the above triviality criterion, an analytic justification of the strategy is likely out of reach. For the sake of completeness, we summarize our observations about the linear behavior in Tables 7.1 and 7.2: Table 7.1 holds the Pearson correlation coefficients $\rho_{\eta,\star}$ between the hit rate η and the hit rate specific means of a second quantity $\star$ ranging over the SHD ΔG, the absolute target estimation error $\Delta\tau$, the relative target estimation error $\frac{\Delta\tau}{|\tau|}$ and the binary target hit y_τ. The first row uses all the means displayed in Figures 7.2 and 7.14, including those in the sparsely populated hit rate bins. Since no weighting is employed, these unrepresentative means disproportionally decrease the scores and should therefore be filtered out by considering only the 6 hit rate bins that contain the most samples, leading to an almost perfect correlation.

	$\boldsymbol{\rho_{\eta,\star}}$			
	ΔG	$\Delta\tau$	$\frac{\Delta\tau}{\lvert\tau\rvert}$	y_τ
Unfiltered	0.80	0.55	0.66	0.82
Filtered	0.99	0.99	0.94	0.99

Tab. 7.1: Correlation coefficients (based on aggregated data): The Pearson correlation coefficient between the hit rate η and other aggregated quantities of interest (cf. Figures 7.2 and 7.14) is displayed. In the unfiltered version, all 14 hit rate specific means entered the calculation (top row). In the filtered version, only the 6 means with a hit rate of $\eta > 0.75$ were considered (bottom row). The unweighted contribution of the sparsely populated regions with a low hit rate disproportionally decreases the scores.

In order to remind us that the observed linear relation holds only for the aggregated measures and not or single runs, we repeat the procedure using singular runs instead of means as input and display the results in Table 7.2. As we have already visualized in Figure 7.1, the downwards trend is hardly visible on an individual level, with the exception of a rather weak correlation

between the hit rate η and the SHD ΔG. Therefore, we cannot make reliable statements about the success of a single analysis based on its hit rate, except for the case of a perfect hit rate where we have observed only few outliers and discussed strategies for avoiding them in Sections 7.4 and 7.5.

	$\boldsymbol{\rho_{\eta,\star}}$			
	ΔG	$\Delta\tau$	$\frac{\Delta\tau}{\lvert\tau\rvert}$	y_τ
Unfiltered	0.64	0.28	0.14	0.30
Filtered	0.62	0.26	0.14	0.29

Tab. 7.2: Correlation coefficients (based on non-aggregated data): The Pearson correlation coefficient between the hit rate η and other non-aggregated quantities of interest (cf. Figure 7.1) is displayed. In the unfiltered version, all 1378 runs entered the calculation (top row). In the filtered version, only the 1303 runs with a hit rate of $\eta > 0.75$ were considered (bottom row).

7.6.3 Comparative benchmarking

Keeping in mind the model-agnostic and knowledge-based nature of quantitative probing, it is possible to combine the strategy with both the model-specific and the domain-agnostic validation strategies presented in Section 5.1.2. Therefore, the author has focussed on establishing the credibility of quantitative probing, as opposed to benchmarking different validation strategies that work together rather than compete in practice. Nevertheless, it would be interesting to compare the effectiveness of the various possible combined validation strategies. The main problem in setting up such a benchmark is given by the drastically different demands that each validation strategy has with respect to the capabilities of the model under consideration: Whereas quantitative probing relies on the ability of the model to predict multiple causal effects, other strategies might not need non-target effects but other characteristic properties of the model. These might include:

- Marginal probabilities over the involved variables: Especially for generative models, such as Bayesian networks, the fitness of the model could be judged by comparing its entailed marginals with the actually observed data.

- Transition probabilities: If the model is to be used in a reinforcement learning setting (cf. Chapter 9), transition probabilities are closer to the intended downstream task than the above marginals.
- Graphical properties: For graphical models, the edge density, number of connected components, maximum number of parents per node and other descriptors of the graph structure are usable for comparing the model to prior expectations.

In order to achieve a fair comparison between validation strategies that test these criteria, the selected model type needs to be able to provide all of them to the validator. A fully-specified causal Bayesian network would fit this description, but this is most likely due to the author's focus on graphical models that has influenced the exemplary selection of the above properties. Benchmarking combinations of validation strategies on Bayesian networks might therefore favor strategies that perform well in this setting but are unsuitable for validating potential outcomes models.

Even if we accept this flaw, we could use different Bayesian networks as candidate models: Besides the parameter choices discussed in Section 7.1.3, we could use randomly generated candidate models or ones that result from causal learning strategies such as the end-to-end analysis. While random models might serve the purpose of uniformly covering the space of models, the validations are more likely to be applied to learned models. Another point of debate is the question on what basis a candidate model is to be evaluated as trustworthy in the ground truth itself: Is it enough to correctly recover the target effect, even if the marginals or the graph structure are incorrect? Extending our Hypothesis 1 in Chapter 7, we can replace not only the quantitative probes as our refutation criterion, but also the recovery of the target effect as our ultimate goal. In a general form, we can then ask how closely certain properties of a causal model are related to each other, not only how closely the quantitative probing hit rate is related to successful target effect estimation.

The author has attempted to tame the complexity of the many involved entities and their interfaces by employing class diagrams from software engineering [93], as shown in Figure 7.15. For simplicity, we only show the highest organization level of a possible benchmark study: A ModelGenerator generates both the true and the candidate model, the latter of which is then validated using the validation strategy under consideration, an instance of the Validator class. In order to determine the desired result of the validation,

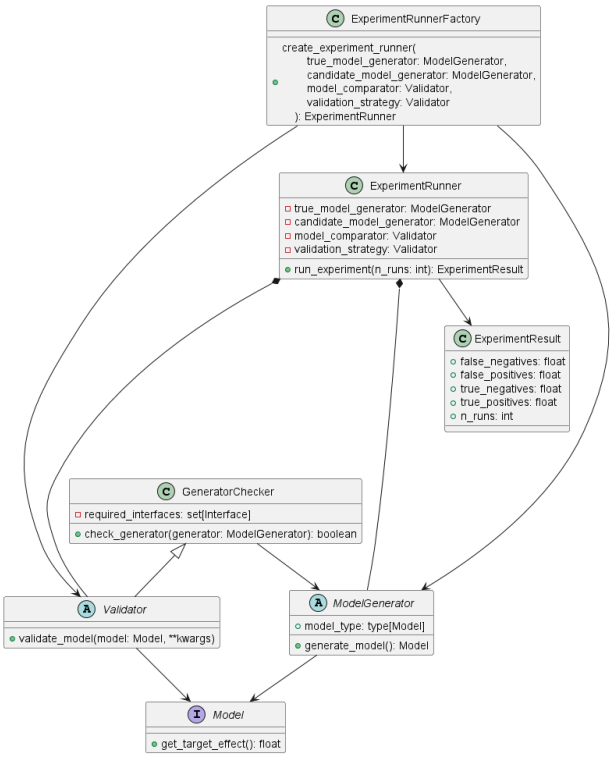

Fig. 7.15: The design for a benchmarking of validation studies involves multiple entities, all of which are to be fleshed out further, in order to create a prototypical implementation.

the two models must be compared using another instance of the Validator class. The GeneratorChecker ensures that the generated models provide all the interfaces that are needed by the Validators, such as the ability to answer queries about non-target effects in the case of quantitative probing. Although the author could not finish the implementation of the presented design due to time constraints, the topic is included in this thesis as a starting point for other researchers who are interested in a general comparative study of causal validation strategies.

Part IV

APPLICATIONS

Despite the author's focus on methods for performing and validating causal analyses, we will complement these theoretical considerations with a short overview of selected applications. The goal is to illustrate how the concepts can be applied to real problems, here shown at the example of industrially manufacturing light-emitting diodes (LEDs). Given that some of the data in the chosen use cases is sensitive and the main contribution of the thesis consists in the presented methodical developments, we will keep this excursion brief and highlight only the main points of each application. Firstly, we apply causal methodology to challenges in optimizing the color point of white LEDs that use phosphor conversion. These applications motivated the methodical developments in Parts II and III. Secondly, we describe ongoing efforts to combine the respective strengths of causal inference and reinforcement learning. The work in Section 8.3 and Chapter 9 was conducted as part of the Deep Thought project, which is a collaboration between ams OSRAM and Economic AI with the aim of studying and implementing novel methods of artificial intelligence for the optimization of dynamic and complex process chains. The Deep Thought project is funded by the Bavarian Joint Research Program (BayVFP) - Digitization (Funding reference: DIK0294/01).

8. LED COLOR POINT OPTIMIZATION

The first two applications of the presented causal inference methods aim at optimizing the color point of LEDs that use phosphor conversion to emit white light. Therefore, we will give a short introduction of the working knowledge that is necessary to understand the context of the analyses, before advancing to the description of the actual applications.

8.1 Background: White LEDs

Before we proceed to the applications, we introduce just enough working knowledge about white LEDs to understand the context of the following use cases. For further reading we recommend Schubert's seminal and comprehensive work on LEDs [94]. The active region of an LED is the *die* made out of semiconductor material. Although semiconductors are themselves a fascinating topic, we will ignore most of the underlying physics and treat the die as a blackbox that emits light according to a narrow wavelength distribution. White LEDs can be produced by manufacturing dice that emit blue light and adding *phosphor conversion material* around or on top of the die. When hit by a photon, some particles in the conversion material emit light following a broader wavelength distribution in the yellow part of the spectrum. Since not all photons hit these particles on their way, the human eye receives a mixture of light from the two wavelength distributions (cf. Figure 8.1), which is perceived as white.

The *CIE 1931* color space chromaticity diagram in Figure 8.2 can serve as a reference system for describing the resulting color in two-dimensional coordinates. By varying the amount of the applied conversion material, the manufacturer can achieve different colors, which results in a *conversion curve* in the color space that connects the color points of pure unconverted and completely converted emission. Snell's law [94]

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{8.1}$$

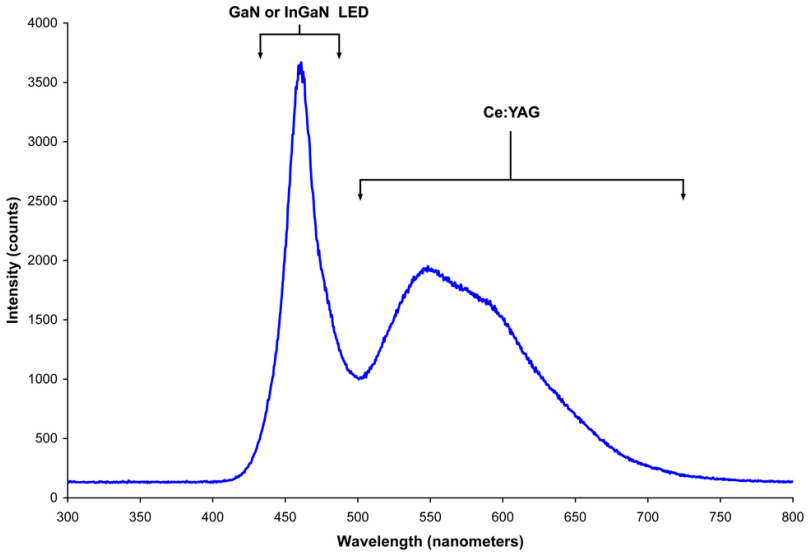

Fig. 8.1: Emission spectrum of a white LED, taken with permission from [95]: The left peak stems from the light that leaves the LED without interacting with the conversion material. The broader right peak in the yellow region of the spectrum represents converted light. Overall, the combination of both contributions is perceived as white by the human eye.

which relates the refractive indices n_1, n_2 of two materials and the incident angles θ_1, θ_2 of a light ray that passes the interface between the materials (cf. Figure 8.3), plays an important role for determining the exact mixture of blue and yellow photons: Depending on the involved refractive indices and angles, unconverted blue light can fail to pass the boundary of the conversion layer. Whenever this occurs, the chances of the photon being converted increase, which shifts the overall color point towards the yellow region of the color space and results in a warmer impression of the emitted white light. In summary, the LEDs under consideration consist of a blue chip and a phosphor conversion layer whose properties including its roughness need to be optimized, in order to achieve a desired white tone.

Due to the complex nature of the color point optimization process, the color of each LED is measured multiple times along the production chain. We will make use of the *inline color control (ICC)* measurement that occurs

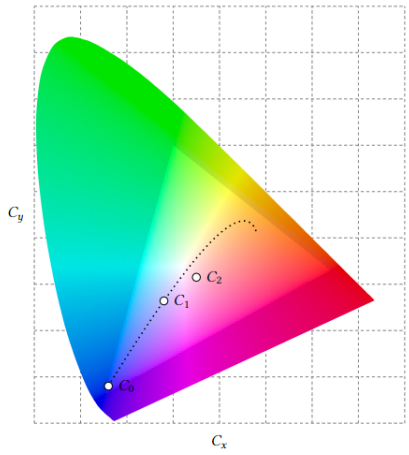

Fig. 8.2: The CIE 1931 color space chromaticity diagram allows to specify colors of interest in a two-dimensional coordinate system. Both coordinates between 0 and 1 are obtained by embedding the normalized three-dimensional response of the short, medium and long wavelength cone receptors of the human eye into the real projective plane [96]. The white point is at $(\frac{1}{3}, \frac{1}{3})$, whereas the regions of blue and yellow emission can be found outside of the center. Exemplary color point measurements before (C_0) and after conversion (C_1), as well as for the final LED (C_2) are included. The dotted conversion curve through C_0 and C_1 depicts the effect of applying a varying quantity of a given conversion material made from a green and a red phosphor component. The white triangle shows the gamut that is reachable by additionally varying the proportions of the two phosphor components in the conversion material. Unlike C_0 and C_1, C_2 does not lie on the conversion curve due to other processes influencing the color point of the LED at final testing. Figure taken with permission from [97].

immediately after phosphor deposition, as well as the *final testing (FT)* measurement that is performed before shipping the finished LED to the customer. The final testing measurement determines whether the LED can be sold and is therefore the figure to be optimized with respect to the envisioned color target. Based on these measurements, the percentage of LEDs that satisfy the customer's specifications within an agreed upon tolerance is referred to as *yield* and should be maximized.

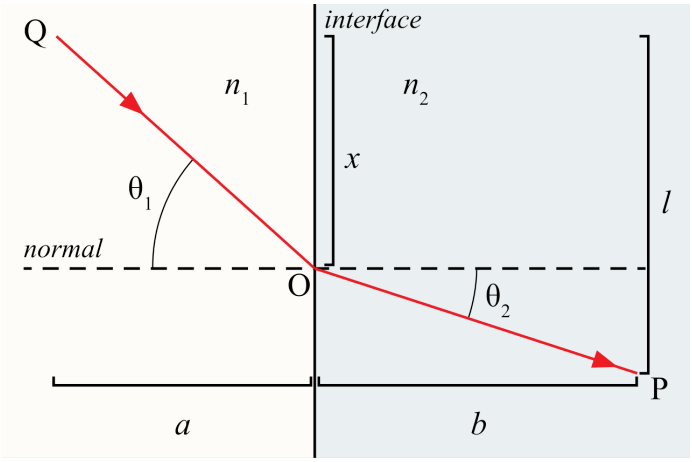

Fig. 8.3: A light ray emanating from point Q in a material of refractive index n_1 hits the interface to another material of refractive index $n_2 > n_1$ and is refracted according to Snell's law to reach point P. Figure taken from [98].

8.2 Color shift analysis

Hitting the color target at final testing is a delicate matter in LED manufacturing: Even if the target is hit perfectly after the aforementioned phosphor conversion step, further process steps are performed between the intermediate color measurement and its counterpart at final testing, which leads to a shift of the emitted spectrum and therefore a change in perceived color. There are two major options of dealing with this shift: On the one hand, we can find its root causes and eliminate them, in order to avoid the shift altogether. On the other hand, we can try to predict the shift under various circumstances and adjust our phosphor conversion target accordingly. Both alternatives require a causal understanding of the underlying mechanisms, such that a causal end-to-end analysis with the cause2e package was performed for a certain type of LED.

A company-internal prestudy suggested that the shift could be partially explained by a computational model if two properties of the conversion material, which we will call A and B, were included as inputs. These observations lead to the conjecture that properties A and B could be used to manipulate the color shift. Furthermore, properties A and B were experimentally confirmed

to have a causal influence on the roughness of the boundary between the conversion layer and the surrounding material. The roughness of boundary layers in LED manufacturing is known to be an important parameter due to its substantial impact on the light extraction efficiency of the device [99]. Process experts were therefore interested in knowing whether A and B were directly responsible for the color shift or only indirectly via the roughness of the conversion material. The latter possibility was supported by the fact that a silicon lens is molded on top of the conversion material in a subsequent process step (cf. Figure 8.4), such that the color point might be calibrated to the wrong boundary conditions at the interface between conversion material and surrounding material: According to Snell's law, the difference between the refractive indices of the conversion material and the surrounding air influences the outcoupling dynamics at the time of the conversion step. Therefore, this difference determines the amount of conversion material that must be applied, in order to reach the color target. However, the finished LED is equipped with a lens such that the refractive index of silicon instead of air determines the actual light outcoupling at final testing. Additionally, the weight of the applied conversion layer is expected to play an important role: For a conversion layer consisting of a fixed material, the weight is directly proportional to the thickness of the layer, which in turn influences the likelihood of a blue photon hitting a phosphor particle on its way through the layer.

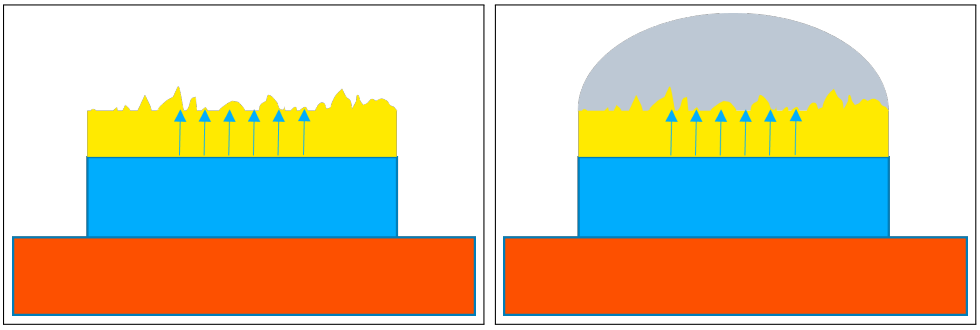

Fig. 8.4: At the phosphor conversion process step (left), the LED consists of a substrate (orange), the active material (blue) and the conversion material (yellow). The silicon lens is molded on top only later (right), such that it cannot be taken into consideration for calibrating the application of the conversion material.

The conducted causal analysis was required to determine not only average

treatment effects, but also natural direct and indirect effects (cf. Section 1.3.2), in order to enable a mediation analysis. Based on the above observations, two main groups of variables were selected for inclusion in the causal model:

- Color point measurements at the inline color control directly after applying the conversion material (ICC_x, ICC_y) and at final testing (FT_x, FT_y) served as the main measurements for detecting the color shift. The coordinates refer to the CIE 1931 depicted in Figure 8.2.
- The phosphor characteristics A and B, as well as the total weight W and surface roughness R of the conversion layer were included to describe the applied conversion material.

Subsequently, minor preprocessing was applied:

- The final testing measurements were normalized by the corresponding inline color control measurements, leading to their replacement by the deviations
$$DFT_x = FT_x - ICC_x$$
and
$$DFT_y = FT_y - ICC_y$$
in order to directly incorporate the color shift of interest. The inline color control measurements were not discarded, since a color point dependent color shift could not be excluded a priori.
- In order to bring the variables to a common scale, each variable X was replaced by its corresponding z-score
$$z(X) = \frac{X - \mathbb{E}[X]}{\sigma_X} \tag{8.2}$$
using the estimators from Equations (1.12) and (1.13).

Afterwards, qualitative domain knowledge was gathered in discussions with an expert. Some knowledge items, such as the constraint that the deviation measurements have no causal children, were obvious to the expert whereas others had to be distilled by iteratively applying causal discovery with the available knowledge and discussing the resulting graph proposals. Over several sessions with the expert, the knowledge graph in Figure 8.5 was distilled, which already eliminated most of the previously possible DAGs before causal

discovery. After applying fast greedy equivalence search, the causal graph in Figure 8.5 was used for the identification of suitable causal estimands for all possible average treatment, natural direct and natural indirect effects between the involved variables. Due to the lack of unoriented edges, no manual postprocessing was necessary.

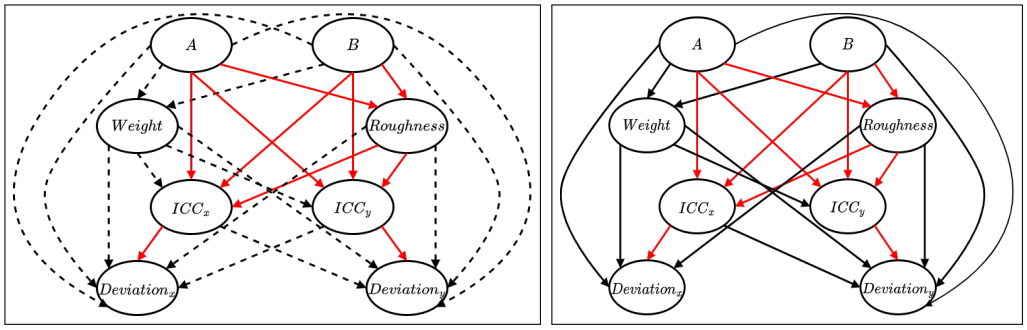

Fig. 8.5: Causal discovery using observational data and qualitative domain knowledge: The knowledge graph (left) indicates edges that are required by domain knowledge in red. Forbidden edges are omitted from the graph whereas the remaining possible edges are drawn dotted. The causal discovery result (right) contains all required edges (red), but only a subset of the possible edges (black) have been selected by the discovery algorithm based on the observational data. The densely connected graph underlines the difficulty of resolving all causal dependencies, which is necessary to arrive at an unbiased estimate of the target effects. Both graphs were created using cause2e's graph processing and visualization capabilities during the study, but redrawn by the author for this thesis, in order to achieve a clearer arrangement of the nodes.

The resulting estimands were then computed via linear regression, which corroborated the aforementioned experts' opinion on the color shift: The full output and exact numbers are not included for confidentiality reasons, but the format was the same as in the exemplary causal end-to-end analysis in Section 4.4. Parameter A had a large influence on the color shift in both directions, but a sizeable part of it was due to its influence on the roughness R, as was evidenced by comparing the pertaining overall, direct and indirect treatment effects. Parameter B turned out to be of minor importance. All the steps, including reading and preprocessing the data, organizing and visualizing the domain knowledge, executing the causal discovery and displaying its

result, identifying and fitting the unbiased regression models, as well as finally summarizing the results in a detailed report, were performed using the cause2e package.

Together with the domain expert, non-target effects were used to validate the model by an intuitive first version of quantitative probing. However, quantitative probing had not been developed as a principled strategy at the time of the study, such that cause2e offered no support for automated validation at the time. Therefore, manual formulation and evaluation of the probes led to an oversight that was only identified later on: The model predicted that an increase of the conversion layer weight would lead to a shift of the inline color control measurement in the blue direction. However, this effect is in sharp contradiction to the principle of phosphor conversion, which aims at shifting the color away from the natural blue color of the LED by adding a yellow peak to the emission spectrum. Consequently, the model could not be trusted and the communication of its possibly corrupted predictions about the target effects was avoided before further harm was caused by potentially changing production processes accordingly. Indeed, problems with the qualitative domain knowledge were identified that hinted at a misspecified graphical model. On the other hand, it is possible that the issue is related to the simple linear model that might have to be replaced by more flexible estimation strategies. In light of the remaining uncertainties about the model and the possible dangers of using it in production without thorough validation, the author decided to focus on developing proper validation strategies for causal models instead of further investigating the color shift problem. This decision resulted in the quantitative probing framework that is described in Part III of this thesis.

8.3 Color rework

The second application is again focussed on reaching a given white color target by combining a blue LED with a phosphor conversion layer (cf. Figure 8.1). After applying the conversion layer, an intermediate color measurement is performed, in order to control the success of the conversion step. In case of an unsatisfactory outcome, additional phosphor can be deposited on selected chips to further shift the color point along the conversion curve that is determined by the fixed conversion material. We will refer to this additional application of conversion material as the *rework step*. Perhaps surprisingly,

an exploratory data analysis showed that reworked chips were less likely to meet the customers' requirements at final testing than their non-reworked counterparts. Should we therefore conclude that the rework step should be removed from the processing chain? After all, it is costly and seems to worsen the outcome instead of bettering it. The attentive reader might already be thinking of Simpson's paradox (cf. Section 1.1), and rightfully so. It can be expected that the problematic chips that were selected for rework had worse chances of fulfilling the requirements, regardless of actually being reworked or not. Therefore, the criterion C that determines the rework decision acts as a confounder that influences both treatment (rework R) and outcome (yield Y) and we must account for it in our analysis. Putting this into proper notation, we should not use the conditional probabilities

$$p(Y|R) \tag{8.3}$$

for a fair assessment of rework effectivity, but rather the do-probabilities

$$p(Y|do(R), C = c) = p(Y|R, C = c) \tag{8.4}$$

where the last equality is justified by the causal diagram in Figure 8.6 and the corresponding causal Bayesian factorization

$$p(Y, R, C) = p(Y|R, C) \cdot p(R|C) \cdot p(C). \tag{8.5}$$

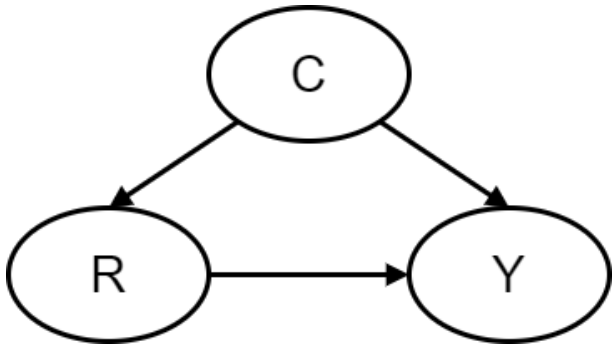

Fig. 8.6: The criterion C that determines the rework decision R simultaneously influences the yield at final testing (Y). Therefore, C is a confounder that must be adjusted for, in order to reach an unbiased causal conclusion.

In order to elucidate the nature of the criterion C, it is worth noting that chips show two sorts of deviations from the desired color point after the first conversion step (cf. Figure 8.7):

1. Along the conversion curve: The deviation can be explained by having applied too little or too much conversion material. Therefore, chips that show signs of insufficient phosphor application can be reworked until they reach the desired point on the conversion curve. On the other hand, chips that suffer from abundant phosphor application cannot be saved, since it is not possible to remove the material after hardening.

2. Orthogonal to the conversion curve: The deviation can be explained by problems that are not related to the conversion itself, provided that the conversion material has been properly crafted. Therefore, these problems cannot be mitigated by a rework step.

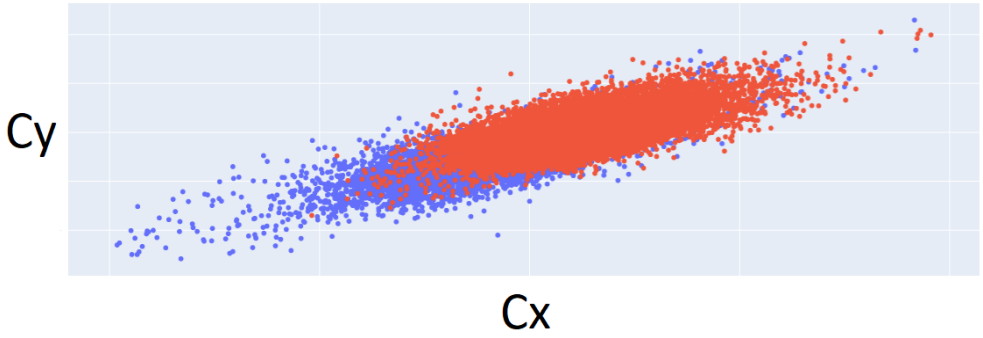

Fig. 8.7: The data lies on an ellipsis whose major axis corresponds to the conversion curve. Variation along this axis is captured by the first component of a principal component analysis. Samples selected for rework are colored blue whereas the remaining ones are colored red. The axes deliberately do not show units for confidentiality reasons.

It was confirmed that the deciders indeed used the deviation along the conversion curve as the single criterion for administering the rework step. Formalizing these observations, the criterion C is given by the first component of a principal component analysis (PCA) of the color measurement after the initial conversion step: The two main directions of variation in the data occurred along and orthogonal to the conversion curve, respectively. Given that the process is sufficiently optimized to produce colors near the target color point, the conversion curve could reasonably be approximated by a linear function, such that a PCA was applicable. The main direction of variation was along this line, such that the first component of the PCA encoded the

desired criterion C numerically. In the spirit of quantitative probing, it was confirmed that the likelihood of a chip being reworked

$$p(R = 1|do(C = c)) = p(R = 1|C = c) \tag{8.6}$$

increased with growing distance from the color target for chips with insufficient conversion material. On the other side of the target color point, chips with excessive conversion material were all unlikely to be reworked. After having distilled the previously hidden criterion C and having validated the model, the data was binned along the C-axis into bins $C_1, ..., C_n$ and the adjusted probabilities from Equation 8.4 were evaluated for the C-bins that contained both reworked and non-reworked chips. This directly enabled the calculation of the corresponding conditional average treatment effects (CATEs, cf. Section 1.3.2). Thereby, the initially negative evaluation of the rework process could be reversed and the benefit of the process was demonstrated numerically. In order to provide a single figure, the ATE was calculated by evaluating the unconditional do-probabilities

$$p(Y|do(R)) = \sum_{i=i_0}^{i_1} p(Y|R, C \in C_i) \cdot p(C \in C_i). \tag{8.7}$$

These were used as a discretized version of the continuous expression

$$p(Y|do(R)) = \int_{c_0}^{c_1} p(Y|R, C = c) \cdot p(C = c)dc \tag{8.8}$$

in order to account for the C-binning. The summation boundaries i_0 and i_1 as well as the integration boundaries c_0 and c_1 enclose the part of the data in which both reworked and non-reworked chips appeared. Again, the ATE confirmed the positive effect of the rework step on the overall yield. Currently, methods from Double Machine Learning [100, 101] are investigated in the aforementioned Deep Thought funded project, in order to build a continuous CATE estimator that avoids the C-binning and includes uncertainty quantification for establishing the significance of the presented effects [97]. Together with a cost estimate for the rework step itself, the goal is to give a robust recommendation for when to rework a chip, which includes reliable estimates of the expected monetary profit.

9. CAUSAL REINFORCEMENT LEARNING FOR PRODUCTION OPTIMIZATION

The final and most ambitious application of causal inference methods was the creation of a reinforcement learning environment for optimizing complex industrial production chains. Although the application to the real use case is still ongoing, a principled framework for causal reinforcement learning as well as a proof of concept on simulated data could be distilled as intermediate results. From a methodological perspective, the ultimate goal behind these efforts is the extension of the causal end-to-end analysis by including a component for decision making.

9.1 Problem statement

The manufacturing of LEDs is a complex sequence of many production processes, such that a variety of choices along the process need to be made in order to arrive at a viable product. These choices include selecting the right machines, optimizing their respective control parameters and choosing the right materials to be used. Given the complex causal mechanisms that govern the interdependences between the process steps, a naive local optimization of each stage leads to suboptimal results: A set of choices for the first step that yields good results at an intermediate measurement can have negative side effects that only become visible later on. Furthermore, some intermediate measurements are not understood well enough to be equipped with a clear target value.

The machine learning community has produced *reinforcement learning (RL)* as an approach for devising strategies (*policies*) that an *agent* should employ to advantageously manipulate its *environment*. In each step, the agent observes the current *state* of the environment and performs an *action*, before receiving a *reward* (cf. Figure 9.1, which is reproduced from [29]). The overall goal consists in maximizing an aggregation of these rewards,

although the exact weighting differs depending on the application case. A proper introduction of RL methods and terminology is vastly out of scope for this thesis, so we refer the reader to the seminal work by Sutton and Barto for an in-depth treatment of the subject [29].

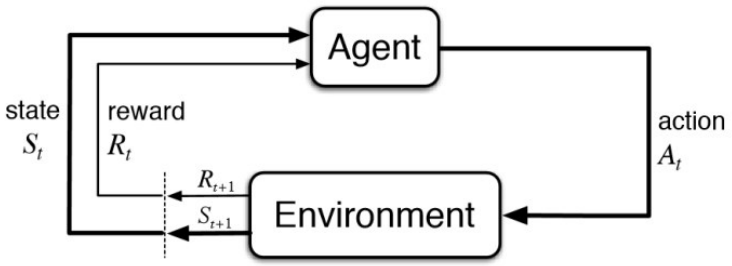

Fig. 9.1: After observing the current state, the agent selects an action according to its policy and thereby manipulates the environment. Consequently, the environment changes its current state and hands out a reward to the agent.

The important takeaway is that a classic RL agent needs an environment that responds to each of its actions by changing its state accordingly and handing out rewards. In some popular problem settings, such as teaching the agent to play video games [102, 103], this can be achieved simply by letting the agent play the original game until its performance has risen to a satisfactory level. Training the agent by direct interaction with the target environment is called *online RL*. In our setting, this strategy cannot be used: Even if the agent was able to learn the perfect policy for controlling the production chain over many runs, by then the company would no longer exist because the agent would sabotage the whole production during the learning phase. Therefore, we need to find a way of training the agent without interfering with the real environment. As a substitute for the latter, we can make use of observational data that has been logged along previous production runs, which the RL community calls *offline RL* [104]. Ideally, we can create a digital twin from the observational data, such that the agent can interact with this surrogate model as if it were the real environment. We call this approach *simulated online learning* and consider it an instance of offline RL.

9.2 Bridging the gap between causality and reinforcement learning

From a high-level perspective, simulated online learning does not differ much from ordinary online RL. The only change is that we need to provide a surrogate model that responds appropriately to the agent's actions. In causal language, an action is an intervention and we are looking for a model that answers interventional queries of the form

$$p(S_{i+1}|do(A_i) = a, S_i = s) \tag{9.1}$$

by sampling from an appropriate interventional distribution given the current state $S = s_i$ and an action $A_i = a$. The attentive reader will recognize the above expression as the covariate-specific intervention from Equation (1.15). The interventional probabilities that describe the behavior of the system under various actions are called *system dynamics* or *transition probabilities* by the RL community. In case of a stochastic policy $\pi(s)$ that makes non-deterministic decisions based on the observed state, we can even use soft interventions as indicated in Equation (1.16). Such a model, as explained in Part II of this thesis, can be built by combining observational data with domain knowledge and learning the causal graph via causal discovery. The generative nature of the model can be achieved by equipping the graph with CPDs, which can again be learned from the observational data, thereby converting it into a causal Bayesian network (cf. Section 2.2.2). A schematic overview of the involved entities is given in Figure 9.2. As an alternative to the simulated online learning approach, *model-based RL* provides *dynamic programming (DP)* algorithms that input the transition probabilities and directly exploit them for deriving an optimal policy [105, 106].

9.3 State of affairs

In order to separate the basic viability of the outlined strategy and the practical success of its implementation for real world data, we will first look at a proof of concept in an idealized setting and then highlight the difficulties that arose when leaving the latter.

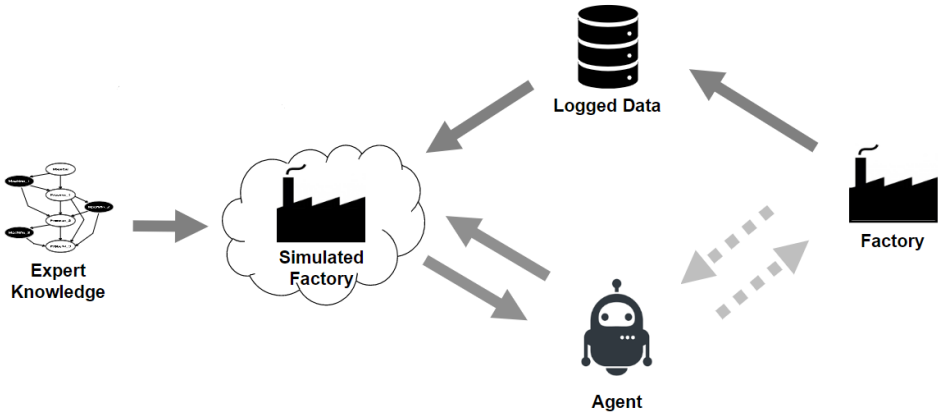

Fig. 9.2: Schematic overview of the simulated online learning strategy: Logged observational data and expert knowledge are combined into a generative causal model that serves as a simulated factory. The agent is trained and validated by interacting with the simulated factory, before it is transferred to the real factory.

9.3.1 Proof of concept on synthetic data

Given that the combination of causal inference and reinforcement learning, both in itself complex fields, was expected to be challenging in practice, the first step was the implementation of a proof of concept with synthetic data. This project was realized as a bachelor thesis by Andreas Marchl, which was co-supervised by the author [107]. A Bayesian network over fully observed binary variables was created as a known DGP using the aforementioned pgmpy Python package for probabilistic graphical models [89]. By sampling from this network, an observational dataset was generated, in order to mimic our logged production data. Assuming that the causal graph was known, CPDs for each variable given its causal parents were fitted using maximum likelihood estimation. The thereby recreated Bayesian network was used to algebraically calculate the transition probabilities (9.1), which were then compared to the known ground truth. After verifying their correctness, the fitted Bayesian network was used as a surrogate model for online RL as described above. In accordance with the real use case where the quality of the LED is judged mainly in a final benchmarking step, rewards were only given at the last step of the agent's trajectory. A variant of the popular model-free online Q-learning algorithm [108] was employed to train the agent,

which lead to near optimal performance on the given task. Additionally, *value iteration* as a model-based exhaustive dynamic programming algorithm [29] was employed and yielded similar results. Near optimality was benchmarked by running value iteration on the known ground truth model and observing the results. The process was repeated for two altered versions of the causal graph:

1. The first alteration consisted in adding further edges to the assumed causal graph without removing any of the true edges. As expected, this lead to slower convergence of the maximum likelihood estimation of the CPDs because the dimension of the CPDs was increased by the addition of irrelevant parents. By increasing the sample size, the same performance as for the true causal model could be achieved. This observation suggests that several causal models can serve as a viable surrogate model for RL, but the true causal model achieves the best sample efficiency.

2. The second alteration consisted in a complete rework of the edge structure, such that all included edges were incorrect. Even for large sample sizes, the resulting Bayesian network could not recover the correct transition probabilities. Consequently, the resulting policy after training the agent fared worse than a random uniform policy. This observation highlights that it is not enough to fit any Bayesian network without regarding the underlying causal structure of the DGP.

The quality of the produced goods for varying causal graphs and RL algorithms is shown in Figure 9.3. A more detailed quantitative evaluation, including pseudocodes and a comparison between true and recovered transition probabilities, can be found in Andreas Marchl's thesis [107].

In summary, the proposed causal reinforcement learning approach leverages the strengths of causal inference by providing a surrogate model of the appropriate complexity as an environment for RL algorithms. By establishing a clean interface between the responsibilities of causality (provide the model) and RL (exploit the model for learning), the concrete implementations of the two constituents can be varied independently of each other in a modular approach. From a high-level perspective, we can see this as an extension of the causal end-to-end analysis that includes a component for decision making.

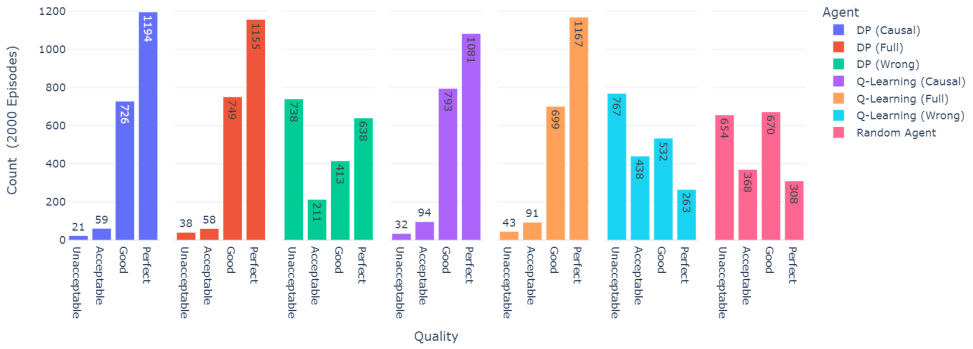

Fig. 9.3: Histograms of the quality states achieved by the agents in the hypothetical production environment. Every Bayesian network (causal, full and wrong) used for training was fitted with 100.000 observations of the environment. The full model is the first above-mentioned alteration and the wrong model is the second one.

9.3.2 Application to real data

Despite the encouraging results of the proof of concept, the strategy could not yet be successfully transferred to the intended application domain of LED production optimization. The major obstacles include:

- The data was too big to fit into RAM, such that currently available implementations of causal algorithms could not be used. Based on the Apache Spark engine for big data processing [109], parts of the causal pipeline had to be reimplemented. The modular nature of causal end-to-end analysis and other causal analysis strategies considerably alleviated this issue because data independent parts, such as the estimand identification from a given causal graph, could be left unchanged.

- Contrary to the binary data in Section 9.3.1, the involved variables were either continuous or categorical. Therefore, the difficulty of finding appropriate functional forms for the CPDs of the desired causal Bayesian network considerably increased. A binning strategy was employed for the continuous variables, in order to reduce all data to the same categorical format. However, the dense causal graph lead to situations where only sparse or even no observational data was available for the estimation of the corresponding CPD entries. Furthermore, it is to be expected that

valuable information is lost by transforming the continuous variables into their binned counterparts that require a larger set of parameters, such that the effective sample size is drastically diminished.

- Validating initial models was again complicated by the nature of the binned data: As discussed in Section 4, reliably assessing causal effects based on categorical input was complicated, such that quantitative probing could not be employed. For this use case, causal effect estimation was of minor importance, because the generative model can in principle just provide the desired interventional distribution by sampling. Therefore, alternative validation techniques were employed, such as a histogram-based comparison between the observed marginal distributions over each variable and their counterparts that were sampled using the CPDs of the Bayesian network.

- Considering the task of the RL agent, namely optimizing the policy with respect to the crucial intervention targets, the available data may not have been suitable: Some of the most important target variables, according to process experts, were not available for the analysis. For other variables, the production data showed insufficient variation because a stable configuration was used for most of the logged production samples. The latter problem might be solved by providing domain knowledge driven functional forms for the CPDs, in order to extrapolate and produce samples from the generative model.

In summary, we see that many problems may be mitigated by having access to appropriate functional forms for the relevant CPDs. On one hand, this can be viewed as further evidence that domain knowledge is an essential pillar of successful causal analysis, and it is reasonable to expect that a subset of the CPDs can be modelled together with domain experts. On the other hand, it shows that there are still too many questions to be answered on the causal side, before we can rely on the causal model as the perfect oracle that was envisioned in Section 9.2 and Figure 9.2. These questions are currently being addressed within the aforementioned Deep Thought project.

Part V

CONCLUSION

In summary, we have addressed two gaps that need to be closed in order to allow human or digital decision makers to exploit causal methods: Following Pearl's graph-based flavor of causal inference, we have constructed a causal end-to-end analysis including data reading and preprocessing, causal discovery, identification of an unbiased estimand based on the do-calculus, estimation of the estimand and reporting of the results. This flexible framework allows us to combine causal algorithms into a holistic strategy for answering interventional queries based on observational data and domain knowledge, and the open-source Python package cause2e [4] is provided as an implementation.

An integral part of such a holistic strategy is the validation of the resulting causal model, which we have addressed by developing quantitative probing [2] as a largely model-agnostic approach that benefits from quantitative expert knowledge. Although an analytic proof for the effectiveness of quantitative probing could not yet be distilled, we have gathered favorable evidence via a thorough simulation study. Limits of the strategy have been identified and discussed extensively at the example of malfunctioning validation runs, and the shortcomings have inspired a guide for practitioners with the aim of avoiding critical scenarios. The open-source Python qprobing package [5] provides researchers with the possibility to further study quantitative probing, in order to answer the remaining open questions.

The methodological contributions have been illustrated by three use cases from the domain of manufacturing light-emitting diodes. While the first two applications have shown the causal end-to-end analysis, quantitative probing and Simpson's paradox at play, the third one has been concerned with the combination of causal inference and reinforcement learning. This attempt at bridging the gap between causality and decision making, which has served as a central motivation for this thesis, unfortunately has shown that there are still plenty of obstacles to overcome. Although the thesis ends on this sobering note, the author is optimistic that the recent surge of interest in causal methods will help establish graph-based causal inference as a valuable tool for researchers of all domains, thereby paving the way for a solution of the remaining challenges.

ACKNOWLEDGEMENTS

At the end of this thesis, I want to thank the people who have made it possible for me to make it through the always interesting, often funny but sometimes also stressful time of being a PhD student. The following list is non-exhaustive and even the included persons would deserve far more praise than I can give here, but I suppose it is better than nothing. Thank you!

- First and foremost, I am deeply grateful for the continuous support of my advisor Prof. Dr. Elmar W. Lang. He has introduced me to the field of machine learning and played an important role in making my bachelor and PhD theses in cooperation with ams OSRAM possible. My gratefulness for his efforts to share his vast experience with me, be it in research questions or otherwise, even in the face of considerable adversities, cannot be overstated.

- Furthermore, I want to thank the other three examiners of this thesis, Prof. Dr. Dieter Weiss, Prof. Dr. Rainer Spang and Prof. Dr. Tilo Wettig for their willingness to offer their expert opinion on the results of my work.

- As this PhD project was a cooperation between Universität Regensburg and ams OSRAM, Dr.-Ing. Maike Stern served as my advisor from company side. She has been a source of inspiration and advice with respect to countless research-related and more practical topics. I am indebted to her for accompanying me through all the stages of this PhD project.

- I want to thank my colleagues at ams OSRAM's data science team (and Alex) who have done a great job at introducing a clueless algebraic topologist to the fascinating world of light-emitting diodes: My fellow PhD students Dr. Heribert Wankerl, Alexander Luce and Philipp Schwarz have shared the experience of growing into a researcher with me and I

thoroughly enjoyed our discussions on topics related to artificial intelligence, machine learning, data science and other buzzwords. Andreas Marchl has been the smartest, most talented and only bachelor student that I have had the joy of advising so far and I hope that he will further explore the possibilities of causal reinforcement learning in his future research career. Dr. Sebastian Imhof has been a great collaborator and fellow explorer in my quest to learn about big data analysis, clean code and DevOps practices, in order to better understand our manufacturing landscape. My boss Dr. Thomas Weig has placed his trust in me to add causal inference to ams OSRAM's toolkit and I am most grateful for having felt this trust throughout the last years. Furthermore, I am indebted to him for hiring me as a data scientist such that I can keep working with all of the above and many other great colleagues at ams OSRAM.

- Causal inference is a fascinating topic, but unfortunately not too well-known as of now. Considering this, I am even more grateful for the inspiring and sometimes humbling conversations with my fellow causality enthusiasts in the Deep Thought project, in our causal inference working group and in the Spang Lab group meetings.

- At the beginning of my studies, I heavily fell in love with the world of pure mathematics. This is mostly due to Prof. Dr. Stefan Friedl who has guided me through the entirety of my bachelor and master studies in mathematics by supervising my theses, helping me organize two semesters of studying in Montpellier and offering his valuable advice on career paths inside and outside of academia.

- I was lucky not only with respect to my teachers and mentors, but also with respect to my fellow students in mathematics and computational science. Discovering new ideas together was great fun and I very fondly remember our time together.

- I want to thank the subset of the above persons who have taken the time to proofread parts of this thesis and made a considerable number of helpful suggestions that improved its quality.

- There is a life outside of work and it can be hard to keep it unaffected by the frequent failures that mark a PhD project. I am forever grateful

to my family and friends who shared their love and friendship with me in these inspiring but challenging times, even without understanding what exactly I was doing.

BIBLIOGRAPHY

[1] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

[2] Daniel Grünbaum, Maike L. Stern, and Elmar W. Lang. Quantitative probing: Validating causal models using quantitative domain knowledge. *Journal of Causal Inference*, 2023. Accepted for publication. Preprint available at https://arxiv.org/abs/2209.03013.

[3] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1934.

[4] Daniel Grünbaum, Maike L. Stern, and Elmar W. Lang. cause2e: A python package for causal end-to-end analysis, 2021. Available at https://github.com/MLResearchAtOSRAM/cause2e.

[5] Daniel Grünbaum, Maike L. Stern, and Elmar W. Lang. qprobing: A python package for evaluating the effectiveness of quantitative probing for causal model validation, 2022. Available at https://github.com/MLResearchAtOSRAM/qprobing.

[6] K. Lorenz. *Die Rückseite des Spiegels: Versuch einer Naturgeschichte menschlichen Erkennens*. R. Piper, 1973.

[7] Voltaire and Burton Raffel. *Candide: or Optimism*. Yale University Press, 2005. Work by Voltaire originally published in 1759.

[8] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[9] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[10] Neil Savage. Game changer. *Commun. ACM*, 55(6):22–23, June 2012.

[11] Peter Hull, Michal Kolesár, and Christopher Walters. Labor by design: contributions of David Card, Joshua Angrist, and Guido Imbens. *The Scandinavian Journal of Economics*, 124(3):603–645, jul 2022.

[12] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017.

[13] Sara Solnick and David Hemenway. The "twinkie defense": The relationship between carbonated non-diet soft drinks and violence perpetration among boston high school students. *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention*, 18:259–63, 10 2011.

[14] Tyler Vigen. Spurious correlations. `https://www.tylervigen.com/spurious-correlations`. Accessed: 2023-03-28.

[15] Thomas Wenzl. Smoking and covid-19 - a review of studies suggesting a protective effect of smoking against covid-19. (KJ-NA-30373-EN-N (online)), 2020.

[16] Gareth Griffith, Tim Morris, Matthew Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma Sharp, Jonathan Sterne, Tom Palmer, George Smith, Kate Tilling, Luisa Zuccolo, Neil Davies, and Gibran Hemani. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature Communications*, 11, 11 2020.

[17] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[18] Wikipedia contributors. Berkson's paradox — Wikipedia, the free encyclopedia, 2023. [Online; accessed 29-March-2023].

[19] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951.

[20] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[21] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

[22] Rogier Kievit, Willem Frankenhuis, Lourens Waldorp, and Denny Borsboom. Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, 4, 2013.

[23] Miguel A Hernán, David Clayton, and Niels Keiding. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3):780–785, 03 2011.

[24] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

[25] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

[26] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

[27] Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

[28] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022.

[29] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[30] Heribert Wankerl, Maike L Stern, Ali Mahdavi, Christoph Eichler, and Elmar W Lang. Parameterized reinforcement learning for optical system optimization. *Journal of Physics D: Applied Physics*, 54(30):305104, may 2021.

[31] Alexander Luce, Ali Mahdavi, Florian Marquardt, and Heribert Wankerl. Tmm-fast, a transfer matrix computation package for multilayer thin-film optimization: tutorial. *J. Opt. Soc. Am. A*, 39(6):1007–1013, Jun 2022.

[32] J. M. Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20(2):164–168, 2003.

[33] P Pallmann, AW Bedding, B Choodari-Oskooei, M Dimairo, L Flight, LV Hampson, J Holmes, AP Mander, L Odondi, MR Sydes, SS Villar, JMS Wason, CJ Weir, GM Wheeler, C Yap, and T Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16:1–15, 2018.

[34] Timothy Ogden. RCTs in Development Economics, Their Critics and Their Evolution. In *Randomized Control Trials in the Field of Development: A Critical Perspective*. Oxford University Press, 09 2020.

[35] Benedicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. arXiv:2011.08047, 2020.

[36] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[37] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[38] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann series in representation and reasoning. Elsevier Science, 1988.

[39] Richard E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, Inc., USA, 1990.

[40] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[41] Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 3–11, Arlington, Virginia, USA, 2012. AUAI Press.

[42] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, page 1219–1226. AAAI Press, 2006.

[43] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 217–224, Arlington, Virginia, USA, 2006. AUAI Press.

[44] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence.

[45] Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, December 2020.

[46] Russell Millar. Maximum likelihood estimation and inference. with examples in R, SAS and ADMB. 07 2011.

[47] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery, 2021.

[48] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[49] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018.

[50] Wikipedia contributors. Directed acyclic graph — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-April-2023].

[51] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[52] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, March 2003.

[53] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[54] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[55] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[56] Chirayu (Kong) Wongchokprasitti, Harry Hochheiser, Jeremy Espino, Eamonn Maguire, Bryan Andrews, Michael Davis, and Chris Inskip. Pycausal v1.2.1, December 2019. Available at https://github.com/bd2kccd/py-causal.

[57] Joseph Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Immé, Ebert-Uphoff, Savini M. Samarasinghe, Elizabeth A. Barnes, and Clark Glymour. Tetrad - a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

[58] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv:2011.04216*, 2020.

[59] PyWhy Developers. Pywhy, 2023. https://github.com/py-why.

[60] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

[61] Joseph 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. *International Journal of Data Science and Analytics*, 3, 03 2017.

[62] Daniel Grünbaum, Maike L. Stern, and Elmar W. Lang. Quantitative probing: Validating causal models using quantitative domain knowledge. arXiv:2209.03013, 2022.

[63] 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.

[64] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. arXiv:2103.04850, 2021.

[65] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society Series B*, 82(1):39–67, 2020.

[66] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. arXiv:2112.13398, 2021.

[67] Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10999–11009. Curran Associates, Inc., 2020.

[68] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. arXiv:2011.15007, 2020.

[69] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[70] Joseph D. Ramsey and Bryan Andrews. A comparison of public causal search packages on linear, gaussian data with no latent variables. arXiv:1709.04240, 2017.

[71] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference*

on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 191–201. PMLR, 09–15 Jun 2019.

[72] Dustin Tran, Francisco J. R. Ruiz, Susan Athey, and David M. Blei. Model criticism for bayesian causal inference. arXiv:1610.09037, 2016.

[73] G.E.P. Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143:383–430, 1980.

[74] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[75] Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms. In Manfred Jaeger and Thomas Dyhre Nielsen, editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 17–28. PMLR, 23–25 Sep 2020.

[76] Amit Sharma, Vasilis Syrgkanis, Cheng Zhang, and Emre Kıcıman. Dowhy: Addressing challenges in expressing and validating causal assumptions. arXiv:2108.13518, 2021.

[77] William H. Jefferys and James O. Berger. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

[78] A. J. Ayer. *The Problem of Knowledge*. Macmillan, New York, 1956.

[79] Enrico Bombieri. The riemann hypothesis. In A. Wiles J. Carlson, A. Jaffe, editor, *The Millenium Prize Problems*, chapter 7, pages 107–128. Clay Mathematics Institute, Cambridge, MA, 2000.

[80] Peter Sarnak. *Problems of the Millennium: The Riemann Hypothesis*. Princeton University & Courant Institute of Math. Sciences, 2004.

[81] Aleksandar Ivić. On some reasons for doubting the riemann hypothesis. arXiv:math/0311162, 2003.

[82] S. Skewes. On the difference $\pi(x) - li(x)$ (i). *Journal of the London Mathematical Society*, s1-8(4):277–283, 1933.

[83] S. Skewes. On the difference $\pi(x) - li(x)$ (ii). *Proceedings of the London Mathematical Society*, s3-5(1):48–70, 1955.

[84] Dijkstra, Edsger W. The humble programmer. *Commun. ACM*, 15(10):859–866, 1972.

[85] Dirk Beyer and Thomas Lemberger. Software verification: Testing vs. model checking. In Ofer Strichman and Rachel Tzoref-Brill, editors, *Hardware and Software: Verification and Testing*, pages 99–114, Cham, 2017. Springer International Publishing.

[86] Gerwin Klein, June Andronick, Matthew Fernandez, Ihor Kuz, Toby Murray, and Gernot Heiser. Formally verified software in the real world. *Commun. ACM*, 61(10):68–77, sep 2018.

[87] 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, oct 2006.

[88] 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.

[89] Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*, 2015.

[90] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[91] Max Planck. Über das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, 309(3):553–563, 1901.

[92] A. Einstein. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt [AdP 17, 132 (1905)]. *Annalen der Physik*, 517(S1):164–181, 2005.

[93] Scott W. Ambler. *UML Class Diagrams*, pages 47–72. Cambridge University Press, 2005.

[94] E. Fred Schubert. *Light-Emitting Diodes*. Cambridge University Press, 2 edition, 2006.

[95] Wikipedia contributors. Light-emitting diode — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-May-2023]; Figure by Deglr6328, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3242448.

[96] Wikipedia contributors. Cie 1931 color space — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-May-2023].

[97] Oliver Schacht, Sven Klaassen, Philipp Schwarz, Martin Spindler, Daniel Grünbaum, and Sebastian Imhof. Causally learning an optimal rework policy, 2023. arXiv: 2306.04223, submitted to the Causal Inference and Machine Learning in Practice workshop at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

[98] Wikipedia contributors. Snell's law — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-May-2023]; Figure by Smedlib, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=60639100.

[99] T. Fujii, Y. Gao, R. Sharma, E. L. Hu, S. P. DenBaars, and S. Nakamura. Increase in the extraction efficiency of GaN-based light-emitting diodes via surface roughening. *Applied Physics Letters*, 84(6):855–857, 02 2004.

[100] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[101] Philipp Bach, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

[102] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

[103] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas

Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–33, 02 2015.

[104] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

[105] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *CoRR*, abs/2006.16712, 2020.

[106] Andrew G. Barto. Reinforcement learning and dynamic programming. *IFAC Proceedings Volumes*, 28(15):407–412, 1995. 6th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Man-Machine Systems 1995, Cambridge, MA, USA, 27-29 June 1995.

[107] Andreas Marchl. Causal reinforcement learning: Offline reinforcement learning with bayesian networks. Bachelor thesis, Ostbayerische Technische Hochschule Regensburg, Regensburg, 2022.

[108] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, May 1992.

[109] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, October 2016.