Causal Discovery from High-dimensional Observational Data

by

Mehrdad Mansouri

M.A.Sc., University of Victoria, 2014
B.A.Sc., Sadjad University of Technology, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Mehrdad Mansouri**
Degree:	**Doctor of Philosophy**
Thesis title:	**Causal Discovery from High-dimensional Observational Data**
Committee:	Chair: Anoop Sarkar Professor, Computing Science
	Martin Ester Supervisor Professor, Computing Science
	Oliver Schulte Committee Member Professor, Computing Science
	Richard Lockhart Examiner Professor, Statistics and Actuarial Science
	Jiuyong Li External Examiner Professor, Computer Science University of South Australia

Abstract

With the rise of digital observational data, there has been an increasing amount of attention to the discovery of causal relations from large datasets. In the last three decades two major approaches have emerged to deal with high-dimensional Causal Discovery; well-known SL methods optimized by bypassing their exponentially growing conditioning tests, and quasi-experimental designs equipped with machine learning algorithms to efficiently search for promising hypotheses. These methods have mainly focused on dealing with the computational complexity and the assumptions made about the data.

In this thesis, the goal is to expand the use of these approaches, by attempting to solve various types of problems that are inspired by real-world applications, and are hard or impossible to solve by the general causal discovery methods. First is the Relational Causal Discovery, in which the prior knowledge of the association between variables can be used to improve accuracy and reliability. Second is the Stratified Causal Discovery, which identifies causes for different subpopulations, and potentially different underlying mechanisms. Third is the Causal Profile Discovery, which specifies the temporal sequence of causes to have the most significant effect. Fourth is the Compound Causal Discovery, which identifies the sets of causes that are only jointly sufficient. The thesis will be concluded by discussing the applications of the proposed methods and how they can be used as platforms for other potential problems.

Keywords: Causal Discovery; High-dimensionality

Dedication

I dedicate this to all those who made this dark universe a little more enlightened, first and foremost, my grandfather, who taught me how to think, and my dog, who taught me how to live. Thank you.

Acknowledgements

I have received a great amount of support and guidance for this. I would like to thank my supervisor Prof. Martin Ester who was very patient and understanding, and always guiding my research in a reasonable direction; Prof. Leonid Chindelevitch whose insights were invaluable in fixing my broken ideas; my colleagues Dr. Sahand Khakabimamaghani, Ali Arab, Rita Vityaz, and Bowei Yuan, who were patient with my flaws and assisted me in the research; our collaborators in the Genomic and Outcomes Database for Pharmacogenomics and Implementation Studies, especially Prof. Bruce Carleton and Prof. Colin Ross, who provided the funding, data, and medical insights, without which my work would have no hope of real-world impact; my friends Dr. Hossein Sharifi and Tong He who humored my pessimistic views on our research; and most importantly my family, Bahman, Gita, and Maral, for supporting me selflessly.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 High-dimensional Causal Discovery Premise

Causal Discovery (CD), as the systematic search of the data for cause-effect relationships between variables, has grown out of the field of Causal Inference, which strives to statistically evaluate whether two particular variables are causally related or not [183]. High-dimensional Causal Discovery (HCD), which addresses the CD problems when the number of potential causes and confounders is larger than to be properly analyzed even by the multivariate CD approaches, is one of the most promising extensions of CD.

Although causality has been historically significant to many fields from philosophy [84] to social [51] and natural sciences [27], capturing causal relations has been ambiguous and challenging. However, in recent years, there has been an increasing amount of attention to CD, especially HCD [2]. There are three main reasons for HCD's increasing relevance:

- With the prevalence of large datasets with many variables and samples, the burden of analysis has shifted from epistemological theorization and controlled experiments to efficient quantitative algorithms.

- Performing sophisticated CD search on a large number of variables is becoming more feasible [111].

- There are many emerging sub-fields and interdisciplinary applications with large data archives but with limited understanding of the underlying causal process [162, 32, 25].

The main challenges of HCD are those rooted in CD that are exacerbated by the high-dimensionality. Specifically, the large number of potential causes results in:

- CD methods which are typically super-exponential [64] to become computationally prohibitively expensive.

- Severe data dredging, which potentially leads to many false discoveries, or missing many legitimate causes that do not have a very high statistical power in the multiple hypothesis testing adjustment.

This thesis discusses the implications of the high-dimensionality challenge, and the various preliminary search and filtering strategies for dealing with it.

Rooted in the aforementioned common interest in causality, a key feature of HCD is that its theories, assumptions, methods, and evaluations are developed from multiple disciplines [66, 6]. This brings in a wide range of approaches that can compete and be combined to great success, but also redundant and overlapping definitions, and assumptions and parameters that are unrealistic for the use outside the local applications in mind [50]. Hence, in this thesis every effort is made to maintain the framework general, consistent and unified; and make the utmost use of all the rich diversity of the different techniques available, while also analyzing and proposing methods as processes with interchangeable blocks.

Problems and methods in this thesis usually assume the values of the variables to be discrete; although with minor changes most methods can be adjusted to continuous data. There are three reasons for this:

- Discrete data can be modeled through Multinomial distribution, which reduces the complexity in training and evaluation of most CD methods.
- Discretization has been shown to improve the prediction accuracy and generality of many CD methods [183, 73, 86].
- Most of the currently available high dimensional data are discrete in nature [135].

1.1.1 Goals

The Objective of this thesis is to extend the applicability of HCD. The current space of HCD development is mainly focused on efficient and accurate methods for the general HCD problem. HCD can benefit from defining and solving specific problems, inspired by reoccurring real-world applications, such as finding causes in active and reinforcement learning settings, exact function between the cause and the outcome, and distinguishing causes that are sufficient from those that are necessary. Contribution of this thesis is formalizing four of these problems that are hard or impossible for the general HCD methods, and proposes methods for solving them that work even in the high dimensional setting, and applies them to real-world problems to demonstrate their applicability and performance:

1.2 Application

Throughout this thesis, we used pharmacogenomics as the baseline application for evaluations with real-world data. In fact, ARISTOTLE and HUME were originally published with the pharmacogenomic applications in mind. The main reasons for this decision are:

Problem	**Method**
Relational Causal Discovery: Discovery when prior knowledge of causal relations or associations between some of the variables is available.	HUME using Network Analysis and Matching
Stratified Causal Discovery: Discovery of causes of specific subpopulations or underlying mechanisms.	ARISTOTLE using Divide & Conquer and Biclustering
Causal Profile Discovery: Discovery of temporal patterns for which causes have the most significant effect.	HEIDEGGER using Graph Traversal and Randomized-Block Design
Compound Causal Discovery: Discovery of combinations of causes whose joint occurrence is required to change the outcome.	HEGEL using Subgroup Discovery and Markov Blanket Assembly

- Pharmacogenomic datasets are usually very high dimensional and with variables of different types, containing millions of genetic, transcriptomics, proteomics and metabolomics biomarkers as well as demographics and clinical information [186], which turns pharmacogenomic data to ideal test ground for high-dimensional models.

- The interactions between the drugs, diseases, and the patient involve complex networks of causal relations at both micro and macro levels [206].

- We were fortunate to be part of the large and multiple disciplinary project, Genomic and Outcomes Database for Pharmacogenomics and Implementation Studies (Go-PGx) [194], which gave us the access to valuable pharmacogenomic data and advice from researchers with different backgrounds.

- There is a large body of knowledge gathered of pharmacogenomic causal relations, including experiment design and cohort studies, in vivo and in vitro controlled tests, and pathological study of causal mechanisms. The are also various high quality aggregation sources, from knowledge bases such as PharmGKB [16] and Kyoto Encyclopedia of Genes and Genomes Pathway [87], to periodically published reviews and metadata analyses that reproduce validity of discoveries of particular studies.

- Pharmacogenomics is an active application that can directly and significantly impact life of patients, and in the last decade many research studies with different approaches, including CD, have been working on it.

- Pharmacogenomics is a flexible application that we can apply the proposed problems for; hence providing a baseline to compare the performance of different introduced or baseline methods across our studies.

Our ultimate hope is that in addition to effective and accurate evaluation of the proposed problems and methods, our CD of the pharmacogenomic datasets may provide insights into the underlying biological mechanisms of these adverse drug reactions (ADR)s and pave the way for personalized medicine models that minimize their effect.

Under the Go-PGx project, we applied the proposed problems and methods to a particular Pharmacogenomic application: discovering genetic variant causes of anthracycline induced cardiotoxicity in children treated for cancer. The non-sparse, relatively small sample size, but high-dimensional and comprehensive patient records of the dataset that we use for this application makes them ideal for HCD.

1.2.1 Anthracycline Cardiotoxicity Dataset

More than half of childhood cancer treatment protocols include an Anthracycline. However, the usefulness of Anthracyclines is limited by asymptomatic cardiac dysfunction and heart failure [49], which makes Anthracycline Cardiotoxicity one of the more impactful applications in pharmacogenomics.

The particular dataset that we used is based on the Canadian Pharmacogenomics Network for Drug Safety (CPNDS) [31]. The CPNDS database has detailed patient records of 434 children mainly treated with a class of cancer drugs known as anthracyclines, 90 of which show cardiotoxicity, from nearly 100000 adverse drug reaction cases and controls from across Canada. Each record includes Imputed (IMPUTE2) germline Single Nucleotide Polymorphism (SNP) profiles, demographics (e.g. gender, age, etc.), relevant medical history (e.g. dosage, cardio-protectant, Vincristine, etc.), prior and concurrent treatments (e.g. drugs, radiation, etc.) and a detailed description of ADRs and their timeline relative to the treatments.

However CPNDS patient records are stored in a free text format, suitable for case by case manual review by medical researchers, but not for quantification and large-scale mining by computer algorithms. Therefore, before any analysis, we needed to systematically and automatically extract the relevant categorical data from the descriptive texts of patient records and map them to statistically significant variables. To achieve this, as the first step, initial tokenization, standardization, and stemming are applied to clinical report entries texts to extract frequent keywords. Those keywords are then query-corrected and matched to DrugBank [97] and ADReCS [29] to find the classified list of valid drugs and reactions of each patient, respectively. Similarly, dbSNP [176] is used to standardize variants. For this, ancestral and variant alleles are turned into two binary biomarkers for each variant locus (position on the genome), representing the four possible values that a variant can take based on paternal and maternal allele combinations:

- Dominant biomarker where homozygous alternate alleles (aa, encoded in the original dataset as 0) are encoded as 1; and heterozygous (Aa, encoded originally as 1) and homozygous reference alleles (AA, encoded originally as 2) are encoded as 0.

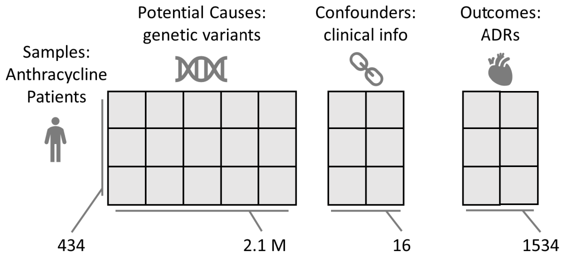


Figure 1.1: **Pharmacogenomic Dataset.** Anthracycline Cardiotoxicity patient samples with clinical and biomarker variables used as the real-world application throughout this thesis.

- Recessive biomarker where homozygous alternate and heterozygous alleles are encoded as 1; and homozygous reference (AA) alleles are encoded as 0.

Furthermore, some of the variants are filtered based on the prior significance and similarity:

- Variants that have low annotation score ($CADD < 1$) are removed.
- Among variants in linkage disequilibrium ($LD > 0.95$), only the one with the highest annotation score is kept and others are removed.

Ultimately, preprocessing of the CPNDS data resulted in categorization of 121 drugs (including two anthracyclines Doxorubicin and Daunorubicin), 32 reactions (including cardiotoxicity), and 2.5 million biomarkers, which we ultimately aggregated into a dataset of 1534 ADRs and 2.1 million statistically significant (based on significance level of 5% which is equivalent to $n_0 = 6$) unique biomarkers for the 434 patients. For uniformity of the dataset type, we turned the continuous variables age and dosages into binary variables based on their quantiles, resulting in a total of 16 confounders (Figure 1.1).

We used the Guideline of Genetic Variants in Anthracycline-induced Cardiotoxicity [7] as the ground truth of the known relations. The Guideline reviewed Anthracycline Cardiotoxicity biomarkers reported in the literature using the Canadian Pharmacogenomics dataset as well as independent reproducibility analyses, and categorized them into strong, significant, and notable level of evidence of association with anthracycline cardiotoxicity in childhood cancer patients, which matches with the biomarkers we expect to observe from the dataset. Although we evaluate the performance of different methods by how well the Guideline set of relations are predicted, this does not mean that the only possibly true relations are Guideline relations, and the other discovered relations that have already passed through strict statistical tests and have a high chance of being undiscovered true causes whose external validity is to be evaluated by other sources.

It should be noted that in addition to Anthracycline cardiotoxicity, we have used Cisplatin ototoxicity as a pharmacogenomic application. However since it was not used for all the problems (for the lack of temporal data and its expected unifactorial causal mechanism), and its results and yet under study, and due to its similarity to Anthracycline cardiotoxicity from HCD prospective, it was omitted from this thesis.

The rest of this thesis is organized as follows. In chapter two, related HCD ideas, methods, and techniques are introduced, and existing definition, assumptions and problems are laid out. In chapters three to six, Relational Causal Discovery, Stratified Causal Discovery, Causal Profile Discovery, and Compound Causal Discovery are discussed in order. In chapter seven, applications and results of the proposed problems and methods are discussed and limitations are future works are reviewed.

Chapter 2

Related Works

2.1 Introduction

In the last three decades, there has been an increasing amount of attention to CD, and HCD in particular, mostly concerned with improving the constraints and validity of the discovered relations [64]. However, the rigorous study of causality from observations dates back to three millennia ago. To understand the mechanism and limitations of CD and especially HCD, it is critical to review the relevant causal inference methods.

2.1.1 Characterizations of Causality

Characterizing the conditions of causality is crucial to CD, but it is especially crucial to HCD because of the implications of the interactions between the large number of variables. Characterization of causality has a deep root in philosophy.

Four Causes

One of the oldest and best-known analysis of causality is Four Causes, introduced by Aristotle, which along with the rest of his natural philosophy became predecessor to the scientific approach two millennia after it [84]. Four Causes broke down the explanation of why an outcome has occurred into (1) the Matter via which the outcome occurs, (2) the Form which when present results in the outcome, (3) the Agent that produced the outcome, and (4) the End purpose behind creating the outcome. Although usually ignored in CD, distinction of four causes is crucial in certain applications [170, 54, 42].

Humean Causality

The philosophical characterization of causality that has arguably the biggest impact on the scientific notion of causality was presented by Hume [79]. Hume proposed the empiricist position that we tend to claim causality based on induction, i.e. one event consistently following another. Hume argued that the condition for establishing these causal relations is the constant association of the cause and then the outcome across time and space, which resembles the criteria used in CD.

Mill's Method

However, the arguably closest philosophical characterization of causality to CD comes from Mill [126]. Mill's methods are a set of five scenarios where causal relations can be identified. Most notable of Mill's methods is the Agreement and Difference, which states that if occurrences and absences of one event only have occurrences and absences of another event in common, one is the necessary cause of another. These considerations not only match the multivariate association analysis of CD, but it also recognizes the ambiguity in the direction of causality of those associations.

2.1.2 Early Applied Approaches

Unlike most subfields of Computer science and Statistics, many of the ideas and methods of CD has grown out of different fields. Although these approaches are not designed to scale with the HCD problems, their contrast shows gaps that needs to be fields in HCD.

Criteria of Causality

Some of the earliest forms of practical causality conditions were first used in Health sciences. Built upon Koch's Postulates [92], Hill attempted to list the sufficient criteria for causality in epidemiology [76]. Hill's criteria include (1) strength of association, (2) reproducibility, (3) temporal precedence of cause, (4) effect scaling with cause, (5) specificity of confounders, and (6) mechanism plausibility. There have been arguments against Hill's criteria including lack of confounders' control, counterfactuals, and reliance on common sense rather than specific criteria [155]. Yet, because of its interpretability and flexibility, Hill's criteria has remained widely used to this day in various public health applications [57].

Temporality in Causality

Panel data is where CD has arguably performed the best. This is because:

- The order in which events occur gives away direction of the causal relation.
- Instances where outcome precedes the potential cause can reveal random and spurious association and hence provide higher degree of confidence in the inferred relations.
- Mapping timeseries to one another can lead to quantified causal relations with time (i.e. beyond a simple cumulative effect) [184].

Granger pioneered the CD of panel data by defining a special type of causality for time-series, Granger Causality, which defines causes of an outcome as those variables that provide statistically significant unique information about future the outcome [66]. The original CD method consisted of F-Test on an Autoregression Model of the outcome and a potential cause to determine whether any of the lagged version of the potential cause is significant in forecasting the outcome.

Granger Causality can be extended to multivariate analysis (vector autoregression) [18], hidden confounders [52], Non-parametric test [69], non-stationary [197], nonlinear relation (transfer entropy) [17], feedback loop (convergent cross mapping) [191] but they may produce misleading results [119]. Although Granger Causality is a weak notion when it comes to intervention, counterfactuals, and discovering underlying mechanisms, it matches HUME's notion of constant conjunction of events, which is reasonably sufficient for applications with limited possibility of interference. This, in addition to its computational simplicity has led to Granger Causality's extensively use in Economics and Neuroscience, resulting in Nobel Prize in Economic Sciences for Granger himself [26].

2.1.3 Hypothesis Testing

Today, statistical hypothesis testing is not only used in many CD approaches, but also the dominant statistical tool in the experimental academic publications; where it is used as a "soft" causal inference to validate hypotheses derived through systemic analysis, or even when no scientific theory exists [100].

Premise

Pearson laid the foundations of hypothesis testing theory [147], where p-value is used as the probability that a given result would occur under a hypothesis. Fisher systematized the significance test, where p-value is used to evaluate whether the null-hypothesis can be rejected [59]. The modern frequentist hypothesis testing has remained mostly unchanged since the major developments in the early 20th century, with the notable exception of the null hypothesis which is now used as a strawman position.

Shortcomings

The is a long list of concerns associated with hypothesis testing, and although mostly are related to its misuse rather than inherent flaw [202], some are applicable to CD. These include:

- There are critical considerations related to the research design such as design assumptions, quality of measurements, original sample size and representativeness, which are usually unknown, misunderstood or ignored by the CD user [93].

- Significance tests can only assess how incompatible the null hypotheses and the data are, hence successfully rejecting the null hypothesis may offer little support for the particular research hypothesis [9].

- Significance test's decision depends on the significance threshold and definition of multiple comparison, which is unavoidably ambiguous [68].
- Interpretation of the results is susceptible to context and various effects such as selection bias, Simpson's paradox, Berkson's paradox, Base Rate fallacy, and Will Rogers phenomenon, which are especially problematic when hypothesis tests are done automatically under the hood of the CD algorithm [85, 143].

2.1.4 Experimental Design

Well-designed experiments are arguably the most reliable source for establishing causal relations. However, unlike CD which attempts to find causality by carefully evaluating potential relations in observational data, Design of Experiments pays the price upfront and facilitate simple hypothesis evaluations by planning how samples are to be assigned to different conditions.

Randomized Controlled Trials

In 1880s Peirce formulated Blinded Randomized Experiments [148], Repeated Measures Design [150] and Optimal Design [149]. In the early 20th century, Fisher popularized Randomized experiments [58], and by the late 20th century Randomized Controlled Trials were recognized as the gold standard for scientific causal inference [124].

Components that contribute to Validity, Reliability, and Replicability of Randomized Experiments include [49, 172]:

- Randomization, which is randomly assigning samples to a treatment group that receives the potential cause, and a control group that receives alternative conditions such as a placebo or no intervention; minimizes selection bias and confounding.
- Blinding, which is concealing influencing information from subjects, experimenters and analysts; minimizes experimental biases that arise from these parties.
- Replication, which is Repeated Measures of the same samples under different conditions; minimizes variations and measurement uncertainty.
- Control, which is carefully selecting participants based on the hypothesis and keeping conditions highly controlled throughout the experiment, minimizes the effect of confounders.

Obstacles and Objections

Randomized Controlled Trials are widely used in many clinical and social settings [124]. However, there are fundamental challenges that limit their applicability and scope, including cost and time to conduct, ethical concerns, and infeasibility for certain hypotheses [83].

Furthermore, in recent year many have raised objections against Randomized Controlled Trials [169]. One of the objections, which is most relevant to CD, is that extensive Control and atypical treatment groups risk External Validity relevance of results to real-world practice [172]. Some of these concerns are addressable via Quasi-Experimental Designs, which will be discussed in Chapter 4.

2.1.5 Discussion

With the increase in availability of large datasets and abundance of computational power, some of the methods for causal inference and CD were modified to HCD problems. Unlike the methods which will be discussed in the following chapters, the previously mentioned approaches lack the option to be extended for HCD; because of certain requirements of HCD:

- Efficient and automated evaluation of potential relations. Hence, methods such as Hill's Criteria that rely on human judgment or extensive analysis for each possible relation, cannot be used for HCD.

- Scalability with a large number of potential causes. The number of potential relations and confounders to those relations grow exponentially with the number of potential causes, which in HCD can cause problems in multiple comparison and computational cost. An example is Mill's Method which requires finding samples with specific matches across all confounders.

- Availability of data and satisfiability of conditions. Hence, approaches based on Granger Causality that require temporal data of variables and Experiment designs which require recording and Factorial Experiment of all potential causes, cannot be expanded to HCD.

2.2 Structure Learning

2.2.1 Overview

Structure Learning (SL), also known as Graphical Causal Models, has been the center of CD developments [184]. The goal of SL is to infer the "flow of causality" between variables of interest, modeled by a causal graph. In the causal graphs, nodes can include the outcomes, potential causes, confounders, unmeasured factors, selection biases, and noises; and each edge represents the hypothesis that there exist interventions on the variable on its head that directly change the distribution of the variable on its tail.

The argument for the relations of the causal graph to be considered causal, and not just mere conditional independences, is that if all other variables are kept constant and the variable on the tail of the edge changes, so will the variable on the head of the edge, but not vice versa [64]. Despite existing objections to the validity of SL discoveries, which will be discussed later on, it remains one of the most reasonable and useful notions of causality in data science.

It should be noted that in the graphical model predecessors of the SL [96, 142], where variables are only defined unrelated iff they are conditionally independent given all the remaining variables, the notion of causality above could not be assumed.

Search Strategies

SL methods usually narrow down the space of possible causal graphs by evaluating the conditional independences between the variables using statistical tests. Traditionally, SL is categorized based on the search strategy into two classes of:

- constraint-based methods, which locally evaluate each possible edge based on the conditional independencies of nodes in its proximity.
- Score-based methods, which globally search for graphs that represent the conditional independences of the data the best. These methods will be discussed in the following sections; and mathematical definitions, pseudocode, and running example are presented at the end of the chapter.

Assumptions

Another aspect that helps to classify SL methods, is the key assumptions usually made for inferring joint probability distribution of a set of variables. Each SL method pays a tradeoff between (1) dropping some of these assumptions about the data, (2) complexity of the algorithm, and (3) informativeness of the output.

Causal Faithfulness assumes that the conditional independences between the variables inferred from the observational data are identical to those implied by the true causal graph, Causal Markov assumption assumes that every variable is independent of its non-descendants conditioned on its parents, and Causal Sufficiency assumes that a direct cause of every two variable in the graph is also in the graph. These assumptions are critical in guaranteeing the validity of the causal structure, locality, and absence of latent confounders and selection variables, respectively [183].

2.2.2 Constraint-based Methods

In its general form, the constraint-based approach can be defined as systematic selecting and orienting the potential edges of causal graphs, based on the tested conditional independencies of that edge. A typical constraint-based method specifically entails:

- Pruning Phase: starting from a complete undirected graph, iteratively checking conditional independency of every valid edge that has not yet proven independent, conditioned on larger and larger subsets of neighboring nodes

- Trimming Phase: deciding the orientation of edges from a set of rules, based on the condition under which the dependency was established.

PC Algorithm

Although the development of Constraint-based Methods can be traced back further down in the past [142], the major methodological breakthrough did not occur until the PC algorithm, which has remained arguably the best-known SL method to this day, and many of the other key methods are developed as an extension to it. PC was first introduced in [182], and its latest version along with other SL algorithms and related concepts were analyzed in [183].

What distinguishes PC from the works before it is that it can efficiently guarantee causal independencies without testing for larger condition sets. Specifically, in its pruning phase, PC uses the fact that in the absence of latent confounders, two variables are causally related if and only if there is no subset of the remaining variables conditioning on which they are independent. The key rule in PC's trimming phase is orienting a chain of triplets into a v-structure when the separation condition is eliminated, which is rooted in the fact that unlike other configurations of a condition in a path between two nodes, in a v-structure, conditioning on the collider makes the other two variables dependent, also known as Berkson's paradox.

Consequently, in addition to search efficiency, under the Causal Faithfulness, Causal Sufficiency, and Causal Markov assumptions, and IDD samples, PC's output is asymptotically correct and maximally informative. Furthermore, because of its simplicity, PC is adaptable to different types of data, distributions, and relations, given reliable independence test methods.

Most of the constraint-based methods consist of adjacency search of the PC-algorithm, and hence inherit some of its problems. For instance, PC is order-dependent in the sense that the generated causal graph can vary based on the order in which conditional independences are checked, especially for high-dimensional data. This problem can be solved by updating deleted edges only when the size of conditioning sets is increased, which also allows the parallelization of independency checks [46].

FCI Algorithm

Unlike the Faithfulness and Markov assumptions which are concerned with the linking the data to the conditional independences and conditional independences to causal relations, and are therefore inherent to the constraint-based SL, Causal Sufficiency is concerned with latent variables, which is a common challenge to be dealt with in causal inference, and data science in general.

Fast Causal Inference (FCI) algorithm is arguably the most important generalization of the PC algorithm, and is designed to alleviate the Causal Sufficiency assumption [183].

In doing so, FCI is able to not only take the effect of unknown confounding and selection variables into account, but also discover the existence of some of them.

FCI achieves this by performing additional conditional independency tests. Specifically, FCI uses the fact that if two variables are independent conditioned on a set of variables, they are also independent conditioned on their separation set. Consequently, in order to evaluate whether two variables are related or not, only their conditional independence given their possible separation sets are needed to be tested.

Hence FCI first runs a pseudo-PC algorithm to estimate an initial skeleton of causal graph and orient available v-structures, in order to narrow down the possible separation sets, and then uses the separation sets and a conservative set of rules in order to find the final skeleton of causal graph and orient as many relations as possible. The major drawback of FCI lies in this second phase, where the space of possible separation sets can become computationally infeasible for large graphs.

FCI can also be modified to output the causal relations in a maximally informative partial ancestral graph, by adding few rules in the second trimming phase, which use the dependency on the descendants of the colliders in v-structures [205].

Really Fast FCI

One of the major drawbacks of FCI and to some degrees PC is their complexity, which can become concerning for high-dimensional problems, and even smaller problems that are not sparse. Various methods have been proposed to speed up the conditional independency tests by check for fewer and smaller conditioning sets.

PC* is the most immediate variation of PC [183], with the goal of avoiding unnecessary computations of PC by ignoring conditioning variables that are likely to be irrelevant to the relations and focusing on eliminating weak edges before they can preoccupy conditional independency checks. Specifically, when evaluating conditional independency between two variables, PC* first conditions on subsets that are adjacent to and on undirected paths between the two variables. By prioritizing the removal of these edges, the search space at later stages of the algorithm will be significantly smaller. However, for large numbers of variables, searching for the paths between nodes becomes ineffective and PC* underperforms even compared to PC.

Anytime-FCI is a modified version of FCI that tries to avoid the exponentially growing possible separation sets by only considering independency tests with conditioning sets of size less than a prespecified cut-off [181]. Despite its simple and restrictive nature, Anytime-FCI can provide relatively similar results to that of FCI and some preliminary guarantee of correctness.

However, the most significant improvement in speeding up FCI is RFCI [47]. By compromising dependency assumptions and informativeness of partial ancestral graph, RFCI uses fewer independence tests and smaller conditioning sets than FCI. RFCI achieves this by substituting FCI's conditional independence tests given possible separation sets, with additional tests on subsets of adjacency sets before orienting v-structures, which are fewer and have smaller conditioning sets than its FCI counterpart. As a result, RFCI can produce causal graphs that are halfway between FCI and PC, but is significantly faster than FCI, and by avoiding larger conditioning sets is even more reliable than FCI for small samples. However, the improvement is not as significant for non-sparse graphs.

2.2.3 Score-based Methods

In its general form, the score-based approach is finding the causal graph by optimizing a properly defined score function. A typical score-based method directly searches over the causal graph equivalence classes for the graphs that fit the joint probability distribution in the data, measured by a score of how faithfully the graph can generate the factors within the distribution. However, since the space of unique graphs for n variables is $O(2n^2)$, the usual score-based methods iteratively add and remove edges or cliques of edges that improve the score the most, which can be searched in $O(n^2)$.

Greedy Search

Greedy Equivalence Search (GES) [40] is arguably the best-known score-based methods, and is still used as its baseline [138]. Like many score-based methods, GES is comprised of two phases of:

- Iteratively adding edges with the best score to an empty graph until the score cannot be improved further.

- Then removing edges that do not increase the score of the constructed graph significantly

However, because inferring whether an edge improves score for the optimal graph is challenging, GES as its name hints, greedily includes and excludes edges based on whether they improve on the score of the existing graph or not.

Score-based methods are generally more robust than constraint-based methods with respect to local errors and violation of assumptions, but scale worse to high-dimensional data. Because of its compromise to heuristic search, GES can achieve speed comparable to the PC algorithm, but the trajectory toward the optimal graph becomes dependent of the relative strengths of the associations between the variables [113].

Another major drawback of Score-based methods is that there is yet no method that can deal with hidden confounding and selection variables in general. Although, under similar assumptions GES asymptotically converges to the same Markov Equivalence Class of PC [64]. GFCI [139] is proposed to alleviate this problem, which combines a robust search of GES to find the skeleton of the partial ancestral graph, and causal sufficiency alleviation of FCI to prune the partial ancestral graph and identify the orientations.

2.2.4 High-dimensional Search

The Fastest general constraint-based and Score-based SL methods are super-exponential and NP-hard respectively, which makes their applicability to HCD very limited. However, the problem of SL methods is not limited to complexity, and various challenges to accuracy also arise when dealing with high dimensional data. There are SL methods that find efficient approximations of large causal graphs by employing specialized heuristics [204], or solve HCD under certain assumptions that specific algorithms can take advantage of.

Sparse Graphs

One of the scenarios where the search space of HCD can be managed is in the sparse causal graphs, which is especially common for high dimensional problems. FCI+ [45], a modification of FCI, is a prime example of SL methods designed to efficiently search for sparse graphs. For graphs bounded by degree k, with respect to the number of independency tests N, FCI+ is polynomial $O(N^2k)$. FCI+ is based on the idea that FCI's time-consuming checks can be broken down by node, where the conditional independencies already considered can guide the search for new node's separation sets. To achieve this, FCI+ substitutes FCI's skeleton construction step with testing for single node additions that change independency between processed nodes.

The tradeoff that FCI+ pays is that it has to reconsider previously rejected edges. As a result, although FCI+ is significantly faster than exponential, it is still infeasible for problems that have variables with potentially large number of edges. However, the more damaging limitation of FCI+ in practice comes from the fact that the conditioning sets for independency tests may become larger than necessary, leading to a loss of statistical power and inaccuracy in the evaluation of relations.

Markov Blanket

Another direction of research is to limit the search space of HCD to local regions in the causal graphs. Markov Blanket which consists of parents, children, and parents of children of the outcome variable, is the minimal set of variables that given them the outcome is independent of all other variables, and therefore ideal for SL localization.

Max-Min Markov Blanket (MMMB) [198] was one of the first methods explicitly designed to find the Markov blanket. What distinguishes MMMB from few earlier works such as [116], is that the number of independency tests it requires depends on the structure of the local causal graph and not the number of variables in the Markov blanket.

MMMB, like most Markov blanket SL methods that followed it [6, 157], uses a two-phase process:

- inclusion phase, where variables are sequentially considered for the Markov blanket, based on a fast liberal heuristic criterion that would accept all variables from true Markov blanket and possibly some non-Markov blanket variables.
- exclusion phase, where because the number of filtered down variables is much smaller and the space of possible Markov blanket graphs is much more limited, it becomes feasible to carefully remove all false positives that entered in the first phase, and then assign the remaining variables to children, parent, and parent of children categories.

In MMPC specifically, the inclusion phase pre-orders sequence of variables to be checked based on their raw association with the outcome, and its heuristic criteria accepts variables with the highest association with outcome, conditioned on the subset of discovered Markov blanket that results in the minimum association. The intuitive reason for this criterion is to choose variables that despite the model's best efforts to make them independent of the outcome, provide an unexplained association with it. In the exclusion phase, MMPC removes variables from the Markov blanket that are independent of the outcome, conditioned on any of its possible subsets of the Markov blanket.

More recent iterations of Markov blanket SL provide additional features, e.g., symmetry correction of exclusions [199], robustness to violation of faithfulness [153], and empirical and theoretical analysis of interpretability of Markov blanket discovery [6].

Two considerations set the Markov blanket approach apart from the rest of the SL methods for HCD:

- Unlike other SL approaches, its computational complexity and number of samples required to maintain statistical significance does not grow exponentially with the number of variables, and can achieve $O(NM^2M)$ for the number of variables N and number of elements in Markov blanket M.
- It can be used as a state of the art feature selection approach [6], and has the potential to merge accuracy of machine learning's classification models with reliability and interpretability of CD.

Trace Method

Trace method is one of the computationally efficient CD methods that is closely tied to SL. What differentiates Trace methods from the typical SL approach is utilizing the asymmetry in factorizing the joint probability distribution, based on the idea that effects conditioned on causes result in simpler factors [185]. This leads to an independency condition involving traces of the transfer matrix mapping causes to outcomes and the covariance matrix of causes. What distinguishes the Trace method is that unlike the other SL approaches, it does not struggle in identifying the direction of causality because the complementary independency condition is strongly violated for anti-causal direction. There have been attempts to extend the Trace method to tackle nonlinearity using Kernel Hilbert Space [38]. However, the most significant drawback of the Trace method is its heavy reliance on the causal sufficiency assumption.

Parallel Computing

One of the more obvious directions for speeding up the SL methods is parallelization. In particular, variations of the PC algorithm, whose independency tests at each level are independent of each other, can very effectively distribute the tests evenly to multiple cores, and update the causal graph only at the end of each level. A good example of SL parallelization is done in [98], where an order-independent stable PC could be run multiple times faster with no downside, by grouping independency tests of the same edges together. A more extreme example can be found in [178] where a simple v-structures discovery algorithm, which runs cubic with the number of variables, was parallelized with minimal overhead. It should be noted that although the parallelization can reduce the run-time of CD algorithms directly, it cannot compensate for the escalated multiple comparison problem of high dimensional data.

FGES

One of the most noteworthy development in adopting SL methods to high-dimensional data is Fast Greedy Equivalence Search (FGES) [157]. FGES is a modification of GES that can also be adjusted for Markov blanket discovery. Although FGES uses some of the discussed techniques, what differentiates FGES is that these techniques are carefully incorporated with how the GES works and how they interact with each other to great effect.

To achieve this, the following techniques are used:

- Superposition: GES score is defined such that the score of each potential causal graph is defined as the sum of scores of its potential edges, and the score of adding a potential edge at an iteration can be updated using its score from previous iterations. As a result, the score of potential edges at each iteration can be estimated efficiently, and the score of part of the graph which is unaffected by the new edge can be updated in no time.

- Parallelization: because of the superposition property, FGES can be parallelized locally in all iterations, with small overhead.

- Partial solution: by selecting a more conservative scoring criterion in the forward search phase, the generated potential causal graphs become sparser, which reduces problem size drastically, at the risk of relations false negatives. (4) weak sufficiency: ignoring the edge between variables that are uncorrelated in the forward search to gain significant speed up, at the expense of potentially violating Markov factorization, which is a tradeoff for sufficiency assumption.

Using these techniques, FGES can solve HCD problems with millions of nodes and edges in a reasonable time, and orders of magnitude faster than similar SL methods. Furthermore, this shows the potential in polishing the existing SL methods such as FCI and GES using simple machine learning and statistical techniques. However, FGES's strong assumptions and suboptimal solutions leave plenty of room for improvement.

2.2.5 Discussion

The maximally informative SL problem has the super-exponential lower bound complexity (59). Hence, to apply SL to high-dimensional data, various methods shifted their focus to limited solutions, such as the local solution of Markov Blanket or partial solution of FGES. However, other challenges that are risen as the result of high dimensionality or in general remain understudied in the SL literature. Some of these challenges related to HCD include:

Independence Test

A typical SL model relies on performing a large number of conditional independency tests for each possible edge, mostly in the form of standard statistical hypothesis tests or BIC score [183]. As a result, some of the challenges of SL are rooted in the repetitive usage of the tests:

- Sensitivity of the output to the threshold or significance level used for the independence tests [72].
- Validity of the statistical test when the form of dependency is unknown or type and distribution of variable varies.
- Error in identifying one edge propagating and affecting the validity of other edges.

Strength of Relations

The output of an SL problem is in the form of a causal graph, and quantifying the strength of each relation and the causal graph as a whole has historically been seen as a secondary problem [146]. Because the output of many of the SL methods includes the Markov equivalent class, and not just a unique directed acyclic graph with directed edges between every cause and effect, specifying the significance of particular edges is even more challenging. As a result, evaluation of statistical significance or average causal effect of particular edges has usually been done by a complementary post-processing analysis, such as intervention calculus [110].

A classical point of disagreement in this regard is whether the measure of significance represents the effect of Intervention or the probability distribution corresponding to the causal hypotheses, in frequentist or Bayesian interpretation [143, 64]. However, the more critical point for high dimensional data is incorporating the effect of data dredging and multiple hypothesis adjustment in the evaluation of the output.

Assumptions

The arguably the most important challenge in applying the SL methods to real-world high dimensional data is due to the strong assumptions usually made in the SL framework, including causal faithfulness and Markov assumptions, and sometimes causal sufficiency, which are usually violated in most realistic scenarios [46].

On top of this, there exist an implicit universal assumption of SL methods, which shows itself in the asymptotical correctness of the output. This assumption is a requirement of a very large sample size which is because a large number of conditional independency tests with multiple arguments demand very high statistical power to provide any meaningful output.

2.2.6 Definitions

Causal Graph

In SL, causal relations between variables is represented by a Directed Acyclic Graph $G = (X, E)$, where $X = \{X_1, \cdots, X_n\}$ is the set of observed variables, and $E = \{X_i \rightarrow X_j \mid i, j \in \{1, \cdots, n\}\}$ is the set of edges with $X_i \rightarrow X_j$ representing that X_i is a direct cause of X_j. To infer the true causal graph, SL methods rely on conditional independency tests represented by $I(X_i, X_j \mid \{X_S, S \subset \{1, \cdots, n\}\})$ that evaluate whether X_i and X_j are independent of each other, when the effect of some of the other variables S is controlled.

Unfortunately, sometimes more than one Directed Acyclic Graph can represent the same set of conditional independencies between the variables. Furthermore, identified hidden confounders cannot be represented on the Directed Acyclic Graph. The set of causal Directed Acyclic Graphs that are equivalent to the same set of conditional independencies is called Markov Equivalent Class. Partial Ancestral Graph is a generalization of the Directed Acyclic Graphs that can represent all of the graphs in a Markov Equivalent Class. Unlike a Directed Acyclic Graph whose edges must have an arrowhead on one side (representing the effect) and tail on the other (representing the cause), each side of a Partial Ancestral Graph can take an arrowhead (not cause), tail (cause), and circle (unknown).

d-separation

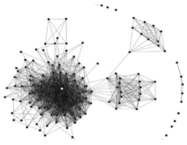
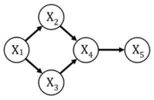
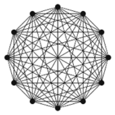
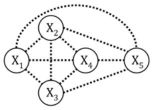
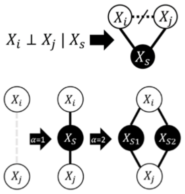
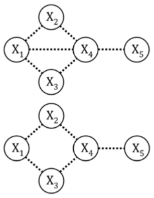
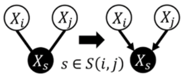
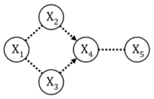
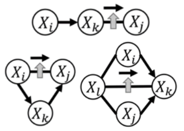
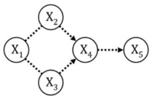
Conditional independencies manifest themselves in the causal graphs as d-separation. A path between a source node and a sink node $\langle X_i, \cdots, X_j \rangle$ consists of distinct nodes linked via directed edges $X_{k-1} \rightarrow X_k \; \forall \; i < k < j$. A node X_k is a collider on a path if it is the common dependent of its neighbors $X_{k-1} \rightarrow X_k \leftarrow X_{k+1}$; and a collider is a v-structure if its adjacent nodes on the path are not directly connected $X_{k-1} \nrightarrow X_{k+1}$. Node X_i is d-separated from node X_j conditional on set of nodes X_S iff all paths between X_i and X_j are blocked by X_S. The path between X_i and X_j is blocked by X_S if at least one of path's non-collider nodes is in X_S or there exists a collider that neither itself nor its descendants are in X_S.

One special case of separation set is the Markov blanket which is the set of nodes $X_{M(i)}$ that block all paths from all other nodes to node a node X_i, effectively specifying the set of variables that conditioned on them the particular target variable is independent of all other variables. Although Markov blanket is not unique, the set of parents, offspring and parents of offspring of a node in the true causal graph comprise minimal Markov blanket set of that node $X_{M^*(i)} = \{Pa(i) \cup Of(i) \cup Pa(Of(i))\}$.

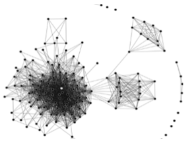
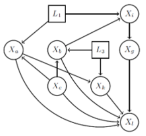
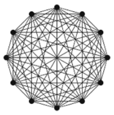
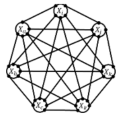
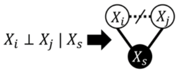
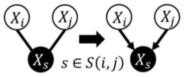
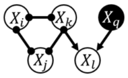
2.2.7 Pseudocode and Example

In the following, the pseudocode and a running example of important SL algorithms are described.

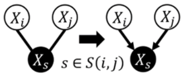
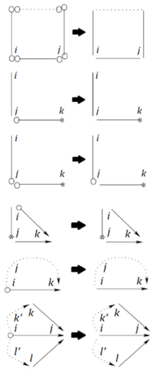
PC Algorithm

Pseudocode	Representation	Example	DAG
Input-Output Given conditional independencies between variables, find the true causal Directed Acyclic Graph under faithfulness, sufficiency and Markov assumptions.		$X_1 \perp X_5 \mid X_4$ $X_1 \perp X_4 \mid \{X_2, X_3\}$ $X_2 \perp X_3 \mid X_1$ $X_2 \perp X_5 \mid X_4$ $X_3 \perp X_5 \mid X_4$	(True Unknown) 
Initialization start with a complete undirected graph of all the observed variables $G = (X, \mathbf{1})$		Complete graph of nodes X_1 to X_5	
Pruning initialize separation set size $\alpha = -1$ while $\alpha <$ max degree of G $\alpha = \alpha + 1$ for each edge $X_i - X_j \in E$ for each subset X_s of neighbors of X_i & X_j of size α if they are conditionally independent: $X_i \perp X_j \mid X_s$ remove edge $X_i - X_j$, record its Sep set $S(i,j) = s$	 $X_i \perp X_j \mid X_s$ X_i X_j X_s α=1 α=2 X_{S1} X_{S2}	$\alpha = 0$: No edges can be d-separated | $\emptyset$ $\alpha = 1$: $X_2 \perp X_3 \mid X_1$, $X_1 \perp X_5 \mid X_4$, $X_2 \perp X_5 \mid X_4$, $X_3 \perp X_5 \mid X_4$ $\alpha = 2$: $X_1 \perp X_4 \mid \{X_2, X_3\}$ * X_4 is the gatekeeper to X_5 and hence is separation set of many edges to X_5	
V-Structure Orientation If $X_i - X_k - X_j$ & $k \notin S(i,j)$ then $X_i \to X_k \leftarrow X_j$		X_4 is not in separating set of $X_2 \nleftrightarrow X_3$ ($S(2{,}3) = \{1\}$), hence it is turned into v-structure	
Other Orientation - If $X_i \to X_k - X_j$ & $X_i \nleftrightarrow X_j$ then $X_k \to X_j$ - If $X_i \to X_k \to X_j$ & $X_i - X_j$ then $X_i \to X_j$ - If $X_i \to X_k \leftarrow X_j$ & $X_i - X_l - X_j$ & $X_i \nleftrightarrow X_j$ & $X_k - X_l$ then $X_k \to X_l$		X_4 is a non-collider in only path between X_2 & X_5, hence $X_4 \to X_5$ is oriented along with $X_2 \to X_4$ * only the unspecified orientation of edges from X_1 does not match the ground truth	

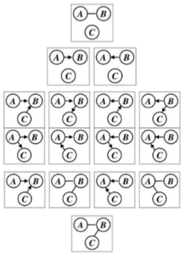
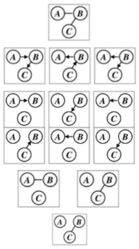
FCI Algorithm

Pseudocode	Representation	Example
Input-Output Given conditional independencies between variables, find the true causal Partial Ancestral Graph under faithfulness, and Markov assumptions.		
Initialization start with a complete partial ancestral graph of all the observed variables $G = (X, \mathbf{1})$ and all edges with Unknown orientation X_i ∘–∘ X_j		
Pruning Repeat Pruning step of PC Algorithm.		$X_i \perp X_l \mid \mathrm{adj}(X_l)$ ...
V-Structure Orientation If X_i ∗–∘ X_j ∘–∗ X_k & $k \notin S(i,j)$ then X_i ∗→ X_k ←∗ X_j	 X_i X_j X_s $s \in S(i,j)$ X_i X_j X_s	X_i ∗–∘ X_k ∘–∗ X_l ...
Sep-set Narrowdown find Possible Sep-set for each node $Q(i)$ as: $q \in Q(i)$ iff ∃ path X_i -- X_q of V-structures & triangles		$Q(i) = X \backslash \{i,k,l\}$ ...

Pruning II
initialize separation set size $\alpha = -1$
while $\alpha <$ max degree of G
  $\alpha = \alpha + 1$
  for each edge $X_i - X_j \in E$
    for each subset s of Sep-sets $Q(i) \cup Q(j)$ of size α
      if they are conditionally independent: $X_i \perp X_j \mid X_s$
      remove edge $X_i - X_j$, update Sep set $S(i,j) = s$
		$X_i \perp X_k \mid \mathrm{adj}(X_l)$
		...

V-Structure Orientation repeat the V-Structure Orientation step.	 X_i X_j X_s $s \in S(i,j)$ X_i X_j X_s	$X_a \to X_l \leftarrow X_k$ ...
Orientation Rules Execute following orientation rules until none applies: 1- If X_i ∗→ X_j ∘–∗ X_k & $X_i \nleftrightarrow X_k$ then X_i ∗→ $X_j \to X_k$ 2- If X_i ∗→ X_j ∗→ X_k & X_i ∗–∘ X_k then X_i ∗→ X_k 3- If X_i ∗→ X_j ←∗ X_k & X_i ∗–∘ X_l ∘–∗ X_k & X_l ∗–∘ X_j then X_l ∗→ X_j 4- If $\langle X_i, \dots, X_j, X_k \rangle$ & X_j ∘–∗ X_k & $X_j \notin S(i,k)$ then $X_j \to X_k$ 5- If X_i ∘–∘ X_k & $\langle X_i, X_l \dots, X_j, X_k \rangle$ & $X_i \nleftrightarrow X_j$ & $X_l \nleftrightarrow X_k$ then $X_i - X_l - \cdots - X_j - X_k$ 6- If $X_i - X_j$ ∘–∗ X_k then X_j –∗ X_k 7- If X_i –∘ X_j ∘– X_k & $X_i \nleftrightarrow X_k$ then X_j –∗ X_k 8- If X_i –∗ $X_j \to X_k$ & X_i ∘→ X_k then $X_i \to X_k$ 9- If $\langle X_i, X_j, \dots X_k \rangle$ & X_i ∘→ X_k & $X_j \nleftrightarrow X_k$ then $X_i \to X_k$ 10- If X_i ∘→ X_j & $X_k \to X_j \leftarrow X_l$ & $\langle X_i, X_{k'} \dots X_k \rangle$ & $\langle X_i, X_{l'} \dots X_l \rangle$ & $X_{k'} \nleftrightarrow X_{l'}$ & then $X_i \to X_j$		$X_a \to X_k \to X_l$ X_i ∘–∘ X_a ...

GES Algorithm

Pseudocode	**Example**
Input-Output Given conditional independencies between variables, find the true causal Directed Acyclic Graph under faithfulness, sufficiency and Markov assumptions	(True Unknown)
Initialization start with an empty graph of all the observed variables $G = (X, \mathbf{0})$.	
Forward Equivalence Search while current graph is being updated $G^{*(t)} \neq G^{*(t-1)}$ for each graph G_i with similar Equivalence Class to G find graphs $\{G_i^+\}$ resulted from edge addition to G_i find unique equivalence classes of found graphs' union go to the graph with the best Score	(Sample Iteration) 
Backward Equivalence Search while current graph is being updated $G^{*(t)} \neq G^{*(t-1)}$ for each graph G_i with similar Equivalence Class to G find graphs $\{G_i^+\}$ resulted from edge deletion from G_i find unique equivalence classes of found graphs' union go to the graph with the best Score	(Sample Iteration) 
Scoring Criterion any function that maps a graph to a scalar with properties: Global Consistency, Local Consistency & Score Equivalence	(Bayesian Criterion: marginal likelihood, regularizing prior) $F_{Bayes} = \log p(D\|G)$ $+ \log p(G)$

2.3 Quasi-Experimental Design

2.3.1 Overview

Quasi-experimental designs (QED) are a family of methods that were developed in order to imitate true experiments when they are not feasible. In the true experiments, the treatment variables are systematically manipulated. Randomized Controlled Trials which had been considered the gold standard experiment for causal inference, randomly select a treatment group, and manipulate it by the treatment, as well as an otherwise similar control group to which no treatment or placebo is applied, and compare their difference. While in general in QEDs either control group or random assignment is missing, most QEDs study the pre-existing groups that have received different treatments [49].

Historically, QEDs were used for causal inference in applications with few potential causes and particular outcomes. However, in recent years machine learning-based search algorithms have made QED applicable to CD for problems with many possible hypotheses and confounding variables. What makes QEDs great baselines for CD methods is that they have both the advantage of (a) CD approaches, such as SL, in working offline with observational data and evaluating many hypotheses on the same population with the only cost being computational and multiple hypothesis adjustments; and (b) having a similar experimental design to those of true experiments, and hence inferring causality in the common scientific sense and not relying on an alternate definition of causality and measures of its strength.

The price that QEDs pay is their high sensitivity to threats to internal validity, such as confounding effects. Hence, the key in using QED in CD is recognizing threats to internal validity in a particular design, and the ways to safeguard it. This is usually achieved by applying as much control as possible and randomly filtering samples or variables, at the cost of statistical power.

The overall process of a QED usually involves (1) deciding QED type based on the hypothesis and threats to internal validity, (2) performing the Assignment, Measurement, and Comparisons, on samples via the experiment design, (3) evaluating the significance of the hypothesis. In the following, the considerations for Assignment, Measurement, and Comparisons are introduced. Next, the threats to internal validity will be discussed. Then the different types of QEDs and the range of hypotheses that they can formulate are presented. This is followed by notes regarding the evaluation of QEDs. Lastly, the state-of-the-art machine learning approaches that use QED for HCD will be discussed.

2.3.2 Building Blocks

QEDs can be categorized based on the operations used in each of the Assignment, Measurement, and Comparison components.

Assignment

Assignment is the critical component of QEDs, where samples are assigned to different groups. Here, to mimic the random manipulation of the treatment variable in randomized controlled trials, QEDs assign samples with different treatment to groups. The general techniques for improving the validity of Assignment are using random selection or instrumental variable for choosing a subset of valid samples in each group, and making sure samples in different groups have similar distribution across all confounders [49].

Based on these techniques, the most common Assignments include:

- Random Assignment, where a fraction of samples from each treatment group are randomly selected to represent their group. This balances the distribution of variables that are not heavily associated with the treatment variables toward their marginal distribution, and hence improves the internal validity, especially robustness to latent confounders.

- Non-random Assignment, where all the samples that match the treatment and control groups are used in the experiment. This maintains the statistical power at the cost of internal validity, and is usually only reserved for scenarios where the population is very small, or the distribution of the treatment is very skewed [30].
- Matching/Stratifying Assignment, where each group selects samples that has similar values of confounders to that of selected samples in other groups. This is achieved by scoring each sample based on its similarity, measured via a function such as Manhattan distance, to the most similar samples in the other groups, and then selecting the top-scoring samples [189].

Because complexity of Matching scales with the number of samples, Matching is usually accompanied by a heuristic search for similar pairs and stable marriage algorithms. Matching is very effective in not only controlling the effect of confounders, but also improving the statistical power by pairing samples, but makes the results more susceptible to latent confounders and more computationally expensive [156].

The other situational but effective Assignments include (4) Cut-off based Assignment, where samples in different groups are selected based on closeness to specific value of the treatment variable, and (5) Masking, where the samples used in the experiment are selected without considering the treatment variable to control the effect of confounders [190]. It should be noted that sometimes a combination of these Assignments can be used to great effect [115].

Measurement

Measurement is collecting the outcome for the samples of identified groups. The key distinction between the different types of Measurements is based on whether multiple time-points for each sample is recorded, whether the outcome before the treatment is recorded, and whether dependent variables or measures of threats to internal validity are recorded throughout the experiment or not. The difference between these Measurement methods only exists if temporal data of the treatment and outcome variables are available, such as time series, timestamps of major changes in the variables, or values of the outcome before and after changes in the treatments.

The common Measurements include [49, 30]:

- Post-Test Only, which only records the value of outcome after treatment. Post-Test Only is the simplest and weakest Measurement and should be only be considered when there is no interference between the variables or the system under the study is not stationary. The only advantage of Post-Test Only when no temporal data is available, is that it avoids mistaking cases where the change in the outcome precedes the treatment.

- Pre-Test Post-Test, where the outcome before and after the treatment is recorded. Adding Pre-Tests improves the validity of the results especially if the time between the Pre-test and Post-test is short enough that the outcome would remain unchanged in absence of intervention and it is unlikely for the confounders to change consistently in between, and long enough for treatment to have its complete effect on the outcome.

- Time Series, where the value of the outcome at multiple points before and/or after the outcome is recorded, preferably with fixed intervals. This not only eases some of the assumptions for Pre-Test Post-Test especially autocorrelation and indirect changes of the outcome, but also allows capture of time response of outcome with respect to the treatment. However, Time Series require more complex Comparisons, and have their own concerns regarding the consistency of the data recording over time.

Comparison

Comparison is defining the treatment group and potentially control groups. The decision for the type of Comparison used mostly depends on the hypothesis, not validity. For the binary treatment, the decision for the groups boils down to whether the non-treatment state is used as the control group or not. For treatments with more than two possible states, each state of the treatments can also be controlled against all other states. Sometimes a placebo or baseline treatment exists, which can be used as the control group for all other states.

The common Comparisons are [30, 156]:

- One-Group/Within-subjects Design, where only the samples that have received the treatment are used in the experiment. The absence of a control group results in many threats to internal validity, which are usually remedied by modeling the effect of confounders within subjects. Hence, One-Group is only recommended when all the samples are at some point exposed to some level of the treatment.

- Non-Equivalent Groups/Case-Control Study, which uses samples with the state of the treatment variable specified in the null hypothesis as the treatment group, and samples with the baseline state of the treatment variable as the control group, and compares the difference of outcome between them. Non-Equivalent Groups is the best-known Comparison and is usually performed on the categorical variables whenever possible. The most common concern against the Non-Equivalent Groups is when the treatment and control groups are not under similar confounding effects. There are also less-common approaches, such as comparison to other data sources, which are used when the external validity is a major concern.

2.3.3 Threats to Validity

The Validity of results of a CD can be broken down into two categories: (1) Internal validity, concerned with whether the results are justified given the experiment, and (2) External validity, concerned with whether the results are generalizable beyond the experiment. External validity depends on how causal mechanisms vary across different subpopulations and context.

As discussed in SL assumptions, many CD methods make heavy assumptions about the distribution of the data reflecting those of reality. But even weaker assumptions used in some QEDs, such as the potential cause having homogeneous effects across different real-world contexts, sacrifice the External validity. Although low External validity is not unique to QEDs, or CD for that matter, due to randomization and Multi-modality of Assignments, depending on the design, QEDs preserve the External validity relatively well [190].

The same cannot be said about Internal Validity and as mentioned in the Overview, threats to Internal Validity are the main concern for using QEDs. Because many of these threats are rooted in alternate explanations, the usual remedy is the inclusion of control groups. Furthermore, most threats are specific to some of the QED designs. Nevertheless, for every specific QED, the possibility of each threat should be checked, or countered. Hence, it is crucial to identify the categorization and solution of the major categories of threats to Internal Validity [49, 30, 190]:

Samples Threats

Some of the most basic and yet most common threats are rooted in the samples. Selection bias is the threat of systematic differences between groups' samples, and therefore can influence the outcome. Selection bias can manifest itself in the change in the distribution of confounders, and hence can be detected by checking for unexpected differences of observed confounders in treatment and control groups. Random Assignment and Matching Assignment can limit the Selection bias's effect. It should be noted that Selection bias is not limited to Case-Control Studies.

History and Maturation are the threat of natural changes of samples between the pre-test and post-test. Maturation and History usually occur when the duration between the observations is long enough that the outcome can drift even in absence of any intervention, and other causes could have affected it in the meantime respectively. History and Maturation can be adjusted for by using a control group Comparison that would be under similar temporal effects, or time series Measurement which allows modeling the autocorrelation of the outcome and effect of other observed causes.

Attrition is the threat caused by dropout or missing observations of samples that can be systematically different from the rest of the population. This can be hidden by the preprocessing procedures commonly used in CD that involve removal, interpolation, or imputation, and should be considered before preprocessing. As for detection, a rule of thumb is that <5% attrition leads to little effect, while >20% poses serious threats. The solution for Attrition in serious cases is limited to Worst-case Sensitivity analysis.

Instruments Threats

Some of the threats are rooted in the measurement and manipulation, which are usually out of control of the observational CD, but are still important to consider as the potential explanation of the outcome. Testing threat occurs when the active measurement of the outcome affects it. Testing threat can even lead to Practice Effect, where multiple observations lead to samples adapting to it, or Sensitization, where multiple manipulations of treatment results in the amplification of its influence on the outcome. Testing threat is common in health and social systems, and are best investigated by studying the manipulation mechanism, and can sometimes be solved via control group Comparison.

Other notable Instrument threats are Instrumentation, which is the measurement criteria or setting change across observations, Low Construct, which is when measurements of variables are inaccurate or noisy, and Artificiality/Demand Characteristics threats which is the change in the measurement setting, criteria, or sample when treatment is applied. Although these threats are very common, they usually cannot be controlled for at the QED level, and are mostly assumed and accounted for in the modeling and evaluation of significance.

Regression to Mean

Regression to Mean is the statistical tendency of outliers in one observation to be less extreme on another observation. Because in many QEDs samples are selected based on their observation results, Regression to Mean is one of the most common threats to their internal validity. Regression to Mean can usually be identified from the QED design, and its impact can be estimated via a control group, which is also very effective in adjusting for it.

Other notable threats to the internal validity of QEDs include: Ambiguous Temporal Precedence, where the order of change in the treatment and the outcome is not known, Demand Characteristics, where characteristics of samples change for the observation, and Experimenter Expectancy Effect, where observer's bias subconsciously influence the observations. However, arguably the most important threats to internal validity occur when two or more of the threats discussed previously to interact with each other, which makes identification and control of them more challenging.

2.3.4 Common Designs

Common QED designs can be built from the Assignments, Measurements, and Comparisons, discussed in the section Components. Although some of these designs are strictly superior to others due to providing higher statistical power or lower threats, discussed in the section Threats to Validity, they might make more significant assumptions about the problem or be limited to specific dataset types. In the following, the common QED designs, their threat, assumptions, and limitations are discussed. QED designs can be represented minimally using sequences of symbols X (Treatment), I (Selector), and O (Observation), with their order specified on a line from left to right, and sometimes in multiple parallel lines for different groups, and bolded letters indicating where the values based on which the significance is estimated [49, 30, 190, 70].

Within-Subjects Designs

Post-Test X O

One-Group Post-Test Only design or One-Shot Case Study is the simplest QED, where the outcome is only observed once after the treatment is applied. Consequently, One-Shot Case Study can be applied to any data, but because of lack of any baseline from a control group or pre-test observations, has the highest threats to validity. Hence, it is only saved for problems with unexpected or terminating events where no other QED designs can be applied, and the baseline of outcome is clear.

Pre-Test O X O

One-Group Pre-Test Post-Test design improves over the One-Shot Case Study by observing the outcome before the treatment. Requiring multiple observations for the same samples limits the applicability of One-Group Pre-Test Post-Test. Although by comparing samples to themselves before and after treatment effect of many threats are controlled, because of lack of a control group, two major assumptions threat the validity. Firstly, it is assumed that outcome would remain unchanged in the absence of intervention, which introduces threats such as History and Maturation. Secondly, it is assumed that any change in the outcome is due to intrinsic change of the outcome, which opens the door for threats, most important among which is the Regression to Mean.

Time Series O O X O O

Interrupted Time Series design is the strongest of the Within-Subjects Designs. Interrupted Time Series compares the change in of the trend of the outcome before and after the treatment, although it can be used for categorical outcomes by using the number of changes in the group instead of the quantitative change of each sample, it is best suited for problems with continuous outcomes. But most crucially, it requires the data to have multiple observations of the outcomes at multiple time points before and after the treatment. In return, the datasets that can provide such observations, face fewer assumption and threats. The major assumption here is that trend would have continued in the absence of the treatment. Hence beyond the Proportionality Bias, Attrition, Testing, and Instrumentation threats, which can be identified relatively easily, the critical design decision is the temporal model for the trend of the outcome, such as ARIMA or nonlinear transformations, and compensating the effect of autocorrelation bias on standard errors. Furthermore, Interrupted Time Series has the unique advantages of distinguishing lasting vs immediate effects, high statistical power, and intuitiveness.

Non-Equivalent Group Designs

Post-Test

$$\begin{array}{cc} X & O \\ & O \end{array}$$

Post-Test Non-Equivalent Groups design or Static Group Comparison is the most used QED design, where the outcome of samples with and without treatment is measured, and compared against each other. This requirement of only one observation per sample, which results in its applicability to non-temporal data, is the reason for the popularity of Static Group Comparison. The key assumption here is that the groups are similar before the treatment and would remain similar in absence of it. This direct control of confounders can act as a double edge sword, in improving the external validity from one side, and missing latent confounders and adding bias by controlling for colliders between treatment and outcome from another.

Various subtypes of Static Group Comparison have been developed to achieve group similarity using techniques such as guaranteeing that confounders are similar across groups (matched design), removing outlier samples (propensity score design), and random subsampling (randomized blocked design) for controlling hidden confounders. However, this shift of the burden of controlling threats to the selection of groups, adds responsibility to the process of designing the QED and ambiguity to the validity of the results [81]. Furthermore, the threat of treatment occurring after the observation and ambiguity in the direction of causality have to be resolved independently.

Pre-Test

$$\begin{array}{ccc} O & X & O \\ O & & O \end{array}$$

Pre-Test Post-Test Non-Equivalent Groups design or Difference in Differences is the evolution of Static Group Comparison with pre-test observations. This controls the Sample threats even further, to the point that Difference in Differences is considered one of the best and yet simplest designs. However, it still suffers from the task of identifying similar groups, in addition to requiring more than one observation per sample. The main assumption of Difference in Differences over Static Group Comparison is that groups would have followed the same trajectories in the absence of treatment.

Switching Treatment

$$\begin{array}{ccccc} O & X & O & & O \\ O & & O & X & O \end{array}$$

Pre-Test Post-Test with Switching Replication design extends the Non-equivalent Group Design even further by using samples multiple times, as both treatment and control. Because of the built-in replication, Switching Replication design not only improves on the History and Regression to Mean threats, and control latent confounders more easily, they can also increase the sample size when the treatment is not sparse. The major assumption is that it is unlikely that an event coincides with both treatments. If the observations are not close enough this assumption might be threatened. The most notable threat is Testing, especially in modified versions of the design in which more than three observations exists.

Furthermore, creating similar groups sometimes requires more than a few observations per sample, which is out of the reach of many datasets. There are other variations of Switching Replication design. For instance, Treatment Removal Design which is a trade-off between Switching Replication and simple Pre-Test Post-Test design in that it does not reuse the control samples as treatment, used for the problems with sparse treatments, Repeated Designs, in which a group is assigned to treatment multiple times, to evaluating the consistency of the treatment effect on the outcome, and Solomon 4-Group design, where the two extra observations are taken from new samples, to provide further control for testing threat.

Natural Variables Designs

X I O
O

Instrumental Variables

Instrumental Variables Analysis has been used for decades beyond the QED and even causal inference has well-established theoretical support and can be used to infer strong causal relations with very low threats to validity. This is achieved by identifying a variable that could possibly only be associated with outcome through the treatment in prior and evaluating causality through that Instrument variable. The Instrument is usually a real-world variable irrelevant to the causal mechanism of the outcome or a dummy randomly assigned variable that happens to be correlated to the treatment [19]. However, because of requiring prior knowledge, its applicability to HCD is situational. Furthermore, the effectiveness of Instrumental Variable is heavily reliant on the relevance assumption, which is Instrument not having a causal effect on the treatment, and exclusion restriction, which is the Instrument not having a direct causal effect on the outcome [28].

I X O
O

Regression Discontinuity

Although Regression Discontinuity is a Pre-Test Post-Test Between Subjects design, similar to the Non-Equivalent Group Designs mentioned above, its power lies in a threshold-based Assignment. In Regression Discontinuity design, samples with a variable just above an arbitrary threshold are used as treatments, and samples just below the threshold as control. This threshold acts as pseudo randomization and the differences between those just on the sides of the threshold is small enough that it does not cause bias in the confounders. Regression Discontinuity is especially robust to Samples threats and results in an unbiased estimate of the causal effect [196].

However, what distinguishes Regression Discontinuity and Instrumental Variables as the QEDs of highest the standard is that they account for differences of hidden confounders between the groups, which other QEDs at best controlled for. One of the main considerations in investigating the validity is that there should be no clustering of samples on sides of the threshold which would suggest manipulation of the treatment. Furthermore, Regression Discontinuity needs a large sample size and precise modeling of the causal function. Because access to a continuous threshold variable used for assigning treatments is situational, especially in high-dimensional observational datasets, the application of Regression Discontinuity is limited to specific problems [35].

2.3.5 Discussion

In recent years, a paradigm has emerged to solve the HCD problems by empowering QEDs with machine learning, which are becoming very popular in social sciences [71, 168, 133]. The new methods consist of a heuristic search for filtering the potential causes from the set of all variables, coupled with an efficient QED or sometimes a Rubin Causal framework [164] for evaluating the validity of those potential causes more rigorously. Technically almost any machine learning method, from clustering to classification, such as feature selection, autoencoding, recommendation, anomaly detection, and transfer learning can be used to guide the QED. But most of the existing methods use unsupervised algorithms in order to avoid high multiple hypothesis adjustment.

One of the most successful of these attempts is the line of work by Athey in [13, 200]. They have introduced a family of nonparametric CD methods based on decision trees. What distinguishes their approach is that it can discover subpopulations with outlier Average Causal Effects and the confidence intervals for each. This is achieved via an adaptive nearest neighbor algorithm in identifying critical variables. Furthermore, this approach overcomes selection bias and effect of hidden confounders by subsampling of the random forests. However, these methods fail to not rely on subspaces with low coverage, and consequently its accuracy plummets at datasets with small population and large number of variables, which is critical in HCD.

Another notable line of work is done by Li [102, 103] in which efficiency is the main focus. In these methods, candidate variables are selected through an association rule mining algorithm, and then tested in a series of retrospective cohort studies which evaluate causality of those candidates. However, these methods suffer from sensitivity to controlled confounders, and an implicit causal sufficiency assumption.

In general, by drastically reducing the number of hypothesis tests, these methods scale to HCD at a much better rate than SL. Furthermore, because of the QED designs and the possibilities of feature extraction and mapping during the search phase, they can adapt to CD problems with specific objectives, constraints, or hypotheses. However, what distinguishes QED-based HCD methods is higher external validity than most SL models and even true randomized experiments. This is because observational QEDs involve real-world treatment patterns instead of artificially controlled settings. In return, they usually suffer from reliability and theoretical guarantee of the results. This is partly because they inherit the threats to internal validity and contamination of magnitude of causal effect by confounders of QEDs they use, and partly due to the heuristic nature of the non-causal search process.

2.4 Conclusion

2.4.1 Progress Made

As discussed, the general causal inference and CD approaches in dealing with HCD problems face problems in computational complexity, multiple hypothesis adjustment and various biases. In the last decade, well known SL methods has been extended to deal with HCD by bypassing the exponentially growing conditioning sets. However, the theoretical guarantee which they offer relies on a list of assumptions that is rarely fulfilled in real world applications. On a parallel line of research, machine learning techniques equipped QEDs against large HCD problems by searching for promising hypotheses. Although QED-based methods are flexible and efficient, they are plagued by validity threats and are reliant on the user to adjust method to the particular HCD problem.

2.4.2 Existing Gaps

The first gap in HCD is with regards to the credibility of results. Overall, HCD methods are becoming more and more efficient but less and less reliable. At its current state, to the best of our knowledge, there are no significant laws, similar to Newton's laws of motion or Hardy-Weinberg principle, or major scientific discoveries, identified by any CD method, in any field, despite the overflow of practitioners and datasets.

If CD, and HCD in particular, is going to have a future outside topics such as explainability and fairness, and be used for scientific knowledge discovery, similar to randomized controlled trials, they need to demand more from the data and less from the assumptions. To achieve this, there is a desperate need for a falsifiable platform that evaluates how well the HCD approaches can replicate existing discoveries in various real-world applications from existing large observational datasets, in presence of nonlinearities, non-stationarity and data corruptions.

The second gap in HCD lies in its applicability. The current space of HCD is mostly focused on variations of similar successful methods outperforming each other in the general HCD, or applying HCD methods to specific applications. This is in contrast to sister subfields such as machine learning and causal inference, where there are various problem types, inspired by real-world applications. In this vein, HCD can benefit from defining and solving specific problems, inspired by reoccurring real-world applications, such as finding combinations of causes whose joint occurrence is required to change the outcome, temporal patterns which causes need to follow, causes of specific subpopulations or underlying mechanisms, causes in active and reinforcement learning setting, exact function between the cause and the outcome, and distinguishing causes that are sufficient from those that are necessary. Most of these tracks can be tackled by adjusting the existing methods or combining them with other well-known algorithms.

Chapter 3

Relational Causal Discovery

As discussed previously, a large enough number of samples is crucial to overcome key challenges of CD such as large conditional independency sets and multiple hypothesis correction. This demand grows nonlinearly with the number of variables, and by the size of HCD problems becomes the key decider of the performance.

On the other side, in many applications the prior knowledge of the relations between some of the variables is available. This opportunity even more pronounced in HCD, where the wide range of variables provides ample opportunity to import various relational data into the analysis. For instance, in bioinformatics problems, the association between different levels of genomic, transcriptomic, metabolic, and clinical phenotypes is created by extensive databases [14].

These prior relations can range from knowledge of causal relations to simple correlations extracted from other datasets. The prior relations not only can help narrow down the set of potential causal relations and increase their posterior confidence, but also help to identify the potential confounders, thus lowering the pressure from multiple hypothesis correction and conditional independency sets respectively.

The incorporation of prior relations is applicable not only as a performance enhancing technique, but also when the known relations are ought to be enforced. The incremental search for the new discoveries based on existing ones is the common mode of real-world knowledge discovery. This is indeed how the modern scientific studies are done; the findings of previous researches motivates the search for potential relations between the variables and mechanisms involved. Information transfer has even been used in other sample-starved subfields of machine learning such as transfer learning and active learning to a great success [208, 175, 174].

Yet, despite the practical and epistemological motivations mentioned above, CD of a partially discovered system of relations has remained understudied. CD methods either cannot use the relational data as prior, such as most QED based methods, or are limited to assuming relations as known causal relations, such as SL methods.

We name this problem of CD using prior relational information Relational Causal Discovery (RCD). To solve RCD, we propose HUME [115]. HUME is also capable of performing on a very large number of variables, and variables of different types. Because of the quadratically large space of possible relations, the main challenge of HUME is to very efficiently guide the search and conditioning for causal relations based on associative relation.

HUME is based on guiding a matching-based QED via network analysis. At its core, HUME uses three ideas. Firstly, it finds the promising candidate relations for the QED, while avoiding the data dredging of checking a large number of potential hypotheses. This is done by utilizing a novel link prediction algorithm on a network of co-occurrences of variables that is also populated by known associations. Secondly, it narrows down a large number of potential confounding variables for each candidate relation, which because of the high-dimensional nature of the problem cannot be ignored. This is done by using proximities in the co-occurrence network to distinguish hard and soft confounders and matching using a novel optimal pairing technique. And thirdly, across different steps of the algorithm, it eliminates the variables that lead to prevalence-wise untestable hypotheses. This is achieved by reverse engineering the number of samples required for the best-case scenario to use the variable.

3.1 Problem Description

RCD takes two inputs:

- A necessary binary sample-variable Data matrix $M \in \{0,1\}^{n \times m}$, where $M_{i,j} = 1$ if the j^{th} potential cause variable is positive in sample i, and $M_{i,j} = 0$ otherwise.
- An optional bounded variable-variable Prior matrix $P \in [0,1]^{m \times m}$, where $P_{i,j} = p$ indicates the evidence for relation between variables i and j, with $p = 0$ for no evidence and $p = 1$ for highest evidence.

And outputs:

- The sets of cause-effect relations between the variables $\{x_i^o \rightarrow x_j\}, i, j \in \{1, ..., m\}$.
- The vector of causal statistical significance p^*, where p_i^* corresponds to adjusted p-value of i^{th} causal relation.

Here, the Prior matrix can aggregate multiple sources and types of relations, and values can be assigned depending on the nature of interactions between the variables and the reliability of relations. For instance, larger values can be assigned to relations that are considered causal, or simply have higher correlation, or there is a higher confidence in source reporting their association. In experiments done with HUME, we will show the real-world example of how the various data types and be aggregated and enrich the Prior.

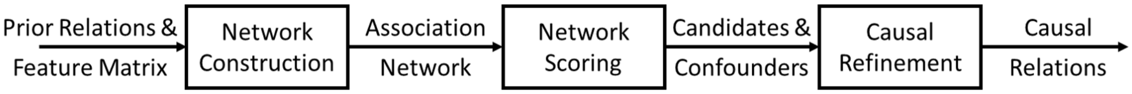


Figure 3.1: **Phases of HUME.** In phase 1, the Association Network is constructed based on Prior matrix and co-prevalence between variables in the Data matrix. In phase 2, the networking scoring equation is used to sort and filter candidate edges in the network. In Phase 3, the QED is used to test the causal significance of candidate relations after controlling the effect of Relational Confounders and multiple hypothesis testing.

3.2 HUME Method

To solve RCD, HUME deploys three phases of analysis using various novel techniques (Figure 3.1).

- First, HUME needs to construct a model that can represent the prior associations between variables, for which we introduce the Association Network.
- Next, since all possible relations cannot be evaluated, HUME needs to recruit a set of top candidate relations, for which we develop a novel link scoring method that efficiently ranks all possible relations based on their associations in the Association Network.
- Ultimately, HUME needs to test each of the candidate relations for causality.
    - Network: Since the number of potential confounders for each candidate relation can be very large, candidate confounders are identified and categorized by priority using the proposed Relational Confounder Identification.
    - Since a large number of confounders can reduce the statistical power significantly, confounders are efficiently controlled via the suggested matched-pairs QED that uses Hungarian algorithm to optimally pair samples based on their similarity in confounders.
    - Then McNemar's test can be used for evaluating significance of each candidate relation from the matched samples.
    - Finally, HUME must control the effect of the multiple hypothesis testing, for which we use the False Discovery Rate (FDR)-based Benjamini-Hochberg correction.

Performing phase 1 and phase 3, in one form or another, seem to be obvious steps for solving RCD. The reasoning behind performing phase 2 is to have an unsupervised control over the trade-off between false positives and false negatives. If we were simply testing every possible relation, even ignoring the computational cost, the number of hypotheses would scales quadratically with the number of variables. To avoid the potential of passing some of

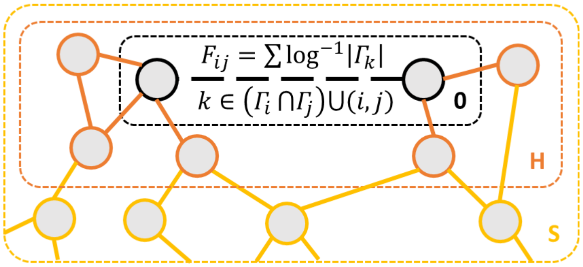

$F_{ij} = \sum \log^{-1}|\Gamma_k|$
$k \in (\Gamma_i \cap \Gamma_j) \cup (i,j)$
0
H
S

Figure 3.2: **Association Network.** Two variables are connected if they co-occur in n_0 samples in or more in the data matrix or have a non-zero value in the Prior matrix. Candidate relations are selected from the edges with the highest score F. Hard Confounders H and Soft confounders S are selected based on their minimum distance from the variables of the candidate relation.

the non-causal potential relations just by random chance, the multiple hypothesis correction would have to be very strict which could potentially reject many of the causal relations.

This is even a bigger problem in the context of HCD, due to the large number of mostly unrelated variables and the relatively small number of samples relevant to each hypothesis. Like many HCD methods, HUME solves this problem by reducing the set of all possible relations to a smaller set of candidate relations, using phase 2 as well as the proposed Prevalence Threshold once before phase 1, and once before phase 2, to further filter variables and relations based on a lower bound for their co-occurrence. In the following, we explain each phase of HUME in more detail.

3.2.1 Network Analysis

Goal of the first and second phase are to model the inputs on a network and then use the network to find promising candidates for causal relations.

Association Network

Association Network $G = (X, E)$ is an weighted undirected graph, were each node represents a variable in the Data matrix $X = \{x_1, ..., x_m\}$, and each edge represents the association between the nodes on its ends $E_{ij} > 0 \mid x_i \sim x_j$ (Figure 3.2). There exists an edge E_{ij} between two nodes x_i and x_j if they either:

- Are considered associated apriori, indicated by non-zero elements in the Prior matrix $P_{ij} > 0$.
- There is evidence of their association, indicated by co-prevalence higher than a specified threshold in the Data matrix $\sum_{k=1}^{n} M_{ik} \cdot M_{jk} > n_0$.

Hence, the edge weight matrix can be neatly calculated as $E = max(P, [M^T M > n_0])$. It should be noted that since E is large, sparse, and each of its elements and rows can be computed independently and efficiently, HUME does not need to store it, and instead can derive them on the fly.

Network Scoring

Now that the Association network is constructed, the next step is to rank the edges. To achieve this, HUME uses a modified version of an unsupervised link prediction scoring scheme. The reasons for using an unsupervised link prediction approach are its well-studied power in measuring the chance of existence of an edge between nodes, minimizing the dependence on learning parameters, keeping the prediction whitebox for the expert's interpretation and intervention, and most importantly, avoiding additional data dredging.

In most unsupervised link prediction methods, a score is assigned to the potential link between each pair of unconnected nodes, and the edges the with highest scores are selected as the predicted links. HUME's score is based on the Adamic-Adar (AA) score [3] which itself is a version of common neighbors score that weights the uniqueness of the shared nodes. The reason for using the AA score is that it not only outperforms other local scores in most scenarios, but also outperforms global heuristics, such as Katz score, in the networks, such as ours, that have short paths between potential nodes [105].

However, unlike the standard link prediction methods that outputs nonexistent edges, HUME expects the network scoring to output a subset of existing edges, since only those edges can represent the relations that have enough supporting samples in the Data matrix to potentially qualify the causality test. To solve this problem, unlike the original Adamic-Adar algorithm, HUME includes the joint neighboring set, and computes the score only between connected variables. Hence, HUME's score function becomes:

$$F_{ij} = \sum_{k \in (\Gamma_i \cup \Gamma_j)} \log^{-1} |\Gamma_k| \tag{3.1}$$

Where S_{ij} is the score of the possible edge between two variable nodes i and j, and $\Gamma(k)$ is the set of neighbors of node k. After computing the score for all existing edges, the edges with the top c highest scores, that have at least n_0 samples with both variables positive, are selected for the next phase as the candidate relations.

3.2.2 QED

Having obtained the candidate relations from the network scoring phase, the objective of the third phase is to only accept the relations that have sufficient evidence for their causality in the data matrix. To the best of our knowledge, QEDs had not yet been used for CD of relational datasets and networks, and hence multiple techniques needed to be developed to apply QEDs to RCD.

3.2.3 Matching

The core of QED is the experiment design model. The experiment design that HUME uses is the matched pairs design. This is because matched pairs design is flexible and efficient in dealing with a large number of confounders and confounders of different types. Matched pairs design, tries to couple every sample that has received the treatment to a sample that has not received the treatment, but is similar otherwise, and then uses the difference between the outcome in these set of pairs to evaluate causality of the hypothesis.

For HUME, Since the candidate relations reported from phase 2 are undirected ($x_i — x_j$) and matched pairs design is asymmetric, we need to assume one of the variables in the relation as the treatment and the other as the outcome and perform the hypothesis test ($x_i \overset{?}{\rightarrow} x_j$), and then swap their place and repeat the test ($x_j \overset{?}{\rightarrow} x_i$).

Relational Confounder Identification

Matched pairs design requires the confounder of every hypothesis to be specified. However, due to the large number of the potential confounders for each candidate relation, it is impossible to specify the confounders of all possible hypothesis without the risk of data dredging, and control for them without the risk of heavy statistical power loss. Therefore, HUME needs to efficiently estimate confounders of every candidate relation. The proposed approach divides confounders of a candidate relation into two sets of hard confounders and soft confounder and uses the proximity between nodes in Association network to distinguish them.

We define the Hard confounders as the potentially confounder variables that have significant association and therefore must be controlled, and Soft confounders as the other relevant variables which we prefer to be paired, if possible, to maximize the similarity between paired samples. Specifically, hard confounders of a candidate relation are defined as the set of nodes in the network that are at geodesic distance of l or less from either of the nodes involved in relation, and its necessary to match them in pairing. Soft confounders on the other hand, are the set of non-hard confounder nodes in the same connected component of network as candidate nodes, and are only preferred to be matched in the pairing. Hence, the set of hard confounders and soft confounders of variables i and j of distance l are defined as:

$$\begin{aligned} D_{i,j}(k) &= \min\{d(i,k), d(j,k)\} \\ H_l\,(i,j) &= \{x_k \in X \mid D_{i,j}(k) \leq l\} \\ S_l\,(i,j) &= X \setminus H_l\,(i,j) \end{aligned} \tag{3.2}$$

Where $d(\cdot,\cdot)$ is the geodesic distance between the nodes in its argument, $D_{\cdot,\cdot}(\cdot)$ is the geodesic distance between the node of its argument and the edge between nodes of its subscript, and is used to identify the set of nodes that are only $\leq l$ node apart from variables in candidate relations. It should be noted that although equation 3.2.3 would be computationally very expensive in general, in practice since number of candidate relations is much smaller than the number of variables and usually the candidate relations share variables, HUME can first identify the unique variables across all candidate relations, and then for each for query the neighbors of side d, before aggregating them for hard confounders, meaning that the number of neighborhood search queries (α) is bounded by: $\alpha < 2 \cdot c \ll m$.

There are multiple reasons for using geodesic distances of the network as the measure of finding confounders. Firstly, this guarantees competing candidate relations to be counted as confounders of each other due to their adjacency in the network. Secondly, any statistically significant confounder would have high enough co-prevalence with at least one of the nodes, inevitably be a neighbor to it in the network, and therefore be selected as a hard confounder. And lastly, the larger set of soft confounders helps to control for the cumulative effect of many nodes that are not direct confounders, but their weak correlation can cascade through chains of nodes, a phenomenon known in network analysis as contagion [43].

Optimal Matching

To match pairs of samples from the treatment and control group, HUME needs to consider many hard and soft constraints implied by the hard and soft confounders respectively. This may lead to heavy statistical power loss for the matching, if the pairings are not done efficiently. We formulate this problem as an optimization problem of assigning treatment samples to control samples with hard confounders as constraints and soft confounders as penalties. A technique used in integer programming is to move constraints to the cost function, and weight them so high that it is guaranteed that violating them cannot be compensated, and using a threshold equal to weight of the hard constraint for the accepted optimal value, so that if satisfying hard constraints are impossible, the assignment be rejected [10]. The optimization problem can be solved efficiently using the Hungarian Algorithm [94].

The overall process of the Relational Confounder Identification and optimization formulation can be combined elegantly. This can be done by computing the minimum geodesic distance of every node from the nodes of the candidate relations, and then computing the sharp inverse sigmoid of the minimum distances, to find weight of every node. Next, for every pair of treatment and control, weights of those nodes that do not match between the two samples are summed as the difference between them. This matrix of difference can be fed to the Hungarian algorithm to find the treatment - control pairs. Therefore:

		$\overline{x_i}$	
		x_j	$\overline{x_j}$
x_i	x_j	a	b
	$\overline{x_j}$	c	d

Table 3.1: Discordant b is the number of pairs of positive treatment and negative control samples, and discordant c is the number of pairs of negative treatment and positive control samples. For instance for the candidate relation $(x_i \rightarrow x_j)$, discordant b for instance, is the number of paired samples where one has the positive candidate cause $x_i = 1$ and the positive candidate outcome $x_j = 1$, and the other one that also has received the positive candidate cause $x_i = 1$ but does not have the positive candidate outcome $x_j = 0$.

$$\begin{aligned} &\min_{U} \sum_{i,j} q_{ij}\, U_{ij} \\ q_{ij} = &\sum_{k} \left(1 - \operatorname{sig}\left(D_{i,j}(k)\right) \cdot I_k(i,j) \quad \forall\, i \in X_{treat} \;\&\; j \in X_{cont}\right. \end{aligned} \tag{3.3}$$

Where U_{ij} is one if sample i from the treatment group is paired with the sample j from the control group, and is zero otherwise, and q_{ij} is the aggregate difference of the confounders in i and j, $sig(\cdot)$ is the sigmoid function, and indicator $I_k(i,j)$ is 1 if treatment sample i and control sample j are different at variable k.

A potential issue is the controversy over controlling potential confounders. Matching based on non-causes, such as colliders, can lead to false inferences [91, 144]. Although some authors argue that score-based/propensity score matching and randomized comparison safeguards against it [161, 163, 165]. both of which are used here.

Statistical Test

The ideal statistical hypothesis test for HUME's paired categorical data is McNemar's test. The standard McNemar's test approximates the p-value by applying Chi-Square test on the Test statistic consisting of difference between the discordants of the contingency table. In HUME's hypotheses however, most of the times the numbers of discordants are low, and thus, the normal distribution approximation of the Chi-Square test is overly liberal. Therefore, HUME uses the McNemar's test with the yates continuity correction [55].

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{3.4}$$

Multiple Hypothesis Correction

Since one hypothesis test is done for each candidate relation, a multiple hypothesis testing approach is required to counteract the chance of some of the large number of hypotheses being accepted by random chance [82]. To compensate for the large type I error caused by the number of test hypotheses, we need to control the data snooping effect while still aiming to discover as many causally significant G-DR relations as possible [56]. To do so, HUME uses the Benjamini-Hochberg procedure [20], a well known FDR approach that controls the expected ratio of false discoveries by sequentially adjusting the significance levels of statistical tests, to contain the overall type I error rate across all predictions. For HUME, the procedure is:

$$\text{Reject } \{\ \text{Hyp}_1, \text{Hyp}_2, \ldots\ \text{Hyp}_t\ \} : t = \underset{i}{\text{argmax}} \left(p^{(i)} \leq i \cdot \alpha_0 / c \right) \tag{3.5}$$

Where α_0 is the maximum acceptable FDR which controls how liberal the causality test can be, Hyp_1 to Hyp_c are the candidate null hypotheses, $p^(1)$ to $p^(c)$ are their corresponding p-values, sorted from smallest to largest, and t the adjusted significance threshold.

In the context of the problem, an FDR based approach is superior to alternative approaches such as Family-wise Error Rate Correction or Closed Testing Principle that set an overly conservative and less relevant goal of containing type I error rates of all hypotheses simultaneously, or Resampling based procedures which rely on tuning test statistics of the system of interdependent variables of the problem.

Prevalence Threshold

Throughout HUME, we used the parameter n_0 to filter variables. In the following, we show the justification and estimation for it. One of the major inefficiencies of CD methods, which becomes more significant in HCD, is evaluation of hypotheses and conditional independencies for which there is not sufficient volume of evidence to make judgement. Although in general it is not possible to find a tight bound for the requirement for the sufficient statistical power, we show that in a modular CD process like HUME, it becomes possible to find a lower-bound for minimum number of relevant samples.

We define the Prevalence Threshold n_0 as the minimum number of samples required to identify the significance of a potential causal relation. Hence, every variable that is not positive in at least n_0 samples and is not negative in at least n_0 samples should be removed. Furthermore, every potential relation whose cause and outcome are not both positive in at least n_0 sample is unfalsifyable and hence should be ignored.

The main strengths of a Prevalence Threshold concept are:

- It can be used at different points of the analysis, from checking the falsifiability of a conditional dependency, to evaluation of a potential causal relation.
- It can reject a variable at the preprocessing, and before any analysis, thus in a HCD setting potentially reducing the dimensionality significantly.

In HUME, we used n_0 three times:

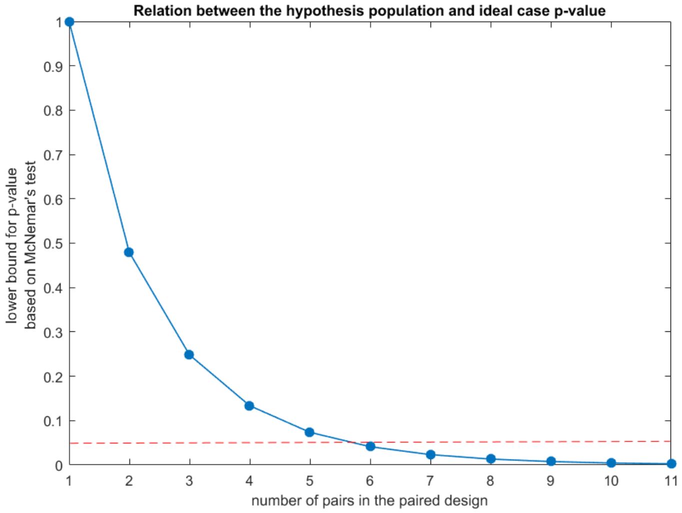


Figure 3.3: **Prevalence Threshold Estimation** Best-case reverse engineering approach is used to decide Prevalence Threshold n_0, a minimum for the number of samples required to pass the QED phase. For a selected significance level α_0, n_0 is the smallest number that its corresponding p-value, derived from ideal scenario of McNemar's test, is less than α_0. For $\alpha_0 = 0.05$ for instance, this correspond to $n_0 = 6$.

- As a prevalence criteria for the variables to filter the Association network nodes X.
- As a co-prevalence criteria for the Data matrix-based potential relations to filter the Association network edges.
- As a co-prevalence criteria for candidate relations to filter the hypotheses being selected via Network Scoring.

It should be noted that n_0 was not used for the QED phase because of the potential of skewing the number of candidates and underestimating the number of tested hypotheses.

Now we can discuss the process of deriving the value of Prevalence Threshold. Prevalence Threshold is designed as a necessary condition for variables and relations that would otherwise fail the QED due to insufficient prevalence of samples supporting them. We, therefore need to compute the upper-bound for prevalence of a variable or relation, where even in best-case scenario, it cannot pass the adjusted statistical test. To do so we need to traverse HUME from the end and reverse engineer its conditions for the best-case scenario:

- In the Multiple Hypothesis Correction, a candidate relation $(x_i \to x_j)$ would have the lowest p-value among all candidates so that the multiple hypothesis testing procedure would not affect its significance level. It should be noted that the Family-Wise Error Rate corrections, would result in a much more strict criteria and hence much larger n_0.

- In the Statistical Test, since the diagonal entries of the Contingency Table have no impact on the McNemar's test, for maximum test statistic, one of the discordant entries (b and c) would be zero and the other would be equal to the total number of pairs $[n/2]$. Hence, by computing the p-values of the the McNemar's test corresponding to different values of the discordant, we can derive that for any given significance level, what minimum number of pairs should be (Figure 3.3).

- In the Optimal Matching, the number of pairs would be equal to the minimum of size of treatment and control sets. This is exactly the condition which defines the Prevalence threshold n_0. Therefore, we have:

$$\begin{aligned} n_0 = & \underset{b}{\operatorname{argmax}}\, p_{MN}(b) \\ \text{s.t.} \;\; & b \in \mathbb{N} \;\; , \;\; p_{MN}(b) \leq \alpha_0 \end{aligned} \tag{3.6}$$

  Where $p_M N(b)$ is the p-value calculated using McNemar's test for contingency table with non-zero value b only at a discordant, $\mathbb{N}$ is the set of natural numbers, and α_0 is the desired significance level.

3.3 Experimental Evaluation

For assess the relevance of RCD and HUME's ability in solving it, we applied HUME to the pharmacogenomics application and dataset discussed in Chapter 1.2, and performed experiments in three folds:

- Performance of each of the phases of HUME are compared against baseline methods to judge the justification of each of the comprising component:
  - Filtering efficiency of Network Scoring based on how high it ranks the true causes among all possible relations.
  - Accuracy of QED independent of filtering based on the p-values it assigns to all millions of possible relations.
- Dependency of HUME accuracy on its parameters is studied to investigate the risk of overfitting and the applicability to other settings.
- The performance of HUME with all phases and the default parameters in finding known causes and new discoveries is measured to test the overall charactristics of HUME, and provide new discoveries for the application.

3.3.1 Filtering Efficiency

To evaluate the effectiveness of the first two phases of HUME in modeling interconnection between relations and filtering them, we have studied how well the Guideline relations

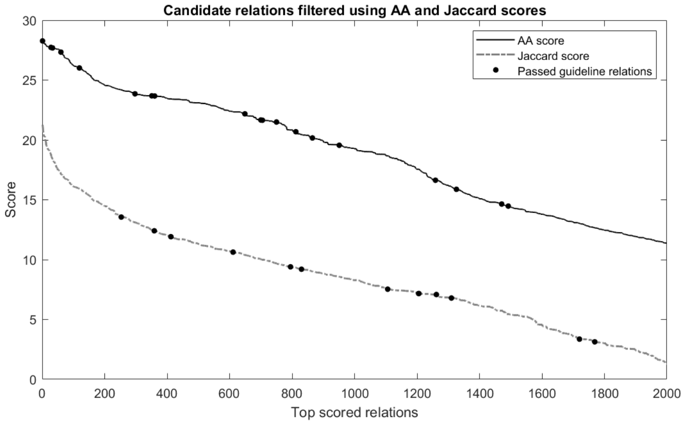


Figure 3.4: **Network Ranking of Guideline Relations.** Guideline relations are ranked very highly among potential relations for both AA score and Jaccard index. AA score significantly outperforming Jaccard index in ranking Guideline relations.

are ranked among all possible potential relations, and how well it can be compared to the intuitive and well known correlation used in the Jaccard index (Figure 3.4), which is commonly used to capture association in network analysis [159].

Modified AA score ranked the Guideline's strong, significant and notable relations are on average in the top 0.01%, 0.06% and 0.27% of top scored links, showing reliability of the HUME's Network Scoring in passing true relations, even after applying a harsh filtering. Moreover, median and minimum ranking of all Guideline relations are 3.02% and 11.57%, which indicates that by proper tuning of the Candidate threshold c, the majority of promising relations can pass, resulting in a potentially low false negative rate.

The respective average and minimum ranking of the Guideline relations for the Jaccard index are 7.38% and 26.52%. These metrics are much higher than the AA score, demonstrating superiority of the measure used by HUME, but also flexibility and potential for using alternative measures. This is an expected outcome since by weighting unique neighbors highly, AA score effectively emphasizes variables that cannot be explained by other confounders, and controls the contribution of potential causes that are common in many cases and would have passed by the large co-prevalence they would have with many other variables. These properties, in addition to the minimum number of tuning parameters involved, make AA score a preferred method for Network Scoring.

The candidate threshold c should be decided based on the size of the Association network. Increasing c indefinitely, in hope of discovering all promising relations, results in the opposite effect of overburdening the QED phase and pruning of many candidate relation by the multiple hypothesis test's compensation for the number of hypothesis. For the pharmacogenomic dataset, Candidate threshold of $c = 2000$ is used, which passes the majority of Guideline relations.

3.3.2 Causality Accuracy

Even by using a conservative Relational Confounder Identification Confounding Distance of $l = 2$, across all candidate relations passing from Network Scoring phase, the median number of confounders and the median number of paired cases is 13 and 24 respectively, indicating the statistical power of the match paired design experiments.

To evaluate performance of the QED phase, independent of how well the candidates are filtered in the network analysis phase, we measured the p-value of all 1.1 million valid relations under HUME's QED. Results show presence of most Guideline's strong and significant relations in top ranked relation p-values, on median being ranked in the top 0.082% of all relations.

To further demonstrate the overall validity of the QED process used by Hume, we studied distribution of p-values of candidate relations (Figure 3.5) evaluated by HUME under default parameters. The overall results show the ability of QED phase in detecting significance of Guideline's strong and significant relation and passing majority of them while preserving a high standard for new relations.

Next, to validate confounding correction phase and the overall matched pair design approach, we compared the distribution of p-values with the default settings to the scenario where key external confounding information about the patients is known (Figure 3.5). Specifically, we provided Hume with age, gender and dosage of patients to be used as hard confounders, with continuous values of age and dosage categorized into 5-quantiles with equal number of patients.

The results show a relative consistency of p-values, especially close to the deciding BH line, which indicate reliability of the selected confounders in simulating a quasi controlled and randomized trial. The only exception are the top ranked candidates, which on average received an order of magnitude higher p-values. This phenomenon is due to the fact that by adding more constraints on the matched design, smaller number of pairs will be available to be matched, and therefore statistical power of the test will be reduced, and less significant p-values will be observed, even despite having the same ratio of cases supporting the relation. It should be noted that in practice such key information about the samples and their level of treatment are not necessarily known, and Hume is able to provide similar evaluation in absence of hidden confounders.

3.3.3 Parameter Tuning

In total, there are three tunable parameters in HUME, namely Prevalence Threshold n_0, Confounding Distance l, and Candidate Threshold c. However, since It is not always easy to set these parameters based on a ground truth set of relations, we would like the performance of HUME to be relatively robust with respect to changes in n_0, l, and c. To test this, we

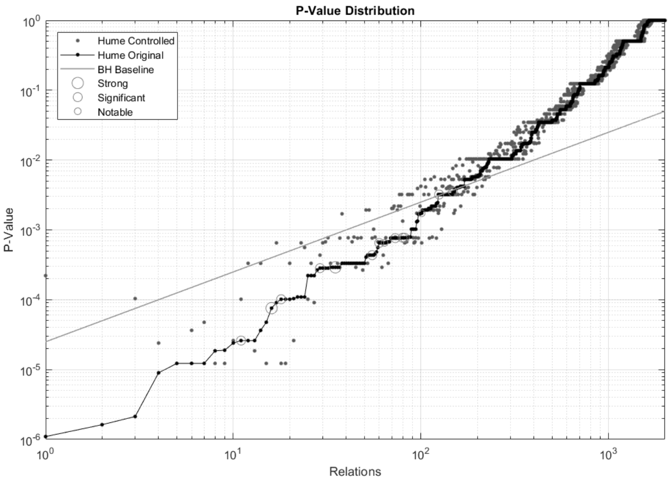


Figure 3.5: **QED Ranking of Guideline Relations.** Distribution of QED p-values of candidate relations on a logarithmic scale. For every relation, the p-value is calculated once only based on confounders discovered automatically in the Confounder Identification step, and once by also including clinical variables of the pharmacogenomic dataset as confounders. Guideline's strong, significant and notable relations are indicated by circles around them proportional to their level of significance. The Curve of p-values is compared to BH line, resulting in approval of the top 153 relations.

varied l between 0 to 3, c in range of 500 to 4000 with 500 unit steps, and n_0 between 5 and 10, corresponding to baseline p-values of 0.0736 and 0.0044 respectively (Figure 3.6).

By increasing Confounding Distance l, HUME considers more confounders for each hypothesis, and therefore controlling for variables associated with Guideline variables. This is evident as by increasing l from 0 to 1, the number of Guideline's strong and significant relations passing the process, increases from 3 to 6. On the other hand, by significant increase of l, pairing becomes harder and therefore the number of paired cases and consequently statistical power reduces significantly. This phenomenon happens when going from $l = 2$ to $l = 3$, as not only the number of passed Guideline's strong and significant relations drops from 7 to 4, but also 39% of the newly discovered relations with $l = 2$, fail to have enough pairs to pass the QED in the $l = 3$ scenario.

By increasing c, more statistically significant relations that do not have sufficiently strong connections in the network will have the chance to be evaluated by the QED phase. We can show this by measuring the number of Guideline relations that fail to pass the network filtering phase, which increases from 3 to, 5, 11 and 26 as Candidate Threshold becomes more strict by decreasing c from 2000 to 1500, 1000 and 500. On the other side, increasing c liberally results in a large list of candidate relations to be tested by the QED phase. This results in a much harsher multiple hypothesis testing step, which in turn threatens rejection of some of the significant relations. This can be noticed when increasing c from

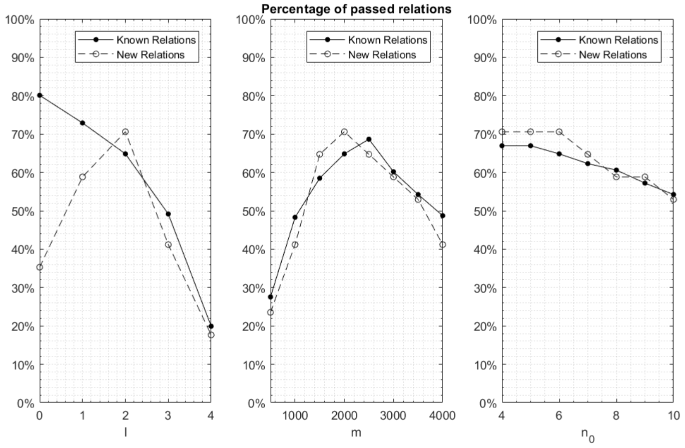


Figure 3.6: **Dependence of Performance on Parameters.** Percentage of Guideline's strong and significant (known) relations and non-Guideline (new) relations discovered for different values of the three tunable parameters of HUME, namely Confounding Distance l, Candidate Threshold m and Prevalence Threshold n_0. Discovery rate of new relations is based on the ratio between the number of non-Guideline relations and union of all non-Guideline relations discovered under all configurations.

2000 to 4000, the number of Guideline's strong and significant relations passing the process decreases from 6 to 2.

The effect of the Prevalence Threshold is more subtle. For the majority of realistic values for n_0, it only eliminates nodes and edges prior to their elimination by the network filtering and QED phases, making the process computationally more efficient. However, for large enough values of n_0, Prevalence Threshold works as an additional correlation test which negatively affects promising relations, similar to the effect of large values for c. This happens for $n_0 = 9$ and 10, resulting in the number of Guideline's passed strong and significant relations the process decreasing from 6 to 4.

Overall, unless extreme values are assigned to parameters, the HUME's performance shows good robustness. This is desirable, as we ideally expect HUME to be responsive to the False Discovery Rate α_0, which control how conservative the passed predictions should be.

As a Guideline, based on the most reliable results of the experiments, we recommend the default values of $l = 1$, $c = 1000$ and $n_0 = 6$ for the problems with similar scale to that of the pharmacogenomic dataset.

3.3.4 Overall Performance

We have tested the overall performance of HUME with all of its phases with default parameters. Among Guideline relations, 2 of 3 strong relations, 10 of 14 significant relations and 2 of the remaining notable relations pass the system. In addition, 153 new relations with significant p-values are discovered.

As discussed in the Chapter 1.2, pharmacogenomic applications have the unique advantage of validating results in different ways. We already discussed HUME's competence in discovering biomarkers that are known to be involved with anthracycline cardiotoxicity in the scientific community. However, there are additional pathway analysis evidence to support significance of the new discoveries. Among the 153 SNPs discovered by HUME, there are several cases of SNPs from the same gene being detected independently.

For instance, HUME's newly discovered SNPs $rs2587895$, $rs10748505$, $rs12078289$, $rs11185112$, $rs10494069$, $rs17014199$ and $rs521721$ all belong to gene $NTNG1$ which is involved in post-translational protein modification and cell adhesion molecules. Similarly, $rs3765483$, $rs16827109$ and $rs12757549$ are from gene $ZMPSTE24$ which is part of the Adipogenesis and Terpenoid Backbone Biosynthesis pathways, and $rs1444307$, $rs11073622$, $rs7161752$ and $rs2173090$ are from gene $AGBL1$ which are also involved in post-translational protein modification.

There are also HUME SNPs that match with the Guideline's strong and significant SNPs' genes, such as $rs17868327$ which similar to $rs17863783$, is from gene $UGT1A6$, and pathways, such as $rs11869821$ which is from ABC pathway similar to many of the Guideline SNPs. These overlaps support functional significance of SNPs discovered by HUME in gene and pathway levels.

3.4 Discussion

Beyond the common pros and cons of the hybrid (machine learning search and QED evaluation) HCD approach discussed in 2.3.5, specifically lack of theoretical guarantee and causal sufficiency assumption, HUME has few unique properties:

- By using the patterns in the prior datasets, as is premised by RCD, HUME is able to evaluate relatively very few hypotheses and which in turn leads to low false positive, low multiple hypothesis correction, and high statistical power relative to the high-dimensional and small sample size nature of the problem.

- HUME is the fastest of all the proposed methods in this thesis as well as the vast majority of the CD methods, including all mentioned in this thesis. HUME owes this to the very simple operations that it performs in each phase, as well as the efficiency at which the Network Scoring can filter causal variables.

- HUME is the only of the proposed methods in this thesis to not require a specific outcome, i.e. it searches for causes of every variable in the dataset, and in this sense is more similar to the established SL CD approach rather than emerging HCD methods. With the increasing interest in the latter, HUME can be modified to only construct the local network around a particular outcome to avoid the search in the whole hypothesis space.

- Although HUME achieved the efficiency goal we had in mind via its effective general design, its technical details have a lot of space for improvements. For instance, Network Scoring completely ignores the combined interactions between the variables, and hence would perform poorly if causal variables have non-factorable effects such as conditional or multiplicative effects. We will explore these challenges in the next chapters. Another example is matched pairs design which is one of the simplest QED, and prune to errors discussed in the Related Works chapter, and due to relative smallness of candidate causes, can be replaced with a more sophisticated QED, with little rise in the computational cost.

- To the best of our knowledge, HUME was also the first CD method applied to ADR SNP discovery. Despite its simplicity, HUME used the rich bioinformatics datasets, to find promising results and inspire further studies [34].

Chapter 4

Stratified Causal Discovery

The existing CD methods compare the potential causes and outcomes across all samples, and hence have the tendency to discover the causes that would try to explain all positive cases in the outcome. However, this approach will not perform well for the causes specific to a particular subset of samples, for instance in scenarios when an outcome can have different causes in different groups of the population or when the outcome has different underlying mechanisms.

For example, the pharmacogenomic application, anthracycline cardiotoxicity, is a multi-factorial adverse drug reaction whose main mechanisms are believed to be the inhibition of Topoisomerase 2β and the reaction of reactive oxygen species [120]. This means that the $rs2229774$ mutation in the *RARG* gene, which influences Topoisomerase 2β, would cause toxicity in some, but not all of the anthracycline patients. Current causal analysis methods may lack the statistical power for detecting the $rs2229774$ as a cause for anthracycline cardiotoxicity, and $rs2229774$ would be dismissed by a current CD method on the whole population, due to its absence in those cases where the ADR is caused by the Reactive Oxygen Species mechanism, as there would be patients without $rs2229774$ in which anthracycline cardiotoxicity would nevertheless occur. See Figure 4.1 for a visual demonstration.

However, to the best of our knowledge, there has been no method capable of discovering causally distinct groups and their corresponding causes in high-dimensional settings. To fill this gap, we name this problem of discovery of causes specific to subsets of samples Stratified Causal Discovery (SCD). Here, strata are the disjoint subsets of the samples with a similar outcome. We show that if the outcome is modified to represent the positive sample of strata instead of all samples, the less prevalent causal mechanisms are more likely to be captured.

We present ARISTOTLE, the first method for solving SCD. ARISTOTLE is a multi-phase algorithm that tackles the above challenges by using a divide-and-conquer scheme that utilizes biclustering for finding the promising strata and candidate causes and QED to identify the stratum-specific causes. ARISTOTLE is also capable of using the prior domain knowledge of confounders and variables' grouping.

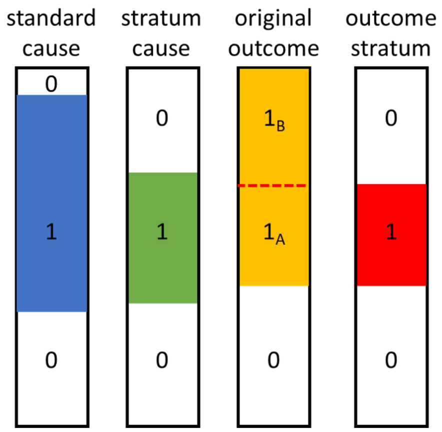


Figure 4.1: **Stratified Causal Discovery Example.** Diagram representing four variables (columns) across different sample (rows). A standard cause (blue) is highly associated with the original outcome (yellow), and can be identified by most measures of association. However, a stratum cause (green) that only affects the outcome in a particular type of samples (1_A), could be missed from simple association with the outcome, and requires comparison against its corresponding outcome stratum (red) to be identified confidently.

4.1 Problem Description

SCD is defined as the problem of identifying the strata with positive outcome such that for each stratum, there exists a set of variables, at statistically significant association with membership in the stratum, after controlling the effect of confounders. This means that for the samples in the stratum, the values of these variables, called stratum causes, is responsible for the positive outcome. Here, SCD follows the strict notion of causality, which expects that the outcome is positive almost always when and only when the variable group is positive.

SCD receives inputs:

- A necessary Outcome vector $Y \in \{-1, 1\}^n$, where $Y_i = 1$ if sample i has the positive outcome, and $Y_i = -1$ otherwise.
- A necessary binary sample-variable Data matrix $M \in \{0, 1\}^{n \times m}$, where $M_{i,j} = 1$ if the j^{th} potential cause variable is positive in sample i, and $M_{i,j} = 0$ otherwise.
- An optional sample-variable Natural Confounders matrix $Z \in \{0, 1\}^{n \times l}$, where $Z_{i,j} = 1$ if Natural confounder j is positive in sample i. Natural Confounders represent variables with apriori expected relation with the outcome that must be controlled as confounders and ignored as potential causes.
- An optional variables clustering Grouping set $G = \{G_1, \ldots G_g \mid G_i \subset \{1, \ldots, m\}\}$, where $i \in G_j$ if variable i belongs to Grouping j. Grouping represent the prior knowl-

edge about the sets of variables that are causes to similar outcomes. We refer to columns of Data matrix that correspond to Grouping set variables as $M^{(1)}, \dots M^{(g)}$.

And outputs:

- Causal Strata $P = \{P^{(1)}, \dots P^{(k)} \mid P^{(j)} \in \{0,1\}^n \;\&\; \forall\, Y_i = 1 \; \exists\,!\; P_i^{(j)} = 1\}$, which is the disjoint partition of outcome-positive samples with distinct causes. In other words, every sample whose outcome is positive, is assigned to a causal group $\sum_{j=1}^{k} P_i^{(j)} = Y_i$.
- For each causal stratum, Strata Causes $\{x_i^o \xrightarrow{P_j} Y\}$, which is the sets of variables causing the outcome in the stratum.
- The vector of causal statistical significance p^*, where p_i^* corresponds to adjusted p-value of i^{th} causal relation.

For instance for the pharmacogenomic application, a target ADR can be used as the Outcome vector, patient genetic panel as the Data matrix [62], clinical and demographic information as Natural confounders, genetic pathways as the Grouping set.

There are two main challenges for SCD:

- Stratification of population thins out the number of samples, while multiplying the number of possible causes by the number of strata, and thus exacerbates the chance of data dredging. This is an even bigger problem for HCD which already suffers from small in comparison to the large number of variables.
- Clustering cannot identify the hidden strata that reflect the underlying causal mechanisms (because of using both causal and non-causal variables), and if are chosen poorly, may very negatively impact validity of the results. In a causal stratification one would expect the samples from the same Stratum (with the same causal mechanism) to have similar Causes positive. This in fact results in a Catch-22, where a good approximation of the Causal Strata is needed to discover the Strata causes, while a good approximation of Strata causes is needed to arrange samples into Causal Strata.

It is worth noting that in addition to the causal variables, the causal strata may also include confounding and conditioning variables, and even non-causal associated variables. What distinguishes a stratum as causal is that there exist variables that cause the outcome for the set of samples in the stratum.

4.2 ARISTOTLE Method

To solve SCD, we propose ARISTOTLE [114]. The overall design of ARISTOTLE is based on a divide-and-conquer scheme that breaks the set of variables into groups and aggregates the significant sample and variable patterns. To do this, ARISTOTLE utilizes supervised biclustering to identifying the promising strata and variables, and matching QED to evaluate the causality of variables with respect to a particular stratum. Initial partitioning of the variables can be performed randomly or based on the prior clustering of Grouping set, or skipped if the problem is not very high-dimensional.

Like most HCD methods, ARISTOTLE's solution to SCD's first challenge is to select a shortlist of candidate causes from the set of variables, based on their association with the effect, which is the Causal Strata in SCD. However, Causal Strata are unknown, which brings us to SCD's second challenge which is to simultaneously stratify the samples and identify the candidate causes. This can be solved using a biclustering that jointly groups variables and samples based on their association, which could also be used to score the variables.

However, due to the high dimensionality, it is infeasible to perform the computationally complex biclustering on all variables simultaneously. Hence, ARISTOTLE has to divide the variables into groups, perform the biclustering to find the candidate causes and strata locally, and then merge the results to form a data matrix with fewer but of higher relevance variables. After the variables are reduced to candidate causes, biclustering can be reused, this time satisfying the second challenge.

For causal inference in presence of the low sample size created by the first challenge, as discussed in the 2, Matching QED, is one of the best approaches when the statistical power is the greatest threat to the validity. It should be noted that unlike the classical CD where the causes are considered based on their adjusted association with the positive outcome, in SCD the causes are considered based only on the outcomes corresponding to the particular stratum. Figure 4.2 illustrates these two types of causes.

It should be noted that although the Stratified Causes are identified using causal inference, the Causal Strata are established only based on association. Hence, a theoretical guarantee for causality of the outputs in general, does not exist.

This completes the reasoning for the design choices of ARISTOTLE: divide-and-conquer for high dimensionality, Matching QED for CD under the first challenge, and biclustering for candidate and stratum discovery under the second challenge. Hence, ARISTOTLE consists of the following five phases (Figure 4.3):

- **Grouping:** break down the variables into Groups, via Grouping set or randomly.
- **Scoring:** Bicluster each Group, and score variables based on their association to its strata.
- **Filtering:** Select the highest scoring variables from each Group, and aggregate them as candidate variables.
- **Stratification:** Bicluster candidate variables to find the Causal strata.
- **Inference:** evaluate causality of each candidate variables for each Causal stratum.

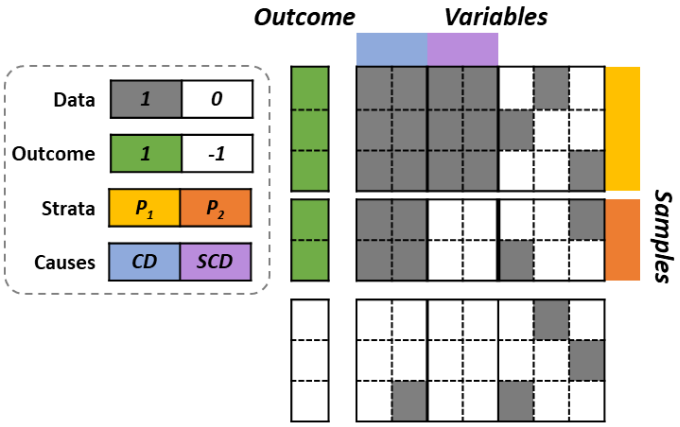


Figure 4.2: **Stratified Toy Example.** A sample dataset illustrating stratum-specific and general causes.

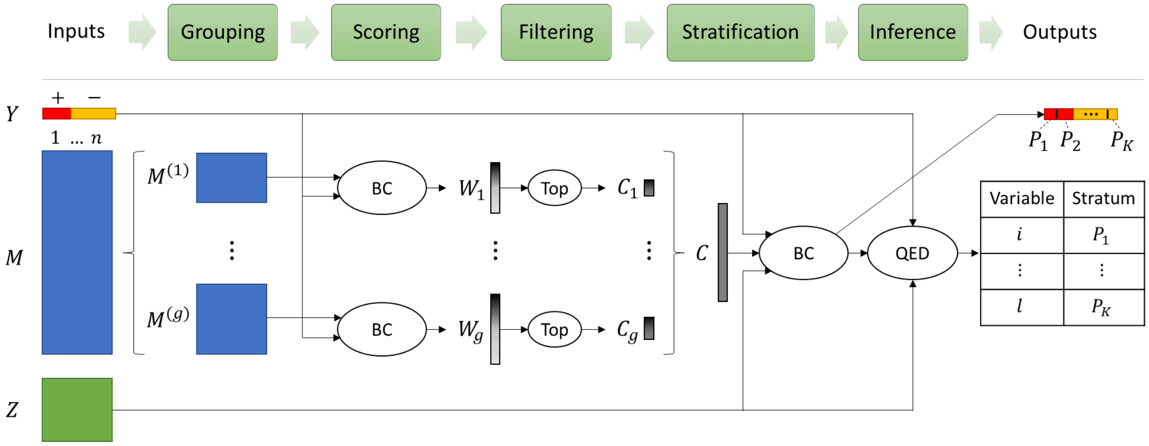


Figure 4.3: **Overview of ARISTOTLE.** The data are shown by rectangular blocks and the methods are shown by ovals. M_1 to M_g are submatrices of the Data matrix M, corresponding to variables in Grouping set G produced in the Grouping phase. W_1 to W_g are the variable scores produced by the biclustering algorithm during the Scoring phase. C_1 to C_g are the sets of the top scoring variables in each of the Groups, selected in the Filtering phase, and aggregated into candidate causes C. Causal Strata P_1 to P_K produced by biclustering of the Stratification phase. QED evaluates the candidate variables in the set C for causality with respect to their corresponding Causal Strata P and confounders Z in the Inference phase and outputs the pairs of Strata Causes and Causal Strata P.

Further details of each step are provided in the following sections.

4.2.1 Grouping

This phase simply groups variables into g possibly overlapping subsets $\{G_1, \dots G_g\}$, with $m_1, \dots$ and m_g variables respectively, either based on the Grouping set, if available, or by partitioning variables randomly. Each group should be small enough to be handled efficiently by biclustering, and at the same time, there should be enough variables for biclustering to be able to estimate the causal strata.

4.2.2 Scoring

This phase, for every Group, biclusters Data matrix of Group's variables, and assign scores to them based on their association $W_i \in \{0, 1\}^{m_i}, i \in \{1, \dots g\}$.

For biclustering and acquiring strata-based variable scores, ARISTOTLE uses SUBSTRA [90]. SUBSTRA is a state-of-the-art Probabilistic supervised biclustering method which takes a Data matrix and Outcome vector as input, and produces three related outputs: Strata of samples, clusters of variables, and variable scores.

SUBSTRA learns these outputs through an iterative approach that simultaneously optimizes two objectives: biclustering quality, and predictive performance. The probabilistic graphical model of SUBSTRA incorporates an information flow between the strata, variable clusters, and the outcome. The unique characteristic of SUBSTRA is that it uses a score of a variable to estimate its ability in identifying the strata that can discriminate between the two classes of the outcome.

In the first step of each iteration, in an approach analogous to leave-one-out cross-validation, the variable scores are modified such that the new assignment probability distribution of each sample over all strata results in a better predictive performance for that sample. In other words, the new scores increase the probability of assignment to strata with outcomes consistent with that of the sample itself. Then, new scores are used for clustering sample again before moving to the next sample. This improves the biclustering towards a better prediction. In the second step of each iteration, the variable clusters are updated given the updated sample strata such that the coherence of the variable values within each bicluster increases. This increases the biclustering quality.

Therefore, by optimizing the biclustering quality and prediction quality simultaneously, SUBSTRA identifies:

- Sample strata, such that samples from the same stratum have similar values in the high-scored variables and the outcome.
- Variable groups, such that the patterns of variables grouped together are different between the strata as well as between the positive and negative outcomes.

ARISTOTLE uses SUBSTRA for three reasons:

- SUBSTRA assumes similar outcomes for samples in each stratum. This is important for consistency with the notion of Stratum causes defined by the SCD problem.
- SUBSTRA learns the variable scores according to their relevance to the outcome, and uses these scores when computing the strata. Hence, the produced strata are related to the significant variables which are associated to specific strata of the outcome.
- The variable scores produced by SUBSTRA indicate the amount of dependency between the variables and the outcome, and can be used for filtering out irrelevant variables and narrowing down the set of candidate causal variables. This variable filtering helps reduce the number of effective hypotheses and avoid the multiple-hypothesis testing penalties, which is the first challenge in SCD. It should be noted that this is only possible because the patterns that cannot reach statistical significance, or are not tested against the outcome, do not need to be taken into account while correcting for multiple hypothesis testing [195, 151]. This assumption is tested in practice in the result section.

To the best of our knowledge, SUBSTRA is the only existing method with the above characteristics.

4.2.3 Filtering

With the variables' scores measured, in this phase, ARISTOTLE locally select the highest scoring variables from each Group, and aggregated them as the set of candidate variables $C = \cup_{i=1}^{g} C_i,\ C_i \subset G_i$. For this task, ARISTOTLE uses an outlier detection algorithm, which selects variables with scores outlier to the distribution of scores in the group. Since biclustering produces score of a variable only relative to the other variables in each group, the scores from two different variable groups are not comparable to each other. Hence, ARISTOTLE applies a feature selection to each group independently.

Specifically, ARISTOTLE uses the scaled Median Absolute Deviation (MAD) [101] to identify a boundary above which the scores are considered outliers. The MAD score for variables can be computed using the formula:

$$MAD_i = \frac{\mu(|W^{(i)} - \mu(W^{(i)})|)}{\sqrt{2} \cdot \text{erf}^{-1}(1/2)} \tag{4.1}$$

where $W^{(i)}$ is the vector of scores of variables from Group i, $\mu(\cdot)$ is the median of its input argument, and μerf is the Gaussian error function. ARISTOTLE decides whether score $W_j^{(i)}$ of variable j in group i is an outlier, if $W_j^{(i)} - \mu(W^{(i)}) > L \times MAD_i$. The parameter L essentially counts the number of standard deviations, and determines how extreme a score should be to be considered an outlier. In standard practice, the value of L is usually selected as an integer between 2 and 5.

There are two reasons for choosing MAD:

- MAD is based on a parsimonious and well-studied equation which requires only one parameter, and adds a minimum level of complexity to the process.
- MAD is extremely robust to heavy-tail distributions, which is common to SUBSTRA's scores.

Stratification

Causal Strata capture the subtypes of the positive outcome from the patterns between the Strata Causes and samples and the best estimation of Strata Causes are candidate variables C, and natural confounders Z, if available. Hence, this phase biclusters samples based on the Data matrix of Candidate variables, the outcome, and natural confounders, resulting in the Causal strata $P = \{P^{(1)}, \ldots P^{(k)}\}$. Once again, due to the aforementioned reasons, ARISTOTLE uses SUBSTRA for biclustering.

4.2.4 Inference

In this phase, ARISTOTLE evaluates causal relation between each candidate variables and each Causal stratum, while controlling for natural confounders and other candidates [125, 115]:

$$X_i \xrightarrow[Z \cup X_{-i}]{?} Y^{(j)} \;\; \forall X_i \in C \;\&\; P_j \in P : Y_k^{(j)} = \begin{cases} 1, & \text{if } k \in P_j \\ 0, & \text{else} \end{cases} \tag{4.2}$$

Where X_i is a candidate variable, $Y_k^{(j)}$ is sample k membership of Causal Stratum P_j, Z is the set of natural confounders, $X_{-i} = C - \{X_i\}$ is the set of candidates variables other than X_i.

The causal inference algorithm takes Data matrix M, candidate variables C, Causal Strata P, and natural confounders Z as input, and outputs pairs of Strata Cause and its corresponding Causal Stratum. It should be noted that each variable can be associated to more than one strata, and each stratum can have more than one causal variable.

For this task ARISTOTLE uses a Matched pairs QED. The matching used in ARISTOTLE is similar to that of HUME with two key differences:

- **Controlling Confounders:** Natural Confounders Z must have identical values in each treatment-control pair. Sum of the differences of other candidate variables of treatment-control pairs should be minimum. For the optimal pairing, ARISTOTLE employs the Hungarian matching algorithm [94] based on the Manhattan distance, differences in the Natural confounders are multiplied by a large value to act as a hard constraint.

- **Multiple Comparison Correction:** Like most methods for controlling the FDR, ARISTOTLE requires the number of true null hypotheses [80]. In ARISTOTLE however, due to the supervised elimination of some of the potential hypothesis during the Filtering phase, and the significant dependency between hypothesis that share candidate causes or stratum, the equivalent number of hypotheses is hard to identify and lies in the wide range from the number of candidates to the total number of variables times the number of strata. Hence, a method for estimating the number of true null hypotheses is used in ARISTOTLE. A well-known procedure for estimating of the number of true null hypotheses is the adaptive Benjamini-Hochberg [21]. Adaptive Benjamini-Hochberg works based on the graphical interpretation of the $q - q$ plot of p-values, which results in a simple stepwise procedure for estimating the number of true null hypotheses.

In summary, in the Inference phase: (1) for each candidate, samples are match paired based on their differences with respect to the candidate variable, confounders, and other candidate variables using the Hungarian algorithm, (2) p-value of each matched pairs hypothesis is calculated using McNemar's test with Yates' correction, and (3) the p-values of candidate variables are analyzed using the adaptive Benjamini-Hochberg method and a subset of them are reported as the causal variables.

4.3 Experimental Evaluation

4.3.1 Experiments on Synthetic Data

Synthetic Stratified Dataset

To properly evaluate the performance of ARISTOTLE under different conditions and with known ground truth, we created the synthetic stratified datasets. In order to faithfully recreate the type of biclusters that exist in real-world applications [104, 48, 160, 39], the following assumptions are made for the generation of the synthetic stratified data:

- variables form clusters and whose members have similar values across samples. The sizes of clusters follow a Binomial distribution. Each cluster is present in a subset of samples, i.e. the variables of the cluster have a value of 1 for those samples but 0 for the others.

- Some of the clusters are causal, meaning that all of the variables in those clusters are strata causes. The remaining clusters are non-causal, meaning that none of their members is causal. The members of causal clusters have value 1 only for a subset of samples with the positive outcome. There is a one-to-one relationship between the causal variable clusters and the strata of samples with positive outcome.

Table 4.1: Parameters of synthetic data generation and their values. The default values are underlined.

Parameter	Value
Number of variables (m)	100000
Number of clusters (f)	2500
Number of groups (g)	100
Variable noise (q_x)	0.05
Fraction of positive samples (p)	0.2
Number of causal variable clusters (g)	2, $\underline{3}$, 4, 5
Outcome noise (q)	0, 0.05, $\underline{0.1}$, 0.2
Number of positive samples (n_+)	75, $\underline{100}$, 125, 150

- On top of the variable clusters, there is another layer of grouping of variables which represents the Groups. These groups are mixtures of different clusters; however, they tend to contain variables from similar clusters, represented by low entropy in clusters values of groups. To achieve this, a process based on the concept of 'rich gets richer' is used, which is introduced in the Barabasi-Albert algorithm [5], originally designed for network simulation. In this process, variables are randomly selected to be assigned to groups, and a variable is assigned to a group with a probability proportional to the fraction of the current variables in the group that belong to the cluster containing the variable.

- Both variables and outcomes contain noise. Noise is added to the variables and outcomes by flipping a randomly selected portion of the entries.

The parameters of synthetic stratified data generation are shown in table 4.1. All parameters are set to be in the same order of magnitude as the pharmacogenomic data and the typical values of omics datasets [134, 104, 24], which also matches with the marginal distributions in [180]. The effect of the last three parameters in this table are evaluated, which are the most significant in HCD, and are expected to have the most significant impact on the performance, namely Number of clusters, Outcome noise, and Number of positive samples. The effect of varying each parameter is investigated by fixing the other two at their default values, shown in the table, and apply the three methods ARISTOTLE, Fuzzy, and RFCI on 100 generated datasets. It should be noted that similar clustered data-generation processes in the literature could not be used [180], because they could not produce the biclustering joint distribution.

Baseline Methods

Two other baseline methods are included for the evaluation purposes:

- **RFCI:** The state of the art general CD method that balance of accuracy and computational complexity, to compare the quality of discovered causes. However, RFCI is still computationally intensive. Therefore, the same divide and conquer approach as in ARISTOTLE was employed for RFCI. Specifically, RFCI is applied to each group separately and then applied again to the union of causal variables selected for different pathways by RFCI. This produces the final set of variables predicted to be causal for the variable.

- **Fuzzy:** A well known supervised clustering method introduced in [2], to compare the quality of strata. Similar to SUBSTRA, this method incorporates supervision into the clustering by producing classification scores for variables, and uses the class label of each point to identify the optimal set of clusters that describe the data, and the obtained clusters are then used to build a fuzzy classifier based on relevant variables identified using Fisher-interclass separability method [44]. All variables are given to this method as input (i.e. no divide and conquer).

Results of Experiments

The results of tests on synthetic datasets are shown in figure 4.4.

Causal Strata: First, accuracy of ARISTOTLE in finding the true causal strata is evaluated, and compared to that of the Fuzzy method. This is measured using the Rand index [78], a well-known method for measuring the performance of a clustering algorithm based on an external gold standard. These results show that the Rand index of ARISTOTLE's strata with respect to the true strata across the different experimental settings is consistently above 0.5 and reaches to 0.8 in most cases.

Somewhat unexpectedly, ARISTOTLE tends to be more successful in scenarios with a larger number of clusters. This is because ARISTOTLE tends to decompose the true strata into subsets, due to the effect of non-causal variables, resulting in smaller Rand index for cases with fewer clusters. As expected, the Rand index decreases with increasing noise, but does so slowly. Most importantly, the Rand index varies little with respect to the number of positive samples, which supports the claim that ARISTOTLE can deal with the first challenge.

Compared to the Fuzzy method, ARISTOTLE achieves a consistently and substantially larger Rand index in all experimental settings. Even though both methods use supervision, this indicates a better incorporation of the supervision information in ARISTOTLE through an effective variable scoring.

Strata Causes: Second, ARISTOTLE performance in finding the causal variables is evaluated, and compare it to RFCI. Since the number of true Strata causes is fixed across the experiments, Precision and Recall of the causes are good indicatives of methods' performance. Overall, Recalls are the sensitive to experiment setting, which is indicative of the

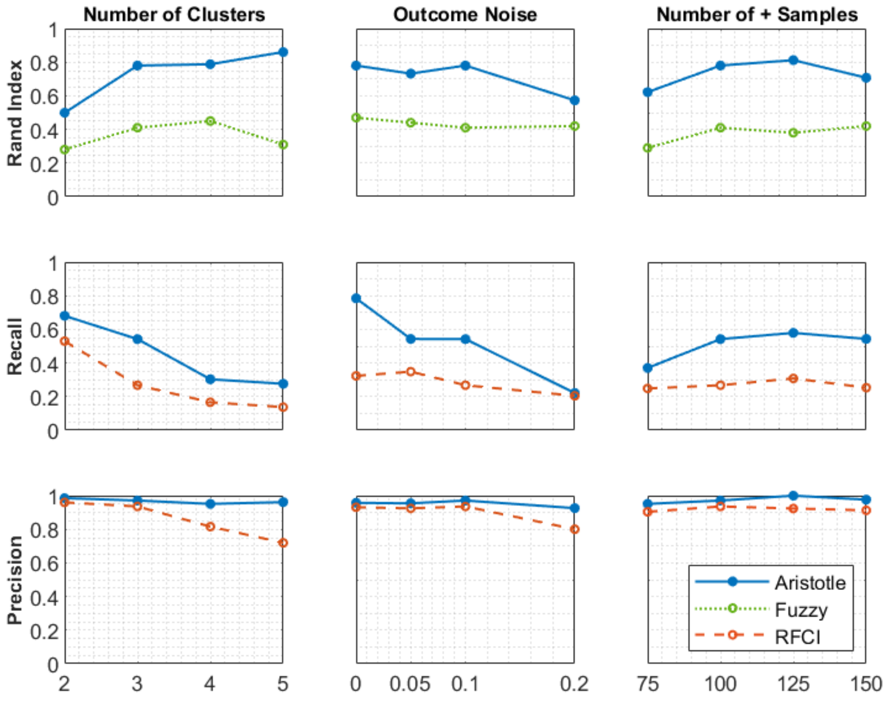


Figure 4.4: **Results of Synthetic Experiments**. Rand Index, Recall and Precision of ARISTOTLE, RFCI, and Fuzzy are measured across 100 experiments with different value of number of clusters, outcome noise, and number of positive samples. There is no value for Rand Index of RFCI since it does not discover strata, and there is no value for Precision and Recall of Fuzzy since it is designed for clustering not CD.

reliance of CD methods on various assumption, and Precisions are much higher and more consistent, which is a common phenomenon for CD methods, due to their conservative nature in compare to predictive models.

ARISTOTLE's baseline Recalls are moderate at above 60%, but decreases significantly with the increase in noise and number of causal clusters. This may be due to the reduction in size of the positive strata, which results in a weaker signal for each causal variable and consequently lower statistical power. For the same reason, increasing the number of positive samples improves the Recall, although to a more moderate degree. ARISTOTLE's Recall consistently outperforms RFCI's, which demonstrates the advantage of SCD over the classical CD that discover causes in the whole population. The gap between the two methods decreases with increasing outcome noise, and both methods perform equally poorly (Recall ≈ 0.2) for 20% outcome noise. However, the difference between ARISTOTLE and RFCI becomes more pronounced for higher number of positive samples. This is because RFCI substituting the FCI's possible separation sets with adjacency sets (to speed up the independency checks) compromises the completeness of the causal graph, and hinders it from achieving higher accuracy.

ARISTOTLE's Precision remains at a high level, around 95%, consistent with the premise of an FDR of 5%, and is always slightly better than RFCI in all the experimental settings tested. Although RFCI achieves a similar Precision for smaller numbers of strata and levels of noise, the advantage of ARISTOTLE increases for the harder problems with more strata and higher noise levels. To conclude, the experimental results on synthetic data show that ARISTOTLE SCD consistently outperforms the well-known methods with classical CD approach.

SCD: Lastly, the role SCD in ARISTOTLE's accuracy is evaluated. To test this, the performance of ARISTOTLE under the default parameters is compared with two alternative settings applied to the candidate variables:

- **No Strata:** Inference without considering the strata, i.e. QED for all positive samples in one stratum.
- **Perfect Strata:** Inference using the ground-truth positive strata as the outcomes.

ARISTOTLE Inference achieves a Precision of 0.93 and a Recall of 0.62. No Strata Inference has a Precision of 0.95 and a Recall of 0.29. The greatly reduced Recall is due to failure of No Strata Inference to detect the causal variables with small stratum. This is because the concentration of the occurrence of causal variable in a small subset of positive samples constituting a stratum. When matching in the QED, this small difference fails to separate the positive strata and matching positive samples reduces the support for the hypothesis. This further supports the claim that stratification of the samples can find the causes which would have been otherwise missed. The reason for the higher Precision of the

ARISTOTLE's QED is that it results in larger p-values in general and, therefore, is more conservative.

The Perfect Strata Inference has a Precision of 1.00 and a Recall of 0.69. This indicates that ARISTOTLE has a very similar performance to the method with the true knowledge of causal strata. This experiment provides evidence that ARISTOTLE can deal with the second challenge effectively. It should be noted that Recall is more important than Precision in the CD setting, because the computational predictions are often later examined experimentally in order to rule out any false discoveries. However, missing an important true causal variable may be more detrimental from a scientific point of view.

4.3.2 Experiments on Real-world Data

As discussed earlier in the introduction, Anthracycline Cardiotoxicity is now believed to be a multi-factorial ADR with multiple underlying mechanisms including the inhibition of Topoisomerase 2β and the action of Reactive oxidation species [120, 63]. Hence, discovering the genetic risk biomarkers of pediatric Anthracycline Cardiotoxicity is a SCD problem.

From the pharmacogenomic application's dataset introduced in the chapter 1.2, we use anthracycline cardiotoxicity as the Outcome vector Y, patient SNPs as the Data matrix M, clinical factors as Natural confounders Z. Because biological pathways are the functional units inside the cell, using them to cluster omics data is commonplace [8, 203]. Therefore, in the experiments, Groups are defined according to SNPs memberships in pathways [88]. Specifically, in the Grouping phase, we mapped the SNPs to genes using the tool introduced in [123]. Then, based on the genes in the Encyclopedia of Genes and Genomes pathways, 1.2 million SNPs were associated with pathways to form the 323 final Groups. The SNPs that did not correspond to any gene belonging to one of the pathways were discarded.

The results of applying ARISTOTLE to the pharmacogenomic application can be evaluated from three standpoints:

- **Results Profile:** By statistical analysis of distribution of the p-values of the discoveries.
- **Literature Review:** By comparing the overlap between ARISTOTLE's discoveries and the known causes, which were deduced from independent medical records and the literature.
- **Systemic Interpretation:** By providing biological interpretation for corresponding genes and pathways of the discovered SNPs.

Results Profile: Figure 4.5 shows the logarithmic $q-q$ plot of the distribution of p-values of the candidate variables computed by ARISTOTLE for the pharmacogenomic dataset, where candidate are sorted in ascending order of p-values. The straight lines indicate what would happen under the null hypothesis for different numbers of hypothesis [80], i.e.

if candidate causes and outcomes were independent and randomly selected. Interestingly, adaptive estimation of the number of true null hypotheses (the purple line) results in almost the same number of hypothesis as the number of candidate variables after filtering phase (the yellow line), and the corresponding lines align. This indicates that the number of the variables passing the filtering of ARISTOTLE is a reasonable estimate of the number of true null hypothesis. The order of magnitude of ARISTOTLE's p-values is similar to those reported in the Guideline [7] and HUME [115] and the distribution of the p-values significantly deviates away from the straight line, i.e. null distribution, especially in its heavy tail which correspond to the significant relations. This indicates that significance of ARISTOTLE's discoveries are comparable to those reported by HUME and Guideline.

Literature Review: We investigate the overlap of ARISTOTLE SCD discoveries with the SNPs discussed in the Guideline, introduced in the chapter 1.2.1, and HUME [115] as estimations of the ground truth knownledge of causal SNPs. 28 candidate variables passed the statistical test at the FDR significance level of $\alpha = 5\%$ and were selected by ARISTOTLE as the SCD causes.

Two of the Guideline's three strong SNPs, namely $rs2229774$ and $rs17863783$, passed the test with p-values of $5.4E-05$ and $8.9E-05$ respectively. $rs2229774$ was also detected by HUME and its validity is carefully studied in [7]. But more importantly, $rs17863783$ was not detected by HUME. This shows the advantage of ARISTOTLE in finding those factors that are causal for one archetype of cases, but would be missed if all cases are counted the same. Looking more closely, $rs17863783$ has only a very strong association with Anthracycline Cardiotoxicity for one of the strata, but does not have enough prevalence in the other strata to be detected by existing causal analysis methods. The Guideline's third strong SNP, $rs7853758$, was missed due to not being included in any of the pathways. However, if it had been tested under the current QED, it would have resulted in a p-value of 0.03, which seems significant in isolation, but would probably not have survived multiple hypothesis testing. It should be noted that because the variable groups and consequently the final results are susceptible to change by inclusion of new pathways and significant SNPs, such posterior evaluations are not valid.

Furthermore, three of the Guideline's fourteen significant SNPs passed the test, namely $rs17583889$, $rs10426377$ and $rs4673$, with p-values of $1.0E-05$, $5.4E-05$ and $1.0E-04$ respectively. Of the remaining eleven significant SNPs, three had fewer cases than the minimum prevalence threshold, four did not correspond to any of the pathways, two did not pass the filtering process, and the two remaining ones did not have sufficiently low p-value to pass the multiple hypothesis testing.

It should be noted that not reporting all of the significant Guideline SNPs is not necessarily an undesired outcome. First, as mentioned earlier, significant relations are not the ground truth, and some were even considered insignificant by the Guideline and HUME. Second, and more interestingly, there can be associations that are considered significant

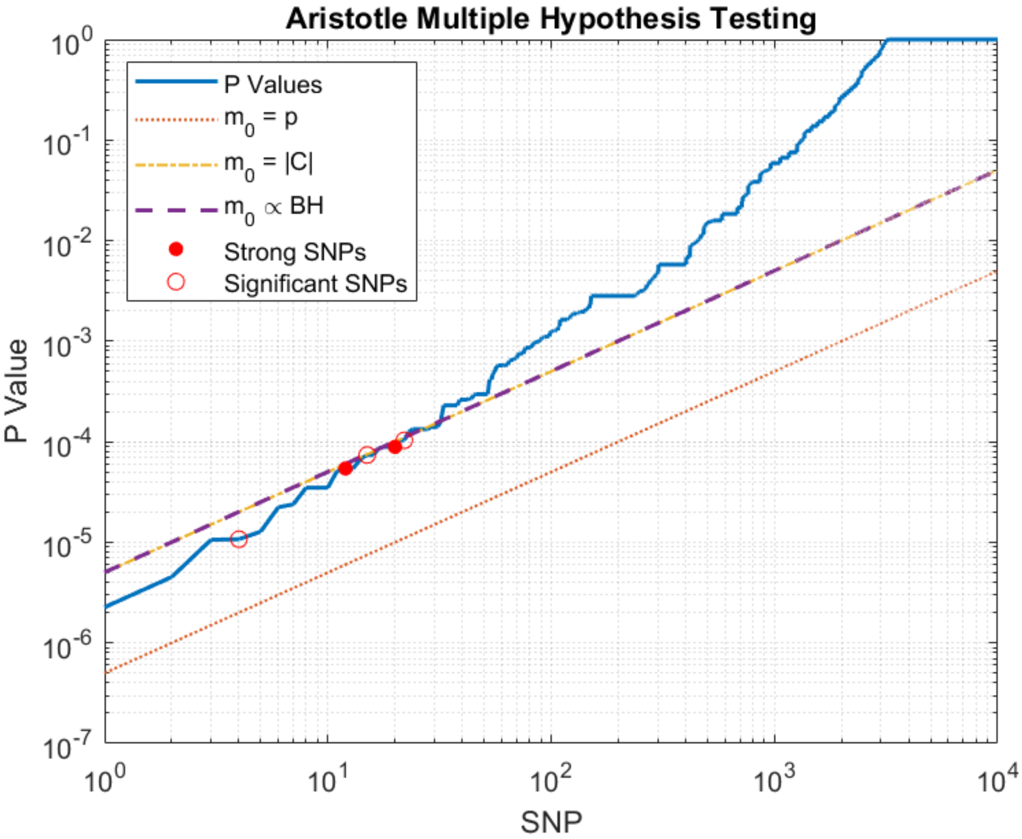


Figure 4.5: **Results of Real-world Experiments**. *q*-*q* plot of *p*-values of candidate variables. The blue curve is the *p*-values of candidate variables, calculated by McNemar's test. The three linear curves show the significance level adjusted by FDR of $\alpha = 0.05$, each based on different assumptions about the number of true null hypotheses. The yellow line assumes that the number of hypotheses is equal to the total number of variables, which gives the most conservative possible adjustment. The orange curve assumes that the number of hypotheses is equal to the number of candidates produced by variable Filtering and used for the statistical test, which gives the least conservative adjustment. The purple line estimates the number of hypothesis according to [21], a method for the estimation of the effective number of hypotheses, which is almost perfectly aligned to the less conservative approach.

for the overall population, but lack a sufficient statistical power or association when the hypothesis is focused on specific stratum [85].

Systemic Interpretation: From a biological standpoint, a significant number of ARISTOTLE's discoveries share the same corresponding genes and pathways. This not only provides evidence for the functional involvement of those genes and pathways in Anthracycline Cardiotoxicity, but also provides further evidence for the validity of ARISTOTLE's results.

12 of the 28 discovered SNPs share a gene with at least one other SNP. Five of these SNPs, namely $rs795887$, $rs6436364$, $rs6756107$, $rs6722420$, and $rs10755042$, are all from the gene $ACSL3$, which is involved in two of the pathways, 1212-Fatty Acid Metabolism and 4146-Peroxisome. Similarly, $rs496179$ and $rs885622$ are both from the $DPYD$ gene, involved in pathway 410-Beta-Alanine Metabolism. SNPs $rs17863783$ and $rs10426377$, which are a strong and a significant SNP from the Guideline, respectively, are both from the $SLC28A3$ gene in pathway 140-Steroid Hormone Biosynthesis.

The association among SNPs is also present at the pathway level. SNPs $rs26848$ and $rs26849$, both in the PGP gene, share three pathways with $rs545253$ in the $MTND4P31$ gene. Similarly, SNPs $rs16972837$ and $rs659517$, both from gene $RYR3$, and SNP $rs607483$, are involved in the same pathways 4371-Apelin Signaling Pathway, 4713-Circadian Entrainment and 5010-Alzheimer's Disease. Moreover, the two $RYR3$ SNPs share pathway 4020-Calcium Signaling Pathway with $rs11869821$. Another example is the SNPS in the $UGT2B7$ gene, $rs7662632$ and $rs4356975$, which have the same pathways 40-Pentose & Glucuronate Interconversions, 53-Ascorbate & Aldarate Metabolism and 830-Retinol Metabolism, as the Guideline's strong SNP, $rs17863783$. Interestingly, $rs11869821$, $rs2271235$, $rs11936348$ and $rs611954$ are each from a different gene, but are from the same pathway.

4.4 Discussion

Evaluation based on synthetic data shows ARISTOTLE SCD is capable of outperforming both CD and biclustering competitors under different settings. Experiments on a real dataset on the pharmacogenomic application indicates that SCD platform can identify new types of causes that were not previously possible by the previous methods. Moreover, ARISTOTLE makes additional predictions that suggest further investigations.

To be scalable to the HCD, ARISTOTLE uses background knowledge for decomposing the high dimensional data into smaller subsets that are easier to handle.

However, ARISTOTLE not only suffers from the causal sufficiency assumption, but it also performs worse than similar methods in terms of validity and power when the causal mechanism is universal.

Another possible limitation of ARISTOTLE is that it consists of five main phases, more than most multi-phase CD methods. This can result in the propagation of error from one phase to another, and suboptimal and ad-hoc solutions. Therefore, a direction for future work would be to reduce the number of phases by providing a method that finds the causal variables and the corresponding strata in an integrated phase.

Finally, and arguably most importantly, although the causal inference evaluates the Stratified Causes, Causal Strata never receive an analysis beyond association. Hence despite ARISTOTLE's accuracy in discovering the true Causal Strata in synthetic and real-world settings, a theoretical guarantee for validity of ARISTOTLE's Strata in general, does not exists.

Chapter 5

Causal Profile Discovery

In many CD problems, such as in biological and medical domains, the patterns of changes of variables over time are significant [107]. For instance, in a pharmacogenomic application, distinguishing the samples exposed to a high level of treatment in a short time from those with continuous moderate treatment, specifying the minimum duration of treatment to be effective, narrowing down the treatment-effect persistence time, finding the most effective frequency of episodic interventions, and identifying and compensating for the impact delay between treatments and outcomes can be invaluable. There are also numerous applications for such CD with temporal dimension, including bioinformatics [107], neuroscience [32], climate studies [122], and economics [117], to name a few.

Temporal CD emerged to identify causal relations between time series, despite challenges such as autocorrelation, impact delay, latent confounding effect, under-sampling, episodic interventions, and computational complexity [112].

Granger causality is one of the pioneering techniques for evaluating and quantifying causal relations between time series based on competence prediction via a Vector Autoregression [66]. Various extensions of the Granger causality have been proposed in the literature. For instance, the LASSO Granger method utilizes the L_1-regularization [12] to learn causal relations in the presence of sparse data. To deal with hidden confounders, sparse plus low-rank networks add an extra layer to explicitly model hidden variables [209]. Gaussian mixture expectation-maximization has been applied to identify Granger causality in the presence of subsampled data[65]. Several other methods are proposed to address nonlinear and higher-order causality, among which are copula-based methods [77] that dissociate the marginal distributions from their joint density distribution to only focus on statistical dependence between variables, but they face causal interpretability challenges. The major limitation of the Granger causality is its reliance on a weaker notion of causality, since the variables that contribute most to predicting the outcome are not necessarily causal.

Another line of research is the adaptation of SL methods for time series. The temporal precedence constraint, i.e., the constraint that the cause should precede the effect, has helped with reducing the search space of structural learning [77, 184]. The major drawback of SL is the computational complexity that grows exponentially with the number of variables, which limits its applicability to HCD.

On the basis of cause-effect asymmetry, Kernel Conditional Deviance for causal inference is applied in [127]. In [173], some of *structural equation models* such as cross-lagged panel models [118], linear panel models [89], and autoregressive cross-lagged models are used for CD in longitudinal data. Moreover, PCMCI and tsFCI, as the extensions of the PC and FCI algorithms respectively, are designed for a temporal domain[167, 53].

Hidden Markov random field regression [106] exploits the variables' covariance to learn their causal relations, but it is usually limited in its assumptions about latent variables. Some variations of this approach are proposed to do multi-task learning and also to deal with dynamic temporal graphs where the graph structure is variant over time [107, 106].

By the same token, some other techniques have targeted specific problems in this context, like convergent cross-mapping to infer causality from noisy time series data [128], Bayesian to handle highly correlated and noisy time series space [136], and decomposed slab-and-spike to learn the causal relations as well as the arbitrary lag among different time series simultaneously from data [179].

Nowadays, in the matter of outstanding progress of *deep learning* methods in discovering hidden patterns in various contexts, some studies have applied deep learning to temporal CD. For instance, to solve individual-level causal effects and estimate unknown latent space summarizing the confounders, Variational Autoencoders is applied in [108]. Also, an attention-based convolutional neural network is applied in [133] to perform temporal CD, as well as predicting one-time series based on the others.

Despite the progress of discussed CD methods, several challenges still remain unresolved. The existing methods, share the common limitation of being only concerned with causality in a restricted sense, as they show solely whether a variable can change the outcome or not, without indicating the causal temporal pattern, which limits the interpretability and applicability of the results.

Furthermore, to the best of our knowledge, none of the mentioned methods are able to take full advantage of the historical information [64] and generally take one of the following approaches; (1) treating the observation fully independent of each other over the span of time, (2) analyzing extracted fragments out of data, i.e. time windows, or (3) estimating the range of lagged effects and handling all measurements separated by no more than the ones in the lag window as a disjoint unit.

For instance, in the pharmacogenomic application, one might want to know the minimum duration of treatment to be effective, treatment-effect persistence, or the time delay between a treatment and its outcome. While some of the methods are flexible in terms of defining the hypothesis [32, 179, 23], they still place the burden of finding the true hypothesis on the user, which might result in a wrong conclusion or data dredging.

Table 5.1: Running example dataset. For the candidate variable x_1, offset $[2\ 0]$, and pattern $[\ 0\ \ 1\ ]$, based on values of green cells, instance $(1,3)$ belongs to treatment and $(2,4)$ to controls. These two instances can be matched based on their corresponding values in blue cells, and the causal effect is assessed by the difference of their yellow cells.

	x_1				x_2				z				y			
id	j_1	j_2	j_3	j_4	j_1	j_2	j_3	j_4	j_1	j_2	j_3	j_4	j_1	j_2	j_3	j_4
i_1	0	0	1		.9	.4	.6		.5	.2	.1		.1	.2	.8	
i_2		0	1	0		.9	.4	.6		.5	.2	.1		.1	.2	.1

This problem can be solved if the time series for the cause to have the most significant impact on the outcome, is also discovered. We refer to this problem as Causal Profile Discovery (CPD).

For solving CPD, We propose *HEIDEGGER*. HEIDEGGER is a multi-phase discrete-time framework that exploits the efficiency of graph-based search of the hypothesis space and the flexibility of QED. HEIDEGGER goes beyond a temporal CD technique, and:

- provides highly interpretable causal relations.
- detects complex temporal treatment-outcome relations.
- is robust with respect to outcome noise.

5.1 Problem Description

CPD's input consists of observations for samples at discrete time points of the outcome y, candidate cause variables $X = \{x_1, \ldots x_m\}$, Natural Confounders $z = z_1, \ldots z_C$:

- A necessary sample-time Outcome matrix $Y \in \mathbb{R}^{n \times t}$, where Y_i^j is the value of the outcome for sample i at time j.
- A necessary sample-variable-time 3-dimensional Data matrix $M \in \mathbb{R}^{n \times m \times t}$, where $M_{i,k}^j$ is the value of the k^{th} candidate variable X_k for sample i at time j.
- An optional sample-variable-time 3-dimensional Natural Confounders matrix $Z \in \mathbb{R}^{n \times C \times t}$, with a similar structure to Data matrix.

The reason for extending the data to non-binary values is that multiple levels are usually necessary to represent the dynamic of temporal data, which is also reflected in the majority of datasets using the discrete-time or digital format.

Table 5.1 presents a running example with two candidate variables, one natural confounder, and an outcome, recorded for a period of 4 time points for two samples.

Output to the CPD is:

- The set of pairs of Causal Variables X^p and Causal Profiles CP.
- The vector of causal statistical significance p^*, where p_i^* corresponds to adjusted p-value of i^{th} causal variable and profile pair.

The goal is to identify the true causal variables associated with the outcome X^*, and their corresponding causal profiles CP^*, such that elements in X^* and patterns in CP^* provide the most significant causal association with the outcome when controlling for the effect of confounders and the other candidate variables.

Each element of CP, corresponding to a candidate variable, represents a hypothesis, where the treatment group receives the treatment in the form of the identified causal profile. We define a causal profile, $cp \in CP$, based on the two following components:

- *Treatment Pattern:* A binary vector $cp_p \in \{0, 1\}^L$, where "1" means receiving the treatment while "0" means not receiving the treatment.
- *Pattern Offset:* A vector integer of time offsets $cp_o \in \mathbb{Z}^L$, where each offset indicates the time offsets before the current time the corresponding element of treatment pattern is applied.

The experiment is considered a binary treatment, with one treatment group, and one control group defined as the *no-treatment* version of the causal profile. That is, the equivalent profile for the control group has the same *Pattern Offset* as the causal profile, but its *Treatment Pattern* is a zero vector. For example, the hypothesis that "if x_i exist at time j and does not exist at time $j-2$ then y at time j changes", corresponds to causal profile (cp_p : [0 1], cp_o : [2 0]), and control profile (cp_p : [0 0], cp_o : [2 0]).

Furthermore, the treatment group and control group consist of instances that at *Pattern Offset* timepoints in the past, match the *Treatment Pattern* values and no-treatment values respectively. In the example above, the instance (i, j) (sample x_i at time j) belongs to treatment group if the value of x_i at time $j-2$ is 0 and at time j is 1, and belongs to the control group if the at time $j-2$ and j is both 0.

The reason for this coding is that the indifference of the treatment value at a certain time-point is distinct from its value. Finally, notation efficiency, an *instance*, referred to as (i, j), is defined as a pair of sample i and time point j.

5.2 HEIDEGGER Method

To solve CPD, HEIDEGGER uses analysis at two levels. At the macro-level HEIDEGGER generates the local search neighborhood in the graph of profiles on the fly and iteratively prunes hopeless profiles to efficiently identify all the promising causal profiles for each candidate cause. At the micro level HEIDEGGER evaluates the significance of each requested

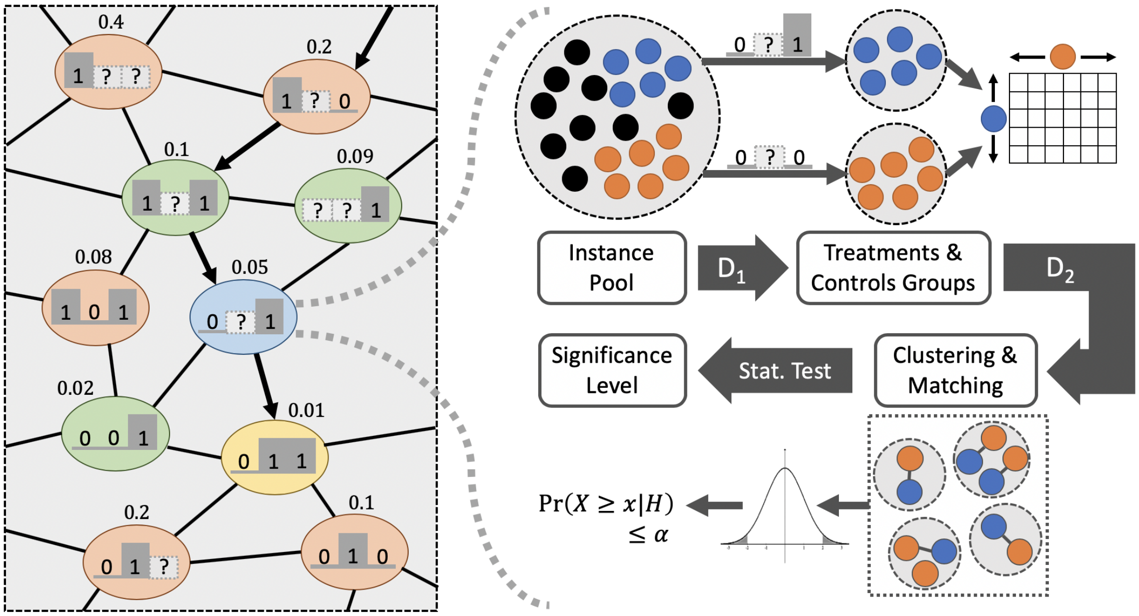


Figure 5.1: **HEIDEGGER Framework.** in each step of the graph search, the most statistically significant of neighbors (green) of the current node (blue) is selected as the current node of the next step. If the current node is more significant than its neighbors (yellow), the run stops. The string shown on each node symbolizes a causal profile, where ? indicates that value the of candidate variable at that time is arbitrary $\in [\,0\ 1\,]$. For each node, to compute the significance level, treatment and control groups are selected from the set of instances, matched via clustering, before computing their test statistic.

causal profile by applying a novel randomized block QED with a flexible matching function (Figure 5.1).

Specifically, for each candidate variable, the hypothesis search space is modeled as a graph, referred to as the Profile Graph, where each node indicates a potential causal profile. Thus, for each candidate variable at a time, HEIDEGGER runs a heuristic search on the profile graph to spot the most significant profile, while it dynamically generates the local neighborhoods in the search space and upon visiting each node, the corresponding causal profile is evaluated using a randomized block QED. Randomized block QED used in HEIDEGGER consists of three stages:

- Identifying the treatment and control groups by comparing the values of the candidate variable of each instance with the causal profile and the no-treatment (control) profile, respectively.

- Clustering the union of treatment and control group instances into blocks, where all instances in a block are similar enough, and randomly matching treatment and control instances within each block.

- Performing a paired statistical test to evaluate whether there exists a statistically significant difference between the outcome of the two paired groups.

It should be noted that the significance level of the candidate variables and their causal profiles are adjusted for multiple hypothesis testing based on the FDR criteria. The detailed description of each phase is presented in the following sections.

5.2.1 Graph Traversal

As aforementioned, the backbone of HEIDEGGER is the traversal of the profile graphs to get to the most significant candidate profile. The algorithm 5.3 describes the main phases of traversing the profile graph. Each node represents a causal profile, and there is an edge between nodes A and B if the causal profile of node A can be converted to the causal profile of node B with the change of a single element. Given this, the shortest path distance between two nodes in the profile graph is equal to their minimum edit distance. The discovery process in Graph is guided by the assumption that the p-values across the network are cohesive, i.e., hypotheses with similar profiles are expected to have similar p-values. HEIDEGGER uses a simple but efficient strategy to build the search space; at every step, it expands the neighborhood of the current hypothesis, and finds the most promising neighboring candidate to be processed by the QED (explained in Section 5.2.2).

To speed up the search, an early pruning mechanism is used on the basis of discarding the causal profiles with permutation entropy above a given threshold. This is because in practice, the causal profiles are expected to have low permutation entropy. For instance, a high entropy causal profile (cp) such as (cp_p : [1 0 1 1], cp_o : [8 5 3 1]) is unreasonable,

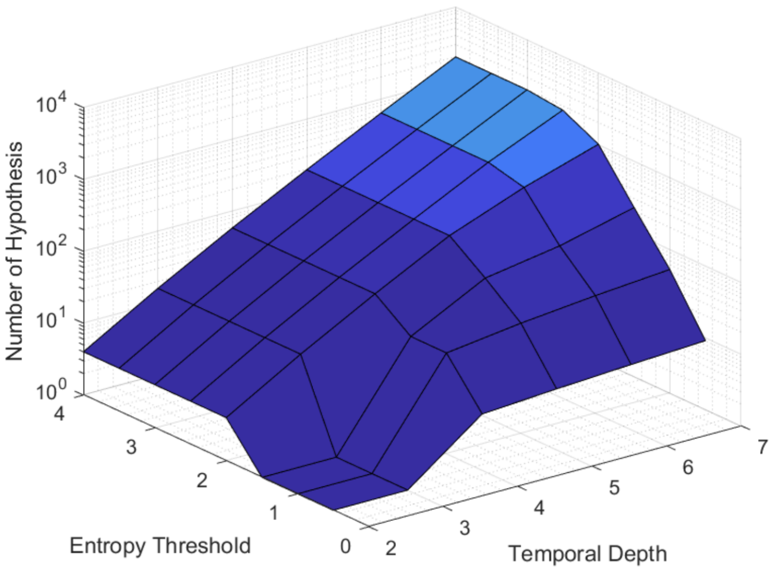


Figure 5.2: **HEIDEGGER's Complexity.** The number of profiles left after pruning with permutation entropy with threshold (θ) for different profile lengths

but a causal profile with a lower permutation entropy such as (cp_p : [0 1] , cp_o : [2 0]) is acceptable.

Graph Traversal

Starting from a random starting node, at each step, the neighboring node that has the most significant p-value is chosen. The search terminates if none of the neighbors of the current node has a better p-value. Having one initialization point is very risky and might lead the search to a non-significant local minimum if the starting point belongs to an isolated island in the search space. Therefore, HEIDEGGER repeats its search process τ times, using different starting nodes, in order to increase the chance of finding the best profile. However, to eliminate redundancy, it records the search history and stops the search process as soon as the current node has been visited in the previous iterations. This early termination is based on the fact that the search path followed from a given hypothesis is deterministic and always leads to the previously discovered local solution. The experimental results confirm that the multi-iteration searching process does not significantly add to the number of visited nodes.

Early Pruning

To avoid applying the expensive QED on all neighboring nodes, HEIDEGGER exploits *Permutation Entropy (PE)* to prune unpromising hypotheses. PE is a robust and non-parametric measure of the complexity of time series that works based on capturing the change of patterns within sub-sequences via entropy of the prevalence of those patterns [74].

Because the value of the time points of causal profile for which we calculate the complexity is limited to only three possible states (0, 1, and ignored), we use a modified version of the PE:

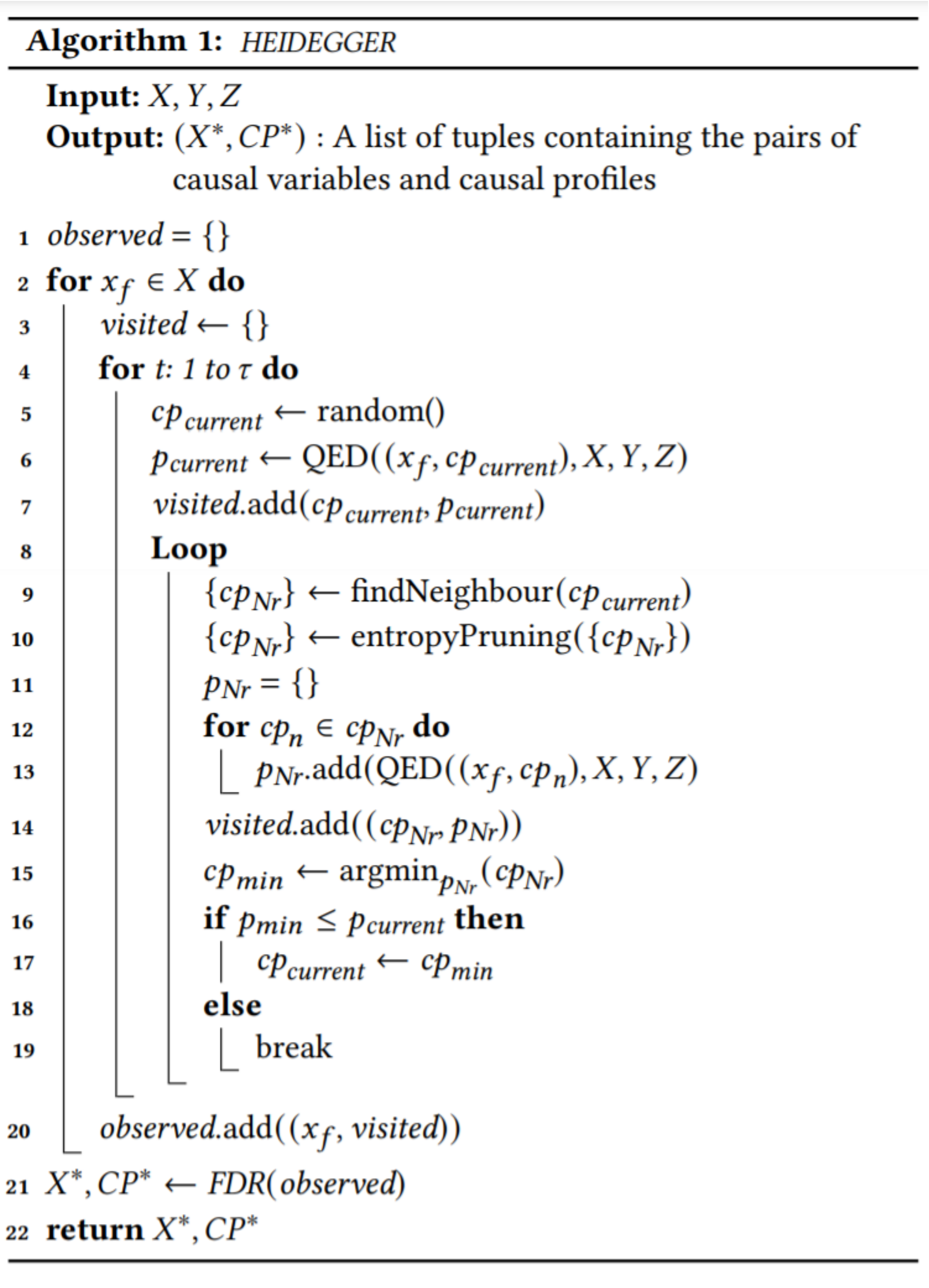
$$PE_{\delta}(\text{cp}) = -d_i \log(d_i) \tag{5.1}$$

Where d_i is the relative frequency of the i_{th} unique subsequence of length δ in the causal profile. HEIDEGGER removes the hypotheses with PE larger than θ, for an adjustable parameter $\theta > 0$. Larger values of θ should be considered when the user is looking for complex patterns. While smaller values of θ significantly reduce the multiple hypotheses penalty, the resulting causal profile is unlikely to be influenced. The reason is that most of the desirable profiles in the real-world follow parsimonious patterns [132]. We note that the size of the complete profile graph is exponential with respect to the maximum length of the causal profiles, which would limit the applicability of the method in terms of the multiple hypothesis penalty and processing time. The explained pruning mechanism reduces the search space to a sub-exponential space, as shown in Figure 5.2. To give insight on the level of pruning, pruning with $\theta = 2$ shrinks the number of hypotheses with the length of 8 or less from 10^8 to almost 10^3. It worth noting that the explained pruning mechanism is independent of the data, and can be performed consistently.

5.2.2 Randomized-Block Design

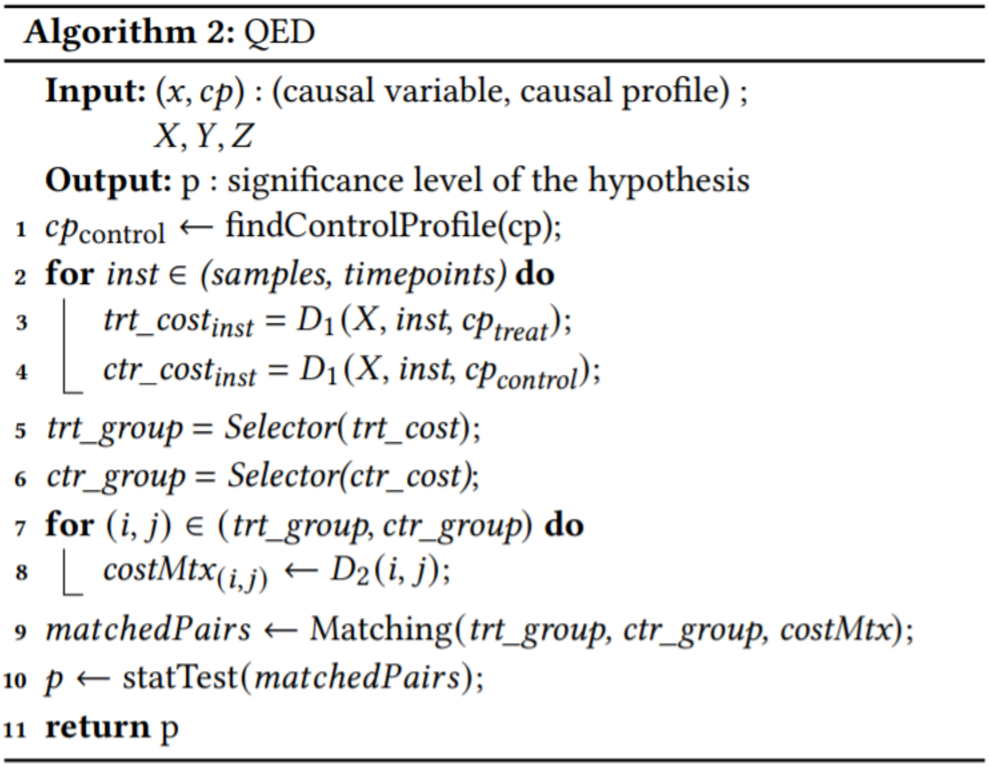
HEIDEGGER needs to evaluate the significance of the adjacent hypotheses to find the most promising profile in the local neighborhood of a current node, which is likely to lead to the global solution. To this aim, HEIDEGGER applies QED to find treatment and control groups for each hypothesis and compute the significance of the difference in the outcome between the two groups.

The specific QED, Algorithm 5.4, used in HEIDEGGER is an extended version of the randomized block design that pairs instances based on their history of confounders, treatments, and the outcome, to guarantee that the effect of these variables is controlled across time. The reason for adopting the randomized block design is its flexibility, lack of constraining assumptions, and its proven effectiveness in similar studies [11], while its major disadvantage is requiring to directly control the confounders which results in loss of statistical power. In HEIDEGGER's case, however, this is compensated by considering combinations of time points and samples as *instances*. This increases the number of effective samples by a factor approximately equal to the number of time points. The process of finding and evaluating the treatment and control pairs is done in the following four main steps:



```
Algorithm 1: HEIDEGGER
Input: $X, Y, Z$
Output: $(X^*, CP^*)$ : A list of tuples containing the pairs of
        causal variables and causal profiles
$observed = \{\}$
for $x_f \in X$ do
    $visited \leftarrow \{\}$
    for $t$: $1$ to $\tau$ do
        $cp_{current} \leftarrow$ random()
        $p_{current} \leftarrow$ QED$((x_f, cp_{current}), X, Y, Z)$
        $visited$.add$(cp_{current}, p_{current})$
        Loop
            $\{cp_{Nr}\} \leftarrow$ findNeighbour$(cp_{current})$
            $\{cp_{Nr}\} \leftarrow$ entropyPruning$(\{cp_{Nr}\})$
            $p_{Nr} = \{\}$
            for $cp_n \in cp_{Nr}$ do
                $p_{Nr}$.add(QED$((x_f, cp_n), X, Y, Z)$)
            $visited$.add$((cp_{Nr}, p_{Nr}))$
            $cp_{min} \leftarrow \text{argmin}_{p_{Nr}}(cp_{Nr})$
            if $p_{min} \leq p_{current}$ then
                $cp_{current} \leftarrow cp_{min}$
            else
                break
    $observed$.add$((x_f, visited))$
$X^*, CP^* \leftarrow FDR(observed)$
return $X^*, CP^*$
```

Figure 5.3: **HEIDEGGER Algorithm 1.**



```
Algorithm 2: QED
Input: $(x, cp)$ : (causal variable, causal profile) ;
       $X, Y, Z$
Output: p : significance level of the hypothesis
$cp_{\text{control}} \leftarrow$ findControlProfile(cp);
for $inst \in$ (samples, timepoints) do
    $trt\_cost_{inst} = D_1(X, inst, cp_{treat})$;
    $ctr\_cost_{inst} = D_1(X, inst, cp_{control})$;
$trt\_group = Selector(trt\_cost)$;
$ctr\_group = Selector(ctr\_cost)$;
for $(i, j) \in (trt\_group, ctr\_group)$ do
    $costMtx_{(i,j)} \leftarrow D_2(i, j)$;
$matchedPairs \leftarrow$ Matching($trt\_group, ctr\_group, costMtx$);
$p \leftarrow$ statTest($matchedPairs$);
return p
```

Figure 5.4: **HEIDEGGER Algorithm 2.**

Identification of Treatments and Controls

To identify the treatment group, HEIDEGGER measures the similarity of each instance to the hypothesis causal profile using a distance function (D_1). Ideally, a zero-distance treatment group is desirable, but enough ideal instances do not always exist. Thus, to introduce a level of tolerance, while preserving the quality of instances, the top κ instances are selected whose distance to the hypothesis profile is smaller than a fixed threshold (λ). For a candidate variable v, the distance D_1 of an instance (i, j) from a causal profile cp is computed as:

$$D_1((i,j), cp, v) = \frac{1}{L}\sum_{l=1}^{L}((1-\beta)(|M_{i,v}^{j-cp_o(l)} - cp_p(l)|) + \frac{\beta}{2}) \tag{5.2}$$

The formula (5.2) can be understood from its boundary values: when the user is fully certain about the prediction of missing values ($\beta = 0$), the formula is simplified to the absolute difference between the causal profile and the instance. When the user is fully uncertain of the prediction ($\beta = 1$), the formula is reduced to 1/2, indifferent toward the value of the instance, as there is a 50% chance to make a mistake. These two boundary scenarios are used to create the linear function (5.2) of the predicted difference term and the prediction certainty.

In the same way, the control group is formed by finding the top k similar instances to the no-treatment profile, using D_1. Also, the missing values of $M_{i,v}^{j}$ are predicted using linear interpolation. Parameter $\beta \in [0,1]$ shows the uncertainty level for the predicted $M_{i,v}^{j}$, and is set to zero for the known entries of $M_{i,v}^{j}$. In the running example, (i_1, j_3) is a treatment instance because $[\ 0\ \ 0\ \ 1\ ]$ has values 0 and 1 at times $j-2$ and j respectively. Similarly, (i_2, j_4) is a control instance.

Matching and Clustering

In order to find similar matches for the QED, we define the distance D_2 between a pair of instances (i,j) and (i',j') as the weighted sum of the difference of all variables across all corresponding time points. The sum of differences consists of three components: (1) the difference between the outcomes, (2) the difference between the confounders, and (3) the difference between the other candidate variables.

$$\begin{aligned} D_2((i,j),&(i',j')) = D_{\text{cnf}}(i,j),(i',j'))+ \\ &D_{\text{trt}}((i,j),(i',j')) + \gamma * D_{\text{out}}((i,j),(i',j')) \end{aligned} \tag{5.3}$$

The reason for computing three separate components is to make sure that a high number of variables in one component does not dominate the variables in the other components. A high weight $\gamma \geq 1$ is used for D_{out} to guarantee that the two matched instances have similar baseline progression of the outcome. The components of the distance are defined as follows:

$$D_{\text{out}}((i,j),(i',j'),v) = \Sigma_{l=0}^{L-1}((1-\beta)(Y_i^{j-l} - Y_{i'}^{j'-l})w(l) + \frac{\beta}{2}) \tag{5.4}$$

$$D_{\text{cnf}}((i,j),(i',j'),v) = \frac{1}{C}\Sigma_{c=0}^{C-1}\Sigma_{l=0}^{L-1}((1-\beta)(Z_{i,c}^{j-l} - Z_{i',c}^{j'-l}) + \frac{\beta}{2}) \tag{5.5}$$

$$D_{\text{trt}}((i,j),(i',j'),v) = \frac{1}{F}\Sigma_{f=0}^{F-1}\Sigma_{f=0}^{L-1}((1-\beta)(M_{i,f}^{j-l} - M'^{j'-l}_{i',f}) + \frac{\beta}{2}) \tag{5.6}$$

Where L is the maximum value of pattern offset of the treatment pattern under evaluation by the QED. Let us define baseline as the time point from which onward the effect of variables is controlled, which for an outcome at time j is $j-L$. $\vec{w} = [1\ ...\ \frac{2}{L-1}\ \frac{1}{L-1}\ 0]$ is a weight vector specifying the importance of each time point in matching the outcome; the points closer to the baseline time point are assigned higher values so that the matched instances have similar values for the outcome at the baseline. For example, in the running example ($cp_p : [\ 0\ 1\ ]$, $cp_o : [\ 2\ 0\ ]$), $j-2$ is the baseline point and $\vec{w} = [\ 1\ 0\ ]$. The matching process in HEIDEGGER's QED consists of the following two steps:

- Assign instances into blocks using a centroid-based clustering method to make sure that the distance between none of the pairs from the same block exceeds a threshold. This provides the opportunity of pairing any two instances from each block without worrying that they significantly differ in any of the observed confounders and selection biases.
- Sample random pairs of one treatment and one control instance from each block without replacement until no more pairing can be made. The randomization controls the effect of hidden confounders and outcome noise.

In the running example, since treatment instance (i_1, j_3) and control instance (i_2, j_4) have identical values for their corresponding confounders, they can be matched in the same block of the QED, where their corresponding outcome of $y_{(i_1,j_3)}$ and $y_{(i_2,j_4)}$ are compared to measure the effect of the potential cause.

Statistical Test

Considering that the values for the outcome are continuous and not necessarily normally distributed, and the instances are paired, the *Wilcoxon* signed-rank test, a non-parametric test for the significance of the difference between the outcome of two instance groups, [61] is used to evaluate the effect of a potential cause. The *Wilcoxon* test for the QED evaluates the null hypothesis that there is no significant difference in the values of the outcome in the treatment and control group because of the candidate variable. This provides HEIDEGGER with one p-value for each hypothesis, i.e. each pair of candidate variable and causal profile.

5.2.3 Multiple Comparison Adjustment

Evaluating a wide range of causal profiles for several candidate variables results in a large number of hypothesis tests. FDR-based *Benjamini-Hochberg* is the most reasonable multiple hypothesis correction criteria in the context of CPD, since it is not overly conservative for the large number of dependent hypotheses and does not rely on the tuning of the test statistics [115].

5.3 Experimental Evaluation

In the experimental evaluation of HEIDEGGER, we aim to answer the following questions corresponding to the claims made:

- What is HEIDEGGER's performance in discovering the causal profiles? To this end, HEIDEGGER's average accuracy in identifying the causal variables with various causal profile configurations is measured. In addition, the distance between the discovered and the known causal profiles is computed.

- Is HEIDEGGER able to detect complex causal profiles? To answer this question, we measure how the accuracy changes as the length of the causal profile is increased.

- Is HEIDEGGER robust with respect to noisy data? To evaluate this property, HEIDEGGER's performance in the presence of different levels of outcome additive noise is measured.

- How strong is HEIDEGGER's search? The internal validity of HEIDEGGER can be influenced by both multiple hypotheses adjustment and the search performance in finding the best causal profile. A study of the Benjamini-Hochberg procedure and comparing it to the scenario with significance levels created by random chance addresses the first property. An analysis of the percentage of nodes in the profile graph search, which lead the search process to the best discovered solution, targets the latter.

For the first three questions which require manipulation and ground truth knowledge of the causal profile, a simulated AC power transmission system is used as the dataset; because (1) transmission systems can be replicated with a very high fidelity [193, 121], (2) fault detection in transmission lines is a real challenging problem [36], and (3) to the best of our knowledge, there is no analytical method for inferring the causal profile of the faults. For the latter questions, where applicability to the typical real-world problems is the key, a longitudinal cognitive health study is used as the dataset. It should be noted that we did not use the pharmacogenomic data as an application because of the insufficient number of time point to perform any serious temporal analysis.

5.3.1 Reproducibility

For all the following experiments, HEIDEGGER's parameters are set to a series of reasonable default values. Hence, there is room for improvement regarding the final results by fine tuning. Default values are used in order to ensure that the results are not tuned for a specific problem, and the performance is generalizable to other applications. These parameter are as follows: Entropy pruning parameters $\theta = 1$ and $\delta = 2$, number of search iterations τ=5, QED matching parameters $\lambda = 0.1$, $\kappa = 2000$, $\beta = 0.3$, and $\gamma = 200$, and statistical test significance level $\alpha = 0.05$.

5.3.2 Power Transmission Datasets

We have generated a series of AC power transmission datasets based on the model created by G. Sybille (Hydro-Quebec) [192]. The model simulates faults at the load terminal of phases of a power transmission system and includes nonlinear elements such as surge arrester, shunt compensation, and complex line impedance. Furthermore, to capture the noise, interference and confounding effects at the level of the distribution system, an additional

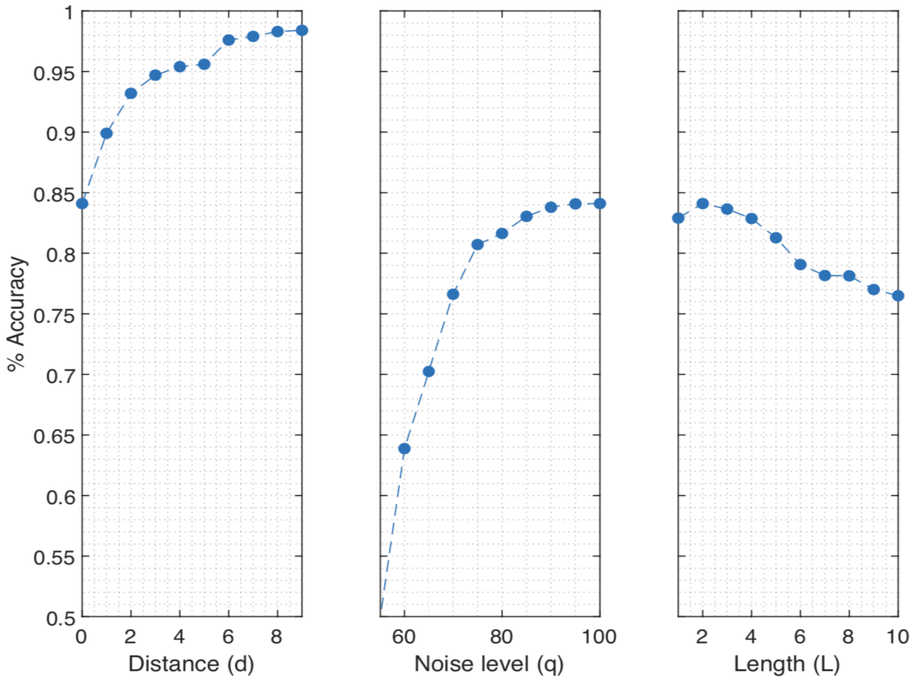


Figure 5.5: **HEIDEGGER Results.** a) percentage of cases with d or less edit distance between the exact causal profile and discovered causal profile. b) Accuracy in discovering the causal profile with respect to the length of treatment profile (L). C) Accuracy in discovering the causal profile with respect to load variation (q).

parallel terminals load is added, with a time-variant impedance sampled for every point of time from a uniform distribution between $q\%$ and 100% of the original load.

All variables are recorded node voltages from the power system. Specifically, the candidate variables and the outcomes are the varistor voltages of shunt inductors and series capacitors respectively, all sampled every half-cycle, for a window of 20 cycles. The causal profile in every simulation is generated using a fault switch occurring during the time window, activated for random intervals. For each causal profile, 50 samples are created by running the simulation with the fault switch, and 50 samples without it. We set the default values of the length of treatment profile (L), load variation (q), and edit distance (d) are set to 2, 100, and 0 respectively, and vary them in Sections 5.3.2 and 5.3.2.

Edit Distance Analysis

Figure 5.5.a presents the percentage of cases with d or less edit distance between the exact causal profile of the transmission system simulation and the discovered causal profile. HEIDEGGER is able to detect the exact treatment pattern and its offset in 84.1% of the cases. The experimental results indicate that HEIDEGGER is able to identify the causal profile with high accuracy.

Table 5.2: 5 most statistically significant hypotheses with the length $L = 10$ in the power transmission dataset

Pattern	Offset	p-value
[0 0 1 1 0 1 1 0 1 1]	[23 22 21 20 13 12 11 8 7 6]	8.72E-15
[0 0 1 0 0 1 0 0 1 1]	[25 24 23 17 16 15 11 10 9 8]	1.57E-14
[1 0 1 0 1 0 1 0 1 0]	[21 20 19 18 6 5 4 3 1 0]	3.25E-14
[0 0 0 1 1 1 1 1 0 0]	[22 18 17 14 13 12 11 10 8 7]	3.33E-14
[0 0 0 0 1 1 1 1 1 1]	[7 9 10 13 15 16 26 17 29 31]	5.02E-14

Robustness to Profile Length

HEIDEGGER's performance in discovering the simulations with various length ($1 \leq L \leq 10$) is shown in Figure5.5.b. The accuracy level in discovering the causal profile is relatively stable around 80% with respect to profile length, but slightly falls off to above 75% for longer profiles, where the differentiation of the pattern offset is more challenging. The results show that HEIDEGGER's accuracy remains acceptable even for long and complex causal profiles. The five most significant discovered hypotheses with the length of 10 and their corresponding p-values, presented in Table 5.2, demonstrate HEIDEGGER's ability to discover non-trivial patterns.

Robustness to Outcome Noise

Similarly, the level of uncertainty in the data is modeled by parameter q, which inserts an additive noise to the outcome. According to the results presented in Figure 5.5.c, accuracy with respect to the small variations in parallel load ($q \gtrsim 75$) is almost unchanged in the proximity of 85%, but rapidly declines as q becomes small enough for the random parallel load to dominate the overall load.

5.3.3 Cognitive Health Dataset

It is still unclear whether changes in lifestyle are effective at reducing cognitive decline and the risk of developing dementia during aging. With this goal in mind, the English Longitudinal Study of Ageing dataset has been collected based on a study on 4091 British adults between the ages of 60 and 85 with no diagnosis of dementia at study onset, across 10 years [187].

In this dataset, ten lifestyle factors of the participants are recorded: vigorous, moderate, and mild physical activity, friendship quality, contact frequency with friends, number of close friends, attendance at church, education class, social clubs, and sport clubs. The main target of this study is assessing the potential impact of lifestyle changes on cognitive function in older adults. Cognitive function is defined as the *memory index*, scored between $0 - 24$.

Candidate Variable	Pattern	Offset	P-value	SR	T	LPR	Itr_1	Itr_2	Itr_3	Itr_4	Itr_5
Vigorous Physical Activity	[1 1]	[1 0]	1.28E-08	1	100	0.49	18	22	34	43	48
Moderate Physical Activity	[1]	[0]	9.24E-13	1	31	0.84	12	22	24	26	28
Mild Physical Activity	[1]	[0]	5.25E-08	1	15	0.40	9	14	15	15	15
Friendship Quality	[1]	[0]	0.004702	1	7	0.43	3	7	7	7	7
Contact Frequency	[1]	[0]	0.006408	1	7	0.57	5	7	7	7	7
Number of Friends	[1 1 1]	[3 2 0]	0.123020	1	15	0.73	10	12	13	14	14
Church Membership	[1 1]	[1 0]	0.087008	0.40	127	0.13	19	30	39	47	61
Education Class	[1]	[0]	0.001289	0.62	37	0.22	16	21	28	29	31
Social Club Membership	[1]	[0]	0.018168	0.94	95	0.44	11	27	35	44	48
Sport Club Membership	[1 1]	[1 0]	6.19E-05	0.86	102	0.44	18	35	43	52	56

Table 5.3: Causal profiles discovered for the cognitive health dataset.

Multiple hypothesis testing correction

As mentioned in Section 5.2.3, the individual p-values are facing the multiple comparisons problem should be corrected using a false discovery rate adjustment. Figure 5.6 shows the curve corresponding to the p-values of all candidate variables and causal profiles, with the blue line depicting the Benjamini-Hochberg adjusted threshold. The distribution of p-values is healthy, with a long linear trail of evenly distributed p-values, as expected from random hypotheses, and ending with few exponentially smaller p-values corresponding to the significant causal variables [1].

A large number of hypotheses tested result in a strict significance threshold. HEIDEGGER was able to find the most promising profiles by evaluating only 24.8% of the total 1270 hypotheses, evaluating only a small portion of possible hypotheses; thus, a reasonable number of promising ones are passed. Among the candidate variables, all levels of *physical activity*, *being a member of sport club*, and *participating in educational class* are the variables that pass multiple hypothesis testing, highlighting their protective role against dementia. These results are consistent with the established findings in the literature, such as [137, 141].

Graph Search Convergence

Convergence of HEIDEGGER to the best causal profile is crucial. To this end, the rate at which the search is able to converge to the most significant profile across different runs is measured as the success rate (Table 5.3). In this table, T presents the number of nodes in the profile graph after pruning, and Leading Point Ratio (LPR) is the ratio of nodes that would direct the search process to the best discovered solution.

The general pattern of the results indicates that HEIDEGGER consistently converges to the most significant hypothesis if LPR is large or the profile graph is small compared to the number of iterations. This is because one of the nodes leading to the most significant profile will most probably be selected as an initialization point. That is why for candidate

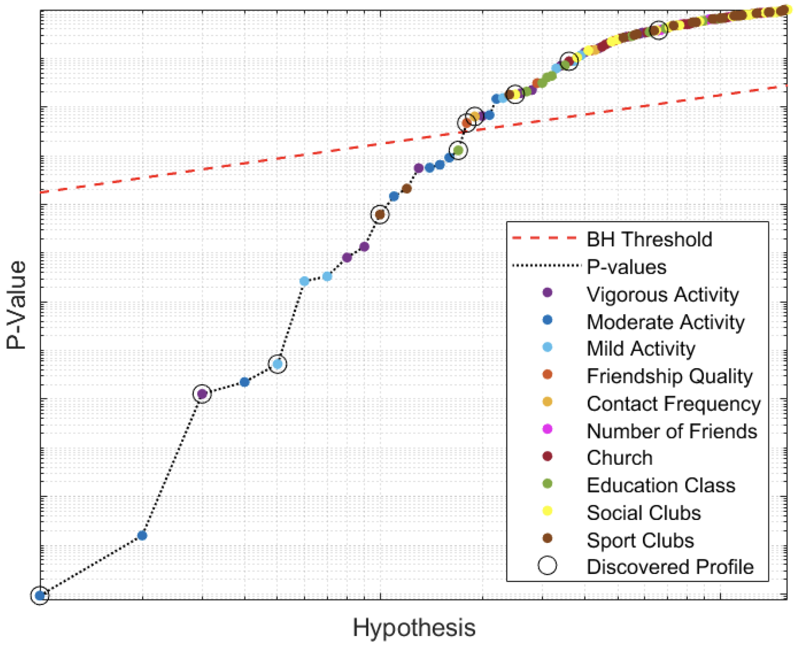


Figure 5.6: **FDR Adjustment.**: q-q plot of significance level of evaluated hypotheses across all candidate variables against *Benjamini-Hochberg* criteria.

variables 7 and 8, HEIDEGGER fails to find the most significant profiles in 60% and 38% of the runs.

Graph Search Efficiency

In Section 5.2.1 we claimed that multiple initializations strategy does not significantly add to the number of hypotheses evaluated. To verify this, we have recorded the number of hypotheses evaluated in each of the five iterations in Table 5.3. These results show that after a few iterations the growth rate of the number of evaluated hypotheses slows down. It should be noted that for candidate variables with no statistically significant profile, the growth rate in the number of evaluated nodes might be higher. This is because if a candidate variable does not have a significant effect on the outcome, p-values of its causal profiles will be random, creating a profile graph which requires many trial and errors to traverse.

Another phenomenon that may affect the graph search efficiency is that a candidate variable may have multiple distinguishable profiles with a statistically significant effect on the outcome. For example, education class might have comparable immediate and long-term effects on the cognitive health, which result in the two equally significant profiles of (cp_p: [1], cp_o: [5]) and (cp_p:[1], cp_o: [0]).

5.4 Discussion

Experimental results demonstrate that despite enormous search space of CPD, HEIDEGGER can deal with the complex causal patterns and noisy data. This is mainly because generating the local search neighborhood in the graph of profiles on-the-fly and early pruning of hopeless profiles enables HEIDEGGER to efficiently hypothesize all the promising causal profiles for each candidate cause.

Despite its efficiency and robustness in solving the Causal Profile Discovery, HEIDEGGER has several key limitations;

- Even though the pruning process shrinks the number of hypotheses, it still grows exponentially, which might lead to 'inflated' multiple hypotheses correction for long causal profiles (> 20). This exponential growth imposed by the graph search mechanism, can result in a high computation cost as well.
- HEIDEGGER is not able to handle non-stationary changes of the relationship between the causal variables and the outcome over time.
- CPD's output can be modified to report an aggregate of all the significant causal patterns. This addresses the limitations discussed in Section 5.3.3 of multiple significant profiles of a causal variable.

However, these limitations are not limited to HEIDEGGER or even CPD, and are common to most temporal CD methods in the literature. Furthermore, due to the modular nature of HEIDEGGER, with invention of new techniques, these problems could potentially be targeted and solved.

Chapter 6

Compound Causal Discovery

In their search, most CD methods assume single cause hypotheses, i.e., that the causal relations exist only between pairs of variables, and ignore the possibility that an outcome could have been caused by only jointly sufficient causes, also known as compound or complex causes. However, this assumption has long been challenged in the philosophical and scientific community as the fallacy of the single cause [154].

Also in practice, many real-world processes involve factors that only act in conjunction with each other. For example, in behavioral genetics, it has been established that a typical human behavioral trait is influenced by many genetic variants, rather than single causes, where each variant is responsible for a small percentage of behavioral variability, known as the 4^{th} law of behavior genetic [33]. Another notable example is in pharmacology, where the interactions between drugs are crucial in adverse drug reactions, and investigating each drug in isolation would be insufficient [129]. The existing CD methods could theoretically be extended to evaluate every combination of the variables as well. However, this is infeasible in most problems of realistic size, because combinations of variables increases the number of potential hypotheses exponentially, which in turn increases the data dredging and computational complexity out of control.

Hence, we name this problem of HCD of sets of variables whose combination is causal Compound Causal Discovery (CCD). Two major classes of the existing methods, structure-learning-based and pattern-mining-based methods, have the potential to solve CCD. However, as we discuss in the following, they can not find compound causes in HCD settings.

SL methods can be adjusted to apply the conditional independency tests on triplets and larger cliques of variables to find compound causes. However, most of the even efficient SL methods, such as Markov blanket discovery [6], barely scale with the magnitude of variables in the modern HCD datasets. For instance, [109] presents MH-PC to discover compound causes, where none of the single variables in the compound should be causal in isolation. This constraint allows the method to discard a number of variable combinations. Similar to some other SL methods, MH-PC tries to control the false positive rate by limiting the maximum size of the conditioning set to a great extent. However, despite the added constraint, the search strategy is still fairly close to an exhaustive search. As a result, the dimensionality of the datasets that MH-PC can search in a reasonable amount of time is in the order of hundreds of variables. The maximum dimensionality of datasets used for the experimental evaluation of MH-PC was 150.

Pattern mining methods [4] use data mining techniques to heuristically find the combinations of variables that provide information about the target variable. These methods are scalable to high-dimensional datasets with hundreds of thousands of variables, such as the pharmacogenomic application [207]. However, they do not perform causality tests and lack controlling for confounders; therefore, they can identify only associations rather than causal relations. In an effort to overcome this limitation, [102, 103], presented the Causal Association Rule mining algorithm (CAR), which finds the candidate compound causes based on their odds ratio, and uses a series of retrospective cohort studies to evaluate the causality of the candidates. While it is computationally efficient, CAR ignores data dredging and makes the causal sufficiency assumption, hence the significance level of its discoveries is not reliable.

To sum up, to the best of our knowledge, there were no method in the literature for the discovery of compound causes in high-dimensional datasets in the presence of unobserved confounders. Such a method should:

- Unlike the SL approach, be able to efficiently search through a large number of potential compound causes.
- Unlike the pattern mining methods, account for hidden confounders and evaluate causality of candidate compound causes.

Hence, we propose HEGEL, a CCD algorithm that scales well to high-dimensional datasets and is robust with respect to noise and hidden confounders. HEGEL is comprised of three phases:

- **Search phase:** a novel subgroup discovery algorithm to find compounds associated with the outcome.
- **Assembly phase:** a SL method to find the causal combinations from the selected compounds.
- **Evaluation phase:** a randomized block QED to test and adjust the significance of the causal compounds.

6.1 Problem Description

CCD receives inputs:

- A necessary categorical Outcome vector Y, where Y_i is the value of the outcome for sample $i \in \{1, \dots n\}$.
- A necessary categorical sample-variable Data matrix M, where $M_{i,j} = 1$ is the value of the potential cause variable $j \in \{1, \dots m\}$ in sample $i \in \{1, \dots n\}$.
- An optional sample-variable Natural Confounders matrix $Z \in \{0, 1\}^{n \times l}$, where $Z_{i,j} = 1$ if Natural confounder j is positive in sample i. Natural Confounders represent variables with apriori expected relation with the outcome that must be controlled as confounders and ignored as potential causes.

And outputs:

- The set of causal compounds C^*.
- The vector of causal statistical significance p^*, where p^*_i corresponds to adjusted p-value of i^{th} causal compound.

A compound C is defined as a logical AND of a set of variable and value pairs that we denote as selectors. A Selector $I(X_j = i)$ for variable X_j and value i, is a boolean function that takes value 1 if j has value i, and 0 otherwise. Formally a compound C over a set of selectors S is defined as follows:

$$C_S = \prod_{(i,j) \in S} I(X_j = i), \quad S \subset \{1, ... \, m\} \times \mathbb{R}^m, \quad I(X_j = i) = \begin{cases} 1, & \text{if } X_j = i \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

Fig. 6.1 shows a running example with three binary variables X_1,X_2, and X_3, which we use throughout this chapter to illustrate the method. Since there are three variables in this example and each has two unique values, 0 and 1, we have 6 selectors, as shown in the figure.

The reason for extending the data to non-binary values is that different values of variables can interact with each other differently. For instance, to model all possible compounds that can be formed from two potential causes each with three possible states with binary variables, not only compounds of size 4 instead of 2 is needed, but also depending on the encoding strategy, some of the mutual exclusivity information may be lost.

The first main challenge in solving CCD stems from two levels of exponential complexity; not only the number of potential compound causes grows exponentially with the number of variables, but also the CD methods themselves are typically super-exponential with respect to the number of potential causes [64]. As a result, the computation cost is prohibitively expensive. In addition, due to the exponential number of hypotheses, many legitimate causes that do not have a very high statistical power would be lost in the multiple hypothesis testing adjustment.

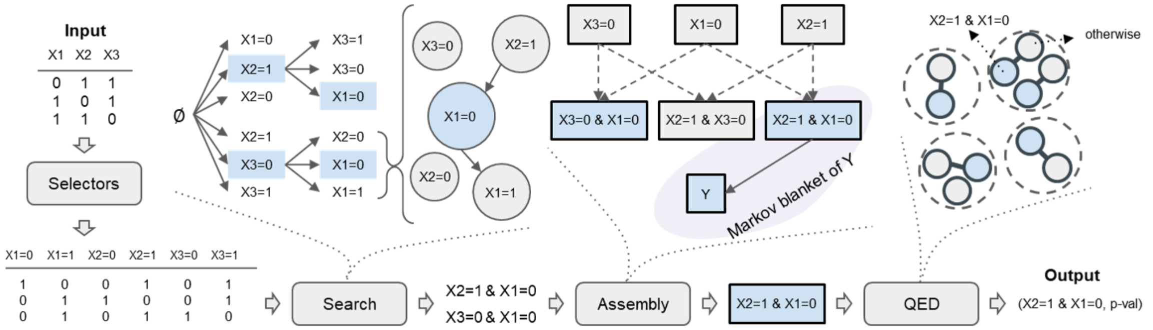


Figure 6.1: **HEGEL Framework.** consisting of search, assembly, and evaluation phases, illustrated by an example.

The second challenge is the threat of hidden confounders, which in addition to screwing the relation between potential causes and outcomes, common in CD, could sabotage the results of CCD even further, by interacting with potential causes in the outcome. Hence, since in practice datasets that satisfy the causal sufficiency assumption are quite rare, especially if they are high-dimensional, CCD methods need to be able to effectively deal with unobserved confounders.

6.2 HEGEL Method

To solve the CCD, we propose HEGEL. HEGEL, in addition to scaling well in high-dimensional settings, does not assume causal sufficiency.

To address the first challenge, in the search phase, HEGEL uses a novel subgroup discovery algorithm with a scoring function tailored to estimate the expectation of a causality criterion for efficiently finding the candidate variables that could be promising in combination with other variables. The choice of subgroup discovery is because filtering is common starting phase in HCD, and subgroup discovery is a data mining approach designed to discover subpopulations within the dataset that are most significant in isolation with respect to a measure of interest [75]. However, (1) causal relation is not usually the target of subgroup discovery methods [152], (2) subgroup discovery methods usually ignore data dredging resulted from extensively checking different variable combinations. Hence, HEGEL needs a subgroup discovery that identifies the set of most significant candidate compound causes while investigating as few compounds as possible.

To address the second challenge, in the Assembly phase, HEGEL breaks down and recombines the candidate compounds obtained from the first phase and then deploys a Markov blanket SL model to find the causal compounds in presence of hidden confounders. The recombination is in order to ensure that the most significant combinations of compounds from the pool of associated candidate compounds are inferred. Markov blanket SL is used because HEGEL is only concerned with relations of the outcomes under no causal sufficiency, and despite the relatively small pool of candidate compounds, the set of all of their possible combinations can become very large. Hence, HEGEL's Markov blanket SL needs to model the recombinations of candidate compounds and deal with hidden confounders and computationally cost effectively.

To produce the outputs expected in CCD, in the Evaluation phase, HEGEL utilizes a randomized block QED that provides the strength of the causal compounds found in the Assembly phase. The reason for choosing randomized block QED is that unlike most SL algorithms, it can provide multiple hypothesis-adjusted and confounder-controlled significance level for causal relations [110, 28].

Figure 6.1 shows the running example progressing through HEGEL phases. In the following, the three phases of HEGEL are described in detail.

6.2.1 Search Phase

Because testing the causality of every possible compound is infeasible and undesirable, HEGEL filters the promising candidate compound causes based on their association with the outcome. However, identifying the most associated among the exponential number of compounds, while comparing as few compounds with the outcome as possible remains a challenge. To solve this problem, HEGEL utilizes a heuristic subgroup discovery method. We introduce a new search strategy, objective function, and constraining criteria for this heuristic search. Algorithm 1 describes the main components of the search phase.

The association between the compounds and the outcome variable is computed by a score function denoted by Q, which guides the subgroup discovery in its search for candidate compounds. The score function can be as simple as accuracy and risk ratio or more complex such as uncertainty coefficient or the coefficient of colligation. Since the selectors and the compounds comprised of them are binary variables, a 2x2 contingency table can encode all the information need to compute the score between a compound and the outcome. There are various association measures applicable to the contingency table. Here, we use the well-known Mathews coefficient which is equivalent to Pearson correlation for binary variables [41]. The Mathews coefficient can be computed as:

$$Q_{Mathews} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\blacksquare}n_{0\blacksquare}n_{\blacksquare 1}n_{\blacksquare 0}}} \tag{6.2}$$

Where $n_{11}, n_{00}, n_{01}, n_{10}$ are the entries for true positive, true negative, false positive, and false negative of the 2x2 confusion matrix respectively, and the black square ($\blacksquare$) indicates their marginal frequencies.

The goal of subgroup discovery is to find the compounds that have the highest score. This is achieved by searching the compound tree that has an empty compound at its root and at each level i has all the compounds of size i, where each compound is connected to those compounds in the next level that are the union of that compound and a selector.

Hence, the first level is comprised of all possible selectors, the second level is comprised of all possible pairs of selectors, and so on. In the running example, level 1 includes compounds of size one such as $X_1 = 0$, and level 2 includes compounds of size two such as $X_1 = 0$ & $X_2 = 1$ which is generated by adding selector $X_2 = 1$ to the compound $X_1 = 0$.

At the heart of every subgroup discovery is the search strategy, which starts with the empty compound root, and one at a time decides which compound node to expand on the compound tree and which selectors to be added. The resulting compounds are constrained based on the depth criteria, i.e., the number of selectors in the compound, denoted by d. This makes sure that tests are not wasted on compounds with insufficient statistical power and external validity.

HEGEL's subgroup discovery uses a heuristic beam search algorithm as the search strategy. Here, at each level of the tree, top-scoring compounds from the previous level, called beam B, are combined with promising selectors to form new compounds whose scores are then computed to see if they can replace the existing beam compounds. The key factor in beam search is the number of compounds maintained in the beam, called beam width β, which for smaller values makes it similar to greedy search and for larger values similar to breadth-first search. In the running example, the beam width is two, and in the first layer, it includes $X_2 = 1$ and $X_3 = 0$. After expanding the compounds from the first layer, the two new generated compounds, $X_2 = 1$ & $X_1 = 0$ and $X_3 = 0$ & $X_1 = 0$, replace the previous compounds in the beam as they have better scores than them.

In the first level, scores of all selectors are computed using the scoring function. Top β selectors with the highest scores are returned as the level 1 beam. This exhaustive search in the first level guarantees high-quality starting compounds, which are crucial in the performance of beam search, and is not expensive in comparison to the next levels, due to the linear number of tests with respect to the number of variables. In the running example, the starting compounds are $X_2 = 1$ and $X_3 = 0$ which form the beam for the first phase.

In the levels after the first level, offspring of the previous level's beam are generated by combining its compounds with the most promising new selectors. For each compound in the beam, u times a promising selector is identified and the score of its combination with the compound is computed and potentially added to the beam. If the score of a newly evaluated compound is better than its parent compound and better than the lowest-scoring compound in the beam, the compound replaces the compound with the worst score. The reason for requiring that a compound has a better score than its parent is that otherwise the variables with a high score could dominate the beam and diversity of the compounds in the beam will be lost. In the running example, to expand $X_3 = 0$, HEGEL evaluates $u = 3$ selectors with the following order: $X_2 = 0 \rightarrow X_1 = 0 \rightarrow X_1 = 1$. Because adding $X_1 = 0$ to $X_3 = 0$ results in the highest score, and the score of $X_3 = 0$ & $X_1 = 0$ is higher than its parent, $X_3 = 0$, we pair $X_3 = 0$ with $X_1 = 0$.

To illustrate the reason why we need to compare the score with the parent, consider an example where compound C has a high score. If we append a selector I to form the compound $C' = C \cup I$, even if C' is not a true cause, the combination could have a high score due to the existing association of C to the outcome. Hence, as a pruning mechanism, if adding I does not improve the score of C, it is discarded.

To select the most promising selectors to add to a compound, a probabilistic approach is utilized. First, for the compound C to be expanded, the fitness value for each selector I is computed using a fitness function, $L_C(I)$. Then, one selector is randomly selected, with the chance of each selector being selected proportional to the corresponding fitness value of that selector. The fitness function consists of two components, the selector's individual score and a term that captures how promising the selector is based on the success of its similar evaluated selectors. The fitness function $L_C(I)$ is computed as:

$$L_C(I) = Q(I) + \alpha \max_{I' \in H} Q_{CI'} \cdot W_{I'I} \tag{6.3}$$

Where H is the set of selectors I' for which the score of $C \cup I'$ is already computed, and $W_{I'I}$ is the similarity between the selectors I and I'. In the running example, the size of the circle representing each compound is proportional to its fitness value. Although $X1_{=}0$ has the highest score, due to the random selection, it is evaluated on the second try.

Early on in the search, because not many Compounds are tested, the selector's individual score dominates the fitness function and assures that in the absence of relative information, high-quality selectors are checked first. However, as the number of tested selectors gets closer to u, the second part of the fitness function becomes significant. This part uses the fact that if a selector I' combined with the compound C can achieve a top score, and selectors I and I' are very similar, score from I and score from I' would be similar, and therefore I has a good chance of also being top scoring. Hence, the first part of fitness works as a global cold start search, and the second part works are a local refined search.

The main strength of the proposed fitness-based selector search to extend beam compounds is avoiding data dredging. Because of the multiple hypothesis testing problem, the significance levels of results of a CD algorithm need to be adjusted based on the number of hypotheses evaluated. The number of hypotheses evaluated includes even the compounds that were excluded after computing their association with the outcome. In HEGEL, to choose the most promising selectors, the score is computed only for u selectors for a given compound, as opposed to all selectors. Therefore, the number of hypotheses tests for each compound of the beam is bounded by u.

6.2.2 Assembly Phase

Once the candidate compounds are obtained from the search phase, in the assembly phase, causal relations between all candidate compounds with the outcome need to be evaluated.

Algorithm 1: Search Algorithm

```
Input: Variables $X$, Depth $depth$, Threshold $\alpha$, Scoring Function $Q$, Beam Width $\beta$, Number of Iterations $u$
Output: $output\_beam$, a set of candidate compounds causes
$selectors = create\_selectors\ (X)$ ;
$m = length\ (selectors)$ ;
$W = create\_similarity\_matrix\ (selectors)$ ;
$beam = heap\_max\ (size = \beta)$ ;
for $selector\ in\ selectors$ do
    $score = Q(selector)$;
    $beam.add\ ((selector, score))$;
$expanded = map()$;
for $d : 1\ to\ depth\ -1$ do
    $output\_beam = beam.copy()$;
    for $C\ in\ beam$ do
        $if(expanded[C] == True)\ continue$;
        $visited[1 \ldots m] = False$ ;
        for $t : 1\ to\ u$ do
            for $j : 1\ to\ n$ do
                $I = selectors[j]$;
                if $visited[j] == False$ then
                    $L_C(I) = Q(I) + \alpha \max_{I' \in H}\ Q_{CI'} \cdot W_{I'I}$
                $L_C(I) = 0$;
            $j = randomChoice(L_C)$;
            $I = selectors[j]$;
            $score = Q(beam_i \cup I)$;
            $visited[j] = True$;
            $new\_beam.add\ (beam_i \cup j), score))$;
        $expanded[C] = True$;
    $beam = new\_beam.copy()$;
return $beam$;
```

However, in addition to these compounds, we also need to evaluate the causality of compounds derived from their selectors as well. To derive such compounds, first, all unique selectors of candidate compound causes are obtained, then, derived compounds are generated by making all combinations of maximum length d of these selectors. For instance, in the running example, the candidate compounds are $X_2 = 1 \& X_1 = 0$ and $X_3 = 0$ & $X_1 = 0$ returned by search phase. Because their unique selectors are $X_2 = 1$, $X_3 = 0$, and $X_1 = 0$, and the maximum depth $d = 2$, the derived compounds are as follows:

- Length 1: $X_2 = 1$, $X_3 = 0$, $X_1 = 0$
- Length 2: $X_2 = 1$ & $X_3 = 0$, $X_2 = 1$ & $X_1 = 0$, $X_3 = 0$ & $X_1 = 0$

To jointly evaluate the original and derived compounds, HEGEL uses a SL method that is highly efficient due to exploitation of the inherent relations between compound causes, as opposed to QED methods which evaluate each candidate compound separately.

SL in HEGEL is based on the FGES-MB [157] which is a modification of the GES algorithm [40] for Markov blankets, and is one of the fastest SL models that is maximally informative and can handle hidden confounders. While learning the causal structure, FGES-MB limits its horizon to only relations within the Markov blanket of the outcome Y to reach efficiency.

FGES-MB starts with an empty graph and iteratively adds the edges that increases the graph score the most. The graph score measures how well the causal graph fits the conditional probability distribution of the variables in the dataset. After adding an edge, the resulting model is replaced with the corresponding Markov equivalence class. This process continues until the graph score can no longer be improved. Then, the edges whose removal improves the graph score the most are iteratively removed until no further edges can thus be removed. In the running example, the true Markov blanket for the outcome Y includes only the node $X_2 = 1$ & $X_3 = 0$.

FGES-MB can solve CD problems with millions of nodes and edges in a reasonable time by using a graph scoring technique with superposition property: graph score of each potential causal graph is defined as the sum of scores of its potential edges. Furthermore, FGES-MB drastically reduces the cost of the computationally expensive edge scoring by leveraging the scores available from the previous iteration using two mechanisms. First, it identifies which edges must have the same score as the previous iteration, skipping their update. Second, for the affected edges, the scores are computed efficiently based on the scores from the previous iteration. As a result, the score of potential edges at each iteration can be estimated efficiently and parallelization can be done with a small overhead.

HEGEL uses a modified version of FGES-MB to test candidate compound causes for causality. In this modified version, as shown in Algorithm 2, each node represents a compound, and edges indicate the potential causal relations. The set of nodes is the power set of maximum length d of unique selectors that candidate compounds from the search phase are made of. This is because the particular grouping of selectors in a compound might change in the light of confounding compounds. In addition, as opposed to the original FGES-MB, HEGEL starts with a pre-configuration of the causal graph, where each compound with l selectors is connected to parent compounds with $l - 1$ selectors. This is true by definition because each subset of a compound can be considered a cause for that compound. Hence, the initialized edges are marked not to be removed by the algorithm. In the running example, HEGEL adds an edge from $X_3 = 0$ to $X_3 = 0$ & $X_1 = 0$ because $X_3 = 0$ is parent and cause of $X_3 = 0$ & $X_1 = 0$. In the Fig. 6.1, the dotted edges indicate the pre-configured edges, and solid edges show the causal relations inferred by the FGES method. Because a causal relation is established between $X_3 = 0$ & $X_1 = 0$ and outcome variable Y, $X_3 = 0$ & $X_1 = 0$ is returned as the output of the assembly phase.

Algorithm 2: Assembly Algorithm

```
Input: Beam $b$, Outcome $Y$, Depth $d$
Output: $causal\_compounds$
$selectors = extract\_unique\_selectors(b)$ ;
$combinations = create\_combinations(selectors,\ d)$ ;
$E = set()$;
for $node1$ in $combinations$ do
    for $node2$ in $combinations$ do
        if $node1 \subset node2$ then
            $E.add((node1, node2))$;
$initial\_graph = construct\_graph(nodes=[combinations, y],\ edges=E)$;
$causal\_graph = FGES(initial\_graph)$;
for $edge$ in $E$ do
    if $edge[1] == y$ then
        $causal\_compounds.add(edge[0])$;
return $causal\_compounds$;
```

6.2.3 Evaluation Phase

The compounds that FGES-MB connects to the outcome can be considered causal. However, their quantitative statistical significance needs to be evaluated by a statistical test that controls the effect of other causes and adjusts for the multiple hypothesis testing. For this goal, a randomized block QED, Algorithm 3, is used which can provide adjusted statistical significance of discoveries and overrides assumptions of FGES in the previous phase. The reason for adopting the randomized block design is its flexibility, lack of constraining assumptions, and its proven effectiveness in similar methods [11]. In particular, randomized block design effectively addresses threats to internal validity, especially selection bias and history. randomized block design groups samples based on their similarity in other potential causes to guarantee that the effect of these variables is controlled. Next, in order to emulate the random assignment, treatment and control samples within each block are paired randomly.

To evaluate each compound using randomized block design, the following steps are performed:

- **Treatment Groups** Partition the samples into treatment group and control group, where the value of candidate compound is one for the treatment group, and zero for the control group.
- **Clustering** Cluster the samples in block based on the possible confounders, i.e., compounds received from phase 2 other than the compound in the evaluation.
- **Matching** Within each block, randomly pair the samples from the treatment group and control group without replacement until no more pairing can be made.
- **Statistical Test** Depending on values for the outcome variable being discrete or continuous, and whether it is normally distributed or not, an appropriate paired difference test (such as Wilcoxon signed-rank test) is performed on the matched treatment control pairs [22]. The test estimates the probability of the null hypothesis that there is no significant difference in the values of the outcome variable in the samples because of the candidate compound.
- **Multiple hypothesis testing** Despite its selective search, HEGEL still evaluates many compounds based on their association to the outcome. Hence, similar to the previously proposed methods, HEGEL uses the well-known *Benjamini-Hochberg* false discovery rate adjustment [20] to evaluate which of the p-values corresponding to candidate causal compounds are statistically significant.

Randomized block design uses two mechanisms for the evaluation of the compounds: a) guaranteeing that samples are similar across treatment and control groups via matching, and b) controlling hidden confounders via random sub-sampling. This direct control of confounders can act as a double edge sword in improving the external validity from one side, and adding bias by controlling for colliders between treatment and outcome from another. However, because all the compounds sent to QED are considered direct causes by FGES which takes all conditional dependencies into account, HEGEL does not have to worry about it.

In the setting with n binary variables and maximum depth of d, the total number of hypotheses is $\sum_{k=1}^{d} \binom{2n}{k}$. Therefore, using a naive evaluation test that is linear with respect to the number of samples m, the time complexity will be $O(mn^d)$.

Algorithm 3: Evaluation Algorithm

```
Input: compounds : list of candidate causal compounds;
Output: adjusted_pVals : p-values of the compounds
$pVals = []$;
for $c$ in compounds do
    $trt\_group = findTreatmentCases(c)$;
    $ctr\_group = findControlCases(c)$;
    for $i$ in $trt\_group$, $j$ in $ctrl\_group$ do
        $costMtx_{(i,j)} = computeDistance(i, j)$;
    $blocks = clusterSamples(costMtx)$;
    $matchedPairs = []$;
    for block in blocks do
        $pairs = randomMatching(block, trt\_group, ctr\_group)$;
        $matchedPairs.append(pairs)$;
    $p = \text{statisticalTest}(matchedPairs)$;
    $pVals.append(c, p)$;
$adjusted\_pVals = adjustPVals(pVals)$;
return $adjusted\_pVals$
```

6.3 Experimental Evaluation

In the experimental evaluation, we aim to answer the following questions corresponding to the claims made in Section 1:

- How effective is HEGEL in solving the CCD problem? And are all the phases necessary for achieving this goal? To this aim, HEGEL's accuracy in identifying the causal variables is measured after each phase of the method. Because we are considering the cumulative results of the three phases, we use SD, $SD + SL$, and $SD + SL + QED$ to denote the result of search phase, search and assembly phases, and the complete HEGEL, respectively.

- How does HEGEL perform in comparison with state- of-the-art CD methods? To answer this question, we compare HEGEL with FCI, the maximally informative SL method which gives asymptotically correct single causes even in the presence of hidden confounders. Since FCI can only detect single causes, to make a fair comparison, we treated its results as successful if all elements (single causes) of compound causes were detected by FCI. We note that detecting components of a compound cause is a much simpler task, which means that the comparison is very favorable to FCI. We also compare HEGEL against CAR, a hybrid CD method based on association rule mining, which has the capability to search for compound causes. For that purpose, we extended the left-hand side of CAR rules to search for compounds, which required no modification to the support odds ratio criteria used by CAR. Although MH-PC seems to be a reasonable baseline as well, because its definition of compound causes is different from CCD,it was not included in the experiments. The default minimum local support was set to 0.01 for CAR, and a a significance level of $\alpha = 0.01$ was used for the statistical tests for all the methods.

- How does HEGEL perform in high-dimensional settings? And can HEGEL perform well for relatively low sample sizes, which are common for some high-dimensional datasets? To evaluate this property, we measure how the accuracy changes as the number of variables and the number of samples are varied.
- How robust is HEGEL against threats to validity? To answer this question, we evaluate the accuracy of the method with respect to the major sources of validity threats: noise and hidden confounders.

6.3.1 Datasets

Answering these questions requires extensive knowledge of ground truth causes and their interaction. Unfortunately, we are not aware of any real-life datasets with known ground truth compound causes. Therefore, we conduct the experiments on synthetic datasets. The following are the steps to create the datasets: 1) a sample-variable matrix of size $m \cdot n$ is generated from independently and identically distributed random numbers, 2) causal selectors are generated by randomly selecting variables and values without replacement 3) single and compound causal compounds are created as AND combination of the causal selectors, 4) The outcome is set to one if disjunction of the causal compounds evaluates to true.

We need to simulate the effect of hidden confounding conditions that are necessary for the cause to affect the outcome. To simulate this effect, we assume that the confounders are satisfied in z percent of the cases; so, the outcome is set to 1 in only z percent of the cases when the value of at least one of the causal compounds is 1. We denote the z parameter as confounder satisfaction. Furthermore, to simulate noise in the data, we flip q percent of the outcome elements to represent the noise in the data.

We set the default values of the number of variables (m), sample size (n), confounder satisfaction (z), and noise ratio (q), to 1000, 500, 0.75, and 0.05, respectively, and vary them for the experiments as follows.

6.3.2 Reproducibility

For all of the following experiments, HEGEL's parameters were set to reasonable default values. Hence, there is room for improvement in HEGEL's performance by fine-tuning. Default values were used in order to ensure that the results are not tuned for a specific problem, and the performance is generalizable to other applications. These parameters are

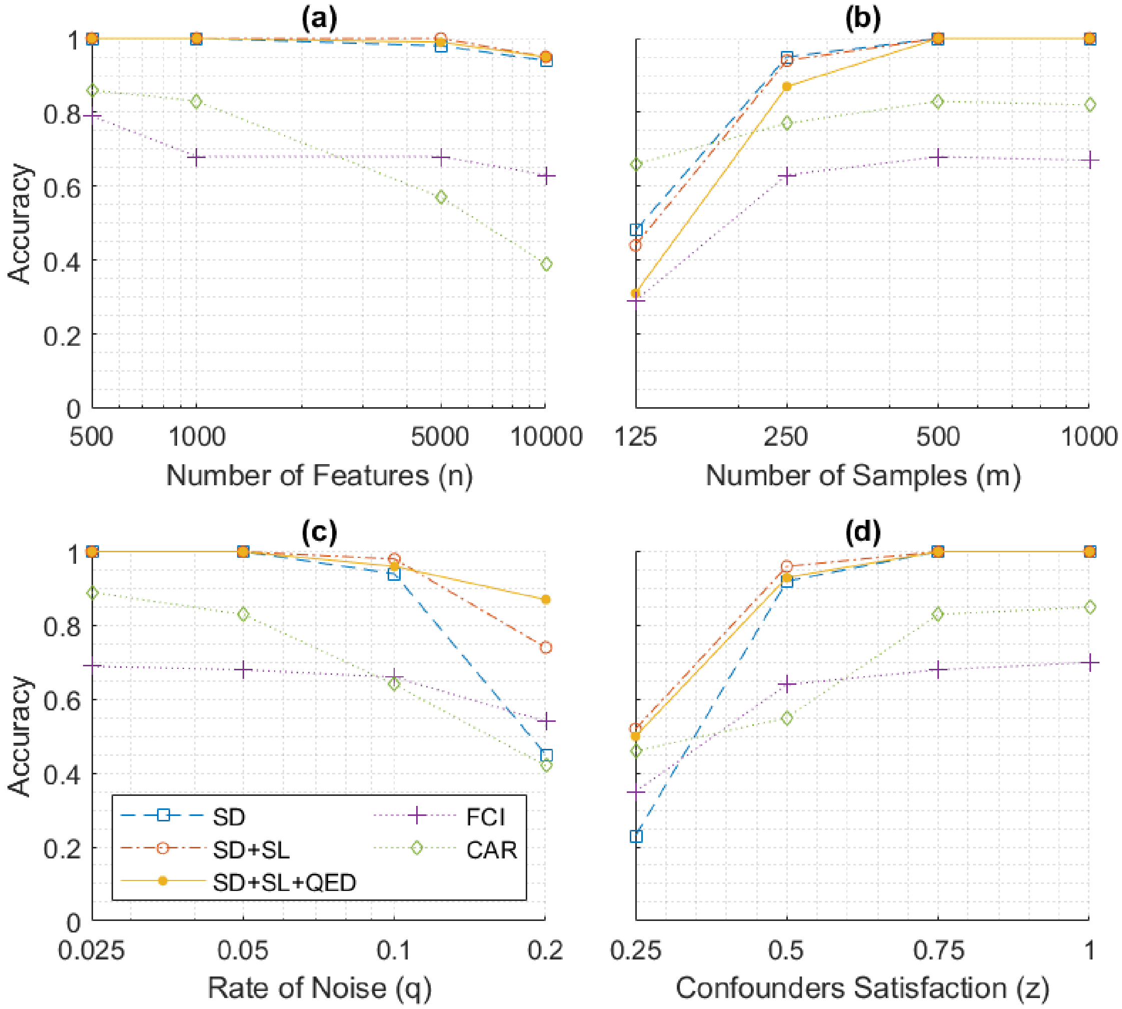


Figure 6.2: **Result of Experiments.** The accuracy of HEGEL, CAR and FCI in different settings. The x-axis in figures a,b, and c is in logarithmic scale.

as follows: heuristic equation weight $w = 2$, beam width $\beta = 10$, depth $d = 2$, and the number of iterations $u = 100$.

6.3.3 Results

In the following, we have devoted separate sections for answering questions 3 and 4. Questions 1 and 2 have been addressed by investigating the performance of FCI, CAR, and the three phases of HEGEL across the experiments for questions 3 and 4. We created 100 different datasets per setting of parameter values, and report the average of the fraction of true causal compounds discovered across the datasets with the same setting.

Sensitivity to the number of Variables and Samples

We generated synthetic datasets with the number of variables from 500 to $10,000$. As shown in Fig. 6.2a, HEGEL maintains a high accuracy for different numbers of variables and consistently outperforms all baseline methods. HEGEL starts with an accuracy of 100% and maintains a high accuracy higher than 94%, even in the case with a very large number of variables. FCI also shows a robust performance, however, the accuracy is inferior in comparison to the HEGEL by a very large margin of around 30%. Interestingly, when the number of variables is small enough, the chance of FCI detecting all components of the causal compounds is higher, possibly because the chance of different selectors having comparable scores by chance becomes smaller. On the other hand, the accuracy of CAR starts at a high value but it decreases significantly for high-dimensional settings. The major contributing factor for this loss of accuracy is the number of false positives from the millions of possible compounds, which are rejected by multiple controlled tests performed by HEGEL and FCI. Although the accuracy of QED and to some degree FGES are sensitive to the larger number of hypothesis, $SD+SL+QED$ (HEGEL) and $SD+SL$ remain accurate thanks to constraints imposed by the search phase: 1) the number of hypotheses evaluated is bounded by beam $\beta \cdot u$ which helps with the QED multiple hypothesis testing problem, and 2) the number of compounds fed to the FGES method is bounded by beam width β, which reduces the risk of data dredging.

To investigate robustness to small sample sizes in a high-dimensional setting, we created datasets with 125 to 1000 samples. Based on Fig. 6.2b, HEGEL has a superior performance compared to the baseline methods when the sample size is not too small. We can observe that the error rate for all methods increases with decreasing the sample size. In the very small sample size experiments, HEGEL ($SD+SL+QED$) underperforms in comparison to SD and $SD+SL$ because QED requires a large enough population to gain statistical power to pass candidate compound causes through the multiple hypothesis testing. Similarly, SD+SL performs slightly worse than SD because more samples are needed to get an accurate estimate for the conditional independency tests used by FGES. Interestingly, even in the subgroup discovery phase where associations are measured we see a degraded accuracy. The reason is that the size of the population, which is an important factor in computing the association, shrinks further as more selectors are added to a compound. This is aggravated further due to the presence of noise and confounders. Consequently, due to the lost true causes in the search phase, SL and QED in the following phases are not able to recover the true causes. Although still worse than HEGEL and its variants, SD and $SD+SL$, FCI shows a moderate accuracy. This can be partly because FCI only needs to discover single causes, which requires fewer samples. On the other hand, CAR manages to outperform HEGEL and FCI in case of a very low sample size, which is reasonable given the odds ratio's robustness to a small sample size.

Robustness to Threats of Validity

To evaluate the robustness to validity threats, we varied the degrees of noise and confounder satisfaction. Fig. 6.2c shows the accuracy of all compared methods for noise levels from 0.025 to 0.2. HEGEL shows the highest level of robustness against noise. Since the correlation measure used in the search phase is sensitive to noise, the performance of SD drops to 45% as the noise level is increased. The second phase, FGES, increases the accuracy of $SD + SL$ by incorporating conditional dependency between variables. As expected, the third phase, QED, provides more robustness against noise, resulting in an accuracy of 87% for $SD + SL + QED$ even in the worst case. It should be noted that FCI is also relatively robust against noise, but it starts from accuracy of 70% compared to 100%. In contrast, CAR starts from a solid accuracy of 89%, but descends to only 42%, possibly because of the high sensitivity of odds ratio to noise.

The accuracy for varying degrees of confounder satisfaction is shown in Fig. 6.2d. We created datasets with confounder satisfaction ranging from 1 to 0.25. It should be noted that a confounding satisfaction of 0.25 severely corrupts the data, sabotaging 75% of positive outcomes. Even at 50% confounding satisfaction all three versions of HEGEL obtain an accuracy of above 91%, outperforming the baseline methods. The search phase is more prone to error at low values of confounder satisfaction due to the asymmetric heavy loss of association because the reduction of confounder satisfaction leads to higher false-negative rates. However, because the components of the true compound causes survive the search phase, they can be recovered through the combination of single causes in the assembly phase. Furthermore, since FGES does not make the causal sufficiency assumption, it can reassemble the causal compounds faithfully, resulting in higher accuracy of $SD + SL$ and $SD{+}SL{+}QED$. It should be noted that FCI is fairly robust against low values of confounder satisfaction, but because of a lower starting accuracy, its performance is much worse than the performance of the HEGEL versions, e.g., 64% compared to 91% for confounder satisfaction of 50%. For CAR, after a critical reduction of confounder satisfaction around z = 0.6, the false negative samples do not affect the results much further, which is consistent with the local support criteria of CAR.

To summarize, the experimental results on a broad range of synthetic datasets show that HEGEL is superior to the state-of-the-art methods FCI and CAR, in almost all scenarios except for a very small sample size. In this case, although the compounds are originally discovered by HEGEL, they did not reach statistical significance after the multiple hypothesis adjustment. It was also observed that each of the three phases of HEGEL played an important role in achieving high accuracy. In particular, the search phase was shown to be crucial in high-dimensional settings by successfully narrowing down the set of candidate compounds, which benefited the following phases as well. Moreover, the assembly phase was shown to be effective at recovering true compound causes by re-assembling selectors falsely paired in the search phase. Finally, in scenarios with a high rate of noise, the robustness of QED helped to refine the true compound causes.

6.4 Discussion

We tried to demonstrate that the single-cause fallacy of CD is unjustified in many scenarios, and CCD should be sought after in many potential applications, such as pharmacogenomics where the interaction between biomarkers may be required to cause an outcome. We also showed that HEGEL is effective in dealing with CCD's two main challenges of exponential complexity of the space of variable combinations, and the exacerbated impact of hidden confounders. This was achieved by respectively, leveraging the inherent relations between variables in the proposed subgroup discovery method, and utilizing SL and QED that take the combinations of compounds and hidden confounders in to account. However, HEGEL has several limitations:

- HEGEL's subgroup discovery method depends on a suitable scoring function to measure the association of candidate compounds with the outcome. Instead of relying on a particular score function, the subgroup discovery can possibly be improved by using an ensemble learning method that automatically selects the best-performing measures of association.

- HEGEL assumes that the compound causes represent the logical AND of variables. HEGEL could potentially be expanded to also consider other logical combinations of variables, such as XOR, without additional multiple hypothesis adjustment cost by adapting the subgroup discovery process to efficiently manage the relevant combinations of variables.

- Although HEGEL demonstrated its value in various synthetic CCD settings against standard CD methods, the theoretical guarantees and expected margins of when a standard CD method fails to find compound causes are yet to be shown.

Chapter 7

Conclusion

The general causal inference and CD approaches face problems in computational complexity, multiple hypothesis adjustment, and various biases. In the last decade, two major approaches have emerged that can deal with HCD.

The well-known SL methods have been optimized by bypassing their exponentially growing conditioning tests. However, the theoretical guarantee which they offer relies on a list of assumptions that are rarely fulfilled in real-world applications. On a parallel line of research, QEDs have been equipped with machine learning algorithms to efficiently search for promising hypotheses. Although QED-based methods are flexible and efficient, they are plagued by validity threats and are reliant on the user to adjust the method to the particular HCD problem.

However, one of the most important gaps in HCD lies in its applicability. The current CD and HCD studies are mostly focused on variations of similar successful methods that improve the computational complexity, reduce the assumptions made about the data, or outperform other methods in accuracy. This is in contrast to sister fields such as machine learning and causal inference, where there are various problem types, inspired by existing applications.

Hence, instead of iterating methods that outperform established HCD approaches by marginal improvements, we established the goal of introducing new promising avenues for research that are practical and aligned with the existing demand outside the computer science circle.

We introduced four problems of CD from high-dimensional observational data and proposed novel methods for solving them. These problems are as much inspired by the theoretical gaps in CD as they are by the real-world applications. The proposed methods use ideas and motifs from a wide range of applied sciences, such as link prediction from social network analysis, subgroup discovery from data mining, and signal flow diagram from signal processing, as well as the known CD methods such as MB-FGES from SL and randomized block design from QED. However, the proposed methods share the common structure of search for promising hypotheses and then validation of those promising hypotheses.

Although the proposed problems and methods are still useful for small causal discovery problems, their necessity becomes apparent in the high-dimensional setting. For instance, if the number of variables is small enough, instead of Compound Causal Discovery, all combinations of variables can be evaluated, instead of Stratified Causal Discovery, variables can be evaluated for all possible subgroups of all subsets of variables, instead of Causal Profile Discovery, all possible profiles of all variables can be evaluated, and instead of Relational Causal Discovery, the small set of existing prior relations of each variable can manually be incorporated as its confounders. However, the results of experiments show that the proposed methods can compete with the state-of-the-art causal discovery methods, even in presence of noise, low statistical power, and hidden confounders.

The analysis of the assumptions and theoretical guarantee of causal inference methods are often made under an asymptotic framework. In such high dimensional settings however, asymptotical analysis is far from realistic, and the theoretical analysis can give be misleading, and is hence less emphasized in this work. Furthermore, all the proposed problems are designed to be solvable by adjusting the existing concepts, which led us to design methods more as a general process with building blocks of methods that are interchangeable with similar ones. Hence, if a user wishes the outputs to have certain properties for the output, the appropriate hypothesis-test layer can be substituted or added. For instance, for a sound and complete compound causal discovery with hidden confounders under causal faithfulness, the FCI algorithm can be used for the evaluation phase of HEGEL; or to find causal profiles with nonlinear relations with the outcome, HEIDEGGER's QED can be substituted with the convergent cross-mapping.

Future Directions

Despite the overflow of practitioners and datasets and tools [60, 158], there been very few major scientific discoveries done by any causal discovery method, in any field [119, 67, 166, 95, 131, 54, 145, 140]. This can be contrasted to the randomized controlled trials for instance, which have been widely adopted by various scientific disciplines for discovering causal relations [188]. This puts decades of research and various technical innovations in the field of causal discovery in a predicament. If causal discovery is rarely adopted for empirical knowledge discovery and is presumably inferior to the predictive models [37, 99] in estimating the outcome, then what is the benefit of causal discovery?

During the experiments on the pharmacogenomic datasets, we faced two of the potential reasons for causal discovery's lack of valuable throughput:

- **Statistical significance:** Causal discovery methods, especially those proposed by the computer science community [MCMB, Li that], rarely bother with systemic data dredging and the loss of statistical power. In our experiments, when the number of variables would reach hundreds of thousands, software such as Tetrad would simply produce hundreds of false negative discoveries for the target variables, with no warning of low confidence. Even when the causal discovery methods provide tool for control, adjust or warn data dredging, such as some of the methods proposed in this thesis, the adjusting thresholds and criteria are usually non-intuitive and distorted by the complex design of the method. For instance, a scientist who would usually evaluate hypotheses with a simple Pearson's chi-squared test, ignore the ones with less than the conventional minimum sample size of 50, and easily adjust their significance by Bonferroni correction, would have a hard time trusting and using toolboxes that spit out a list of significant relations whose length is adjustable by a set of unknown parameters.

- **Complex systems:** With the expansion of our knowledge of the outside world and the increase in sophistication of our fingerprint, the relations that we are interested in discovering are more and more becoming part of larger systems with properties such as non-linear time variant interactions, emergence, or feedback loops [15]. However, the causal discovery methods usually do not stray away to much from assuming that the causal relations are linear, stationary, unidirectional, one to one, and under various strong statistical assumptions. As a result, causal discovery methods would have no chance in discovering more general laws, such as maxwell equations. This is in contrast to fields such as system identification, where specifying the exact nonlinear differential equations that relate the variables is the norm [171]. For example, in our experiment, the expectation that each of the SNPs that influence the adverse drug reaction would have an association with the reaction that can be described independent of other causes, is absurd in principle.

These deficiencies showed themselves in our experiments. For example, the well know methods such as FCI and LiNGAM [177] would identify very different relations based on addition or removal of few samples, and sometimes discover obviously spurious relations such as $Age \rightarrow SNP \rightarrow Sex$. These issues are even more relevant to high dimensional causal discovery, including the methods proposed in this thesis, and could compound issues of the already plagued by publication bias and selection bias research community.

Two possible solutions to these issues are:

- Demanding more from the data and less from the assumptions. Providing strict conservative guidelines for causal discovery methods and embracing the garbage in, garbage out attitude would help to establish causal discovery's reputation in the scientific community not as 'just a fancier additive regression', but rather the best possible analysis for reputable experiments.

- Using HCDs in meta-analysis. Replicating existing discoveries not only could help evaluating which of the causal discovery methods are reliable and based on sensible model and assumptions, it can also be invaluable in triangulation as well as shedding new lights regarding the discovered relations.

These issues and potential solutions are also discussed for the broader replication crisis [201, 130, 82], and addressing them demand a shift in attitude in design and usage of methods. Only time will tell how useful high dimensional causal discovery will be for knowledge discovery.

Bibliography

[1] Ahmed Abbas, Xin-Bing Kong, Zhi Liu, Bing-Yi Jing, and Xin Gao. Automatic peak selection by a benjamini-hochberg-based algorithm. *PloS one*, 8(1), 2013.

[2] Janos Abonyi and Ferenc Szeifert. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, 24(14):2195–2207, 2003.

[3] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[4] Charu C Aggarwal, Mansurul A Bhuiyan, and Mohammad Al Hasan. Frequent pattern mining algorithms: A survey. In *Frequent pattern mining*, pages 19–64. Springer, 2014.

[5] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[6] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

[7] Folefac Aminkeng, Colin JD Ross, Shahrad R Rassekh, Soomi Hwang, Michael J Rieder, Amit P Bhavsar, Anne Smith, Shubhayan Sanatani, Karen A Gelmon, Daniel Bernstein, et al. Recommendations for genetic testing to reduce the incidence of anthracycline-induced cardiotoxicity. *British journal of clinical pharmacology*, 82(3):683–695, 2016.

[8] Muhammad Ammad-ud din, Suleiman A. Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 2016.

[9] David R Anderson, Kenneth P Burnham, and William L Thompson. Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management*, pages 912–923, 2000.

[10] Andreas Antoniou and Wu-Sheng Lu. *Practical optimization: algorithms and engineering applications.* Springer Science & Business Media, 2007.

[11] Barak Ariel and David P Farrington. Randomized block designs. In *Handbook of quantitative criminology*, pages 437–454. Springer, 2010.

[12] Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM, 2007.

[13] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[14] Jane PF Bai and Darrell R Abernethy. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annual review of pharmacology and toxicology*, 53:451–473, 2013.

[15] Yaneer Bar-Yam. General features of complex systems. *Encyclopedia of Life Support Systems (EOLSS), UNESCO, EOLSS Publishers, Oxford, UK*, 1, 2002.

[16] Julia M Barbarino, Michelle Whirl-Carrillo, Russ B Altman, and Teri E Klein. Pharmgkb: A worldwide resource for pharmacogenomic information. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 10(4):e1417, 2018.

[17] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

[18] Adam B Barrett, Lionel Barnett, and Anil K Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, 2010.

[19] Larry M Bartels. Instrumental and" quasi-instrumental" variables. *American Journal of Political Science*, pages 777–800, 1991.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[21] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

[22] Josette Bettany-Saltikov and Victoria Jane Whittaker. Selecting the most appropriate inferential statistical test for your quantitative research study. *Journal of Clinical Nursing*, 23(11-12):1520–1531, 2014.

[23] Hans-Peter Blossfeld and Götz Rohwer. Causal inference, time and observation plans in the social sciences. *Quality and quantity*, 31(4):361–384, 1997.

[24] A Bonin, E Bellemain, P Bronken Eidesen, F Pompanon, C Brochmann, and P Taberlet. How to track and assess genotyping errors in population genetics studies. *Molecular ecology*, 13(11):3261–3273, 2004.

[25] Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1435–1444, 2016.

[26] Steven L Bressler and Anil K Seth. Wiener–granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.

[27] Časlav Brukner. Quantum causality. *Nature Physics*, 10(4):259–263, 2014.

[28] Van Butsic, David J Lewis, Volker C Radeloff, Matthias Baumann, and Tobias Kuemmerle. Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, 19:1–10, 2017.

[29] Mei-Chun Cai, Quan Xu, Yan-Jing Pan, Wen Pan, Nan Ji, Yin-Bo Li, Hai-Jing Jin, Ke Liu, and Zhi-Liang Ji. Adrecs: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic acids research*, 43(D1):D907–D913, 2015.

[30] Donald T Campbell and Julian C Stanley. *Experimental and quasi-experimental designs for research*. Ravenio Books, 2015.

[31] BC Carleton, RL Poole, MA Smith, JS Leeder, R Ghannadan, CJD Ross, MS Phillips, and MR Hayden. Adverse drug reaction active surveillance: developing a national network in canada's children's hospitals. *Pharmacoepidemiology and drug safety*, 18(8):713–721, 2009.

[32] Sezen Cekic, Didier Grandjean, and Olivier Renaud. Time, frequency, and time-varying granger-causality measures in neuroscience. *Statistics in medicine*, 37(11):1910–1931, 2018.

[33] Christopher F Chabris, James J Lee, David Cesarini, Daniel J Benjamin, and David I Laibson. The fourth law of behavior genetics. *Current directions in psychological science*, 24(4):304–312, 2015.

[34] Wan-Chun Chang, Reo Tanoshima, Colin JD Ross, and Bruce C Carleton. Challenges and opportunities in implementing pharmacogenetic testing in clinical settings. *Annual Review of Pharmacology and Toxicology*, 61:65–84, 2021.

[35] Duncan D Chaplin, Thomas D Cook, Jelena Zurovac, Jared S Coopersmith, Mariel M Finucane, Lauren N Vollmer, and Rebecca E Morris. The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2):403–429, 2018.

[36] Kunjin Chen, Caowei Huang, and Jinliang He. Fault detection, classification and location for transmission lines and distribution systems: a review on the methods. *High voltage*, 1(1):25–33, 2016.

[37] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[38] Zhitang Chen, Kun Zhang, and Laiwan Chan. Nonlinear causal discovery for high dimensional data: A kernelized trace method. In *2013 IEEE 13th International Conference on Data Mining*, pages 1003–1008. IEEE, 2013.

[39] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[40] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[41] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48, 2010.

[42] Alberto Chong and Cesar Calderon. Causality and feedback between institutional measures and economic growth. *Economics & Politics*, 12(1):69–81, 2000.

[43] Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

[44] Krzysztof J Cios, Witold Pedrycz, and Roman W Swiniarski. Data mining and knowledge discovery. In *Data mining methods for knowledge discovery*, pages 1–26. Springer, 1998.

[45] Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

[46] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[47] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[48] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[49] Thomas D Cook, Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA, 2002.

[50] Elena-Ivona Dumitrescu and Christophe Hurlin. Testing for granger non-causality in heterogeneous panels. *Economic modelling*, 29(4):1450–1460, 2012.

[51] Emile Durkheim. *The rules of sociological method: and selected texts on sociology and its method*. Simon and Schuster, 2014.

[52] Michael Eichler. *Causal inference in time series analysis*. na, 2012.

[53] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

[54] Markus I Eronen. Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59:100785, 2020.

[55] Morten W Fagerland, Stian Lydersen, and Petter Laake. The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, 13(1):1–8, 2013.

[56] Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research*, 17(4):347–388, 2008.

[57] Kristen M Fedak, Autumn Bernal, Zachary A Capshaw, and Sherilyn Gross. Applying the bradford hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging themes in epidemiology*, 12(1):1–9, 2015.

[58] Ronald Aylmer Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936.

[59] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[60] Xiaoyu Ge, Vineet K Raghu, Panos K Chrysanthis, and Panayiotis V Benos. Causalmgm: an interactive web-based causal discovery tool. *Nucleic acids research*, 48(W1):W597–W602, 2020.

[61] Edmund A Gehan. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224, 1965.

[62] Nils Gehlenborg, Seán I O'donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, et al. Visualization of omics data for systems biology. *Nature methods*, 7(3s):S56, 2010.

[63] Carrie Anna Geisberg and Douglas B Sawyer. Mechanisms of anthracycline cardiotoxicity and strategies to decrease cardiac damage. *Current hypertension reports*, 12(6):404–410, 2010.

[64] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[65] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906, 2015.

[66] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[67] Greta Grassmann. New considerations on the validity of the wiener-granger causality test. *Heliyon*, 6(10):e05208, 2020.

[68] Sander Greenland, Judea Pearl, and James M Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

[69] R Scott Hacker and Abdulnasser Hatemi-J. Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application. *Applied Economics*, 38(13):1489–1500, 2006.

[70] Margaret A Handley, Courtney R Lyles, Charles McCulloch, and Adithya Cattamanchi. Selecting and improving quasi-experimental designs in effectiveness and implementation research. *Annual review of public health*, 39:5–25, 2018.

[71] Hossein Hassani, Xu Huang, and Mansi Ghodsi. Big data and causality. *Annals of Data Science*, 5(2):133–156, 2018.

[72] David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Computation, causation, and discovery*, 19:141–166, 1999.

[73] Paul Helman, Robert Veroff, Susan R Atlas, and Cheryl Willman. A bayesian network classification methodology for gene expression data. *Journal of computational biology*, 11(4):581–615, 2004.

[74] Miguel Henry and George Judge. Permutation entropy and information recovery in nonlinear dynamic economic time series. *Econometrics*, 7(1):10, 2019.

[75] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525, 2011.

[76] Austin Bradford Hill. The environment and disease: association or causation?, 1965.

[77] Meng Hu and Hualou Liang. A copula approach to assessing granger causality. *NeuroImage*, 100:125–134, 2014.

[78] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[79] David Hume. *A treatise of human nature*. Courier Corporation, 2003.

[80] Yi-Ting Hwang, Hsun-Chih Kuo, Chun-Chao Wang, and Meng Feng Lee. Estimating the number of true null hypotheses in multiple hypothesis testing. *Statistics and Computing*, 24(3):399–416, 2014.

[81] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[82] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[83] S Claiborne Johnston, John D Rootenberg, Shereen Katrak, Wade S Smith, and Jacob S Elkins. Effect of a us national institutes of health programme of clinical trials on public health and costs. *The Lancet*, 367(9519):1319–1327, 2006.

[84] Lindsay Judson. Aristotle's physics: A collection of essays. *Clarendon Press*, 1991.

[85] Steven A Julious and Mark A Mullee. Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481, 1994.

[86] Segun Jung, Yingtao Bi, and Ramana V Davuluri. Evaluation of data discretization methods to derive platform independent isoform expression signatures for multi-class tumor subtyping. *BMC genomics*, 16(S11):S3, 2015.

[87] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[88] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 01 2000.

[89] Ronald Kessler and David F Greenberg. Linear panel models. *New York: Academic*, 1981.

[90] Sahand Khakabimamaghani, Yogeshwar D Kelkar, Bruno M Grande, Ryan D Morin, Martin Ester, and Daniel Ziemek. Substra: Supervised bayesian patient stratification. *Bioinformatics*, 35(18):3263–3272, 2019.

[91] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

[92] Robert Koch. An address on cholera and its bacillus. *British medical journal*, 2(1236):453, 1884.

[93] Alexander Krauss. Why all randomised controlled trials produce biased results. *Annals of medicine*, 50(4):312–322, 2018.

[94] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[95] Vincenzo Lagani, Sofia Triantafillou, Gordon Ball, Jesper Tegner, and Ioannis Tsamardinos. Probabilistic computational causal discovery for systems biology. In *Uncertainty in Biology*, pages 33–73. Springer, 2016.

[96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[97] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

[98] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

[99] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[100] Johannes Lenhard. Models and statistical inference: The controversy between fisher and neyman–pearson. *The British journal for the philosophy of science*, 57(1):69–91, 2006.

[101] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[102] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–27, 2015.

[103] Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu. Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):257–271, 2016.

[104] Maxwell W Libbrecht, Oscar L Rodriguez, Zhiping Weng, Jeffrey A Bilmes, Michael M Hoffman, and William Stafford Noble. A unified encyclopedia of human functional dna elements through fully automated annotation of 164 human cell types. *Genome biology*, 20(1):180, 2019.

[105] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[106] Yan Liu, Jayant R Kalagnanam, and Oivind Johnsen. Learning dynamic temporal graphs for oil-production equipment monitoring system. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2009.

[107] Yan Liu, Alexandru Niculescu-Mizil, Aurelie Lozano, and Yong Lu. Temporal graphical models for cross-species gene regulatory network discovery. *Journal of bioinformatics and computational biology*, 9(02):231–250, 2011.

[108] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[109] Saisai Ma, Jiuyong Li, Lin Liu, and Thuc Duy Le. Mining combined causes in large data sets. *Knowledge-Based Systems*, 92:104–111, 2016.

[110] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[111] Chris A Mack. Fifty years of moore's law. *IEEE Transactions on semiconductor manufacturing*, 24(2):202–207, 2011.

[112] Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

[113] Subramani Mani, Peter L Spirtes, and Gregory F Cooper. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*, 2012.

[114] Mehrdad Mansouri, Sahand Khakabimamaghani, Leonid Chindelevitch, and Martin Ester. Aristotle: Stratified causal discovery for omics data. In *Submitted to BMC Bioinformatics*, 2021.

[115] Mehrdad Mansouri, Bowei Yuan, Colin JD Ross, Bruce C Carleton, and Martin Ester. Hume: large-scale detection of causal genetic factors of adverse drug reactions. *Bioinformatics*, 34(24):4274–4283, 2018.

[116] Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1999.

[117] Abul MM Masih and Rumi Masih. Energy consumption, real income and temporal causality: results from a multi-country study based on cointegration and error-correction modelling techniques. *Energy economics*, 18(3):165–183, 1996.

[118] Lawrence S Mayer. On cross-lagged panel models with serially correlated errors. *Journal of Business & Economic Statistics*, 4(3):347–357, 1986.

[119] Mariusz Maziarz. A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105, 2015.

[120] John V McGowan, Robin Chung, Angshuman Maulik, Izabela Piotrowska, J Malcolm Walker, and Derek M Yellon. Anthracycline chemotherapy and cardiotoxicity. *Cardiovascular drugs and therapy*, 31(1):63–75, 2017.

[121] MF McGranaghan, RC Dugan, and WL Sponsler. Digital simulation of distribution system frequency-response characteristics. *IEEE Transactions on Power Apparatus and Systems*, 1(3):1362–1369, 1981.

[122] Marie C McGraw and Elizabeth A Barnes. Memory matters: a case for granger causality in climate variability studies. *Journal of Climate*, 31(8):3289–3300, 2018.

[123] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1):122, Jun 2016.

[124] Marcia L Meldrum. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4):745–760, 2000.

[125] Olli S Miettinen. The matched pairs design in the case of all-or-none responses. *Biometrics*, pages 339–352, 1968.

[126] John Stuart Mill. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation.* Longmans, green, and Company, 1889.

[127] Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems*, pages 6986–6994, 2018.

[128] Dan Mønster, Riccardo Fusaroli, Kristian Tylén, Andreas Roepstorff, and Jacob F Sherson. Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence. *Future Generation Computer Systems*, 73:52–62, 2017.

[129] Debashree Mukherjea and Leonard P Rybak. Pharmacogenomics of cisplatin-induced ototoxicity. *Pharmacogenomics*, 12(7):1039–1050, 2011.

[130] Marcus R Munafò and George Davey Smith. Robust research needs many lines of evidence, 2018.

[131] Kevin Murphy, Saira Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Citeseer, 1999.

[132] Volker Nannen. A short introduction to model selection, kolmogorov complexity and minimum description length (mdl). *arXiv preprint arXiv:1005.2364*, 2010.

[133] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

[134] Alexey I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, 73(11):2092–2123, 2010.

[135] Jeremy K Nicholson. Reviewers peering from under a pile of 'omics' data. *Nature*, 440(7087):992–992, 2006.

[136] Bo Ning, Subhashis Ghosal, Jewell Thomas, et al. Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis*, 14(1):1–28, 2019.

[137] Sam Norton, Fiona E Matthews, Deborah E Barnes, Kristine Yaffe, and Carol Brayne. Potential for primary prevention of alzheimer's disease: an analysis of population-based data. *The Lancet Neurology*, 13(8):788–794, 2014.

[138] Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.

[139] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379. PMLR, 2016.

[140] Mohammed Ombadi, Phu Nguyen, Soroosh Sorooshian, and Kuo-lin Hsu. Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resources Research*, 56(7):e2020WR027251, 2020.

[141] World Health Organization et al. Risk reduction of cognitive decline and dementia: Who guidelines. In *Risk reduction of cognitive decline and dementia: WHO guidelines*, pages 401–401. World Health Organization, 2019.

[142] Judea Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271, 1988.

[143] Judea Pearl. Myth, confusion, and science in causal analysis. *Statistics in Medicine*, 2009.

[144] Judea Pearl. Remarks on the method of propensity score. *escholarship*, 2009.

[145] Judea Pearl. Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference*, 6(2), 2018.

[146] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[147] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[148] Charles S Peirce. A theory of probable inference. *American Psychological Association*, 1883.

[149] Charles S Peirce. Note on the theory of the economy of research. *Operations Research*, 15(4):643–648, 1967.

[150] Charles Sanders Peirce and Joseph Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3, 1884.

[151] Leonardo Pellegrina and Fabio Vandin. Efficient mining of the most significant patterns with permutation testing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2070–2079. ACM, 2018.

[152] Leonardo Pellegrina and Fabio Vandin. Efficient mining of the most significant patterns with permutation testing. *Data Mining and Knowledge Discovery*, 34:1201–1234, 2020.

[153] Jose M Pena, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, 2007.

[154] Thomas F Pettigrew and Miles Hewstone. The single factor fallacy: Implications of missing critical variables from an analysis of intergroup contact theory1. *Social Issues and Policy Review*, 11(1):8–37, 2017.

[155] Carl V Phillips and Karen J Goodman. Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerging themes in epidemiology*, 3(1):1–7, 2006.

[156] Paul Price. *Research Methods in Psychology, 2nd Canadian Edition.* BCcampus, 2015.

[157] 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(2):121–129, 2017.

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

[159] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

[160] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[161] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[162] Ellen V Rothenberg. Causal gene regulatory network modeling and genomics: Second-generation challenges. *Journal of Computational Biology*, 26(7):703–718, 2019.

[163] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

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

[165] Donald B Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

[166] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

[167] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.

[168] Babak Salimi, Corey Cole, Dan RK Ports, and Dan Suciu. Zaliql: causal inference from observational data at scale. *Proceedings of the VLDB Endowment*, 10(12):1957–1960, 2017.

[169] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W Green, and Cate D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2):155–161, 2007.

[170] Glenn N Saxe, Alexander Statnikov, David Fenyo, Jiwen Ren, Zhiguo Li, Meera Prasad, Dennis Wall, Nora Bergman, Ernestine C Briggs, and Constantin Aliferis. A complex systems approach to causal discovery in psychiatry. *PloS one*, 11(3):e0151174, 2016.

[171] Johan Schoukens and Lennart Ljung. Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, 39(6):28–99, 2019.

[172] Daniel Schwartz and Joseph Lellouch. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases*, 20(8):637–648, 1967.

[173] James P Selig and Todd D Little. Autoregressive and cross-lagged panel analysis for longitudinal data. *Handbook of developmental research methods*, 2012.

[174] Hossein Sharifi-Noghabi, Shuman Peng, Olga Zolotareva, Colin C Collins, and Martin Ester. Aitl: Adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*, 36(Supplement_1):i380–i388, 2020.

[175] Aman Sharma and Rinkle Rani. Be-dti': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. *Computer methods and programs in biomedicine*, 165:151–162, 2018.

[176] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[177] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[178] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2):163–192, 2000.

[179] Du Sizhen, Song Guojie, Han Lei, and Hong Haikun. Temporal causal inference with time lag. *Neural Computing*, 30:271–291, 2018.

[180] Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, 47(5):550–554, 2015.

[181] Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

[182] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[183] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000.

[184] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

[185] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23:1687–1695, 2010.

[186] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[187] Andrew Steptoe, Elizabeth Breeze, James Banks, and James Nazroo. Cohort profile: the english longitudinal study of ageing. *International journal of epidemiology*, 42(6):1640–1648, 2013.

[188] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544, 2004.

[189] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[190] Elizabeth A Stuart and Donald B Rubin. Best practices in quasi-experimental designs. *Best practices in quantitative methods*, pages 155–176, 2008.

[191] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

[192] Gilbert Sybille. Use of surge arresters in transmission system. accessed: 11-November-2019.

[193] Gilbert Sybille and Hoang Le-Huy. Digital simulation of power systems and power electronics using the matlab/simulink power system blockset. In *2000 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 00CH37077)*, volume 4, pages 2973–2981. IEEE, 2000.

[194] Reo Tanoshima, Amna Khan, Agnieszka K Biala, Jessica N Trueman, Britt I Drögemöller, Galen EB Wright, Jafar S Hasbullah, Gabriella SS Groeneweg, Colin JD Ross, Bruce C Carleton, et al. Analyses of adverse drug reactions–nationwide active surveillance network: Canadian pharmacogenomics network for drug safety database. *The Journal of Clinical Pharmacology*, 59(3):356–363, 2019.

[195] Robert E Tarone. A modified bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

[196] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

[197] Hiro Y Toda and Taku Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1-2):225–250, 1995.

[198] Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003.

[199] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[200] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[201] Sebastian Wallot and Damian G Kelty-Stephen. Interaction-dominant causation in mind and brain, and its implication for questions of generalization and replication. *Minds and Machines*, 28(2):353–374, 2018.

[202] Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

[203] Changwon Yoo and Gregory F Cooper. Discovery of gene-regulation pathways using local causal search. In *Proceedings of the AMIA Symposium*, page 914. American Medical Informatics Association, 2002.

[204] Changhe Yuan and Brandon Malone. An improved admissible heuristic for learning optimal bayesian networks. *arXiv preprint arXiv:1210.4913*, 2012.

[205] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[206] Peng Zhang, Lin Tao, Xian Zeng, Chu Qin, Shangying Chen, Feng Zhu, Zerong Li, Yuyang Jiang, Weiping Chen, and Yu-Zong Chen. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *Briefings in bioinformatics*, 18(6):1057–1070, 2017.

[207] Qingrun Zhang, Quan Long, and Jurg Ott. Apriorigwas, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. *PLoS computational biology*, 10(6):e1003627, 2014.

[208] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[209] Mattia Zorzi and Alessandro Chiuso. Sparse plus low rank network identification: A nonparametric approach. *Automatica*, 76:355–366, 2017.