Research Article

Daniel Grünbaum*, Maike L. Stern, and Elmar W. Lang

Quantitative probing: Validating causal models with quantitative domain knowledge

https://doi.org/10.1515/jci-2022-0060
received September 08, 2022; accepted June 05, 2023

Abstract: We propose quantitative probing as a model-agnostic framework for validating causal models in the presence of quantitative domain knowledge. The method is constructed in analogy to the train/test split in correlation-based machine learning. It is consistent with the logic of scientific discovery and enhances current causal validation strategies. The effectiveness of the method is illustrated using Pearl's sprinkler example, before a thorough simulation-based investigation is conducted. Limits of the technique are identified by studying exemplary failing scenarios, which are furthermore used to propose a list of topics for future research and improvements of the presented version of quantitative probing. A guide for practitioners is included to facilitate the incorporation of quantitative probing in causal modelling applications. The code for integrating quantitative probing into causal analysis, as well as the code for the presented simulation-based studies of the effectiveness of quantitative probing are provided in two separate open-source Python packages.

Keywords: causal model validation, causal inference, causality

MSC 2020: 62D20

1 Introduction

The topic of causal inference has been a focus of extensive research in statistics, economics, and artificial intelligence [1–6]. Methods of causal inference allow predicting the behaviour of a system not only in situations where we passively observe certain evidence but also when we actively intervene in the data generating process (DGP). Ruling out confounding in such predictions previously required performing costly and possibly harmful randomized controlled trials [7]. Causal inference, on the contrary, provides the methodology to infer the effect of hypothetical actions from passively collected observational data. Additional assumptions can be encoded in graphs [1] or conditional independence statements [3]. While correlation-based prediction techniques can be validated using the well-known train/test split [8], the challenge of validating causal models is still largely unsolved. Without causal model validation, predictions about the behaviour of a system under interventions from purely observational data cannot be relied upon, due to our uncertainty about the validity of the underlying model. This gap in the methodology needs to be filled, before decision makers can leverage the powerful methods of causal inference, in order to complement, enhance, or even replace the costly current method of randomized controlled trials. In this article, we present quantitative probing, a novel, and largely model-agnostic approach for validating causal models, which applies the logic of scientific discovery [9] to the task of causal inference by using domain-specific refutation checks on non-target effects.

*** Corresponding author: Daniel Grünbaum**, ams OSRAM Group/University of Regensburg, 93055 Regensburg, Germany, e-mail: daniel.gruenbaum@ams-osram.com
Maike L. Stern: ams OSRAM Group, 93055 Regensburg, Germany
Elmar W. Lang: Department of Physics, University of Regensburg, 93053 Regensburg, Germany

Both, the code for integrating quantitative probing into causal analysis and the code for simulation-based studies of the effectiveness of quantitative probing are provided in two separate open-source Python packages [10,11].

The article is structured as follows: Section 2 shortly reviews the train/test split as the classical method for validating correlation-based statistical models, and explains why it is not possible to directly transfer this general and model-agnostic method to the field of causal inference. Section 3 briefly clarifies the notion of a causal model and introduces causal end-to-end analysis as a model that generalizes other types of causal models. Section 4 presents related approaches to the validation of causal models and the incorporation of domain knowledge in causal modelling. Section 5 synthesizes the idea of quantitative probing from the observations in the previous sections, and relates the approach to the established method of scientific discovery. The concept is illustrated using Pearl's well-known Sprinkler example [1]. Furthermore, assumptions are explicitly stated, to define the current scope of the quantitative probing approach. Section 6 provides simulation-based evidence for the effectiveness of quantitative probing as a validation technique. Special cases, where the method fails to detect misspecified models, are investigated in detail, to identify limitations and future enhancements. Section 7 provides a guide for practitioners with the aim of facilitating the incorporation of quantitative probing in causal modelling applications. Section 8 concludes the article by summarizing the main points and proposes concrete questions for future research.

2 The role of the i.i.d. assumption in model validation

The predominant strategy for validating correlation-based statistical learning methods is based on the crucial i.i.d. assumption [8]: 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. The two parts of the term i.i.d. have important consequences for how we train or validate machine learning models.

The independence assumption justifies the factorization of the model likelihood into its marginals, which enables commonly used error metrics of independently summing up computed prediction errors of the model for each sample. The assumption of identical distribution enables the use of the train/test split for model validation. If we train the model on a subset of the available data (the *labelled training set*), we can evaluate its performance on the remaining data points (the *labelled test set*). The test set is statistically identical to the new data, 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.

Note that the train/test split treats the underlying model as a black box, which makes it applicable to a wide range of predictive models. Although, with *causal models*, observational data may very well follow the i.i.d. assumption, the concept of test samples is diametrically opposed to the task of causal inference: We want to predict what happens under certain interventions, but interventions inevitably change the distribution from which the samples are drawn.

3 Types of causal models

To formulate and evaluate validation strategies for causal models, we need to specify more clearly what is meant by a causal model. Throughout the article, we will use a very general definition of a causal model: We call everything a causal model that answers interventional queries, i.e. that estimates probabilities of the form $p(y|\mathrm{do}(x = v))$ to calculate the *average treatment effect (ATE)*

$$\tau = p(y = 1|\mathrm{do}(x = 1)) - p(y = 1|\mathrm{do}(x = 0)) \tag{1}$$

from observational data. For the remainder of the article, we will restrict our studies to these ATEs in the binary data setting. Extensions to other types of causal effects such as conditional ATEs [12], controlled/natural direct effects, and natural indirect effects (NIEs) [1] are possible, but not necessary to illustrate the central ideas of our validation strategy. In the same spirit, we refrain from leaving the binary data setting, although all the arguments readily transfer to more general discrete, categorical, and continuous datasets. It is worth highlighting that a causal graph is not a full causal model in the sense of the above definition: Although it can be used to identify a given causal effect, the final estimation of the effect is not possible without additional inputs, such as conditional probability distributions (CPDs) in the case of a causal Bayesian network, or observational data together with a fixed estimation algorithm.

To consider a general scenario without any restrictive assumptions about the underlying DGP, we introduce the following graph-based causal model type that covers many simpler types of causal analysis as special cases.

By *causal end-to-end analysis*, we mean the following procedure for a given dataset and a given target effect.

(1) We pre-process the data by deleting, adding, rescaling, or combining variables.
(2) We pass qualitative 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 qualitative domain knowledge.
(4) We post-process 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 the target effect and other effects of interest by applying the do-calculus to the causal graph.
(6) We estimate the estimands by a method of our choice.

Although we will use the causal end-to-end analysis as an exemplary modelling strategy for illustrating the concept of quantitative probing in the remainder of the article, we stress that the validation strategy is not tied to this algorithm and can be used to validate any causal model that can predict multiple causal effects.

4 Related work

In recent years, others have already tackled the question of how to validate causal models. Furthermore, the use of domain knowledge in causal modelling is widespread, especially in graph-based causal inference. In this section, we will give an overview of existing approaches and their limitations, before the next section will detail how the method of quantitative probing synthesizes ideas from both research fields.

4.1 Causal model validation

One common approach is to accept that the model is likely to have flaws. If these can be identified and bounded, we can still use this knowledge to obtain error bounds around the estimates of the model in the spirit of a sensitivity analysis [13–20]. These methods are easy to implement, but they depend on the specifics of creating the causal model. An additional drawback is that they do not tell us how close our model is to the correct one. A similar problem is faced by causal model selection algorithms [21–24] where the goal is the selection, mostly by cross-validation techniques, of the best causal model among a set of candidates: Even the best model among the candidates could still perform poorly on the actual prediction task, given that it is only evaluated relative to the competing models (cf. Section 7.3.3). Therefore, we focus on validation strategies that aim at ensuring the fitness of one candidate model.

The first such stream of research [25] takes the algorithm used to create the causal model and applies it to simulated data. If the simulation environment can also generate data for the interventional scenarios of interest, we can then compare the predictions of the model to the simulated ground truth. However, this approach relies on some critical assumptions: First, for the simulations to be realistic, we need to have sufficient knowledge of the DGP. But then any causal analysis would be obsolete. Second, we evaluate the performance of a surrogate model instead of the true one, and both are only linked via the algorithm to generate them. This link is quantified in ref. [26] by the use of a Taylor expansion on the space of DGPs: The error of the candidate model for the task of interest is estimated by its error on a simulated DGP, corrected by the product of an influence function and the distance between the real and the simulated DGPs. However, the problem of estimating the latter distance still remains.

Therefore, 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 [27] has been developed based on posterior predictive checks [28,29]. 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 calculated to summarize properties of the data generated from it, using a suitable prior. The model is then evaluated by a comparison of the calculated discrepancy functions 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.

Karmakar and Small [30] propose to make causal theories as elaborate as possible, in order to manually derive falsifiable statements that can serve as refutation tests. The testable parts of the theory are to be verified using observational studies whereby confounding cannot be excluded, but bounded using the aforementioned techniques from the field of sensitivity analysis. Methods for pooling the evidence are discussed from a theoretical perspective, such that the merits and limits of the strategy can be assessed reliably. Besides the manual process of crafting the elaborate theory, the major drawback of this approach is the need to gather data for the additional observational studies, which is not always feasible in practical applications.

In out-of-sample causal tuning [31], 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 ref. [31], 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 also by any model in the same Markov equivalence class. We will build on the point of view that all variables in a the graphical model can be used for refutation checks and extend it to the interventional setting.

in ref. [32], 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 [9], 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 the 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.

4.2 Exploiting domain knowledge in causal discovery

Incorporating domain knowledge in the modelling process is an established concept in graph-based causal inference. In causal discovery, the subfield of causality that is concerned with inferring the causal graph from observational or interventional data, many algorithms provide an interface for specifying edges that are known to be present or absent from the causal graph. While the exact interface is not always described explicitly, most works follow the *ATE* given by Meek in ref. [33] who discusses the challenge of finding causal explanations that are consistent with both the data and the domain knowledge. Meek's strategies can be readily applied to constraint-based causal discovery, where independency constraints on the data are used to reverse engineer features of the causal graph [4,34], by providing a scaffold that restricts the search space. Similarly, score-based algorithms based on Chickering's greedy equivalence search [35] can directly use the domain knowledge to limit the set of possible edges in each add or removal step.

A comprehensive survey of causal discovery strategies, including constraint-based, score-based, and structural equation model (SEM)-based methods can be found in ref. [36]. The authors provide practical guidelines where they recommend the direct incorporation of knowledge about present or absent edges in the search procedure. They suggest to exploit other types of domain knowledge to create a synthetical DGP with similar properties to the original problem set, which can then be used for benchmarking similar to the concept outlined in ref. [25]. In ref. [37], SEM-based discovery is discussed in greater detail for different types of functional models that have been shown to be identifiable from data in previous work [38–42]. The authors suggest using SEM-based discovery to find all causal models consistent with the available domain knowledge as a selection criterion, but do not elaborate on the nature of the knowledge. In ref. [43], causal discovery algorithms are summarized into five groups, before an extensive experimental benchmark study is carried out to highlight the respective strengths and weaknesses of the different approaches.

The aforementioned 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 absence or presence of an edge from node A to node B in the causal graph is equivalent to the controlled direct effects of variable A on variable B being all zero or not, respectively. Knowledge about other types of causal effects such as the ATE, the NIE or the conditional average treatment effect cannot be used in the aforementioned procedures. 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 of effects being either nonzero or zero but also includes knowledge about the effect strength on the full spectrum of real numbers.

5 Quantitative probing

The goal of this work is to establish quantitative probing as a general, largely model-agnostic validation framework that provides powerful problem-specific refutation tests by using quantitative domain knowledge.

5.1 The logic behind quantitative probing

Validation strategies often assume that we want to predict a single *target (causal) effect*, once the causal graph has been specified. We will refer to both the treatment variable and outcome variable of the target causal effect as *target variables* for ease of notation. All the other variables in the dataset, which we will call *non-target variables*, usually are either ignored or treated as confounders, based on the structure of the causal graph. In either case, they are taken care of by methods like the do-calculus, and we do not have to inspect

their quantitative causal relationships with any of the target or non-target variables. However, not taking into account these *non-target effects* means wasting our domain knowledge, just as we can pass parts of the causal graph to a causal discovery algorithm as a form of qualitative domain knowledge, we can use our expectations about selected non-target effects as supplementary quantitative domain knowledge. Analogously to communicating that we expect an edge between two variables in the causal graph, we can specifiy that we expect a certain non-target effect to be close to a given value. As outlined in Section 4.2, there is a key difference between passing qualitative and quantitative domain knowledge. The qualitative domain knowledge can be explicitly accepted as an input by a causal discovery algorithm, meaning that the procedure actively uses the knowledge in recovering the correct causal graph from the observational data [33]. For quantitative knowledge, on the other hand, it is not even clear where to pass these desiderata about the causal model and its predictions. Should we pass them to the estimation procedure and use them for hyperparameter tuning? Or could it be that the hyperparameters are correct, even when the model fails to reproduce the expected effect? Think, for instance, of a failure that can be attributed to a misspecification of the causal graph in the preceding discovery step. It is even conceivable that the expectations have not been met because of a mistake in data preprocessing, and changing the causal discovery or estimation steps will lead to an inferior estimate of the target effect. Such considerations make it clear that the quantitative knowledge can hardly be tied to a specific step in the end-to-end causal analysis. However, we can always use the knowledge in the following way: If we are sure that a given non-target effect must come close to a specific value, but our analysis fails to reproduce this outcome, something must have gone wrong. We do not know exactly what it is, but we cannot exclude that the same error has also affected the estimation of our target effect. Such a failure to reproduce our expectations should therefore diminish our confidence in our estimate of the target effect. Conversely, if we specify many different expected non-target effects and all of them are reproduced by our analysis, our confidence in the estimation of the target effect by the very same analysis should increase. As previously mentioned in the discussion of causal validation via refutation tests [32], such a strategy is in line with the general logic of establishing scientific theories [9]. We can probe a candidate model by comparing its estimates of non-target effects against previously stated expectations about their values. If no inconsistencies arise, our trust in the model increases. Consequently, we will refer to the specified non-target effects as *quantitative probes* or simply *probes* and to the presented validation strategy as *quantitative probing*.

5.2 Step-by-step description

Summarizing this line of reasoning, quantitative probing can be described as the following easy-to-implement stepwise procedure, which is illustrated in Figure 1.

(1) Probe selection and specification: Select a number of non-target causal effects (quantitative probes) whose true precise or approximate values can be specified by quantitative domain knowledge.
(2) Modelling: Build a causal model that is able to predict both the target effect and the quantitative probes.
(3) Probe prediction: Use the causal model to predict the values of the quantitative probes.

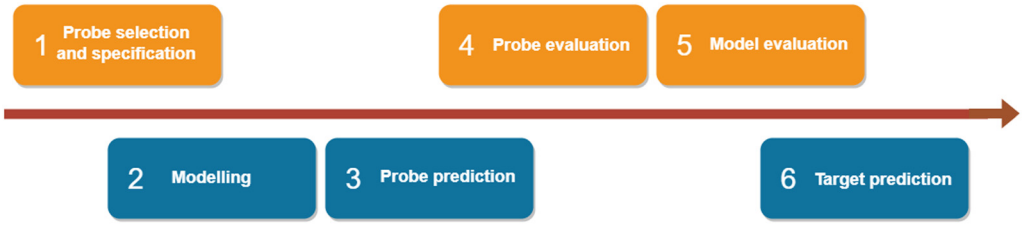

Figure 1: The six steps of a causal modelling workflow that uses quantitative probing as a validation strategy. Steps that are part of the validation itself are coloured in orange, whereas steps that are performed by the modelling algorithm are marked in blue.

(4) Probe evaluation: For each quantitative probe, evaluate whether the predicted value is in accordance with the initially specified quantitative domain knowledge.
(5) Model evaluation: Based on the previous step, decide whether the candidate model is to be trusted to predict the target effect correctly.
(6) Target prediction: If the candidate model has been judged as trustworthy, use it to predict the target causal effect, to answer the original causal question of interest.

Note that probe prediction and target prediction can be carried out simultaneously if this yields computational benefits, as long as the estimate for the target effect is only used in the case of a successful model evaluation. In contrast to the methods presented in Section 4.2, quantitative probing allows the use of any type of causal effect, as long as the candidate model is able to predict it and the domain expert can provide knowledge about it. A drawback is that the knowledge cannot be used directly for model creation, but only for a post hoc analysis of the validity of a given model. It is also worth highlighting that the procedure makes no assumption about the internals of the candidate model, as long as it is formally able to predict the probes and the target effect. This makes quantitative probing a largely model-agnostic strategy that can be used for a broad variety of causal models, and obliterates the need to change the validation strategy whenever the modelling strategy is altered.

Due to its simplicity and close relation with Popper's well-known ideas, domain-specific refutation checks are already used implicitly by researchers and practitioners [24,27,32,44]. However, the focus often lies on comparing the target effect to the domain expert's expectations, which results in circular reasoning: If the domain knowledge about the target effect is sufficiently precise to validate causal models, then one could directly use the domain knowledge to answer the causal query without building the model first. Quantitative probing circumvents this issue by focusing on non-target effects, which can be provided by domain experts even in situations where the target effect cannot be evaluated without causal modelling. Furthermore, the effectiveness of the strategy has never been evaluated, such that it is unclear whether these domain-specific refutation checks can actually detect incorrect causal models. For the remainder of this article, we will present simulation-based evidence for the effectiveness of the proposed validation strategy and provide guidance for practitioners on how to apply it in real-world application scenarios.

5.3 Sprinkler example

Before we step into the technical discussion, let us illustrate the presented motivation for quantitative probing using the well-known sprinkler example [1]. Suppose that we are interested in estimating the ATE of activating a garden sprinkler on the slipperiness of our lawn. Estimating this target effect from observational data is the reason why we are performing the causal end-to-end analysis. For this hands-on example, we generate data using pgmpy, an open source Python package for probabilistic graphical models [45]. The subsequent analysis is performed using cause2e, an open source Python package for causal end-to-end analysis [10]. The data consist of $m = 10{,}000$ samples, each holding values for $n = 5$ variables:

– What was the season on the day of the observation?
– Was the sprinkler turned on the day of the observation?
– Was it raining on the day of the observation?
– Was the lawn wet on the day of the observation?
– Was the lawn slippery on the day of the observation?

All of the variables are binary, except for the season variable. For simplicity, we also binarize the season variable in a preprocessing step: We encode "Winter" as zero, "Spring" as one and discard the observations that were made in summer or autumn.

5.3.1 Probe selection and specification

After pre-processing, the involved variables can no longer change and we want to leverage our quantitative domain knowledge about them by creating two quantitative probes:
- We expect that turning on the sprinkler will make the lawn wetter, so we expect the ATE of “Sprinkler” on “Wet” to be greater than zero.
- We expect that making the lawn wetter will also make it more slippery, so we expect the ATE of “Wet” on “Slippery” to be greater than zero.

Note that these two probes do not directly imply specific edges in the causal graph, as the influence could also be mediated via one of the other variables.

5.3.2 Modelling (correct)

In addition to our quantitative domain knowledge, we can now specify qualitative domain knowledge about the underlying causal graph. For demonstrative purposes, we choose a configuration that enables the causal discovery algorithm to recover the causal graph that was used for generating the data:
- We forbid all edges that originate from “Slippery.”
- We forbid all edges that go into “Season.”
- We forbid the edges “Sprinkler” → “Rain” and “Season” → “Wet.”
- We require the edges “Sprinkler” → “Wet” and “Rain” → “Wet.”

Note that this configuration still leaves nine edges whose presence in the graph is neither required nor forbidden, but needs to be decided by the causal discovery algorithm. If we run fast greedy equivalence search [46], an optimized version of the standard greedy equivalence search [35], as a causal discovery algorithm, we see that we recover the true causal graph from the DGP (cf. Figure 2). Together with the observational data and the choice to use the do-calculus and linear regression for identification and estimation, respectively, this constitutes a causal model in the sense of Section 3.

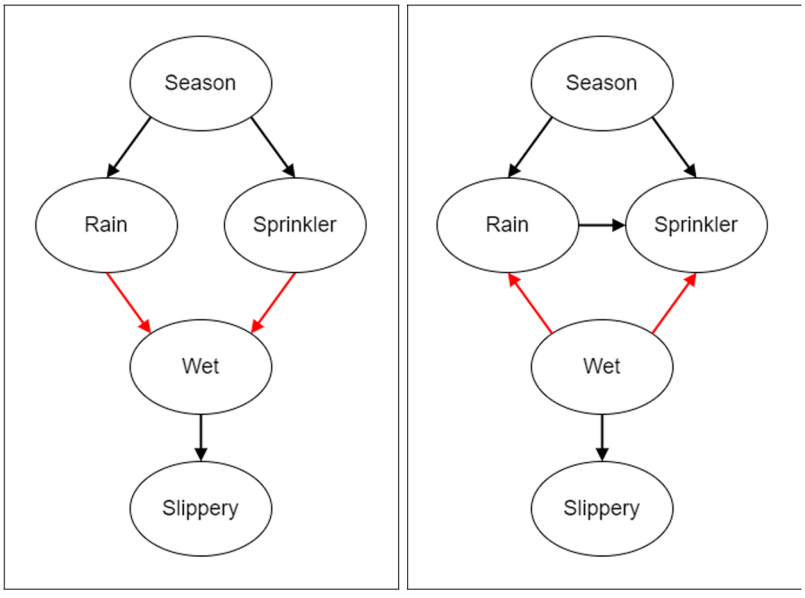

Figure 2: The recovered graphs correctly (left) or incorrectly (right) represent the DGP. Edges prescribed by the respective sets of domain knowledge are marked in red, and the rest of them has been added by the causal discovery algorithm.

5.3.3 Prediction and evaluation (correct)

In the next step, we estimate the target effect and our quantitative probes from the causal graph and the data via linear regression. We recover a target effect of 0.52, but in principle, we have no idea whether this is a reasonable estimate. However, the analysis shows that the ATE of "Sprinkler" on "Wet" is 0.62 and the ATE of "Wet" on "Slippery" is 0.81. Therefore, our expectations about the known causal effects have been met by the predictions of the model, which increases our trust in the model and therefore also in the estimate of the target effect.

5.3.4 Modelling (flawed)

As a contrast, let us see how our probes help us detect flawed causal models: For this purpose, we intentionally communicate wrong assumptions about the causal graph to the causal discovery algorithm. For each edge that we have required in the aforementioned example, we now require the corresponding reversed edge. The forbidden edges are left unchanged. These flawed inputs lead to a flawed causal discovery result, as can be seen in Figure 2 (right).

5.3.5 Prediction and evaluation (flawed)

If we now use the incorrect graph and the data to estimate the three ATE's via linear regression, we observe a target effect of 0 because there is no directed path from "Sprinkler" to "Slippery." The ATE of "Wet" on "Slippery" is again 0.81 in accordance with our expectations. However, the ATE of "Sprinkler" on "Wet" is 0, even though we expect it to be a positive number. We can use this failed validation as evidence that something in our end-to-end analysis must have gone wrong, and therefore, the estimate of the target effect should not be trusted either. In practice, this could lead us to re-examine the qualitative domain knowledge that we have passed, which would then result in a causal discovery result that is closer to the true causal graph. Another possibility would be the examination of the estimation strategy that was used to compute the numerical estimate from the statistical estimand and the data. In our case, this would not lead to a correct estimate of the target effect because the error in the model lies in the graph, not in the estimation procedure. A third possibility would be a re-examination of our quantitative expectations, e.g. if we notice that our expectations were actually about the natural direct effect instead of the ATE. In any case, formulating our explicit quantitative knowledge about certain causal effects in the model, followed by an automated validation at the end of the end-to-end analysis, serves as a valuable evaluation step that can help us detect modelling errors. Note, however, that the two initial quantitative probes lose their effectivity for validating the resulting adapted candidate model, as they were already used for building it (cf. Section 7.3 for practical guidelines).

5.4 Assumptions

In the previous example, we have made several implicit assumptions that are necessary to leverage the power of quantitative probing for causal model validation.

(1) We need to have quantitative knowledge about some causal effects between the observed variables, otherwise we cannot validate the corresponding expectations after the analysis. However, note that these expectations can be stated with any desired precision: We can demand an effect to be simply non-zero, to be positive, to be above a certain threshold, or even to be situated within a narrow neighbourhood of an exact target value.

(2) Estimating the probes should not be excessively complicated to perform in addition to the estimation of the target effect. In our example, we could reuse the causal graph that we had already constructed for target

effect estimation, to identify unbiased statistical estimands for the probes. The estimation itself only required fitting linear regression models and reading off the respective coefficients for each probe. In a setting where the estimation of the probes is more costly, the benefit of using quantitative probing could be overshadowed by the required additional effort.

(3) The parts of the DGP that are responsible for the target effect must be in some way related to the parts that are responsible for the probes. An example would be that all variables in the target effect and in the probes belong to the same connected component of the causal graph. Otherwise, our model might be perfectly accurate for the component that holds all the probes but flawed in the separate component that produces the target effect.

6 Simulation study

In this section, we provide experimental backing for the concept of quantitative probing as a method of validating causal models.

6.1 Simulation setup

To have access to both a ground truth and easy parameterization of the experiments, we chose a setup consisting of the repeated execution and evaluation of the following parameterized simulation run:

(1) Choose n (number of nodes), p_{edge} (edge probability), m (number of samples), p_{hint} (hint probability), p_{probe} (probe probability), and $\varepsilon_{\text{probe}}$ (probe tolerance).

(2) Draw a random directed acyclic graph (DAG) with n nodes $x_1, \ldots, x_n$. 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. If not, repeat the procedure.

(3) 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 $[0, 1]$.

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

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

(6) Randomly choose a non-trivial target effect and $p_{\text{probe}} \cdot n^2$ (rounded down) other treatment-outcome pairs that will serve as quantitative probes. By non-trivial, we mean that there exists a directed path from the treatment to the outcome in the causal graph, because otherwise any causal effect is trivially zero.

(7) Calculate the corresponding ATEs for the target effect and the probes from the fully specified causal Bayesian network, to obtain a ground truth.

(8) Run a causal end-to-end analysis, using the m observational samples and the qualitative domain knowledge, and report the discovered causal graph, the estimate of the target effect, and the hit rate for the quantitative probes. The hit rate is defined as the proportion of probes that have been correctly recovered by the analysis. To account for numerical errors and statistical fluctuations, we allow an absolute deviation of $\varepsilon_{\text{probe}}$ from the true value for a probe estimate to be considered successful.

(9) 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.

In this article, we report the results for experiments with $n = 7$ nodes, an edge probability of $p_{\text{edge}} = 0.1$, $m = 1{,}000$ samples per DAG, $p_{\text{hint}} = 0.3$, meaning that we suppose that we know 30% (rounded down) of the correct causal edges, $p_{\text{probe}} = 0.5$, meaning that we use half of the possible causal effects as quantitative probes, and $\varepsilon_{\text{probe}} = 0.1$, meaning that we consider probe estimates successful if they deviate no more than 0.1 from the

true value on an additive scale. As in the previous example, the causal discovery was performed using fast greedy equivalence search [46], and all ATEs were estimated using linear regression. Our hypothesis is that an end-to-end analysis that results in a high hit rate for the quantitative probes is more likely to have found both the true causal graph and the true target effect.

6.2 Used software

The programmatic implementation of the simulation relies on several open-source Python packages: At the beginning of the pipeline, the networkx package [47] was used for sampling DAGs based on the edge probability p_{edge}. The pgmpy package [45] enabled us to build a probabilistic graphical model from the given DAG by adding random CPDs for each node. The resulting model was then used to sample data for creating the observational dataset, as well as for obtaining the true ATEs by simulating data from interventional distributions. The causal end-to-end analysis and the validation of the quantitative probes were executed with the help of cause2e [10]. All plots were generated using Matplotlib [48]. To make the setup reusable, easily parametrizable, and open for extension by other researchers, an open-source Python package for quantitative probing [11] was developed around the aforementioned software setup (Section 9).

6.3 Results

To support the aforementioned hypothesis, we plot the aggregated results of 1,386 runs. Initially, 2,200 runs were performed, but 814 of them were aborted because of problems in the modelling process that would have required manual intervention. Figure 3 shows the information for every single run as a separate data point: The x-coordinate indicates the hit rate of the run in each of the four plots, whereas the y-coordinate describes varying quantities of interest:

- The plot to the upper left shows the absolute difference between the estimated and the true value of the target effect.
- The plot to the upper right shows the relative difference $|\frac{\hat{\tau}-\tau}{\tau}|$ between the estimated value $\hat{\tau}$ and the true value τ of the target effect. Note that no division by zero happens, since all target effects were chosen to be non-trivial. However, eight runs were excluded because the true causal effect was so close to zero that numerical instabilities occurred during the calculation of the relative estimation error.
- The plot to the lower left shows the structural hamming distance [49] between the true and the discovered graph. This includes both edges that are present in only one of the graphs as well as reversed edges.
- The plot to the lower right is the only aggregation: It shows the absolute frequencies of the hit rates over all runs.

As we can see, the data in the upper plots do not seem to support our hypothesis at all: We expect to see the points following a trend from the upper left (few successful probes, large estimation error) to the lower right (many successful probes, small estimation error), but there seems to be no downward trend. On the contrary, Figure 3 (upper left) even shows a higher number of significant estimation errors when the hit rate is high. At least the second concern can be resolved by a look at the hit rate frequencies in Figure 3 (lower right): Almost all of the runs show a hit rate of at least 0.8. The high ratio of successful probes is likely connected to the fact that many of the randomly chosen treatment-outcome pairs in the probes are not connected by a directed path in the true graph. Even if the discovered graph is not completely correct, it is sufficient not to introduce a directed path by mistake, to estimate the probe successfully. Therefore, the higher number of significant relative estimation errors in the plot can be explained by the higher number of data points for the corresponding hit rates. Similarly, the aforementioned apparent lack of a downward trend is simply due to

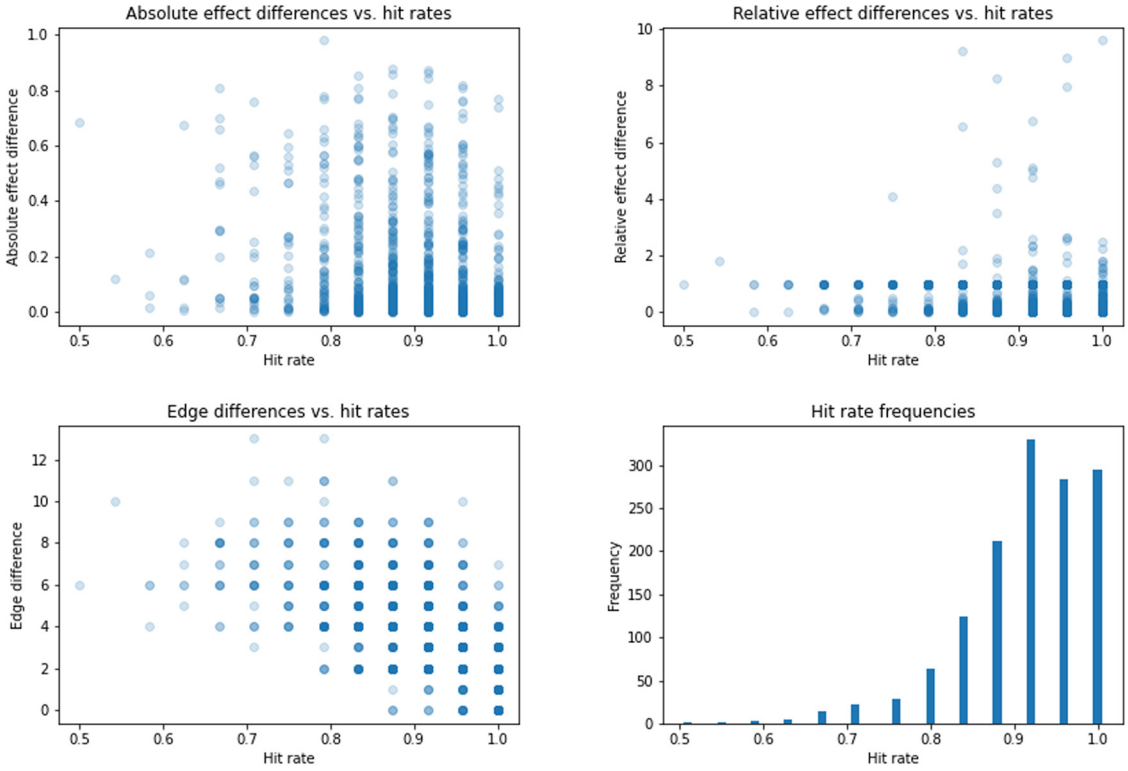

Figure 3: The top row shows plots of the absolute (left) and relative (right) differences between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the structural hamming distance between the true causal graph and the causal discovery result against 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.

visualization problems caused by the high number of data points. To resolve this issue, we replace the many single data points for each hit rate by one data point whose y-coordinate is the mean value over the plotted quantity for the given hit rate. The results are shown in Figure 4: Since the plots are derived from the same runs as before, the hit rate histogram in Figure 4 (bottom right) is unchanged. The other three plots now show the expected downward trend for the regions that contain a sufficient number of samples, indicating that runs with a higher hit rate performed better at estimating the target effect and recovering the causal graph from data and domain knowledge. In a first approximation, it even seems that the relationship between the hit rate and the other variables is linear in the higher hit rate region that holds most of the runs, but a theoretical foundation for this observation could not be established. As the results are produced by an empirical study and therefore subject to sampling variability, we include Figure 5 to account for the lack of uncertainty measures in Figure 4. The downward trend in means, medians, and quartiles is still clearly visible. A notable exception can be observed in the relationship between relative effect differences and hit rates: 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.

6.4 Outlier analysis

The results indicate that the probability of having recovered both the true causal graph and the correct target effect from observational data and domain knowledge increases with the amount of correctly estimated quantitative probes, i.e. the hit rate. However, the presented evidence only supports this in a probabilistic manner and even a perfect estimation of all probes does not guarantee a successful causal analysis. This

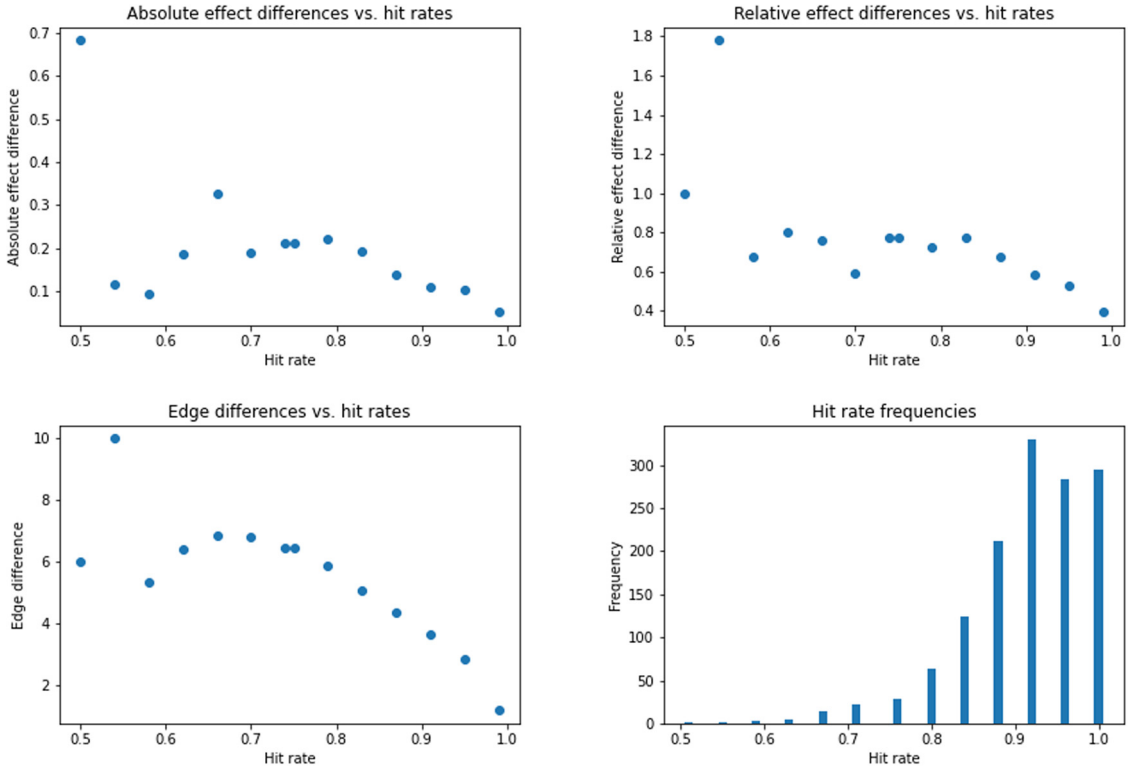

Figure 4: Aggregated results (means only): The top row shows plots of the mean absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the mean structural hamming distances between the true causal graph and the causal discovery result against the hit rate (left), as well as a histogram of the observed hit rates (right).

contrast is reflected in the aforementioned plots: On the one hand, the mean errors in Figure 4 approach 0, as the hit rate approaches 1. On the other hand, the non-aggregated plots in Figure 3 show numerous data points with a hit rate of 1 and considerable errors in both graph and target effect recovery. More precisely, our data contain 15 runs that simultaneously have a perfect hit rate of 1 and an absolute estimation error of at least 0.2. To understand the thereby evidenced limitations of the quantitative probing approach, we look at some of these outliers more closely.

6.4.1 Connectivity

Consider the run depicted in Figure 6: The target effect was the ATE of x_3 on x_5, which has been incorrectly estimated to be 0 instead of 0.5, although all of the probes have been correctly estimated. An examination of the graph structures immediately explains the phenomenon. The true graph and the discovered graph are identical, except for one edge between x_3 and x_5, which has been reversed by the causal discovery algorithm. This is not surprising, as the two structures are Markov equivalent and could only have been discerned by passing the correct orientation of the edge as domain knowledge. However, the only edge included in the domain knowledge is the edge between x_1 and x_6 (red). As a result, all of the probes have been estimated correctly, leading to a perfect hit rate of 1. Of course, the perfect performance in one connected component of the causal graph has no benefits for a task that relates only to the other component of the graph, such as the estimation of our target causal effect. These findings illustrate that it is not only important to correctly recover the probes but also important to select helpful probes in the first place. Given that we have not enforced any connectivity constraints during DAG generation, it is plausible that the proposed validation technique has encountered problems in graphs with multiple connected components.

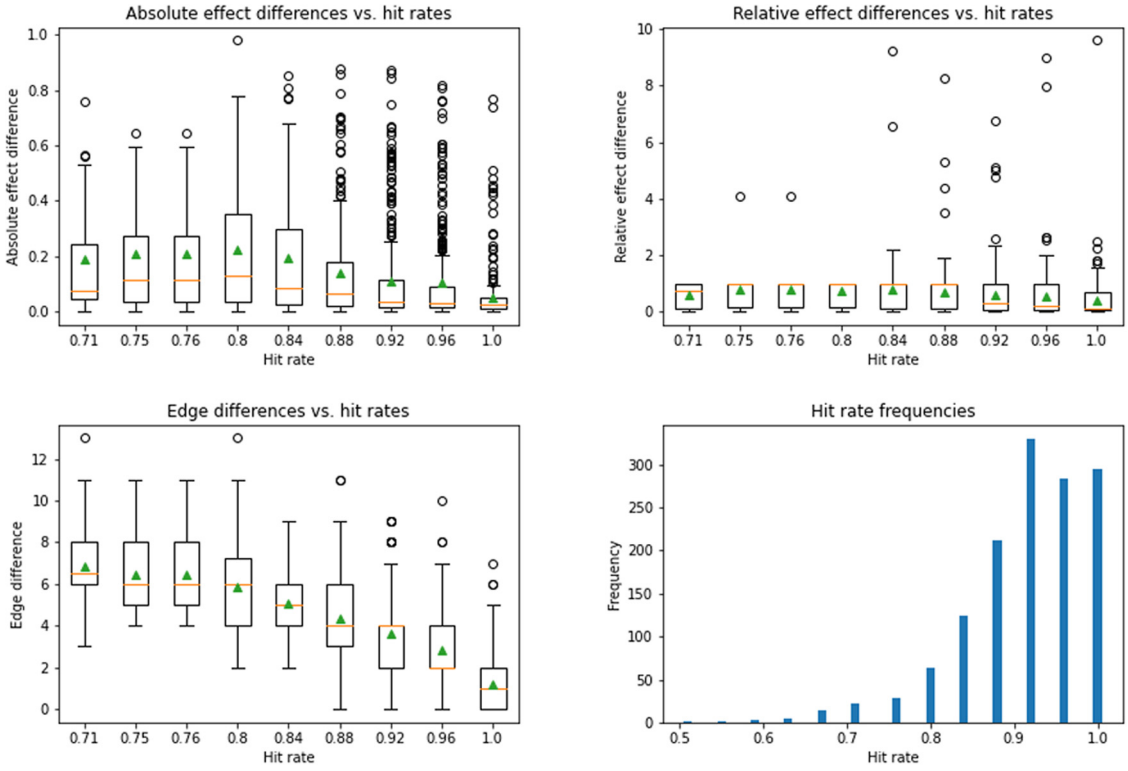

Figure 5: Aggregated results: The top row shows boxplots of the absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a boxplot of the structural hamming distances between the true causal graph and the causal 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.

In order to confirm these explanations, we filter out all experiment runs where the true causal graph consists of more than one connected component, and recreate Figures 3 and 4 from the reduced dataset of 656 runs. The resulting Figures 7–9 indeed show fewer deviations from the bottom right, indicating that a sizeable proportion of the runs where our validation approach failed were linked to the problem of disconnected graphs. If we apply the aforementioned outlier filter, we are left with only four runs with perfect hit rate and an absolute estimation error of over 0.2. In practice, this suggests that we should choose probes that we suspect to be in the same component of the causal graph as the target variables. It seems plausible that even within one component, the probes that are closer to the target effect will be more useful for judging the correctness of the model.

6.4.2 Probe coverage

In order to understand the remaining outliers, we offer two explanations: The first one is related to the probe coverage. Given that in most applications, it is unrealistic to assume that the analyst knows all causal effects except for the target effect, we have selected only 24 of the $7 \cdot 7 = 49$ causal effects as probes. This suggests that some of the erroneous analyses in the outlier runs could have been captured by increasing the number of probes, as illustrated in Figure 10: The estimation of the ATE of x_5 on x_6 has yielded a result of 0 instead of the true value 0.28, although all the probes have been correctly estimated. A closer look at the 24 probes reveals that the ATE of x_3 on x_6 has not been used as a probe. This probe could have detected the incorrect graph, since its effect is trivially zero in the discovered graph, but non-zero in the true graph.

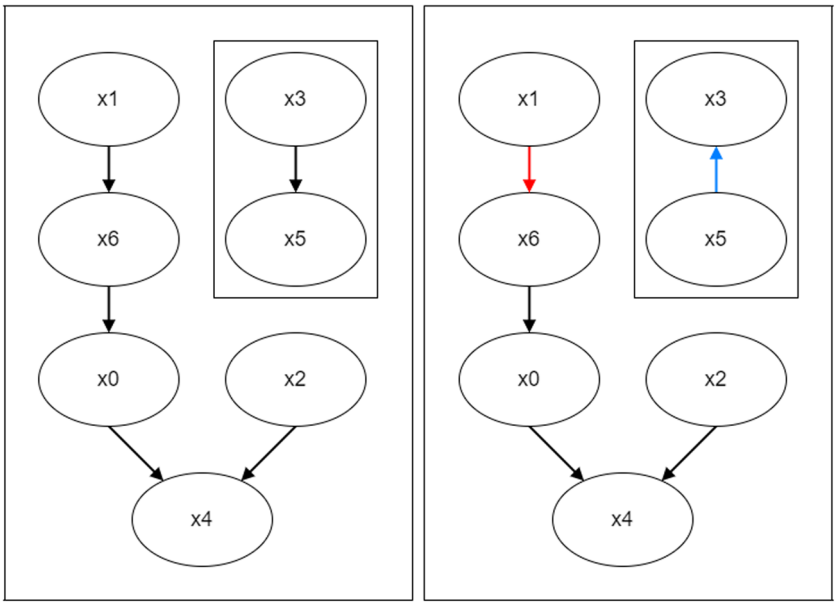

Figure 6: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_3 and outcome x_5 (surrounded by small box). In the discovered graph, the red edge $x_1 \to x_6$ has been required by domain knowledge. The edge $x_3 \to x_5$ (blue) has been oriented incorrectly.

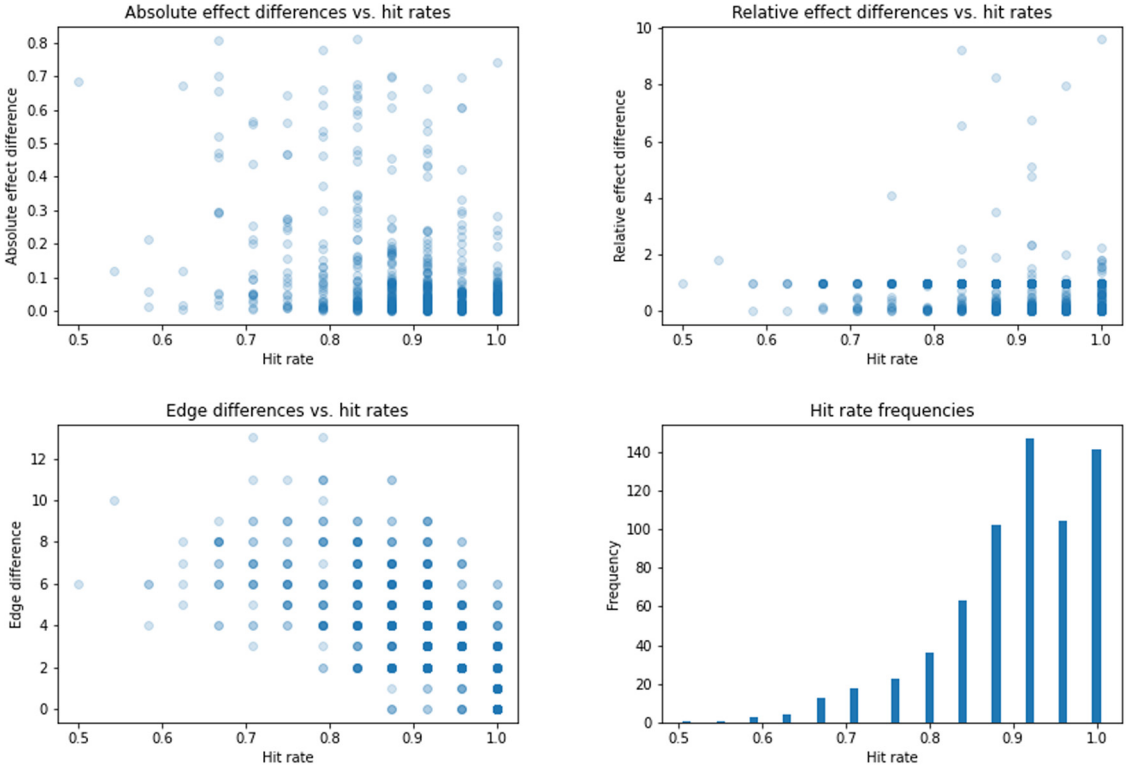

Figure 7: Results for all runs with connected causal graphs: The top row shows plots of the absolute (left) and relative (right) differences between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the number of structural hamming distances between 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.

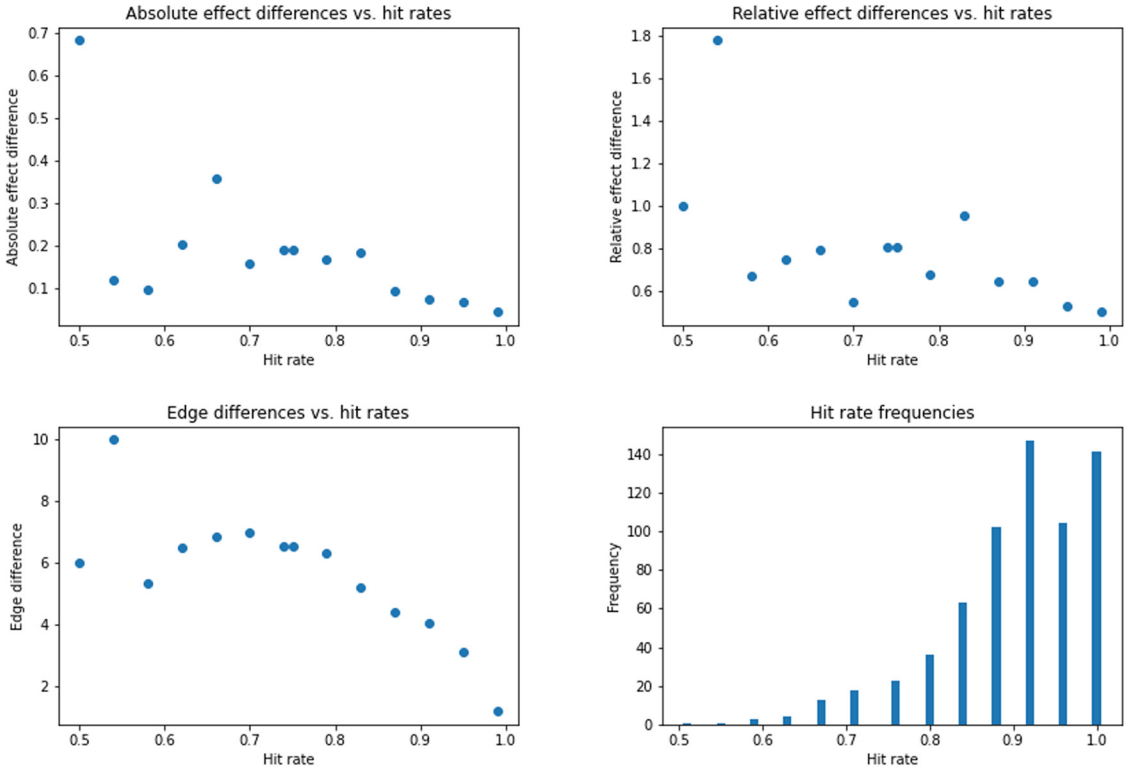

Figure 8: Aggregated results for all runs with connected causal graphs (means only): The top row shows plots of the mean absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a plot of the mean structural hamming 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).

6.4.3 Probe tolerance

The second explanation focuses on the definition of the hit rate, which is clearly a deciding factor for marking a run as an outlier. The hit rate is high if many probes have been correctly estimated, and "correctly" means that the estimate has to lie within some reasonable bounds around the true effect. While this seems straightforward, the problem lies in the definition of the bounds: Should we specify an absolute error margin that applies to all of the probes? Or should we specify a relative margin that depends on the size of each of the probe effects? Using an absolute margin of $\varepsilon_{\text{probe}} = 0.1$ to both sides, as we did in our experiments, can be a good fit for a true effect size of 1, but it might be dangerous for a true effect size of 0.001 (underreject) or 1,000 (overreject).

To illustrate this line of thought, we look at the run in Figure 11: The ATE of x_1 on x_2 was erroneously estimated to be 0 instead of −0.24, although all of the probes were estimated correctly. In this case, it is remarkable that the ATE of x_4 and x_2 was one of the probes. This probe is trivially 0 in the discovered graph, as there is no directed path from x_4 to x_2. In the true graph, however, it can only vanish in the degenerate case. Indeed, the true effect is 0.07. Given our generic bounds determined by $\varepsilon_{\text{probe}} = 0.1$, the incorrect estimate of 0 falls within the acceptance interval $[-0.03, 0.17]$ and the error goes unnoticed. This could have been avoided by specifying proper bounds for each of the probes, a task whose feasibility in practice depends on the available domain knowledge. It is worth noting that all the inspected outliers either show a true or an estimated target effect that is trivially 0. The absence of more subtle estimation problems where the directed path exists in both, the true and the recovered graph, but the estimation is jeopardized by a difference in backdoor paths, is probably due to the low number of variables in the generated DAGs. However, quantitative probing is applicable without any modifications to more complicated graph structures.

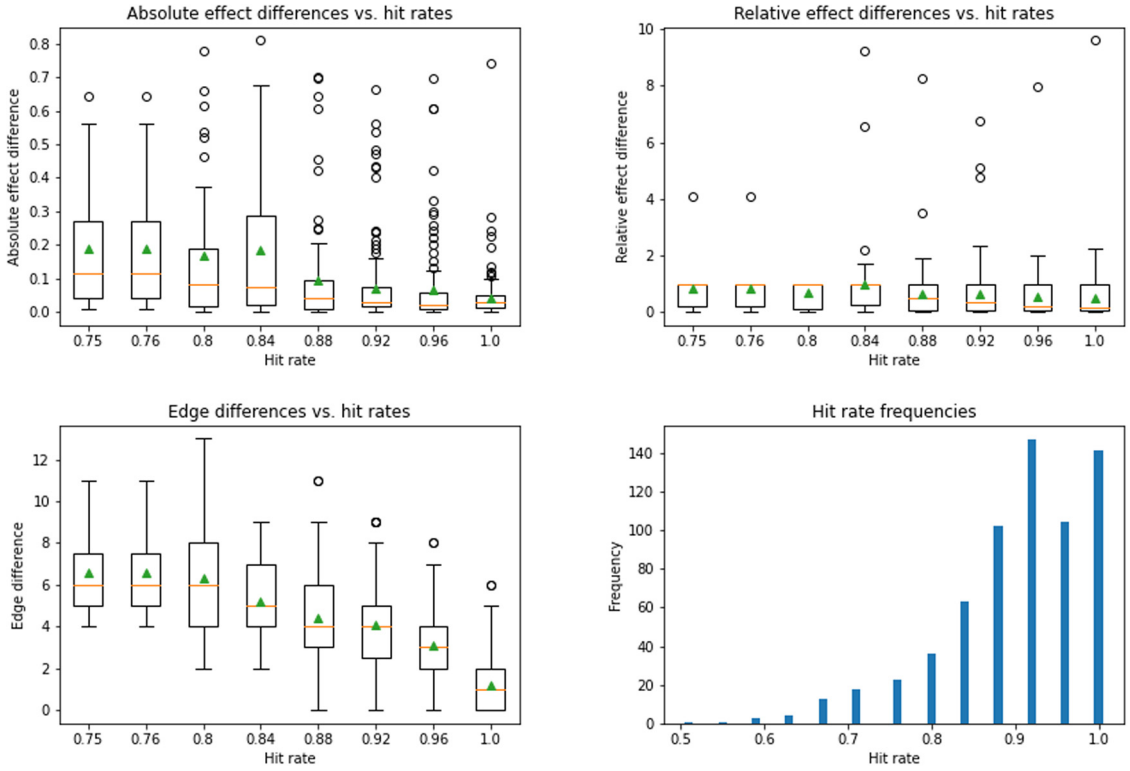

Figure 9: Aggregated results for all runs with connected causal graphs: The top row shows boxplots of the absolute (left) and relative (right) difference between the true target effect and the estimated result against the hit rate. The bottom row shows a boxplot of the structural hamming distances between 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.

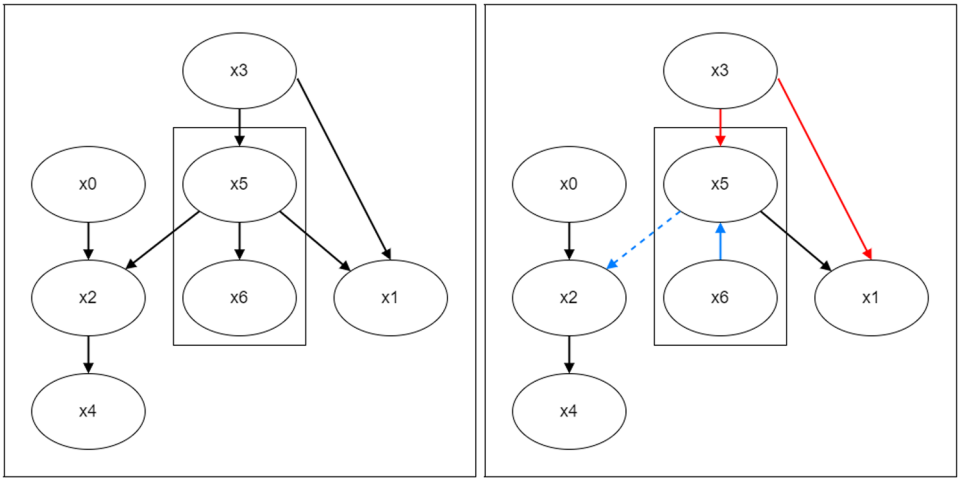

Figure 10: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_5 and outcome x_6 (surrounded by small box). In the discovered graph, the red edges $x_3 \to x_5$ and $x_3 \to x_1$ have been required by domain knowledge. The edge $x_5 \to x_6$ (blue) has been oriented incorrectly and the edge $x_5 \to x_2$ (blue dotted) is missing.

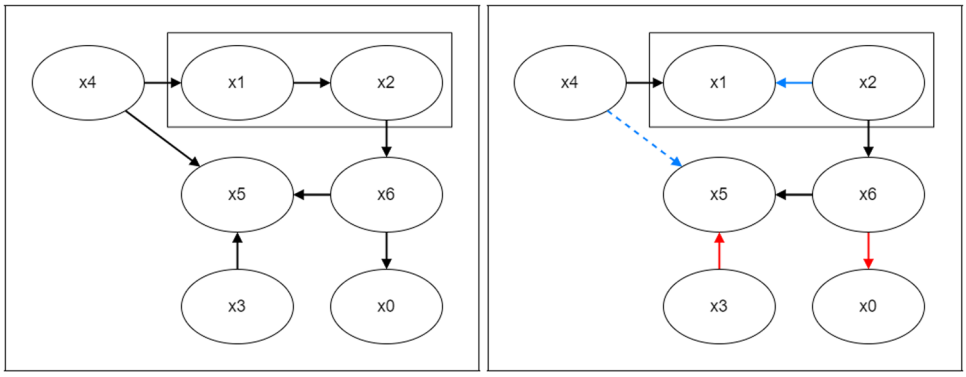

Figure 11: Plot of the true (left) and discovered (right) causal graphs for an outlier run with treatment x_1 and outcome x_2 (surrounded by small box). In the discovered graph, the red edges $x_6 \to x_0$ and $x_3 \to x_5$ have been required by domain knowledge. The edge $x_1 \to x_2$ (blue) has been oriented incorrectly and the edge $x_4 \to x_5$ (blue dotted) is missing.

7 Guidelines for practitioners

The previous sections have shown that quantitative probing is an effective validation strategy, but we have also seen some obstacles for practitioners that are willing to apply the method. Therefore, we provide guidance on how to avoid these pitfalls when implementing the validation-related steps 1, 4, and 5 depicted in Figure 1.

7.1 Probe selection and specification

Probe selection is mainly driven by the available quantitative domain knowledge, since no specification of a probe's desired value is possible otherwise. In general, the discussion of Figure 10 indicates that it is beneficial to provide a high number of probes, as each probe increases our chances of finding discrepancies between the prediction of the model and the domain knowledge. However, as we have stated in Section 5.4 and seen in Figure 6, it is not enough to amass as many probes as possible to make the validation step as effective as possible: Probes about variables that are not sufficiently closely related to the variables in the target effect can be of limited use for filtering out misspecified causal models. The notion of closeness will depend on the choice of model and needs to be judged jointly by the modeller and the domain expert. In scenarios with few variables, the connected components of the causal graph can likely be identified by the domain expert, even if the precise orientation of the edges is not known *a priori*. Section 5.2 points out the widespread circular reasoning of comparing the target estimate itself to domain experts' expectations. In light of the aforementioned discussion, we must acknowledge that quantitative probing, by focussing on non-target effects, only resolves this issue at the cost of incurring another debt: Contrary to the aforementioned practice, we do not need to assume any knowledge about the strength of the target effect, such that the logic behind the proposed validation strategy is no longer circular. However, we need to know the DGP behind the target effect sufficiently well to identify suitable probes for assessing the quality of a causal model with respect to its capability of estimating of the target effect. Although this new challenge did not constitute a major obstacle for most runs in our simulation study, at this point, we cannot yet reliably gauge its difficulty in real applications.

7.2 Probe evaluation

In practical applications, it is unreasonable to expect a perfect estimation of a probe, given that only a finite amount of data is used for the modelling and prediction steps. Therefore, it is necessary to specify bounds within that we deem the probe recovery successful. The discussion of Figure 11 underlines the critical role of

these bounds for the effectiveness of the validation concept. We expect that domain experts favour broader specification of domain knowledge items, such as the nonnegativity of a given probe, over precise point estimates. In these cases, the appropriate bounds are given by 0 and $\pm\infty$, and the issue disappears on one side. However, even in the frequent case where the domain expert specifies that a probe should have a value of 0, it is necessary to specify bounds around this point estimate. As our simulation study has shown, it is dangerous to choose a uniform value for all error tolerances. We strongly recommend practitioners to carefully select the tolerances individually for each quantitative probe. Appropriate values must be provided in cooperation with the domain expert who can judge which deviations can be considered reasonable based on the properties of the application domain. In addition, the sample size contributes to the statistical precision of the estimates and should therefore be used to decide on the tolerance levels. Not only the overall number of observational samples that was available for model construction should be taken into account but also other factors that determine the effective sample size, such as the expected density of the underlying DAG or the complexity of the involved estimators, must not be ignored either.

7.3 Model evaluation

Our simulation study shows that, on average, models with a higher success at recovering the probes perform better at the original task of correctly estimating the target effect. However, the practitioner's next steps after determining the hit rate still depend on the circumstances and goals of each individual modelling task. We also briefly discuss to which extent quantitative probing can be employed outside of the original validation context to perform model adaptation and selection.

7.3.1 Setting the acceptance threshold

It is up to the practitioner to decide which threshold on the hit rate should be used as a cut-off value for judging the fitness of the model. If the domain expert is absolutely sure that a correct model should recover all probes within the prespecified tolerances, then only a perfect hit rate of 100% should lead to accepting the candidate model. If there are doubts about the truthfulness of some domain knowledge items, then these probes should either be removed right from the start of the analysis, or the hit rate threshold for accepting the candidate model should be lowered accordingly. When following the latter route, the threshold should also be adjusted for the overall number of probes: Assuming fixed tolerance levels, the hit rate will be lower in scenarios with more probes because of statistical uncertainties. While the aforementioned work by Karmakar and Small [30] could derive concrete guidelines for combining and evaluating evidence from multiple tests in the context of their framework, our limited understanding of the statistical theory underlying quantitative probing precludes us from doing so. Therefore, we suggest inspecting each failed probe manually and deciding together with a domain expert whether the discrepancy should be attributed to statistical fluctuations or taken as evidence against the validity of the model under consideration.

7.3.2 Adapting rejected models

In Figure 1, a successful model evaluation is followed by the prediction of the target effect. But how should practitioners proceed if the evaluation refutes the candidate model? In principle, it is possible to learn from the shortcomings of the candidate model and design a new version that performs better at probe recovery. However, this new version can no longer be validated using the same probes, as this would constitute an instance of overfitting to the probes. In analogy to the train/test split in correlation-based machine learning, practitioners can possibly circumvent the issue by splitting the available quantitative probes into two sets: The first set serves to validate and adapt the candidate model in an iterative procedure, thereby contributing to the

improved quality of the final model, but being unsuitable for validating the final model due to possible overfitting. The second set is not used until the final model has been derived and can then be employed in a last unbiased validation step. We did not provide a simulation study for this scenario, since it is unclear how the adaptation of the model could be automated in a way that allows the generation of sufficiently many runs to statistically assess the effectivity of the concept. While enabling a safe adaptation process of the candidate model, a practical obstacle in keeping a hold-out set of probes for final validation is the often limited amount of domain knowledge that prohibits a further split.

7.3.3 Quantitative probing for model selection

It is possible to probe not only one, but multiple candidate models that have been built independently from each other, to select the one with the highest hit rate for further downstream tasks. This procedure is consistent with the logical foundation outlined in Section 5: Popper's notion of competing theories [9] corresponds to the multiple candidate models that compete for being selected as the best model. However, practitioners need to keep in mind that the best model of a given candidate set does not necessarily need to be a good model. If an ill-suited candidate set is chosen for a problem, the best of the candidates can still perform poorly in the application task. Without further simulation-based or theoretical backing, we therefore cannot safely recommend to perform model selection by quantitative probing in cases where no additional validation of the selected model is possible.

8 Conclusion and outlook

In summary, we have introduced the method of quantitative probing for validating causal models. We identified the additional difficulties in validating causal models when compared to traditional correlation-based machine learning models by reviewing the role of the i.i.d. assumption. In analogy to the model-agnostic train/test split, we proposed a validation strategy that allows us to treat the underlying causal model as a black box that answers causal queries. The strategy was put into context as an extension of the already existing technique of using refutation checks, which is in line with the established logic of scientific discovery, by exploiting domain-specific knowledge. After illustrating the motivation behind the concept of quantitative probing using Pearl's sprinkler example, we presented and discussed the results of a thorough simulation study. While being mostly supportive of our hypothesis that quantitative probes can be used as an indicator for model fitness, the study also revealed shortcomings of the method, which were further analyzed using exemplary failing runs. Finally, the results of the simulation study were complemented by a guide for practitioners with the aim of facilitating the incorporation of quantitative probing in causal modelling applications.

To conclude this article, we revisit the assumptions from Section 5.4 together with the ideas from Sections 6.4 and 7, to identify topics for future research.

(1) Is there a way to explain the seemingly linear association between the hit rate and the edge/effect differences in Figure 4? Answering this question would provide a solid theoretical backing for the quantitative probing method, in addition to the simulation-based evidence presented in this article.
(2) Can we quantify the "usefulness" of each employed quantitative probe for detecting wrong models? An example would be to plot the results for a single fixed hit rate, but each data point could represent runs that used only specific probes, e.g. only those whose variables lie within a maximum distance from the target variables. What other criteria could make a specific probe useful? Answering this question would aid practitioners in eliciting specific quantitative knowledge from domain experts.
(3) Given the probe coverage issue in Figure 10, is there a way to model how the effectiveness of the method depends on the number of used quantitative probes? Answering this question would aid practitioners in gauging the amount of required quantitative domain knowledge for model validation.

(4) Given the probe tolerance issues in Figure 11, how much more useful do the probes become for detecting wrong models if we narrow the allowed bounds in the validation effects? How can we determine ideal bounds to avoid overreject and underreject? Answering this question would provide important guidance to both researchers and practitioners, as the bounds must be actively chosen by the investigator in each application of quantitative probing.

Similarly to how qualitative knowledge by domain experts can assist causal discovery procedures in recovering the correct causal graph from observational data, we believe that quantitative probing can serve as a natural method of incorporating quantitative domain knowledge into causal analyses. By answering the aforementioned questions, the proposed validation strategy can evolve into a tool that builds trust in causal models, thereby facilitating the adoption of causal inference techniques in various application domains.

9 Code and data availability

All of the aforementioned results can be reproduced by running two notebooks in the companion GitHub repository of this article, which also contains the open-source qprobing Python package [11]. The qprobing package relies on the open-source cause2e Python package for performing the causal end-to-end analysis, which is hosted in a separate GitHub repository [10].

Acknowledgment: We are grateful for the support of our colleagues at ams OSRAM and the University of Regensburg. Sebastian Imhof helped sharpen the idea of using known causal effects for model validation during multiple fruitful conversations. Other validation approaches for causal models were compared in several meetings of our causal inference working group [50]. The idea of different degrees of usefulness for different types of quantitative probes was brought up in a discussion with members of the Department of Statistical Bioinformatics at the University of Regensburg. Heribert Wankerl deserves special credit for proof-reading the article. Finally, we thank three anonymous reviewers for their thoughtful advice that inspired substantial improvements of the manuscript.

Funding information: The authors state no funding involved.

Conflict of interest: The authors state no conflict of interest.

References

[1] Pearl J. Causality. 2nd ed. Cambridge, UK: Cambridge University Press; 2009.
[2] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Amer Stat Assoc. 1996;91(434):444–55. https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.
[3] Rubin DB. Causal inference using potential outcomes. J Amer Stat Assoc. 2005;100(469):322–31. doi: 10.1198/016214504000001880.
[4] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. 2nd ed. Cambridge, Massachusetts: MIT Press; 2000.
[5] Peters J, Janzing D, Schölkopf B. Elements of causal inference - foundations and learning algorithms. Adaptive computation and machine learning series. Cambridge, MA, USA: The MIT Press; 2017.
[6] Holland PW. Statistics and causal inference. J Amer Stat Assoc. 1986;81(396):945–60. doi: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.
[7] Kendall JM. Designing a research project: randomised controlled trials and their principles. Emergency Med J. 2003;20(2):164–8. https://emj.bmj.com/content/20/2/164.
[8] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd edn. Springer series in statistics. Heidelberg, Germany: Springer; 2009. doi: 10.1007/978-0-387-84858-7.
[9] Popper KR. The logic of scientific discovery. London: Hutchinson; 1934.
[10] Grünbaum D. Cause2e: A Python package for causal end-to-end analysis; 2021. https://github.com/MLResearchAtOSRAM/cause2e.

[11] Grünbaum D. Qprobing: a python package for evaluating the effectiveness of quantitative probing for causal model validation; 2022. https://github.com/MLResearchAtOSRAM/qprobing.
[12] Abrevaya J, Hsu YC, Lieli RP. Estimating conditional average treatment effects. J Business Econ Stat. 2015;33(4):485–505. doi: 10.1080/07350015.2014.975555.
[13] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. JNCI: J Nat Cancer Institute. 1959 Jan;22(1):173–203. doi: 10.1093/jnci/22.1.173.
[14] Jesson A, Mindermann S, Gal Y, Shalit U. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding; 2021. https://arxiv.org/abs/2103.04850.
[15] Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. J R Stat Soc Ser B.2020;82(1):39–67. https://EconPapers.repec.org/RePEc:bla:jorssb:v:82:y:2020:i:1:p:39-67.
[16] Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V. Long story short: omitted variable bias in causal machine learning; 2021. https://arxiv.org/abs/2112.13398.
[17] Veitch V, Zaveri A. Sense and sensitivity analysis: simple post-hoc analysis of bias due to unobserved confounding. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems. Vol. 33. Red Hook, New York: Curran Associates, Inc.; 2020. p. 10999–1009. https://proceedings.neurips.cc/paper/2020/file/7d265aa7147bd3913fb84c7963a209d1-Paper.pdf.
[18] Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika. 1987 Mar;74(1):13–26. doi: 10.1093/biomet/74.1.13.
[19] Rosenbaum PR. Sensitivity to hidden bias. New York, NY: Springer; 2002. p. 105–70. doi: 10.1007/978-1-4757-3692-2_4.
[20] Rosenbaum PR. Sensitivity analysis in observational studies. Hoboken, New Jersey: John Wiley & Sons, Ltd; 2014. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06358.
[21] Rolling CA, Yang Y. Model selection for estimating treatment effects. J R Stat Soc Ser B (Stat Methodol.). 2014;76(4):749–69. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12043.
[22] Schuler A, Baiocchi M, Tibshirani R, Shah N. A comparison of methods for model selection when estimating individual treatment effects; 2018. https://arxiv.org/abs/1804.05146.
[23] Kyono T, Bica I, Qian Z, van der Schaar M. Selecting treatment effects models for domain adaptation using causal knowledge; 2021. https://arxiv.org/abs/2102.06271.
[24] Dwivedi R, Tan YS, Park B, Wei M, Horgan K, Madigan D, et al. Stable discovery of interpretable subgroups via calibration in causal studies. Int Stat Rev. 2020;88(S1):S135–78. https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12427.
[25] Neal B, Huang CW, Raghupathi S. RealCause: realistic causal inference benchmarking; 2020. https://arxiv.org/abs/2011.15007.
[26] Alaa A, Van Der Schaar M. Validating causal inference models via influence functions. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on vol. 97 of Proceedings of Research. PMLR; 2019. p. 191–201. https://proceedings.mlr.press/v97/alaa19a.html.
[27] Tran D, Ruiz FJR, Athey S, Blei DM. Model criticism for bayesian causal inference; 2016. https://arxiv.org/abs/1610.09037.
[28] Box GEP. Sampling and Bayes' inference in scientific modelling and robustness. J R Stat Soc. Ser A. 1980;143:383–430.
[29] Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Stat Sinica. 1996;6:733–807.
[30] Karmakar B, Small DS. Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. Ann Stat. 2020;48(6):3283–311. doi: 10.1214/19-AOS1929.
[31] Biza K, Tsamardinos I, Triantafillou S. Tuning causal discovery algorithms. In: Jaeger M, Nielsen TD, editors. Proceedings of the 10th International Conference on Probabilistic Graphical Models. vol. 138 of Proceedings of Machine Learning Research. PMLR; 2020. p. 17–28. https://proceedings.mlr.press/v138/biza20a.html.
[32] Sharma A, Syrgkanis V, Zhang C, Kıcıman E. DoWhy: addressing challenges in expressing and validating causal assumptions; 2021. https://arxiv.org/abs/2108.13518.
[33] Meek C. Causal inference and causal explanation with background knowledge. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 403–10.
[34] Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. J Stat Software. 2012;47(11):1–26. https://www.jstatsoft.org/index.php/jss/article/view/v047i11.
[35] Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res. 2003 Mar;3:507–54. doi: 10.1162/153244303321897717.
[36] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genetics. 2019;10:524. https://www.frontiersin.org/article/10.3389/fgene.2019.00524.
[37] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Appl Inform. 2016;3(3):1–28.
[38] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-gaussian acyclic model for causal discovery. J Machine Learn Res. 2006;7(72):2003–30. http://jmlr.org/papers/v7/shimizu06a.html.
[39] Hoyer PO, Shimizu S, Kerminen AJ, Palviainen M. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. Int J Approximate Reason. 2008;49(2):362–78. Special section on probabilistic rough sets and special section on PGM'06. https://www.sciencedirect.com/science/article/pii/S0888613X08000212.
[40] Peters J, Janzing D, Schölkopf B. Identifying cause and effect on discrete data using additive noise models. In: Teh YW, Titterington M, editors. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. vol. 9 of

Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR; 2010. p. 597–604. https://proceedings.mlr.press/v9/peters10a.html.
[41] Peters J, Mooij JM, Janzing D, Schölkopf B. Identifiability of causal graphs using functional models. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. UAI'11. Arlington, Virginia, USA: AUAI Press; 2011. p. 589–98.
[42] Zhang K, Hyvärinen A. On the identifiability of the post-nonlinear causal model. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI '09. Arlington, Virginia, USA: AUAI Press; 2009. p. 647–55.
[43] Heinze-Deml C, Maathuis MH, Meinshausen N. Causal structure learning. Ann Rev Stat Appl. 2018;5(1):371–91.
[44] Shimizu S, Inazumi T, Sogawa Y, Hyvärinen A, Kawahara Y, Washio T, et al. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. J Mach Learn Res. 2011 Jul;12(null):1225–48.
[45] Ankan A, Panda A. Pgmpy: Probabilistic graphical models using python. In: Proceedings of the 14th Python in Science Conference (SCIPY 2015). Austin, Texas: SciPy; 2015.
[46] Ramsey J, Glymour M, Sanchez-Romero R, Glymour C. A million variables and more: the fast Greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int J Data Sci Anal. 2017 March;3:121–9.
[47] Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using network. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. p. 11–5.
[48] Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5.
[49] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn. 2006 Oct;65(1):31–78. doi: 10.1007/s10994-006-6889-7.
[50] Grünbaum D. Causal inference working group; https://gitlab.com/causal-inference/working-group/-/wikis/home.