G-RIPS Sendai 2024

Fujitsu Group

Final Report

Authors:
John Forde[1]
Gaspar Mendez[2]
Akane Okubo[3]
Daniel Quigley[4]
Renji Sakamoto[5]

Mentors:
Fabiana Ferracina[+]
Jorge Gutierrez[*]
Hiroyuki Higuchi[*]

[1] Florida Atlantic University
[2] Tohoku University
[3] Tokyo University of Science
[4] University of Wisconsin-Milwaukee
[5] The University of Tokyo
[+] Academic Mentor,
[*] Fujitsu Industry Mentors.

August 8, 2024

風が吹けば桶屋が儲かる
Kaze ga fukeba, okeya ga moukaru.
`When the wind blows, the barrel-makers pro t.'

Japanese, Traditional

For want of a nail the shoe was lost.
For want of a shoe the horse was lost.
For want of a horse the rider was lost.
For want of a rider the message was lost.
For want of a message the battle was lost.
For want of a battle the kingdom was lost.
And all for the want of a horseshoe nail.

English, Traditional

Contents

1 Problem statement

`A is the cause of B', an otherwise innocuous statement, has important philosophical, technical, and non-technical consequences. Arti cial Intelligence (AI) is ubiquitous and permeates academic, industrial, and personal domains; the use of AI to identify and process the extent to which A causes B introduces interdisciplinary problems of interpretation, validation, and ethical consideration from domains such as (but certainly not limited to): mathematics; computer science; logic; philosophy; social science.

Causality (from the Latin *causa* `cause; reason') is the generic relationship between an e ect B and the cause A that gave rise to it [32]. The discovery of causal relationships (that is, discovering *that* A is the cause of B) is non-trivial. One approach: given a dataset, derive (read, estimate) a *single* causal structure [77, 81, 96]. Another approach: nd *all* important combinations from the dataset, and infer causal relations under those conditions [57, 61]. Real-world data, however, is often at least as simple as a Persian carpet: perhaps beautiful, but quite complex/complicated in its materials, patterns, and fashion made.

Complicated relationships of the form `A is the cause of B' quickly become even more so when the dataset is large; sifting through datasets using conditions is likewise large and sometimes unwieldy. The challenge, then, is: how to extract useful information from single causal graphs and across multiple graphs *e ectively* from such relationships, and how to explain these relationships *e ectively*.

This project contributes to the analysis and understanding of `A is the cause of B'. This is motivated by the fact that causal graph structures are inherently complicated objects ('spaghetti' graphs), and comparing between graph objects is likewise di cult; see Figure 1.

Figure 1: Example of `spaghetti' graphs.

This is done relative to three main design pillars: (1) convincingness; (2) variety; (3) discoverability.

De nition 1.1 (Convincingness). *Extent to which a suggested model/explanation matches (or exceeds) a user's expectation; claim: convincingness is* at least *related as explainability; shorter, simpler explanations may be considered convincing because they are easy to understand (Occam's Razor), but Hickam's Dictum may counter this notion (at least in the medical domain) [13, 25, 59]*

De nition 1.2 (Variety). *A set of unique `equally good' explanations/models (i.e., subsets of the Rashomon set)*

De nition 1.3 (Discoverability). *Extent to which there exist unexpected causal relations between features and outcomes; when an explanation is not convincing to a user, it is not always because of a bad explanation, it might be a new nding that the user never noticed*

De nition 1.4 (CVD). *Abbreviation for `convincingness, variety, and discoverability'; Figure 2 is a graphical representation of these components*

- *CV - abbreviation for `convincingness and variety'*
- *CD - abbreviation for `convincingness and discoverability'*
- *VD - abbreviation for `variety and discoverability'*

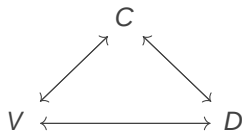

Figure 2: CVD triangle

An elaboration on Figure 2. We work with the following assumptions: convincingness goes at least as explainability; variety goes at least as the Rashomon set; discoverability goes at least as uncovering new or unexpected causal relationships. Consider how they interact:

- CV - we explore by explanation of elements of the Rashomon set with similarities
- CD - we explore by understanding how and why these causalities came to be through independent or directed evaluation of evidence
- VD - we explore by choosing elements from the Rashomon set that have new or unexpected results

From medicine to manufacturing, modern organizations and corporations across a variety of elds are becoming increasingly reliant on comprehensive data analysis to drive decision-making. However, to identify optimal and actionable strategies, a clear understanding of causal relations among data attributes is crucial. Software tools that elucidate causal relations are particularly of interest for observational datasets, as experimentation

to con rm results are often impossible, unethical, or illegal. For instance, it is not feasible to do a randomized controlled trial to test the e ect of implementing a new computer science course requirement in primary schools on the growth of the tech industry in a given area [12].

However, building causal models based on observational data (indicators of educational changes and vocational growth) over time is a much more pragmatic approach; causal models based on observational data *cannot* intervene to check if manipulated the cause changes the e ect [58]. Fortunately, advancements in the causal discovery algorithm space have made it possible to visualize and interpret causal relations among data attributes. However, a widely used approach, DirectLiNGAM, typically produces a single graph after many assumptions are applied. This presents a narrow view of a dynamic situation that can change based on certain conditions.

To circumvent this limitation, novel Wide Learning™ technology has made it possible to uncover causal relations for di erent combinations of attribute values, providing a decision-maker with the ability to explore causality in much greater detail [33]. However, while it is important to have many causal graphs to understand how causality changes under particular conditions, the variety and complexity of graphs are di cult to comprehend and compare.

There is currently no standard tool for the visualization and analysis of sets of causal graphs produced by Wide Learning™. Thus, the focus of this study will be to apply cogent design principles in the creation of a user interface that allows a decision-maker to understand and discover the causal relations in a dataset under di erent conditions.

Table 1: Potential applications of WL causal discovery.

Decision-Maker	Sample Objective
Biopharmaceutical Firm	Determine causality between lung cancer resistance and genes to inform novel immunotherapy R&D.
Chemical Manufacturing Company	Understand causal relations among catalysts and atoms to develop new ammonia synthesis methods.
Real Estate Developer	Rank geographical and demographic attributes that in uence median house value in order to guide development strategy.
Food and Beverage Manufacturer	Analyze determinants of co ee quality to develop a new QC protocol.

2 Background

The basic idea of symbolic AI is to learn, process, understand, and describe the world, its entities, and their relationships according to a formal language and by logical reasoning [49]. Symbolic AI is `high level' (human-readable/interpretable) [28]. Contrast this with deep learning, whose architectures otherwise obfuscate human-interpretability of their results and processes, among other inherent issues such as computability, bias, and

explanation [6, 51]. Two symbolic AI models relevant to this project are explored here: LiNGAM and Wide Learning™.

2.1 Causal discovery

A Pearlean description of causality is a **Structural Equation Model** (SEM) [64, 67, 80].

De nition 2.1 (Structural Equation Model). *A Structural Equation Model (SEM) is a quadruple of the form*

$$M_{SEM} = \langle U; V; F; P \rangle$$

where:

- *Set* $U = \{u_1; u_2; \dots; u_n\}$ *of* **exogenous variables** $V \setminus U = \emptyset$*, representing factors outside the model that a ect relationships within the model that are not caused by endogenous variables* V
- *Set* $V = \{v_1; v_2; \dots; v_n\}$ *of* **endogenous variables** $V \cap U$*, representing observed variables; each* V_i *is functionally dependent on a subset of* $U \cup V \setminus \{v_i\}$ *called the parents* PA *of* v_i
- *Set* $F = \{f_1; f_2; \dots; f_n\}$ *of* **functions** *over the variables such that each* f_i *determines the value of* $v_i \in V$ *by* $v_i = f_i(PA(v_i); u_i)$
- *Joint probability distribution* $P(U) = \prod_i P(u_i)$

This quadruple builds a **directed acyclic graph** (DAG), where $U; V$ are sets of nodes of the graph, and directed edges indicate causal relationships between nodes. **Causal discovery** is the recovery of these representations. Causal models of this form, and in causality more generally, need to account for three kinds of questions [8, 66]: prediction; intervention; counterfactual.

Acyclicity enforces that there is no directed path from a variable to itself; a feature, after all, is not its own cause and effect.

1. **Predictions**: will the sidewalk be slippery if we nd the sprinkler o ?
2. **Interventions**: will the sidewalk be slippery if we make sure that the sprinkler is o ?
3. **Counterfactuals**: would the pavement be slippery had the sprinkler been o given that the sidewalk is not slippery and the sprinkler is on?

Predictions are solved via deductive inference; interventions and counterfactuals, however, are solved via some interference with the model. This interference follows from the do operator $do(X = x)$, which simulates physical interference with some variable by deleting certain functions from the model and replacing them with a constant. These interventions induce a `submodel' M_x: given some model M with set of variables $X \subseteq V$, we say that M_x is obtained from M by replacing F with $F_x = \{f_i | v_i \notin X\} \cup \{(X = x)\}$; that is, we delete from F all f_i that correspond to elements of X and replace them with $X = x$. This may be interpreted as asking: `Whether $Y = y$ would hold, had X been x?'.

Interventions and counterfactuals also a ect the joint probability distribution P [64].

1. Post intervention: $P_M(y|\text{do}(x)) = P_{M_x}(y)$
2. Post counterfactual: $P_x(u_i) = P_{M_x}(u_x)$

So far, we have said nothing about the actual data with which we are making causal inferences and discoveries. At most, we have included the notion of probability, which `may be used to represent our uncertainty about the value of unobserved variables in a particular case, or the distribution of variable values in a population' [34]. We are interested in, therefore, the extent to which features of a causal structure can be identified from their respective probability distributions *in addition* to our assumptions and observations.

2.1.1 LiNGAM

[68] identify an interesting behavior and usefulness of the SEM, in that if $u_i \in U$ are *probabilistically independent*, that is, if $P(u_i)$ and $P(u_j)$ for $u_i, u_j \in U$ and $i \neq j$ do not influence each other, then the probability distribution on V satisfies the **Markov Condition** (MC): for any variable v_i, v_i is independent of all non-descendants of v_i given $PA(v_i)$; that is, the parents of v_i `block` v_i from other variables of the model, except descendents of v_i.

If we want to be technically verbose and indulge a little in the mathematics, MC may be stated in any of the following three equivalent (though with different perspectives) ways [65]:

1. The parents of v_i `block` v_i from other variables of the model, except descendents DE of v_i:

$$\forall v_i \in V, s_i \quad V \setminus DE(v_i), P(v_i | PA(v_i) \wedge s_i) = P(v_i | PA(v_i))$$

2. Once we know the conditional probability distribution of each variable given its parents, we compute joint distributions over all of the variables:

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i | PA(v_i))$$

3. For $X, Y \in V$ and $Z \subseteq V \setminus \{X, Y\}$, Z d-separates X and Y if every path $\{X_1, X_2, \ldots, X_n\}$ from a variable in X to a variable in Y contains at least one variable X_i such that:

   (a) X_i is a collider and no descendants of X_i is in Z

   (b) X_i is not a collider, and X_i is in Z

   If Z d-separates X and Y, then:

$$P(X, Y | Z) = P(X | Z) \quad P(Y | Z)$$

However, simply given a set V and a P on V, without further assumptions, we simply cannot identify the *unique* causal structure; at best, only the *set* of all causal structures (the set of DAGs) can be identified. This set is the **Markov equivalence class**.

Given an SEM $M = \langle U, V, F, P \rangle$, we must impose additional assumptions so as to not overgenerate Markov equivalent classes. What are those assumptions? At minimum, we can impose that V be either: (1) discrete; (2) continuous. If V is:

1. Discrete, and we make no assumptions about F, then we can do no better than Markov equivalence classes
2. Continuous, in which $v_i = \sum_j c_j v_j + u_i$ (j is over indices of $PA(v_i)$) and P assigns gaussian distribution to each u_i, then we can do no better than inferring Markov equivalence classes

Therefore, models such as LiNGAM [76] and other non-linear models [35, 99] impose additional assumptions/constraints to uniquely identify the causal structure from an otherwise overgenerated set.

Definition 2.2 (LiNGAM model). Together with a set of particular assumptions, a LiNGAM model is a tuple of the form:

$$M_{LiNGAM} = \langle D, SEM \rangle$$

where D is a dataset of the form $D = \{(x_i, y_i)\}_{i=1}^{N}$ and SEM is a Structural Equation Model as defined in Definition 2.1.

The assumptions mentioned in Definition 2.2 required for a LiNGAM model are given below; additionally, assumptions for other non-linear models are likewise given. Note that different species of LiNGAM models may have more assumptions than are listed here.

1. LiNGAM model assumptions [76]:
   - V is continuous
   - f_i is linear: $v_i = f_i(PA(v_i), u_i)$
   - P over U is not gaussian
   - U is probabilistically independent in P
2. Non-linear models assumptions [35, 99]:
   - f_i is nonlinear; additive u_i: $v_i = f_i(PA(v_i)) + u_i$ [35]
   - U is probabilistically independent in P
   - f_i, g_i are nonlinear; additive u_i: $v_i = g_i(f_i(PA(v_i)) + u_i)$ [99]
   - U is probabilistically independent in P

With the exception of five specific cases that cannot be fully specified by [99], the correct DAG is identifiable.

Given the above assumptions, knowing only the probability distribution on two variables, we can infer whether one causes the other. If the above assumptions are imposed, then the correct DAG on V can be uniquely determined by the induced probability distribution P on V.

2.1.2 Wide Learning

Wide Learning —(WL) is a symbolic classification machine learning model [61] for causal discovery developed by the Artificial Intelligence Laboratory, Fujitsu Limited, Kawasaki, Japan that generates representations that explain causal relationships between features; as a classification model in machine learning (ML), it is an extension of logistic regression [57].

De nition 2.3 (Wide Learning — model). WL is a model M tuple of the form

$$M_{WL} = \langle D; S; \{I; H\} \rangle$$

where D is a dataset of the form $D = \{(x_i; y_i)\}_{i=1}^{N}$, S is a subset of the powerset of indexed features, and $\{I; H\}$ is an evaluation (either mutual information or entropy) to determine which combinations of features contribute best in a condition.

Or, more abstractly: WL is a model M tuple of the form

$$M_{WL} = \langle \text{features}; \text{combinator}; \text{evaluation} \rangle$$

where features is a feature set of data, combinator is a process of nding all possible combinations of features from the feature set, and evaluation is a metric for determining which features contribute best.

WL works in the following way (assuming a classi cation task): given a dataset $D = \{(x_i; y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^n$ is assigned to one of m classes and $y_i \in \{0; 1\}^m = \{(y_1; \ldots; y_m) \in [0; 1]^m \mid \sum_{i=1}^{m} y_i = 1\}$, nd all possible combinations of x_i that `contribute' to y. Once the model has learned which are the appropriate combinations of features that `contribute' to y, apply the model to novel data as you would usually with a classi cation ML model.

Extant available examples of WL assume a binary classi cation task of the form $y_i \in \{0; 1\}$; in principle, expanding to n-ary classi cation should follow similarly: $y_i \in \{0; 1\}^m$.

The outline for WL is described below:

1. For features x_S, where $S \subseteq \{1; 2; \ldots; n\}$ and $|S| \leq K$, we have for a subset S, $x_S = \{x_{S_1}; x_{S_2}; \ldots; x_{S_{|S|}}\}$, where S_j are indices in S. Note that K is chosen to limit the number of features that contribute to the causal graph; K is typically maxed to 4 (see Section 6.1).

2. Exhaustively combining features follows as: let $\wp_K(\{1; 2; \ldots; n\})$ be the power set of indices restricted to sets of size at most K such that each subset $S \in \wp_K$ de nes a combination of features $x_S = \{x_{S_1}; x_{S_2}; \ldots; x_{S_{|S|}}\}$. Then for each combination, compute a relevance metric to measure the extent to which those features `contribute' to y.

Mutual information checks the reduction in uncertainty of the a liation of a variable randomly drawn from data, if we know the identity of the variables with which it interacts; entropy only checks the information intrinsic to the distributions of the random variables separately [17, 91].

3. To check the extent to which a combination `contributes` to y, we may measure via mutual information or entropy [57]; the former measures the information that one random variable holds about another, while the latter measures the information content of a single random variable (an elaboration is given after this outline).

   (a) Assuming mutual information:

$$I(x_S; y) = \sum_{x_s \in Values(x_S)} \sum_{y \in \{0;1\}^m} P(x_S = x_s; y) \log \frac{P(x_S = x_s; y)}{P(x_S = x_s)P(y)}$$

   Note that to determine if the distribution of x_S di ers at all signi cantly across any of y, we check which combinations exceed some critical value from which we retain such combinations of x_S that do so. χ^2 checks for this di erence.

(b) Assuming entropy:

We compute the entropy H for each feature value for individual information content, which gives an illustration of the variability of each feature. Note that k counts m class labels; because WL checks combinations of features, we calculate the entropy of subsets x_S for $S \subseteq \{1, 2, \ldots, n\}$:

$$H(x_S) = \sum_{x_s \in Values(x_S)} P(x_S = x_s) \log(P(x_S = x_s))$$

$$H(y) = \sum_{k=1}^{m} P(y = k) \log(P(y = k))$$

Conditional entropy then checks the information that a feature x contains about y. This gives the uncertainty of that y given knowledge of x.

$$H(y|x_S) = \sum_{x_s \in Values(x_S)} \sum_{k=1}^{m} P(x_S = x_s, y = k) \log(P(y = k|x_S = x_s))$$

The difference $I(x_S) = H(y) - H(y|x_S)$ gives the information gain from knowing x_S, which, in turn, tells us how x_S 'effects' the uncertainty of y.

4. Whether we evaluate by mutual information or by entropy, we export the values into a graph structure; the features that contribute to some y define a condition, a directed graph representation of causality in WL. For a condition in WL, weights w_S are associated with $v \in V$, and follow from statistically significant combinations labeled from the feature set x_S; directed edges w_{ij} follow from conditional dependencies such that (v_i, v_j) exists if:

   Recall, a directed graph is a tuple of the form $G = \langle V, E \rangle$, where V is a set of nodes (vertices) $v \in V$ and E is a set of edges (v_i, v_j).

   - assuming mutual information: $I(x_{S_j}, y|x_{S_i}) > I(x_{S_j}, y)$ and $S_i \subset S_j$ (in which 'knowing' x_S improves predictiveness of x_{S_j} regarding y); the edge weights follow from the predictive power gained from x_{S_j} when x_{S_i} is known.
   - assuming entropy: $H(y|x_{S_i}) > H(y|x_{S_i}, x_{S_j})$ and $S_i \subset S_j$ (in which 'knowing' x_S improves predictiveness of x_{S_j} regarding y); the edge weights follow from the predictive power gained from x_{S_j} when x_{S_i} is known.

5. Once a set of conditions is learned, we check it against novel data, and assign classification accordingly.

   (a) We must convert the linear combination of values with an interaction term (which we assume is linear, but leave as a more general form) into a probability. We assume this is done via a sigmoid activation function, which essentially combines the contributions from nodes deemed relevant into the appropriate probability:

$$P(y = 1|x) = \left(\sum_{S \subseteq \{1,2,\ldots,n\}, |S| \leq K} w_S \phi(x_S) \right)$$

(b) When predicting the feature combinations and classification of a novel dataset, we choose some arbitrary threshold, where:

$$P(y = 1 | x_{new}) = \left(\sum_{S \subseteq \{1,2,\ldots,n\}, |S| \leq K} w_S \, (x_{S;new}) \right)$$

Threshold for binary classification is taken to be 0.5; for m classes of classification, thresholds may be decided accordingly.

(c) If $P(y = 1 | x_{new}) \geq 0.5$, then classification is $y = 1$, and if $P(y = 1 | x_{new}) < 0.5$, then classification is $y = 0$ (or vice versa, depending on the classification task and design).

A few remarks on the choice of mutual information and entropy (note that the choice of the former follows from first principles in deriving and describing the model, while the latter follows from a reference in [57]). The following may be found in standard references on information theory [3, 17, 50, 82]. At their most elementary, mutual information checks interdependencies and interactions between variables; entropy checks intrinsic properties of data distributions, identifying features with high variability or unpredictability. Mutual information is useful in feature selection (it checks the relationship(s) between variables), hence its inclusion here; entropy is useful in feature filtering (it checks each feature alone for how much inherent information it contains, such that (for example) features with low entropy (less variability; more predictable) might carry less information and thus could be less useful).

Laconically: mutual information cares about relationships between variables; entropy cares about information content and diversity within single variables.

2.2 Rashomon sets

The Rashomon effect, named for the Japanese film Rashomon (1950; dir. Akira Kurosawa), is a phenomenon in which there exists a multitude of different descriptions in a class of functions giving about the same error rate [9]. The Rashomon set is the set of these all almost-optimal models.

The Rashomon effect helps us understand that there is not just one `best' explanation for the data, but many diverse equally predictive models. Almost all current algorithms return only one model which might be complicated. However, it is important for practitioners to find a simpler model. As in [9], if we try to reduce the number of variables to make the model simple and conduct the linear regression, then we can find totally different models with about equal accuracy. If the Rashomon set is large, the Rashomon set could contain many accurate and simple models, and the learning problem becomes simpler. On the other hand, a harder learning problem emerges in the case of few deep and narrow local minima. We need several techniques to calculate the full Rashomon set because both the hypothesis set and the Rashomon set are too large in general. For example [95] use analytical bounds to prove that large portions of the search space do not contain any members of the Rashomon set and it permits memory-efficient storage and easy indexing of the Rashomon set's members.

In what follows, we define the Rashomon set. Let $D = \{x_i, y_i\}_{i=1}^n \subseteq X \times Y$ be a training data set, where $x_i \in X$ are inputs and $y_i \in Y$ are outputs. We consider a hypothesis space F and a loss function $\ell : Y \times Y \to \mathbb{R}^+$. We use the notation $l(f; x_i, y_i) := \ell(f(x_i), y_i)$ to take $f \in F$ explicitly as an argument. By using the loss function, we can define the

(empirical) risk as $L(f) = \frac{1}{n} \sum_{i=1}^{n} l(f; x_i, y_i)$. Then we have the following definition of the Rashomon set.

Definition 2.4 (Rashomon set). For each $\theta > 0$, the (empirical) Rashomon set is defined as follows:

$$R_{set}(F, \theta) := \{f \in F : L(f) - L(\hat{f}) + \theta\},$$

where $\hat{f}$ is a empirical risk minimizer for the training data set S.

Here is a question: Can we take an accurate-yet-simple model from the Rashomon Set?

To answer this problem, we can use the notion of the Rashomon ratio [75]. The Rashomon ratio is the ratio of the volume of the Rashomon set to the volume of the hypothesis space. Since both the Rashomon set and the hypothesis set are huge, we cannot calculate it directly in general. On the other hand, we can estimate it by sampling from decision trees of bounded depth. In [75], the bounding depth is chosen to be seven. In most cases, a large Rashomon ratio implies that a group of machine learning models perform similarly and generalize results [75]. See the example below for an illustration of this behavior (test accuracy is the evaluation metric).

Figure 3: Illustration of similar model performance when Rashomon ratio is large [75].

2.3 Stochastic and statistical aspects

SEM is a statistical analysis method for modelling and validating relationships between variables proposed as hypotheses; DirectLiNGAM[77] is one such method, which exploits non-Gaussianity to allow the causal order of variables and their connection strengths without prior knowledge of the network structure. A linear acyclic model can be constructed that specifies the causal order of the variables and the strength of their connections. This section provides some background on the goodness of fit of the models created by DirectLiNGAM. It is considered that examining the goodness of fit of a model can provide the user with not only convincing but also discoverability.

There are various goodness-of- t indices for models: the CFI, RMSEA, AIC, ^{2}, BIC, GFI, AGFI and NFI. Of these, the rst three are mentioned here.

- CFI - the comparative t index.
- AIC - Akaike information criterion.
- RMSEA - the root mean square error of approximation.

DirectLiNGAM is implemented in the LiNGAM package, which is open source and therefore available to everyone. To check the model t, the evaluatemodelL t function in the package was used. This function is available bysemopy packagehttps://gitlab.com/georgy.m/semopy in python. The following equations are taken from a paper[40],

$$\text{CFI} = 1 \quad \frac{^2_m \quad df_m}{r \; ^2_b \quad df_b}$$

$$\text{RMSEA} = \quad \frac{^2 = df \quad 1}{n \quad 1}$$

$$\text{AIC} = 2(k \quad L):$$

According to the paper[40], n is a number of data samples,$F(\hat{})$ is a value that objective function attains at optimum, $^2 = nF(\hat{})$, ^2_m is a ^{2} statistics for the target model, ^2_b is a ^{2} statistics for the baseline model, wheredf_m is df of target model and df_b is df of baseline model,k is a number of parameters andL is a value of a likelihood function. Here df is a degree of freedom metric, such as$df = \frac{k(k+1)}{2}$ m, where k is a number of observed variables andm is a number of parameters. See [40] for details.

2.4 Label-sample rate

We do not want to let users overwhelmed by the amount of information and various conclusions from causal graphs, which can be sometimes opposite. Therefore, we consider ordering causal graphs by some indices. It may reduce feeling overwhelmed by amount of information and various conclusions from causal graphs.

In what follows, we propose graph ordering by the index rate. LetX be a condition with sample number n and label 1 number Q, where label 1 number is the number of samples whose binary output variable is equal to 1. The rate of the conditionX is de ned as $\frac{Q}{n}$. The index rate could help with graph ordering. However, there is a problem, i.e., the population size can in ate the index to be very high if the sample size is low, or vice-versa. To deal with this problem, we propose other alternative modi ed rates. The rst one is de ned as follows.

$$R_1(X) = \frac{Q}{n} \quad \frac{1}{n}; \tag{1}$$

where > 0 is an optional coe cient. Here, $\frac{1}{n}$ is the penalty term. As the sample number gets lower,$R_1(X)$ tend to be smaller than the normal rate. On the other hand, it is controversial that the penalty imposed on the rate is completely proportional to $\frac{1}{n}$.

If the number of sample is enough, we do not have to impose any penalty on the rate. This is the reason why we have another alternative rate.

$$R_2(X) = \begin{cases} \frac{Q}{n} & (n \quad) \\ \frac{Q}{n} \quad \frac{n}{\quad} & (\quad n < \quad) \\ 0 & (n < \quad) \end{cases}; \tag{2}$$

We choose and in the de nition as follows; if n , then n is enough large to create an accurate model, and if n < , then the consequent causal graph is not reliable at all. (,) can be (100, 1), (720, 72) [69], and so on, depending on the situation.

condition	sample	label 1	R (rate)	R_1	R_2	R R_1	R R_2
1	5039	3243	0.643	0.623	0.643	0.019	0
2	3787	2519	0.665	0.638	0.665	0.026	0
3	675	372	0.551	0.402	0.512	0.148	0.038
4	342	222	0.649	0.356	0.270	0.292	0.378
5	188	121	0.643	0.111	0.115	0.531	0.528
6	80	65	0.812	-0.437	0.010	1.250	0.802
7	72	72	1.000	-0.388	0	1.388	1.000

Table 2: The behavior of the alternative rates with = 100, = 720, = 72. If n is larger than , R_2 is equal to the existing rate. If n is smaller than , R_2 is totally equal to 0.

Although we proposed the alternative rates, they are not incorporated into our interface. This is because we decided to overcome the problem of low population by merely controlling the hyperparameter s_{min} , which is the minimum value of sample size involved in the generation of each condition. We mention this approach in case one may make use of these rates in the future research.

2.5 Graph hierarchies

After the application of the causal discovery tool (as detailed below in Section 3), we are presented with many possibly visually confusing causal graphs. These `spaghetti' graphs obfuscate immediate understanding, and, therefore, may harm any of CVD. An example of such `spaghetti' graphs is given in Figure 4.

Figure 4: Some example `spaghetti' causal graphs

Navigating these graphs leads to the following natural questions:

1. How do we disentangle `spaghetti' graphs?
2. How can we discover which graph is more informative or useful?
3. Is there a principled and well-de ned way (read, data-agnostic way) with which to compare and probe graph structures for useful, interesting, or otherwise unexpected causal relationships and information?

A possible solution is to simply turn this analysis into a graph theory problem. The main idea is to recover structural hierarchy and in uences such that we may have a measurement for the e ective description and management of in uence and prediction of causal behavior; ideally, we input such graph structures and output some numerical grade with which to compare other graphs, and thereby discover structures according to some property or measurement. This material follows and is adapted from: [5, 56], with additional material and notation from [15, 16, 47, 54, 70, 83, 93].

Graph study is typically classi ed into three levels: local; community; global. Our principle tool of choice is graph hierarchies [5, 56], adapted to causal structures; graph hierarchies are equipped with measurements at each level, in both the forward (down) and backward (up) directions: the forward (down) direction quanti es forward dynamics (i.e., given causes, what are the e ects?); the backward (up) direction quanti es backward dynamics (given e ects, what are the causes?).

1. Microscopic - local structure (individual vertices)
    - hierarchical levels (HL)
    - in uence centrality (IC)
2. Mesoscopic - groups or communities

- hierarchical di erence (HD)

3. Macroscopic - global structure
   - hierarchical incoherence (HI)
   - democracy coe cient (DC)

Each metric follows from HL, which builds into HD, from which IC, HI, and DC are derived. Figure 5 gives a cartoon that illustrates how each metric is computationally related.

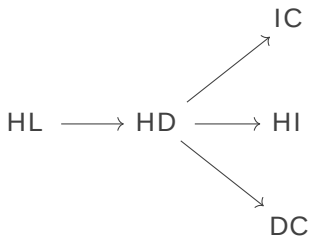

Figure 5: Relationships between hierarchical metrics

To help illustrate and explain these metrics, consider the following analogy. Suppose we are interested in making some wine. It stands to reason that we might arrange our various ingredients on our shelves in such a way that the most frequent ingredients or the most important ingredients are on the middlemost shelf that is easiest for us to reach; on the periphery of that shelf we might have our less frequent ingredients or less important ingredients, and on the higher and lower shelves we might likewise have less frequent or important ingredients. Graph hierarchical metrics measure give a mathematical description of the ingredients on our shelf, one relative to the other. In this way: HL ranks features (ingredients) based on their in uence and importance; HD compares the in uence or importance di erence between features; HI measures the variability in our features; DC assess the extent to which all of the features are fairly distributed according to important, to frequency; IC measures the impact of a feature on an outcome.

It is, after all, fun to imagine [24].

Some preliminary. A is the adjacency matrix, $G = (V, E)$ is a graph with vertices $v \in V$ and edges $(i, j) \in E$ with associated weights w_{ij}. $So(G)$ is the set of all source vertices; $Sk(G)$ is the set of all sink vertices.

Table 3: Relevant de nitions for hierarchical structures

weighted in-degree		weighted out-degree	
$d_i = \sum_j w_{ij}$	for vertex i	$_i = \sum_j w_{ij}$	for vertex i
$d = (d_1, d_2, \ldots, d_n)$	is vector	$= (_1, _2, \ldots, _n)$	is vector
$L = \mathrm{diag}(d) \quad A$	is Laplacian	$= \mathrm{diag}( ) \quad A$	is Laplacian

De nition 2.5 (Hierarchical levels). Hierarchical levels (HL) assign grades or labels to vertices based on how far they are from sources $V \in So(G)$ or sinks $V \in Sk(G)$, in which the HL vector follows from the minimum Euclidean norm $\|x\|$ under the constraint that x minimizes $\|L^T x - d\|_2$ or $\|^T x - \|_2$.

Forward: $g := \operatorname{argmin}_{x \in T} \|x\|_2$, where $T = \operatorname{argmin}_{x \in R^n} \|L^T x - d\|_2$

Backward: $:= \operatorname{argmin}_{x \in S} \|x\|_2$, where $S = \operatorname{argmin}_{x \in R^n} \|^T x - \|_2$

Di erence: $h = \frac{1}{2}(g -)$

HL assigns a scalar value to each vertex representing its position or rank within the overall graph structure, and compares vertices based on their connectivity and in uence, either incoming (in-degree) or outgoing (out-degree).

De nition 2.6 (Hierarchical di erences). Hierarchical di erences (HD) assign grades or labels to edges via di erences in HL.

Forward: $FHD_{ij}(G) = f g_j - g_i g$

Backward: $BHD_{ij}(G) = f_i - _j g$

HD evaluates the di erence in HL between connected vertices, indicating directionality and magnitude of in uence by directly comparing the HL of two connected vertices, checking the relative in uence one vertex has over another.

De nition 2.7 (In uence centrality). In uence centrality (IC) measures the extent to which a vertex is an in uencer of the graph by characterizing how signi cant the vertex is.

Forward: $_f(G; j) = 1 \quad \text{mean}(FHD(G; j)) = 1 \quad \frac{\sum_{i:(i,j) \in E} w_{ij} \, FHD_{ij}(G; j)}{\sum_{i:(i,j) \in E} w_{ij}}$

Backward: $_b(G; j) = 1 \quad \text{mean}(BHD(G; j)) = 1 \quad \frac{\sum_{i:(i,j) \in E} w_{ij} \, BHD_{ij}(G; j)}{\sum_{i:(i,j) \in E} w_{ij}}$

IC determines the degree to which a vertex acts as a source of in uence (forward centrality) or resists in uence (backward centrality) by checking each vertex's HL relative to those of its neighbors, either upstream or downstream by measuring the weighted average of HD for a given vertex. Note that if j has no in- or out-going edges, then the mean tends to 0. A positive $_f(G; j)$ indicates that j is an in uencer, with HL higher than those from which it receives in uence; a positive $_b(G; j)$ indicates that j is resistant to in uence, with HL lower than those it in uences.

De nition 2.8 (Democracy coe cient). Democracy coe cient (DC) measures the extent to which in uencers are being in uenced, and checks for relative `democractic' and `autocratic' behavior.

Forward: $_f(G) = 1 \quad \text{mean}(FHD(G)) = 1 \quad \frac{\sum_{(i,j) \in E} w_{ij} \, FHD_{ij}(G)}{\sum_{(i,j) \in E} w_{ij}}$

Backward: $_b(G) = 1 \quad \text{mean}(BHD(G)) = 1 \quad \frac{\sum_{(i,j) \in E} w_{ij} \, BHD_{ij}(G)}{\sum_{(i,j) \in E} w_{ij}}$

DC checks for equitability and uniformity of in uence and control distribution among vertices by comparing their average HD across all edges to a baseline of zero. If (G) ! +1, then G is more `democractic', in that there is a more equitable distribution of in uence (more variables have a say); if (G) ! 0, then G is more `autocratic', in that there is a less equitable distribution of in uence (fewer variables have a say).

De nition 2.9 (Hierarchical incoherence). Hierarchical incoherence (HI) measures how neatly the graph structure is partitioned into levels.

$$\text{Forward:} \quad {}_{f}(G) = [\text{var}(\text{FHD}(\ G))]^{\frac{1}{2}} = \frac{\sum_{(i,j)\,2\,E} w_{ij}\ (\text{FHD}_{ij}(G) \quad \text{mean}(\text{FHD}(\ G))^{2})}{\sum_{(i,j)\,2\,E} w_{ij}}^{\frac{1}{2}}$$

$$\text{Backward:} \quad {}_{f}(G) = [\text{var}(\text{FHD}(\ G))]^{\frac{1}{2}} = \frac{\sum_{(i,j)\,2\,E} w_{ij}\ (\text{BHD}_{ij}(G) \quad \text{mean}(\text{BHD}(\ G))^{2})}{\sum_{(i,j)\,2\,E} w_{ij}}^{\frac{1}{2}}$$

HI checks for variability or inconsistency in HD across G by evaluating the spread or dispersion of HD from their mean value, indicating the extent of consistency of in uence or control among nodes. If (G) ! + 1 , then G is more `incoherent', in that there are more disparities in hierarchical levels, and less uniform and equitable distribution of in uence; if (G) ! 0, then G is more `coherent', in that there are more uniform structure and minimal di erences in hierarchical levels.

Observe that both DC and HI are measurements of global structures; that is not to say that we are not interested in community and local structures (indeed, HD, a community measurement, is at least necessary for the calculations for global structures HI and DC), but that if we want to grade graphs according to some metric as a rst step, then global metrics are a natural choice. In particular:

1. Democracy coe cient 2 [0; +1)
   - Quanti es in uence: democratic (more variables have a say) orautocratic (fewer variables have a say)
2. Hierarchical incoherence 2 [0; + 1)
   - Calculates incoherence : coherent (neat partition into levels) or incoherent (messy partition into levels)

Once armed with these metrics,we plot for each adjacency matrix generated by the model one against the other. This gives an interpretable distribution of how causal graphs behave according to global metrics. Upon inspection of the causal graphs themselves however, it may be to the user that these remain quite `spaghetti'; however, the chef knows the composition of his spaghetti, and we can query the chef (i.e., the model) about the behavior of the spaghetti. Doing so gives us the distribution space in Figure 6.

LiNGAM uses DOT to plot graphs [20, 27]; this is in no small part a motivation for having interactable graphs in our interface, so we can directly in uence and explore the structures themselves, regardless of how `spaghetti'-like they are.

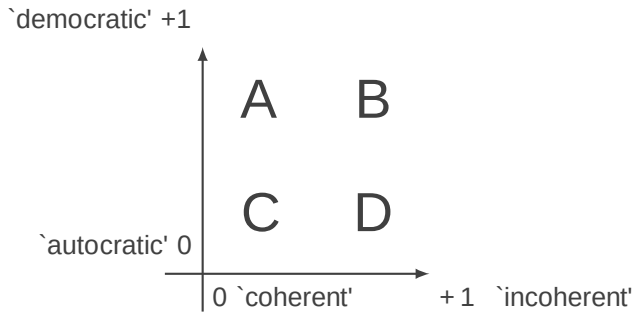

Figure 6: Democracy coe cient and coherence metrics

We may interpret the regions in some greater detail for both the forward and backward directions. For the forward direction:

Region A: high DC and low HI: all vertices have approximately the same HL; G is in uenced by a large percentage of its vertices

Region B: high DC and high HI: general balance in in uence or control across the graph, signi cant variability in individual in uence levels among vertices. Multiple nodes occasionally exert high in uence; no single vertex dominates consistently

Region C: low DC and low HI: distinct hierarchical levels; G is controlled by a small percentage of its vertices

Region D: low DC and high HI: few highly in uential vertices with a high degree of variance in their level of in uence. Control is concentrated but inconsistently exercised among a few nodes (some instability, irregular, nonlinear dynamics)

G is `maximally hierarchical' if DC and HI are 0; vertices can be grouped in `layers', all vertices in a layer have the same HL, HL of two layers di er by an integer, there can only be edges from one preceding layer to the layer succeeding it.

Suppose we generate a collection of causal graphs with our model, using, for example, the wine dataset with the following settings: correlation coe cient threshold $r = 0.2$; maximum number of items to combine as sample selection criteria $k = 2$; sample selection conditions with less than this value are excluded $s_{min} = 300$. For the forward direction, plotting the democracy coe cient against the incoherence metric gives the distribution in Figure 7.

Figure 7: Forward democracy coe cient and forward coherence metrics

Observe the cluster of graphs in region C. We may now con dently extract graph structures whose relationships are coherent and autocratic from this region. If we are interested in more complex spaghetti, 16 would be our chosen plate; if we love democracy and coherence, then 27 would be our plate.

The interpretation of the regions in Figure 6 changes if we are instead interested in querying the chef for the backward direction.

Region A: high DC and low HI: resistance to in uence is uniformly distributed; most vertices show a similar capacity to resist external in uences; autonomy is broadly maintained across vertices, stable and consistent internal resistance dynamics

Region B: high DC and high HI: resistance to in uence is generally balanced across the network; degree of resistance varies widely among the vertices

Region C: low DC and low HI: few vertices signi cantly resist in uence, while most others do not; resistance consistently maintained across interactions; clear hierarchical bottlenecks or control points that are highly resistant (gatekeepers or decision-makers)

Region D: low DC and high HI: few vertices have a high capacity to resist in uence, but is highly variable; unpredictable dynamics in how control or resistance is exercised; indicate key nodes or groups sporadically assert control or resistance

We may have a similar spaghetti dinner for the backward direction as we did the forward direction, except now we are interested in the other direction. Again, we generate a collection of causal graphs with our model using the wine dataset with the same settings as above; we recover the distribution in Figure 8.

Figure 8: Backward democracy coe cient and backward coherence metrics

Observe the cluster of graphs in region C. These are now relative to the backward direction, in which we might query the chef about causes for e ects rather then e ects from causes as we would in the forward direction.

Three representative (read, prototypical) spaghetti graphs representing regions A, C, and D in the forward direction are given in Figures 9, 10, and 11, respectively. In the wine quality dataset, no spaghetti graphs occupy region B, hence the absence of graphs in this region.

Figure 9: Adjacency matrix 27: prototypical region A (forward)

Figure 10: Adjacency matrix 42: prototypical region C (forward)

Figure 11: Adjacency matrix 16: prototypical region D (forward)

In the backward direction, intuition for the visualizations is generally lost, though the math reliably returns hierarchical analyses. Three representative spaghetti graphs representing regions A, C, and D in the backward direction are given in Figures 12, 13, and 14, respectively. Again, in the wine quality dataset, no spaghetti graphs occupy region B, hence the absence of graphs in this region.

Figure 12: Adjacency matrix 13: prototypical region A (backward)

Figure 13: Adjacency matrix 4: prototypical region C (backward)

Figure 14: Adjacency matrix 27: prototypical region D (backward)

Graph hierarchies present one possible solution with which to navigate these `spaghetti' graphs, in which we input such graph structures and output some numerical grade with which to compare other graphs, and thereby discover structures according to some property or measurement. Our causal discovery interface does just that; in particular, the interface measures the adjacency matrices given by our causal discovery engine for incoherence and democracy, and sorts them accordingly. We may now query the chef (i.e., our model) about causal relationships that might be more or less incoherent, or have higher or lower democracy.

If we ask, for example, what are some variables that give us a non-linear spaghetti, then we can find the graphs in the regions of B and D in Figure 6; if we ask for a more linear and democratic spaghetti, we find the graphs in region A; and so on. More directly, however, our interface sorts the graphs according to incoherence, and a general heuristic is to chose that with the lowest incoherence (i.e., highest coherence), for ease of interpretability, directly related to the CVD criteria.

3 Interface development

In designing the interface to promote convincingness, variety, and discoverability, we have devised a set of guiding design principles, as follows:

- Minimize `overwhelmingness': provide users with subsets information at each stage
- Adopt a storytelling approach: build on the understanding gathered at each stage
- Offer guided and unguided discovery: allow exploration and interaction with models

ˆ Incorporate accessibility features: color, navigation, responsiveness

ˆ `Don't think, feel!': accept inspiration from seemingly unrelated concepts

These principles are not only aligned with the project scope and objective, but with human nature; it has become important to consider how knowledge is processed and understood in order to produce a useful causal discovery tool. In the sections that follow, a description of each feature in the interface, as well as the rationale behind the inclusion of each, is presented alongside images of each as they appear on the interface.

This interface was primarily developed using the Dash, Plotly, and NetworkX Python integrations, hosted on a local Flask server. This makes it simple to test, modify, and share. WL technology has not yet been integrated into the tool, though its outputs are used throughout for discovery and analysis.

3.1 Steps to improve CVD

Consider Figure 2:

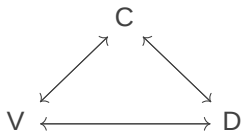

Figure 15: CVD triangle

Since the general objective of this project is to provide the user a result that is convincing, has variety and supplements discoverability. The key actions to evaluate, and improve the convincingness, variety and discoverability of the results shown to the user, are the following:

1. We suggest that as a way to improve the three metrics evaluated, some preliminary testing of the data will be carried out by the interface, as a preventive measure to ensure that the results of the model are not incorrect, or biased, or inconsistent [46, 77]. In accordance with DirectLiNGAM literature [77], in order for the model to hold, the assumptions of the model have to strictly hold. Although it is comprehensible that not all data will hold these assumptions, the user should, at least, be noti ed in a non invasive manner about the possible adverse e ects of the data used on the outcome of the model. In this way, as seen on Figure 15, we intend to improve CD. The following preliminary assumptions and their respective tests would be implemented by the interface, and if the dataset does not comply with the assumptions, the user should decide if the adverse consequences of the assumptions being broken are relevant to his research:

   (a) Large sample size assumption - If the sample size is not large enough, the interface should notify the user. Some of the consequences of low sample size are erroneous directed edges and bias in the estimators.

(b) Non-Gaussianity assumption - If the errors of the treatment variables are Gaussian, it means that in between variable relationships could be mistaken. With Non-Gaussianity, when we regress according to the direction of causality, the explanatory variable, and the residual become independent, which means the relationship is causal. In other words, Non-Gaussianity enables us to determine the direction of any edge in the causal graph and choose one speci c model from the Markov equivalence class, i.e., it makes DirectLiNGAM identi able. Conversely, if the errors are Gaussian, the direct edges cannot nd the appropriate causal directions. Appropriate Gaussianity tests include:

i. Kolmogorov-Smirnov Test, Shapiro-Wilk Test, or the Anderson-Darling Test

ii. Histogram test - Ideally, objective statistical tests would be suitable for the interface, but since the above statistical tests results can be biased if the sample size is too large, histograms are usually used. The user would ideally see a histograms of the errors and, alongside a clear explanation, decide if the errors are non Gaussian from the shape of the histogram.

(c) No Multicollinearity assumption - Since variables should be independent from each other, we need the treatment variables to not be correlated, because correlated variables can lead to wrong estimates and reduced precision of estimators. If this is the case and the following tests fail, the user should know about possible solutions to this problem, such as the removal of one of the highly correlated variables depending on its relation to the outcome.

i. Correlation matrix test - If two or more of the variables are highly correlated (0:8 or higher, in the scale of 0 to 10), the user should know and understand the problems with this.

ii. Variance In ation Factor (VIF) Test - this estimates how much the variance of a regression coe cient is in ated due to multicollinearity.

Additionally, if these statistical tests show that some important assumptions are broken, we plan on suggesting some users about di erent ways of improving the or correcting the data, in order to improve the results, while also conserving their integrity.

2. To improve the discoverability of the output of the interface, without harming variety and convincingness, we observe that the amount of conditional causal graphs displayed may be too many for some users, leading them to feel overwhelmed by the amount of information and various conclusions (which can be sometimes opposite) from this graphs. As stated before, we intend to have the user decide the parameters that a ect the amount of conditional causal graphs that are presented (minimum sample size, correlation coe cient and maximum number of items to combine as sample selection criteria), but also, we propose the addition of a new index number that can help with the hierarchical ordering of the conditions. With the conditional causal graphs being ranked, we could provide the user with the most relevant, the most precise, or the one that its backed by the largest sample size, depending on what the user desires at the time.

   The origin of this proposal is the need to lter the quantity of conditional causal graphs displayed to the user. Additionally, although the results output already has

a variable that can possibly be used to rank the conditional causal graphs, we found out that this index could have a variety of issues.

Currently, the index variable rate is the division of the sample size for which the binary output variable is ful lled, (in the example of the wine dataset, the amount of observations that result in high quality), divided by the sample size for which the condition of the causal graph is ful lled. Although this can be a useful index to rank the conditions provided, we found some potential issues with this measure. First, in the numerator, the user can arbitrarily input the value that is considered ful lled. Second, and most important, the population size can in ate the index to be very high if the sample size is low, or vice-versa. This could be an issue when the condition given is obvious to the user, and lead to redundant conditional graphs.

For example, if it was a proven fact that alcohol percentage is a predictor of wine quality, and we have a conditional graph for observations with alcohol level higher than 5%, we would soon come to realize that this conditional graph might not very useful, since wines generally have a higher alcohol percentage. However, with the current variable rate , the amount of the sample observations of high quality wine and this alcohol percentage is very high, and this would lead us to have a high priority causal graph, although this information is redundant to most users. Conversely, if there is a very low sample size for a condition that positively a ects quality, the rate can become in ated, giving priority to a conditional causal graph mostly because it has a low sample size.

Understanding that low sample high rate conditional causal graphs can provide insights to the user, as long as a high rate is maintained, we plan on reporting these in a non-invasive way, so that the aspect of low sample size trends is not lost, with the goal of not hindering discoverability. This would not be treated as the main nding of the graph, but as a secondary non-invasive piece of information that could signal some important trends in a cluster of the observations.

In summary, the second step and current goal is to create a weighted index that e ciently ranks the most useful conditional causal graphs for the user, giving importance to sample size, score, and statistical signi cance in a balanced manner. This way, we can provide the user with only non-redundant conditional causal graphs, with the purpose of not overwhelming the user with too much information, without harming discoverability. In terms of the Figure 15, we will attempt to improve upon convincingness and discoverability, by not overwhelming the user, without harming discoverability by still presenting low sample results and without hindering variety too much by still presenting various relevant conditional causal graphs.

3. We would like to demonstrate the utility of the interface as measured by the degree of CVD through a survey. Additionally, ideally we would like this survey to be carried out as a feedback to optimize the output. In this way, we could eventually transform this survey into a two way communication tool for the interface and the user. Thus, depending on the characteristic evaluated, for each one of the three characteristics, we add some prototype questions to which the user responds according to: strongly disagree or strongly agree.

(a) Convincingness -According to De nition 1.1, some appropriate evaluations [72] to gauge how convincing the resulting causal graph will be required from the users; some example statements which the user can evaluate might be:

   i. The results showcased on this causal graph are persuasive
   ii. The results showcased on this causal graph are compelling
   iii. The results showcased on this causal graph are plausible

(b) Variety - To inspect variety, as we de ned it in De nition 1.2, the following provisional statements are going to be asked to be evaluated:

   i. The multiple resulting causal graphs were equally helpful to understand the data
   ii. The multiple resulting causal graphs were enough to understand the data
   iii. The multiple resulting causal graphs gave diverse points of view on the data

(c) Discoverability - To study discoverability, which we de ned in De nition 1.3, the user will be asked to evaluate statements similar to the following:

   i. The icons, colors, typography, sizing, whitespace, and contrast of the results presented were appropriate
   ii. I found new information about the data presented because of the graphs showcased

3.2 Data visualization

The ultimate goal in the development of this interface is to guide the user toward discovery while minimizing overwhelmingness. Therefore, prior to introducing causal models, it is helpful to understand the nature of their dataset rst. For this reason, the rst page of the interface is a data visualization dashboard that provides the user with data cleaning suggestions, descriptive statistics, pairwise correlation scatterplots (and heatmaps), and distribution histograms. With this basic understanding of the existing relations across features, the user will be better prepared to draw valid conclusions from their casual models.

Figure 16: The rst page presents data cleaning tips, raw data, and descriptive statistics.

Figure 17: Pairwise correlations can be viewed as scatterplots or as a heatmap.

Figure 18: The histograms provide a visualization of the data distribution for reference.

3.3 Prior knowledge survey

Although algorithms such as LiNGAM highlight causal relationships among features in a dataset and promote discovery, there exists the opportunity for users to apply their prior knowledge such that the results are more convincing. For instance, if a causal graph of the wine dataset shows `quality' as the overall parent node, this graph is not useful for discovering the causes or in uences on quality. Thus, applying prior knowledge, such as setting `quality' as a sink, or outcome node, drives the model in a direction that promotes discoveries that are aligned with a user's hypothesis. However, if a user chooses not to apply prior knowledge, they can explore new trends and view unbiased models, which also promotes discovery.

In the interface, the `Prior Knowledge Survey' tab instructs a user as to how they can apply prior knowledge. There are three options:

- `Default: -1': -1 indicates that the user does not have prior knowledge to con rm or deny a causal relationship, creating greater reliance on the model results to interpret causality.
- Add `0': 0 is used to denote that feature y is caused by feature x. This is especially useful for specifying sink variables, such as the outcome of the dataset.
- Add `1': 1 is used to set source variables, essentially informing the model that variable x causes variable y. This is useful for communicating the actionability of particular features.

In the interface, the user is presented with instructions as to how they can complete the prior knowledge matrix. After selections are made, the nal matrix can be visualized and exported as a .csv le for use as an input into the Conditional Casual Discovery engine. The causal graphs that are produced should be inspected to ensure that the relations speci ed in the matrix are respected by the model and hold true across all graphs. See the gures below for an illustration of this section of the interface.

Figure 19: The user can select 0, -1, or 1 in each pairwise dropdown to indicate relations.

Figure 20: After the prior knowledge matrix is con rmed, a downloadable csv is presented.

3.4 Model builder canvas

The `Model Builder' tool allows users to interact with causal graphs and build, or discover, a highly interpretable and convincing model that they can utilize for strategic decision-making. The underlying design principle for this module is that humans learn best through interaction. Therefore, transforming the models from static images of causal

graphs to interactive ones where nodes can be rearranged, formatted, added, and removed works well to reduce overwhelmingness and improve discoverability. The structure of this section is as follows:

1. Upload les: the user uploads the results dataframe and all of the models in the form of adjacency matrices, so that they can analyze a variety of causal graphs and reference other relevant data regarding each condition.

2. Hierarchical metrics and scatter plot: this section allows the user to visualize and compare the hierarchical metrics for each model (speci cally the democracy coe - cient and incoherence score). The multiple views of this data will enable a user to discover the overall distribution of in uence that particular features (or combinations of them) have on the outcome.

3. Discovery options dropdown: this dropdown populates with options that represent the previously uploaded adjacency matrices. The selected condition will populate the subsequent Discovery tabs, as well as the Summary and Statistical Validity sections.

4. Unguided discovery: Another approach to discovery is to allow a user to view complete, unedited models and learn how they compare to each other. To do this, the `Conditional model explorer' provides multiple views of uploaded adjacency matrices:

    - Dynamic view: this section provides an interactive representation of the graph via an HTML iframe. The user can move nodes to di erent positions, ideally so that they create a more instructive view. Physics can be enabled as well, an extension of the interactive features on this graph that may not have immediately obvious e ects on interpretability, but is aligned with the design principle that interaction promotes learning. Additionally, the node colors correspond to the ux (relative ratio of inputs to outputs) and the edge line weights and line styles are proportional to connection magnitude and direction, respectively.
    - Static view: this representation has the same visual features as the dynamic view, but lacks the physics and interactive features. However, it does o er the aforementioned color coding and edge styling, which is an improvement over the default black and white graphs produced by LiNGAM. This view can also be saved as an image, which allows for simple and quick sharing or sorting of results.

5. Guided discovery: upon selection of a model, the parent nodes will appear in the canvas. From here, the selection of a parent will then populate the children for it, and so on. This allows the user to explore individual branches of a graph in a stepwise fashion, without being overwhelmed by the entire causal graph at once. Additional features that provide visual cues about the model include line weighting by connection strength and line style as indication of positive / negative connections.

    - Path summary: as the user selected nodes, they are highlighting a causal path that is saved on the left hand side. This helps the user track their discoveries.

- Undo button: this button deletes the most recent tier of child nodes so that a user can make changes. This supports the notion that discovery is not always a forward-looking process, it takes iteration and involves backtracking.
- Edge weight threshold slider: this allows the user to lter out connections that are potentially weak in order to de-clutter the space and focus their attention on the most in uential causal relationships.
- Child node preview: this feature allows the user to hover over parent nodes to see which children are associated with it. This control allows users to select paths based on certain hypotheses or prior knowledge, perhaps by deciding to click on more actionable features or by trying to nd the shortest path to the outcome.

6. Summary: this feature allows the user to view additional quantitative details to add to the convincingness of their discoveries. After selecting a model, the user will be presented with a `Recipe' for improving their outcome, by showing the range, average, and mode of values for each feature that had strong in uences on the outcome (for a given condition). This helps a user quickly understand the di erences in the data that fuels the creation of each condition / model. The data is also displayed in the form of a boxplot directly below the table.
7. Statistical validity: this section displays the CFI, AIC, and RMSEA indices to indicate the statistical validity of the model. This will allow the user to evaluate the extent to which the results are accurate or sound enough for one to draw reliable conclusions.

3.5 Insights and analysis (work in progress)

This section focuses on the step beyond data analysis, and the objective is to help user drive strategic decision-making based on the models they have explored. This is less of an exact science, so suggestions for approximate improvement will be provided. Speci cally, improving the dataset itself as well as sorting / ltering out models that are seemingly less signi cant. Some quantitative data here is designed to improve convincingness, and the primary form of communicating information is through text, such that one can easily understand recommendations rather than relying on interpretation of extremely complicated model sets. Finally, the user can identify models that contain the most actionable features, with the hope of having this tool be readily used in the real-world for a variety of applications.

Figure 21: Clicking on the parent node of a model reveals the child nodes.

Figure 22: Selected child nodes up to the outcome are added on the left to visualize the overall path.

Figure 23: The user can select a model from the dropdown, presenting the dynamic view that enables them to interact with and move the nodes of the graph.

Figure 24: The static graph provides color coding and line styling to communicate key information.

Figure 25: The scatterplot and accompanying table communicate key information about the hierarchical properties of each model.

Figure 26: The last page of the app allows for users to make decisions after analyzing the data and the generated models.

3.6 Statistical validity

LiNGAM is a huge package and has many built-in tools for extensive validation. The evaluations against the model t indicator are described here.

The model t was examined using the evaluatemodel_t function in the package, as mentioned in Section 2.3. DoF (degree of freedom), DoF Baseline, ^{2}, ^{2} p-value, ^{2} Baseline, CFI, GFI, AGFI, NFI, TLI, RMSEA, AIC, BIC and LogLik (log likelihood) are returned for every single model. The model here is the linear acyclic model constructed by LiNGAM for a given data, i.e. expressed in the form of an adjacency matrix used to create the graph shown in Figure 4. Evaluations of the model t are conducted for each of the constructed adjacency matrices. Creating a condition according to a certain criterion creates a model for it. Thousands of conditions can be created by adjusting some parameters . This can be a source of annoyance to the user, but it can also help to get an overview of the model t evaluations.

More than 2000 models corresponding to the conditions were generated by adjusting the parameters ($r = 0:2; k = 4; s_{min} = 300$), and the model t was examined for each of them. The table 4 shows the correlation coe cients for the model t indices.

Table 4: Correlation coe cients with model t evaluations

	CFI	GFI	AGFI	NFI	TLI	RMSEA	AIC	BIC	LogLik
CFI	1.00	1.00	1.00	1.00	1.00	-0.66	0.64	0.47	-0.64
GFI	1.00	1.00	1.00	1.00	1.00	-0.66	0.64	0.47	-0.64
AGFI	1.00	1.00	1.00	1.00	1.00	-0.66	0.64	0.47	-0.64
NFI	1.00	1.00	1.00	1.00	1.00	-0.66	0.64	0.47	-0.64
TLI	1.00	1.00	1.00	1.00	1.00	-0.66	0.64	0.47	-0.64
RMSEA	-0.66	-0.66	-0.66	-0.66	-0.66	1.00	-0.96	-0.71	0.97
AIC	0.64	0.64	0.64	0.64	0.64	-0.96	1.00	0.81	-0.99
BIC	0.47	0.47	0.47	0.47	0.47	-0.71	0.81	1.00	-0.76
LogLik	-0.64	-0.64	-0.64	-0.64	-0.64	0.97	-0.99	-0.76	1.00

DoF, DoF Baseline, ^{2}, ^{2} p-value and ^{2} Baseline are omitted due to space constraints and their use in calculating other values. From this table, what information should be provided to the user is decided. Providing all of it would improve explainability, but could confuse the user. However, one is not enough. Therefore, we decided whether the information is necessary or not based on the strong positive correlations.

According to the table 4, correlation coe cients among CFI, GFI, AGFI, NFI and TLI are 1. Then, we choose CFI for representation. The correlation coe cient of LogLik and RMSEA is .97, so RMSEA is chosen. Also, AIC and BIC seem to be strongly and positively correlated, then AIC is chosen for representation.

Figure 27: CFI, RMSEA and AIC correlation coe cients. This gure will be used as an explanation of how to interpret statistical evaluations.

Table 27 is what we will use as the explanation of the model t evaluations to the user. RMSEA and AIC are strongly and negatively correlated, CFI and AIC are positively correlated, and CFI and RMSEA are negatively correlated. Generally, the higher the CFI, the lower the RMSEA and the lower the AIC, the better the model is judged to be. However, there are a few places where the gure does not t into this general understanding, and users need to be informed about this.

Correlation coe cients of hierarchical and statistical evaluations are also calculated. However, no correlations are found. Figure 28 is describing the correlation of graph hierarchy and statistical values.

Figure 28: Correlation coe cients of hierarchical and statistical values. The `back' and `ford' mean backward and forward respectively.

There appears to be almost no correlation between the statistical values and the values relating to the graph hierarchy. As the correlations between values relating to the graph hierarchy are illustrated in the diagram in the previous section, it would not be surprising if there was no correlation with the statistical values. Possible uses for these values include.

1. selecting graphs according to the graph hierarchical values
2. displaying the statistical values corresponding to the chosen graph
3. give an interpretation of the graph hierarchical and statistical values.

Displaying statistical values should be just an option, as these values may be a source of confusion for some users; it will help users who want to know the values of CFI, AIC and RMSEA or interpret the model statistically.

3.7 Accessibility

An interface should be designed with an emphasis on accessibility. The World Wide Web Consortium (W3C) published the Web Content Accessibility Guidelines (WCAG) [11] and outline four key principles (POUR model) to follow when creating any web-based or electronic content:

1. Perceivable - Information must be available to users in ways they can perceive with their senses, using assistive technologies as necessary
2. Operable - Components must work with both keyboards and assistive devices
3. Understandable - Content needs to be clear and limit ambiguity
4. Robust - Documents must maximize compatibility with both current and future technologies like screen readers

Therefore, some design principles should be followed, including:

- Documentation at least in HTML, not just LaTeX; machine readability is lacking for .pdf files generated by LaTeX, and will often fail; tools for accessibility as outlined above are not well-supported [48, 89]
- Color choice should account for options for vision accessibility needs [1, 18, 39, 42, 60]; additionally, between light and dark mode, choice of background color and text relative to each other follows from [11]: `the visual presentation of text to have a contrast ratio of at least 4.5:1'
  - If white background, then dark gray text (not black): hexcode 404040; rgb (64, 64, 64)
  - If dark background (not black: use hexcode 121212; rgb (18, 18, 18)); then off-white text: hexcode ededed; rgb (237, 237, 237).
  - Choice of colorblindness color palettes to account for:
    - Protanopia - affects the ability to distinguish red and green
    - Deuteranopia - red-green color blindness that affects the green cone pigments in the eye
    - Tritanopia - affects how people see blue and yellow
- Assuming reading direction: hierarchical important information should be at head of reading direction (Tufte design principles: see any of [84, 85, 86, 87, 88] for discussion of such arrangements)
  - If LR/RL reading direction, then no footnotes; use marginnotes
  - If TB/BT reading direction, then no marginnotes; use footnotes

This presents, however, a mostly nebulous (though not unimportant!) starting point with which to design the causal discovery interface. In the same way that the enigmatic `user' was analyzed in Section 3.8, so, too, must the content and context of accessibility be understood here. In its current state, this project is primarily focusing on visual accessibility needs; future work, however, will need to accommodate and account for the varying other degrees of accessibility design, of which will be laconically presented here.

Following [7, 53], we de ne the following.

De nition 3.1 (Inclusive design). Inclusive design is a design methodology that enables and draws on the full range of human diversity.

De nition 3.2 (Accessibility). Accessibility refers to the qualities that make an experience open to all; it is a professional discipline aimed at achieving an experience open to all.

[53] disclaims the following, however, to draw a distinction between accessibility and inclusive design; laconically, the former is an attribute, the latter is an action to take:

> An important distinction is that accessibility is an attribute, while inclusive design is a method. And while practicing inclusive design should make your products more accessible, it's not a process for meeting all accessibility standards. Ideally, accessibility and inclusive design work together to make experiences that are not only compliant with standards, but truly usable and open to all

This is similar to the very related space of diversity and inclusion: the former is a noun, a substantive abstraction, while the latter is an active action to be taken [2, 26, 31, 52].

Accommodating accessibility needs broadly follows two (possibly overlapping) design spaces: (1) disability as personal attribute; (2) disability as context dependent. Both spaces occur (at least) at points of interaction between a person and social experience, and physical, cognitive, and social exclusion or limitation are results of otherwise mismatched interactions. As a technology company (Fujitsu) and as academics working on a design interface therein (G-RIPS), it is our responsibility to know how designs a ect these interactions and create mismatches. Crucially, however, disability is not a personal health condition; rather, it is a mismatched human interaction which may be permanent, temporary, or situational [53, 92].

De nition 3.3 (Permanent disability). Permanent disabilities are conditions that persist over time and signi cantly impact how individuals interact with (digital) content.

De nition 3.4 (Temporary disability). Temporary disabilities are impairments that arise from injuries and illnesses that a ect users' otherwise abled abilities for a certain period of time.

De nition 3.5 (Situational disability). Situational disabilities are barriers or impedances that arise due to environmental or situational factors that a ect users' otherwise abled abilities for a certain period of time.

[53] presents a useful representation of these accommodation needs, with some prototypical examples, reproduced in Figure 29.

Figure 29: `Persona Spectrum' [53]

Actionable design that accommodates all the spaces represented in Figure 29 is beyond the scope of the current project; however, e ort has been made to at least accommodate the spaces for visual accessibility needs, including the features described in the design principles above. Much work in this space, and in the remaining spaces of Figure 29, invite continued development.

To help equip future work in this design space (perhaps for future G-RIPS projects, see Section 4. In brief, the guidelines provided by [11] are a natural place to start, and are written to promote the best practices of web design and development, including accommodations for individuals with permanent, temporary, and situational disabilities, including, but not limited to:

- Contrast ratios for foreground and background elements
- Full navigation by a keyboard alone, integrated with assistive technologies (AT)
- Captions and tagged elements for multimedia and machine readability

Note that many countries have laws that mandate at least some extent of accessibility standards [14, 63]; compliance mitigates risk and demonstrates a commitment to social responsibility and inclusivity.

- Functionality that uses multipoint or path-based gestures can be operated with a single pointer

3.8 The user

So far we have made reference to the enigmatic `user'. Indeed, what is `convincing' to one user may not be so for another; what is `discovery' to one user may not be so for another; what is `variety' to one user may not be so for another. To start, let us consider `convincingness'.

Convincingness is related to the notion of explainability discussed in [4]Explainability in AI is directly related to the audience for which the explanation (read, convincing) is given; the cognitive skills and pursued goal of the audience (read, users of the model) have to be taken into account jointly with the intelligibility (the characteristic of a model to make a human understand its function (i.e., how the model works) without any need for explaining its internal structure or the algorithmic means by which the model processes data internally [55]) and comprehensibility (the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion [4]) of the model in use.

> Given an audience, an explainable artificial intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

We should be careful, however, to not conflate `explainable' with `interpretable', which [4] also disclaims. We will use these terms as in Definition 5.1 and Definition 5.2, informed by the above citations. While these overlap, note the following: interpretability is passive, while explainability is active. Interpretability refers to the extent to which a human can understand the cause of a decision made by a model; explainability refers to the methods or techniques used by a model to clarify or justify its internal functions or outputs. Convincingness follows from interpretability and explainability.

We claim, therefore, that convincingness is at least related to explainability, and ideally related to both interpretability and explainability; the audience, therefore, is a critical component of this. Possible target audiences and motivations for these for each domain follows:

- Domain experts
  - Motivation: trust the model, gain scientific knowledge
- Users affected by model decisions
  - Motivation: understand situation, verify fair decisions
- Scientists/developers
  - Motivation: ensure/improve product efficiency, research
- Managers/executive board
  - Motivation: assess regulatory compliance, understand corporate AI applications, protect/implement assets

- Regulatory entities/agencies
  - Motivation: certify model compliance with legislation, audits

It is not productive to assume that CVD (see Definition 1.4) apply in the same way to each domain; indeed, at the very least, even motivation for AI models is not the same across domains! Therefore, the eponymous `user' of the interface should have options available to them according to their priorities and domain.

We must be careful, however, to not conflate the following. `User' is ambiguous, even granted the above delineations for domain. `User' here may well represent both the user proper of the model (called *agent*) as well as the audience (called *patient*) with or for whom the agent is working and/or presenting. Each of the agent and the patient may have differing domain expectations and needs for CVD. This project, then, is as much social (sociological and anthropologic) as it is `hard' scientific or mathematical. This relationship between agent, patient, and model is an extension of Agency Theory [23, 29, 45, 71].

A general agency model is defined in Definition 3.6.

Definition 3.6 (Agency model). An agency model is a quadruple of the form:

$$M_{\text{agency}} = \langle M; a; p; i \rangle$$

where M is the model being implemented or designed, a is an agent using or implementing a model M, p is the patient or audience with or for whom the agent is working and/or presenting, and i is the instrument to use or access the underlying model M.

This definition is an adaptation of the fuller model quintuple: $M = \langle M; a; p; t; i \rangle$, where we have combined t (tool or instrument) and i (implement or medium) into a composite form relevant to this project.

To make explicit the importance of the relationship between the agent and the patient, Figure 30 gives a pictorial representation of how the components of the model defined in Definition 3.6.

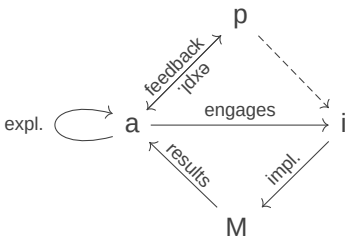

Figure 30: Agency model representation

The agent engages (designs/uses) the instrument for implementing a model, which then gives results to the agent. The agent must understand (explain) to themselves what the model is giving, and then explains those to the patient; the patient provides feedback to the agent, and the loop continues until convergence.

If the agent and the patient are the same, the representation still holds. Note that the patient rarely interacts directly with M, but may do so through i, in which case explanation will either filter through a to p, or pass through a directly to p, in which case a and p may be taken to be the same (as far as the representation is concerned); hence,

the dashed connection from p to i. In our project, we assume the role of a, and both take $a = p$ and $a \neq p$.

The most important interpretation of the agency model adapted to our project is the social aspect of the communication between the agent and the patient; the agent and the patient may be indexed with different sets of contextual aspects: backgrounds, assumptions, contexts, expertise, etc.. Therefore, having a component in our project that accounts for this social dynamic is important. While convincingness is the driving motivation for this aspect of the project, any of CVD is influenced by sociological and anthropological aspects. To account for this, having multiple layers of possible explanations of the interface (varying from exactly technical to less technical) is at least relevant to convincingness; discovery and variety will likewise vary according to the extent to which the backgrounds, assumptions, contexts, and expertise of the domains of agents and patients are aligned.

Finally, a note about the so-called `paradox of choice'. The feasibility of a choice from the Rashomon set is exactly an exercise in the *paradox of choice*: when having a lot of options does not make us happier but instead makes it tougher to decide, and may induce stress or regret upon making a choice [74]; what follows is a decrease in the motivation to choose, to commit to a choice, or to make any choice at all [73].

Generally, however, the results of experimental tests in the paradox of choice are inconclusive; while we should be alright proceeding without much regard to paradox of choice, simply mentioning it would suffice as an element of the interface. [73] concludes (emphasis added):

> Although strong instances of choice overload have been reported in the past, direct replications and the results of our meta-analysis indicated that adverse effects due to an increase in the number of choice options are *not very robust*: The overall effect size in the meta-analysis was virtually zero... The meta analysis further confirmed that `more choice is better' with regard to consumption quantity and if decision makers had well-defined preferences prior to choice.

This last emphasis may be important: what are the defined preferences prior to exploring causal models and prior to the presentation of the Rashomon set? That is, presumably, when presenting a dataset for analysis of causality, the user is bringing a set of preconceived preferences for model and outcome, and so may be biased in that regard.

We may work with this in the following way: setting personal criteria for decision making and choice, limiting the number of options considered, and offering some feedback to the user to increase confidence in choice, which may look like practicing gratitude to the user as a responsive feedback for the choice made rather than focusing on alternatives. It is, in general, bad design to frontload `too many' options and settings to a user; quantifying what is `too many', however, is a problem itself.

As a user evaluates the options in the Rashomon set, explaining and/or disclaiming the effects of the paradox of choice may help users recognize when they may be overwhelmed with choices, and so indicate to the user that they need proactive steps to otherwise streamline their decision-making in that instant.

4 Future directions and discussion

In this section we will discuss some future endeavours that our team would consider implementing into the project with the goal of improving convincingness, discoverability, and variety. We acknowledge that the short amount of time is a big restriction for this project, but we are hopeful that the following measures can improve the nal delivery in a great manner. Additionally, in this section we will attempt to relate this future measures to multiple key variables that were measured on the project.

4.1 Prior knowledge and identi cation

Although in our nal delivery we implemented the concept of prior knowledge to the user interface, there are still some concepts in the literature that were left out because of time restrictions. In summary, in general literature about the LiNGAM, prior knowledge is a very powerful tool to use together with the model. Adding prior knowledge to the model can help tremendously with the coherence of the graphs, as well as the overall strength of the model.

According to [77], although DirectLiNGAM requires no prior knowledge on the structure, more e cient learning can be achieved if some prior knowledge on a part of the structure is available because then the number of causal orders and connection strengths to be estimated gets smaller. We also believe that by implementing prior knowledge in a more generalized way can help with CVD.

In causal graphs, the unrecognized presence of unmeasured variables can lead to wrong conclusions about causal relationships, and a way to deal with this is through the use of identi cation. If the causal relationships are not adequately identi ed, our model will presented with various biases. We believe that further studies and experimenting with the conditions already generated by the working Wide Learning–algorithm to deal with this biases can improve upon convincingness, by making believable graphs, as well as taking advantage of DAG terminology to identify real causal paths. We believe that with the current tools, more things can be done for identi cation, which is to unblock causal paths of the graphs, and interpretability.

There are three fundamental structures in Directed Acyclic Graphs (DAG) [21], the rst one is called a chain, a chain is when there is a causal relationship between U to W, and V stands between them as a mediator (U V W), using prior knowledge to condition V will block the causal path.

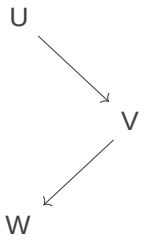

Figure 31: Chain

The second one is called a fork (U V W), where V acts as a confounding common cause, conditioning on V will block the non causal path. Failure to condition on a common cause is called Confounding Bias

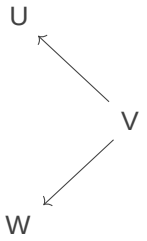

Figure 32: Fork

The third one is called a collider (U V W), where controlling for V will open up a non causal path. Failure to condition on a common e ect is called Selection Bias.

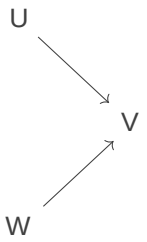

Figure 33: Collider

If the collider variable V had a descendant and its not conditioned, that is also called Selection Bias. This is one of the most common issues with causal graphs [22].

Suppose we have 4 variables; S is maternal smoking, L is low birthrate, U is malnutrition and Y is neonatal mortality. The problem is the following. Maternal smoking is associated with both low birthweight and higher mortality at birth. One key issue, among low birthweight babies (less than 2.5kg), maternal smoking is associated with lower neonatal mortality. Usually, in our project and whenever we don't possess prior information we can easily draw the conclusion that smoking can in fact be bene cial for low birthweight babies, but there is another interpretation:

Figure 34: Working example

As mentioned before, conditioning on a collider by stratifying (in this case Low Birthrate) will open up a non-causal path. Even if S Y is positive, the association between them given L = 1 can become negative. This is a problem for interpretability of the graphs. If maternal smoking and malnutrition both cause low birthweight, then low birthweight infants whose mother did not smoke are likely malnourished (and vice versa). This is a self induced selection bias caused by conditioning on a collider, something that frequently we ignored in our conditional causal graphs.

There is also the concept of Back Door Path, in which there is a direct connection between two variables apart from a chain connection, this can indicate confounding bias.

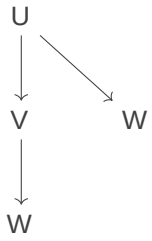

Figure 35: Back Door Path

In DAG literature, conditioning can mean adjusting, restricting, stratifying or matching. The average total causal e ect of U on W is said to be identi able if it is possible to purge all non-causal associations from the observed association between U and W such that only the causal association remains. More often than not, this is something that we took as an interchangeable equivalent for interpretability of the graphs, which is directly related to the convincingness of them. However, we did not consider this approach in our measures to simplify the results for the user to aid in discoverability and reducing overwhelmingness.

In statistical modeling, Confounding Bias is the failure to condition on a common cause andSelection Bias is mistakenly conditioning on a common e ect. Conditioning on a descendant of a common e ect or collider also induces an association between the colliders parents, another type of selection bias. Some ground rules for identi cation are related to conditioning on colliders and whether an analyst can recognize this. There are three main rules [22]:

1. Conditioning on any non-collider on a path blocks the path.
2. Not conditioning on at least one collider (or its descendants) blocks the path.
3. Not conditioning on any non-colliders and conditioning on all colliders (or at least one descendant of each collider) on the path opens the path.

The reason these de nitions are important is because during the span of this project we worked mostly on conditional causal graphs. For the sake of simplicity we mostly ignored these biases and how to deal with them, however, although we understand that the main focus of the project is not focusing on prediction, if actionable measures were to be taken, it is an interesting line of research that we did not contemplate, in summary, working with mostly conditional graphs is likely to induce selection and confounding bias because of the conditioning on numerous causal graphs can lead to making wrong predictions and estimation on the dataset outcomes. For the sake of simplicity, we used the di erent conditions as a mean to aid discoverability, but we did not contemplate that they could be harming convincingness or explainability by accidentally inducing selection bias or confounding bias to common e ects or common causes. The adverse e ects of these biases are exacerbated when we consider data for which the user does not have expertise on.

In summary, we consider that causal discovery graphs shine when looking to understand the causal structure of a dataset and deriving the implications of a model. However, if we were to make some improvements to the interface, we would implement the previously stated selection and confounding bias identi cation methods in a generalized way, letting the user know about the potential issues of conditioning on certain variables, in relation to them being mediators, confounding common causes or colliders. This has to do with the next key recommendation

4.2 Model supplementation

Related to the previous recommendation, endogeneity is also when the e ect of an independent variable on a dependent variable can't be causally interpreted because it includes omitted causes leading to biased estimates. In usual regression analysis, researchers identify correlational associations and assume causality, and although some analysis address endogeneity through Instrumental Variables methodology, these usually don't address latent confounding variables, only speci c causal relationships between selected sets of variables.

DirectLiNGAM is useful to nd a linear causal relationship from one variable to another, but it assumes causal su ciency, its very normal for causal su ciency assumption to not be held in observational data, so it is common to relax this assumption.

In some applications of LiNGAM, in order to not only model the causal structure of the dataset, some authors use other models that relax LiNGAM assumptions to con rm that there is unconfoundedness or exogeneity. [98] use an extension of the PC algorithm called Fast Casual Inference (FCI) from [79] to relax this assumption, if there is no U W causal relationship but there is U$ W in the resulting causal graph, this might imply the existence of unobserved variables, knowing this fact and doing something about this can enhance model building and accuracy, instead of accepting unconfoundingness, the user can know about the existence of an unobserved common cause and take measures. This

would provide another tool for assumption holding detection to prove causal su ciency through the detection of endogeneity.

Additionally, another way to implement other modeling techniques to causal graphs is through regression estimates for interpretability, which is related to convincingness and the causal discovery part of the causal graphs. A way to do this is through linear regression after identi cation. DAGs can help determine which coe cients estimate causal e ects, and check if they have causal e ects in the rst place. Our proposition is to add a linear regression model builder to the interface that deals with all causal and non-causal paths from the treatment to the outcome, depending if they are mediators or colliders, then we can better decide which variables are appropriate to control in a predictive model. Ideally, the user would use the interface and tune the model to obtain a causal graph, then, with the causal discovery of the database he could or would be able to elaborate a linear regression model, looking at the causal graph would be relatively easier to determine which paths contribute to each regression coe cient. We can know which exact paths in a causal graphs contribute to each coe cient of a linear regression, and in what way. This would solve one of the major issues of just using linear regressions, which is not completely understanding causal mechanisms.

4.3 Other recommendations

1. Bootstrapping - For a future interface, we rst recommend the use of bootstrapping methodology to improve the convincingness of the results is desired, bootstrapping software can give the key causal relationships that are supported by the most the sample in a easy to understand manner. A better reporting of bootstrapping in the interface can make the results more convincing.

2. Learning by playing - Secondly, the concept of learning by playing (inspired by the 2024 Exploring the Feeling of Causality Exposition), although we implemented this concept in our interface, we can improve upon this by adding the interface to other interactive tools with touch screen such as a PC with a touchscreen, an iPad / tablet, or smartphone, the graph can be manipulated through `playing'; it is a hands-on approach versus using the mouse on the computer. We believe that adding this to an interface can improve convincingness, discoverability and variety. Additionally, we believe that something like VR integration for advanced data manipulation and interaction can help all three of these variables. Another idea related to this is the use of Generative AI to create a theme for the interface, improving the design of it depending on the theme of the database, making it more engaging for the user.

3. Generalization of the model - Third, a generalization of the model used in the interface, especially for the causal graph generation. At rst we wanted to generalize the results, then we drifted from this goal because of time constraints. In the future, we think that a correlation study between the metrics to order the graphs we introduced (hierarchical ordering, statistical strength, new rate variable) and the di erent values of the parameters used can help us generalize what default parameters should look like. Some interesting lines of research can be the relationship between coherence vs Rate or goodness of t tests or population size or background knowledge input.

4. Literature Review - Fourth, a thorough literature review of causal graph implementation and interpretation of the numerous applications in order to better understand the standard practices in research for numerous topics. How are researchers speci cally utilizing this methodology to draw conclusions? As DirectLiNGAM grows in popularity, what are the standard practice values of statistical reliability by research topic and what are the standard values for di erent disciplines and di erent databases?

References

[1] Apple. Human Interface Guidelines - Accessibility. 2024.url : https://developer.apple.com/design/human-interface-guidelines/accessibility (visited on 06/25/2024) (cit. on p. 41).

[2] Apple. Inclusion & Diversity . 2024. url : https://www.apple.com/diversity/ (visited on 07/25/2024) (cit. on p. 42).

[3] D. Applebaum. Probability and Information: An Integrated Approach . Cambridge University Press, 2008. isbn: 9781139473255url : https://books.google.co.jp/books?id=szYZEd6-NvEC (cit. on p. 12).

[4] Alejandro Barredo Arrieta et al. Explainable Arti cial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI2019. arXiv: 1910.10045 (cit. on p. 44).

[5] Giorgio Ausiello and Luigi Laura. \Directed hypergraphs: Introduction and fundamental algorithms|A survey". In: Theoretical Computer Science 658 (2017), pp. 293{306. issn: 0304-3975.doi : https://doi.org/10.1016/j.tcs.2016.03.016. url : https://www.sciencedirect.com/science/article/pii/S0304397516002097(cit. on p. 16).

[6] Peter W. Battaglia et al. \Relational inductive biases, deep learning, and graph networks". In: CoRR abs/1806.01261 (2018). arXiv:1806.01261. url : http://arxiv.org/abs/1806.01261 (cit. on p. 7).

[7] Aditya Bikkani. Situational Disabilities: A Hidden Challenge for Accessibility. https://aeldata.com/situational-disabilities . 2023. (Visited on 07/25/2024) (cit. on p. 42).

[8] Alexander Bochman and Vladimir Lifschitz. \Pearl's Causality in a Logical Setting". In: Proceedings of the AAAI Conference on Arti cial Intelligence 29.1 (2015).doi : 10.1609/aaai.v29i1.9411 . url : https://ojs.aaai.org/index.php/AAAI/article/view/9411 (cit. on p. 7).

[9] Leo Breiman. \Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: Statistical science 16.3 (2001), pp. 199{231 (cit. on p. 12).

[10] Ruichu Cai et al. \SELF: Structural equational likelihood framework for causal discovery". In: 2018 (cit. on p. 61).

[11] Alastair Campbell et al. Web Content Accessibility Guidelines (WCAG) 2.2. 2023. url : https://www.w3.org/TR/WCAG22/ (visited on 06/25/2024) (cit. on pp. 41, 43).

[12] Kyunghyun Cho. A Brief Introduction to Causal Inference in Machine Learning . 2024. arXiv: 2405.08793 [cs.LG] . url : https://arxiv.org/abs/2405.08793 (cit. on p. 6).

[13] Ciara Conduit et al. \Occam's Razor and Hickam's Dictum: A Rare Case of Synchronous Solid and Hematological Malignancies and Transformed EGFR-Mutated NSCLC". In: Journal of Thoracic Oncology 11.11 (2016), pp. 131{133issn: 1556-0864. doi : 10.1016/j.jtho.2016.06.027 . url : https://doi.org/10.1016/j.jtho.2016.06.027 (cit. on pp. 5, 62).

[14] 101st United States Congress. Americans With Disabilities Act of 1990. Pub. L. No. 101-336, 104 Stat. 328. 1990 (cit. on p. 43).

[15] Bernat Corominas-Murtra et al. \On the origins of hierarchy in complex networks". In: Proceedings of the National Academy of Sciences 110.33 (July 2013), pp. 13316{13321. issn: 1091-6490. doi: 10.1073/pnas.1300832110 (cit. on p. 16).

[16] Michele Coscia. \Using arborescences to estimate hierarchicalness in directed complex networks". In: PLoS ONE 13.1 (2018), e0190825. doi: 10.1371/journal.pone.0190825 (cit. on p. 16).

[17] Thomas M. Cover and Joy A. Thomas. \Entropy, Relative Entropy and Mutual Information". In: Elements of Information Theory. John Wiley & Sons, Ltd, 2001. Chap. 2, pp. 12{49. isbn: 9780471200611. doi: https://doi.org/10.1002/0471200611.ch2 (cit. on pp. 10, 12).

[18] F. Crameri, G.E. Shephard, and P.J. Heron. \The misuse of colour in science communication". In: Nature Communications 11.5444 (2020). doi: 10.1038/s41467-020-19160-7 (cit. on p. 41).

[19] Ziqiang Cui et al. Treatment-Aware Hyperbolic Representation Learning for Causal E ect Estimation with Social Networks. 2024. arXiv: 2401.06557 [cs.LG]. url: https://arxiv.org/abs/2401.06557.

[20] John Ellson, Emden R. Gansner, and Eleftherios Koutso os. \Graphviz and Dynagraph { Static and Dynamic Graph Drawing Tools". In: 2003. url: https://api.semanticscholar.org/CorpusID:1529542 (cit. on p. 19).

[21] Felix Elwert. \Graphical causal models". In: Handbook of Causal Analysis for Social Research. Springer Netherlands, 2013, pp. 245{273 (cit. on p. 47).

[22] Felix Elwert. \Introduction to Directed Acyclic Graphs (DAGs) for Causal Inference". In: SAMSI program on Data Science in Social and Behavioural Sciences (2021) (cit. on pp. 48, 49).

[23] Mustafa Emirbayer and Ann Mische. \What Is Agency?" In: American Journal of Sociology 103.4 (1998), pp. 962{1023. issn: 00029602, 15375390 (cit. on p. 45).

[24] Richard Feynman. Fun to Imagine | Using physics to explain how the world works. English. HTML5 Uploader 1.6.3. Available online: Internet Archive. Internet Archive, 1983 (cit. on p. 17).

[25] Marta Freixa et al. \Occam's razor versus Hickam's dictum: two very rare tumours in one single patient". In: Oxford Medical Case Reports 2019.5 (May 2019), omz029. issn: 2053-8855. doi: 10.1093/omcr/omz029. url: https://doi.org/10.1093/omcr/omz029 (cit. on pp. 5, 62).

[26] Fujitsu. Diversity, Equity & Inclusion. 2024. url: https://www.fujitsu.com/global/about/csr/diversity/ (visited on 07/25/2024) (cit. on p. 42).

[27] Emden R. Gansner. The DOT Language Online. 2002. url: http://www.research.att.com/~erg/graphviz/info/lang.html/ (cit. on p. 19).

[28] Marta Garnelo and Murray Shanahan. \Reconciling deep learning with symbolic arti cial intelligence: representing objects and relations". In: Current Opinion in Behavioral Sciences 29 (2019). Arti cial Intelligence, pp. 17{23. issn: 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2018.12.010 (cit. on p. 6).

[29] A. Giddens. The Constitution of Society: Outline of the Theory of Structuration. Outline of the Theory of Structuration. University of California Press, 1984. isbn: 9780520052925 (cit. on p. 45).

[30] Amos Golan and John Harte. \Information theory: A foundation for complexity science". In: Proceedings of the National Academy of Sciences 119.33 (2022), e2119089119. doi: 10.1073/pnas.2119089119. url: https://www.pnas.org/doi/abs/10.1073/pnas.2119089119.

[31] Google. Belonging. 2024. url: https://about.google/belonging/ (visited on 07/25/2024) (cit. on p. 42).

[32] Ruocheng Guo et al. \A Survey of Learning Causality with Data: Problems and Methods". In: CoRR abs/1809.09337 (2018). arXiv: 1809.09337. url: http://arxiv.org/abs/1809.09337 (cit. on pp. 4, 60).

[33] Hiroyuki Higuchi. \Enhancing explainability of causal discovery AI". In: (2024). url: https://www.mccs.tohoku.ac.jp/g-rips/open/2024/pdf/fujitsu_project_2024.pdf (cit. on p. 6).

[34] Christopher Hitchcock. \Causal Models". In: The Stanford Encyclopedia of Philosophy. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University, 2024 (cit. on pp. 8, 60).

[35] Patrik Hoyer et al. \Nonlinear causal discovery with additive noise models". In: Advances in Neural Information Processing Systems. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc., 2008 (cit. on p. 9).

[36] Patrik O. Hoyer and Antti Hyttinen. \Bayesian discovery of linear acyclic causal models". In: 2009, pp. 240{248 (cit. on p. 61).

[37] Aapo Hyvarinen and Stephen M. Smith. \Pairwise likelihood ratios for estimation of non-Gaussian structural equation models". In: Journal of machine learning research : JMLR 14 Jan (2013), pp. 111{152 (cit. on pp. 61{63, 65).

[38] Aapo Hyvarinen et al. Independent component analysis. Springer, 2009 (cit. on p. 61).

[39] IBM. Design - Visual. 2024. url: https://www.ibm.com/able/toolkit/design/visual/ (visited on 06/25/2024) (cit. on p. 41).

[40] Anna A. Igolkina and Georgy Meshcheryakov. \semopy : A Python package for Structural Equation Modeling". In: Structural Equation Modeling: A Multidisciplinary Journal, 27:6, 952-963 (2020). doi: https://doi.org/10.1080/10705511.2019.1704289 (cit. on p. 14).

[41] Hiroaki Iwashita et al. \E cient Constrained Pattern Mining Using Dynamic Item Ordering for Explainable Classi cation". In: CoRR abs/2004.08015 (2020). arXiv: 2004.08015. url: https://arxiv.org/abs/2004.08015 (cit. on p. 65).

[42] Bernhard Jenny and Nathaniel Vaughn Kelso. \Color Design for the Color Vision Impaired". In: Cartographic Perspectives 58 (2007), pp. 61{67. doi: 10.14714/CP58.270. url: https://cartographicperspectives.org/index.php/journal/article/view/cp58-jenny-kelso (cit. on p. 41).

[43] Yan Jia et al. \Extrapolation over temporal knowledge graph via hyperbolic embedding". In: CAAI Transactions on Intelligence Technology 8.2 (2023), pp. 418{429. doi : https://doi.org/10.1049/cit2.12186 .

[44] Uday Kamath, Kenneth Graham, and Mitchell Naylor. Applied Causal Inference Independent, 2023 (cit. on p. 60).

[45] C. Knappett and Lambros Malafouris. Material Agency: Towards a Non-Anthropocentric Approach. Jan. 2008. isbn : 9780387747101doi : 10.1007/978-0-387-74711-8 (cit. on p. 45).

[46] Samory Kpotufe et al. \Consistency of causal inference under the additive noise model". In: International Conference on Machine Learning. PMLR, 2014, pp. 478{486 (cit. on p. 27).

[47] David Krackhardt. \Graph Theoretical Dimensions of Informal Organization". In: Computational Organization Theory 89 (June 1994) (cit. on p. 16).

[48] Eric Larson and Isabel Vogt. \Making Accessible Documents Using LaTeX". In: Notices of the American Mathematical Society (2023). doi : https://dx.doi.org/10.1090/noti2606 (cit. on p. 41).

[49] Guilherme F. Lima, Rodrigo Costa, and Marcio Ferreira Moreno. \An Introduction to Arti cial Intelligence Applied to Multimedia". In: CoRR abs/1911.09606 (2019). arXiv: 1911.09606. url : http://arxiv.org/abs/1911.09606 (cit. on p. 6).

[50] D.J.C. MacKay. Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003. isbn : 9780521642989 (cit. on p. 12).

[51] Gary Marcus. \Deep Learning: A Critical Appraisal". In: CoRR abs/1801.00631 (2018). arXiv: 1801.00631. url : http://arxiv.org/abs/1801.00631 (cit. on p. 7).

[52] Microsoft. Diversity and inclusion. 2024. url : https://www.microsoft.com/en-us/diversity/default?msockid=0cd3871ac5ff6f463b8f93b8c4856ef8 (visited on 07/25/2024) (cit. on p. 42).

[53] Microsoft. Inclusive Microsoft Design. 2016. url : https://inclusive.microsoft.design/tools-and-activities/Inclusive101Guidebook.pdf (visited on 07/25/2024) (cit. on pp. 42, 43).

[54] Enys Mones, Lilla Vicsek, and Tamas Vicsek. \Hierarchy Measure for Complex Networks". In: PLoS ONE 7.3 (2012). Ed. by Stefano Boccaletti, e33799issn: 1932-6203.doi : 10.1371/journal.pone.0033799 (cit. on p. 16).

[55] Gregoire Montavon, Wojciech Samek, and Klaus-Robert Muller. \Methods for interpreting and understanding deep neural networks". In: Digital Signal Processing 73 (2018), pp. 1{15. issn : 1051-2004.doi : https://doi.org/10.1016/j.dsp.2017.10.011 . url : https://www.sciencedirect.com/science/article/pii/S1051200417302385 (cit. on p. 44).

[56] Giannis Moutsinas et al. \Graph hierarchy: a novel framework to analyse hierarchical structures in complex networks". In: Scienti c Reports 11.1 (2021). issn : 2045-2322.doi : 10.1038/s41598-021-93161-4 . url : https://doi.org/10.1038/s41598-021-93161-4 (cit. on p. 16).

[57] R Muto et al. \Predicting oxygen requirements in patients with coronavirus disease 2019 using an arti cial intelligence-clinician model based on local non-image data". In: P. Front. Med. (2022). doi : doi:10.3389/fmed.2022.1042067 (cit. on pp. 4, 9, 10, 12, 62, 65).

[58] M. Nauta. Temporal causal discovery and structure learning with attention-based convolutional neural networks 2018. url : http://essay.utwente.nl/76360/ (cit. on p. 6).

[59] Thomas A. Newman et al. \Occam's Razor versus Hickam's Dictum". In: Annals of the American Thoracic Society 14.11 (2017). PMID: 29090995, pp. 1709{1713. doi : 10.1513/AnnalsATS.201701-087CC. url : https://doi.org/10.1513/AnnalsATS.201701-087CC (cit. on pp. 5, 62).

[60] David Nichols. Coloring for Colorblindness. 2021. url : https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40 (visited on 06/25/2024) (cit. on p. 41).

[61] Kotaro Ohori et al. \Wide Learning Technology to Provide Trust Through Knowledge Discovery". In: Fujitsu Scienti c & Technical Journal 56.1 (2020), pp. 46{51 (cit. on pp. 4, 9, 62, 65).

[62] Pramod Kumar Parida, Tshilidzi Marwala, and Snehashish Chakraverty. \Altered-LiNGAM (ALiNGAM) for solving nonlinear causal models when data is nonlinear and noisy". In: Communications in Nonlinear Science and Numerical Simulation 52 (2017), pp. 190{202. issn: 1007-5704. doi : https://doi.org/10.1016/j.cnsns.2017.04.018 (cit. on pp. 62, 64).

[63] European Parliament. Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services 2019 (cit. on p. 43).

[64] Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000. isbn : 9780521895606 (cit. on p. 7).

[65] Judea Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988. isbn : 9780080514895 (cit. on p. 8).

[66] Judea Pearl. \The causal foundations of structural equation modeling". In: Handbook of Structural Equation Modeling (2012). Ed. by R. H. Hoyle, pp. 68{91 (cit. on p. 7).

[67] Judea Pearl. \The Do-Calculus Revisited". In: CoRR abs/1210.4852 (2012). arXiv: 1210.4852. url : http://arxiv.org/abs/1210.4852 (cit. on p. 7).

[68] Judea Pearl and Thomas S. Verma. \A theory of inferred causation". In: Logic, Methodology and Philosophy of Science IX. Ed. by Dag Prawitz, Brian Skyrms, and Dag Westerstahl. Vol. 134. Studies in Logic and the Foundations of Mathematics. Elsevier, 1995, pp. 789{811. doi : https://doi.org/10.1016/S0049-237X(06)80074-1 (cit. on p. 8).

[69] Peter Peduzzi et al. \Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates". In: Journal of clinical epidemiology 48.12 (1995), pp. 1503{1510 (cit. on p. 15).

[70] Erzsebet Ravasz and Albert-Laszlo Barabasi. \Hierarchical organization in complex networks". In: Physical Review E 67.2 (2003). issn: 1095-3787. doi: 10.1103/physreve.67.026112 (cit. on p. 16).

[71] John Robb. \Beyond agency". In: World Archaeology 42.4 (2010), pp. 493{520. issn: 00438243, 14701375 (cit. on p. 45).

[72] J. Thomas Rosemary, Judith Mashtho , and Nir Oren. \Can I In uence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale". In: Frontiers in arti cial intelligence . Vol. 2. Studies in Logic and the Foundations of Mathematics. 2019, p. 24. doi: https://doi.org/10.3389/frai.2019.00024 (cit. on p. 30).

[73] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M. Todd. \Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload". In: Journal of Consumer Research 37.3 (2010), pp. 409{425. issn: 0093-5301. doi: 10.1086/651235. url : https://doi.org/10.1086/651235 (cit. on p. 46).

[74] Barry Schwartz. The Paradox of Choice Harper Perennial, 2004 (cit. on p. 46).

[75] Lesia Semenova, Cynthia Rudin, and Ronald Parr. \On the existence of simpler machine learning models". In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022, pp. 1827{1858 (cit. on p. 13).

[76] Shohei Shimizu et al. \A Linear Non-Gaussian Acyclic Model for Causal Discovery". In: J. Mach. Learn. Res. 7 (2006), pp. 2003{2030. issn: 1532-4435 (cit. on pp. 9, 62, 63).

[77] Shohei Shimizu et al. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. 2011. arXiv: 1101.2489 (cit. on pp. 4, 13, 27, 47, 61{63).

[78] Maarten van Smeden et al. \Sample size for binary logistic prediction models: beyond events per variable criteria". In: Statistical methods in medical research 28.8 (2019), pp. 2455{2474.

[79] Peter Spirtes. \An Anytime Algorithm for Causal Inference". In: Proceedings of the Eighth International Workshop on Arti cial Intelligence and Statistics . Ed. by Thomas S. Richardson and Tommi S. Jaakkola. Vol. R3. Proceedings of Machine Learning Research. Reissued by PMLR on 31 March 2021. PMLR, 2001, pp. 278{285. url : https://proceedings.mlr.press/r3/spirtes01a.html (cit. on p. 50).

[80] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, Prediction, and Search Vol. 81. 1993. isbn: 9781461276500. doi: 10.1007/978-1-4612-2748-9 (cit. on p. 7).

[81] Peter Spirtes and Kun Zhang. \Causal discovery and inference: concepts and recent methodological advances". In: Appl Inform (Berl) (2017). doi: 10.1186/s40535-016-0018-x (cit. on p. 4).

[82] James V. Stone. \Information Theory: A Tutorial Introduction". In: CoRR abs/1802.05968 (2018). arXiv: 1802.05968. url : http://arxiv.org/abs/1802.05968 (cit. on p. 12).

[83] Ala Trusina et al. \Hierarchy Measures in Complex Networks". In: Physical Review Letters 92.17 (2004). issn: 1079-7114. doi: 10.1103/physrevlett.92.178702 (cit. on p. 16).

[84] E.R. Tufte. Beautiful Evidence. Graphics Press, 2006. isbn: 9780961392178 (cit. on p. 41).

[85] E.R. Tufte. Envisioning Information. Graphics Press, 1990. isbn: 9781930824140 (cit. on p. 41).

[86] E.R. Tufte. Seeing with Fresh Eyes: Meaning, Space, Data, Truth. Graphics Press, 2020. isbn: 9780961392192 (cit. on p. 41).

[87] E.R. Tufte. The Visual Display of Quantitative Information. Graphics Press, 2001. isbn: 9781930824133 (cit. on p. 41).

[88] E.R. Tufte. Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, 1997. isbn: 9781930824157 (cit. on p. 41).

[89] Jim R. Tyson. Accessible documents from LaTeX. 2020. url: https://blogs.ucl.ac.uk/digital-education/2020/07/22/accessible-documents-from-latex/ (visited on 06/25/2024) (cit. on p. 41).

[90] UCLA. A PRACTICAL INTRODUCTION TO FACTOR ANALYSIS: CONFIRMATORY FACTOR ANALYSIS. 2024. url: https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis-confirmatory-factor-analysis/ (visited on 07/09/2024).

[91] Shinsuke Uda. "Application of information theory in systems biology". In: Biophysical Reviews 12 (2 2020). doi: 10.1007/s12551-020-00665-w (cit. on p. 10).

[92] Yana Vasileva. How Accessibility Can Improve Your Reach: The Case of Situational Disabilities. Series: Digital Experience. 2024. url: https://www.progress.com/blogs/how-accessibility-improve-reach-situational-disabilities (visited on 07/25/2024) (cit. on p. 42).

[93] Alexei Vazquez, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Large-scale topological and dynamical properties of the Internet". In: Physical Review E 65.6 (2002). issn: 1095-3787. doi: 10.1103/physreve.65.066130 (cit. on p. 16).

[94] Zijie J. Wang et al. "TimberTrek: Exploring and Curating Sparse Decision Trees with Interactive Visualization". In: 2022 IEEE Visualization and Visual Analytics (VIS). IEEE, 2022. doi: 10.1109/vis54862.2022.00021. url: http://dx.doi.org/10.1109/VIS54862.2022.00021.

[95] Rui Xin et al. "Exploring the whole rashomon set of sparse decision trees". In: Advances in neural information processing systems 35 (2022), pp. 14071–14084 (cit. on p. 12).

[96] Alessio Zanga and Fabio Stella. A Survey on Causal Discovery: Theory and Practice. 2023. arXiv: 2305.10032 (cit. on pp. 4, 60).

[97] Yan Zeng et al. "A causal discovery algorithm based on the prior selection of leaf nodes". In: Neural Networks 124 (2020), pp. 130–145. issn: 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.12.020 (cit. on pp. 61, 62, 64, 65).

[98] Jiang Zhang et al. Dynamical Reversibility and A New Theory of Causal Emergence. 2024. arXiv: 2402.15054 [cond-mat.stat-mech]. url: https://arxiv.org/abs/2402.15054 (cit. on p. 50).

[99] Kun Zhang and Aapo Hyvarinen. On the Identifiability of the Post-Nonlinear Causal Model. 2012. arXiv: 1205.2599 (cit. on p. 9).

5 Appendix: minimum working de nitions

Some working de nitions and commentary are listed here.

De nition 5.1 (Interpretability). A passive process;extent to which a human can understand the cause of a decision made by a model

`Interpretability' and `explainability' are often used interchangeably, but are technically distinct things; where one is used in this project, the other is implied, unless stated so otherwise.

De nition 5.2 (Explainability). An active process; methods or techniques used by a model to clarify or justify its internal functions or outputs

De nition 5.3 (Causality). Generic relationship between an e ect and the cause that gives rise to it [32]

De nition 5.4 (Causal model). Representation of causality that makes predictions about the behavior of a system; entails the truth value, or the probability, of counterfactual claims about the system; it predicts the e ects of interventions; it entails the probabilistic dependence or independence of variables included in the model [34]

De nition 5.5 (Structural causal model). Relationships between variables expressed as deterministic, functional relationships; probabilities are introduced through the assumption that certain variables are exogenous latent random variables; set of equations that describe all causal relations in a system; abbreviated SCM; also known as (nonparametric) structural equation models (SEM)

De nition 5.6 (Causal inference). The task of quantifying the impact of a cause on its e ect [96], including the e ects of intervention; formal process which allows us to measure causal e ects from data [44]

De nition 5.7 (Causal discovery). General process of learning graphical structures with a causal interpretation [96]; discovery may refer to the recovery of a set of structures (Markov equivalent classes), or the recovery of the unique causal structure

De nition 5.8 (Intervention). Action or activity the variable unit is subjected to; sets the value of that variable by a process that overrides the usual causal structure, without interfering with the causal processes governing the other variables [34]

De nition 5.9 (Counterfactual). Unobserved outcomes that would have occurred for each individual had they been assigned to a di erent treatment [44]; proposition in the form of a subjunctive conditional [34]

De nition 5.10 (Confounding variable). A variable that is positively or negatively associated with both the dependent variable and an independent variable [44]; latent (hidden) variables are not measured directly (observed variables and mathematical inference uncover the existence and relationship of latent variables)

6 Appendix: comparing complexity

Here we give a brief outline and discussion of model complexities. The Rashomon set includes LiNGAM and WL models; the extent to which the complexity of each model affects CVD may vary by user. Complexity of model types, therefore, permeates CVD. Here, we give an overview of the models we are exploring, and analyze their relative complexities and the high-level nature of data with which they work that impact complexity, and, therefore, CVD.

6.1 Model complexity

There are three main types of LiNGAM models:

1. Independent Component Analysis (ICA) models
2. Score (Bayes) models
3. Root models

ICA- and score-based models are essentially function optimization problems, and so inherit the issues therein (sensitivity to initial values and gradient-based traps in local optima rather than global optima [38]). ICA-based models show the full structure without pre-specifying causal ordering on the variables. Score-based models construct maximum likelihood estimation functions for scoring causal relationships [36], and global optimization follows from a local statistical significance metric [10].

Root-based models [37, 77] are non-parametric estimators that directly estimate causality in a finite number of steps by identifying a root node and performing a data updating process to determine causal ordering [97]; consequently when sample size is not adequate, there may not be enough information for the updating process, and when data is high-dimensional, more data updating processes are necessary. However, when the sample size is small compared to the dimensionality, convergence is guaranteed.

The data updating process follows from the fact that the selected node must affect others, and before moving to the next node, it must check that their influences on other nodes are removed.

Deviant from the three main types of LiNGAM models is a leaf-based model [97], in which priority is given to leaf nodes rather than roots nodes. The strategy is as follows: leafs do not affect others (otherwise, we would violate the MC discussed above); therefore, such models may directly estimate a causal ordering in a sort of `bottom-up' way without data updating processes, since removing leaves iteratively does not affect the causal structure elsewhere. Mechanically, identifying leaf nodes follows from computing the regression of a variable on another variable (pairwise regression), as opposed to a variable on all other variables.

Computational complexities of various LiNGAM models and WL are given in Table 5, with their respective references, as well as the data for which each model is better applied (relative to dimensionality and sample size; see Appendices Section 6 for more details). Note that n is the sample size, p is the dimensionality (number of variables), M ($<< n$) is the maximal rank found by low-rank decomposition used in kernel-based independence measures (used throughout LiNGAM models), and K is the length of variable combinations in WL conditions.

See the Appendices Section 6 for a description of each model.

Table 5: Comparison of model complexities

ID	Model	Complexity	Data	Citation
A	ICA-LiNGAM	$O\ np^3 + p^4$	high n; low p	[76, 77]
B	DirectLiNGAM	$O\ np^3M^2 + p^4M^3$	high (in nite) n; low p	[77]
C_4; C_2	ALiNGAM	max: $O\ np^4$; min: $O\ np^2$	high n; low p	[62]
D	GPL LiNGAM	$O\ np^2$	low n; high p	[97]
E	Pairwise-LiNGAM	$O\ np^2 + np^3$	low n; high p	[37, 97]
F_K	Wide Learning	$O\ np^K$	low K ; high n; low p	[57, 61]

n, p, and M are determined by data; K is a tuneable parameter that signi cantly a ects the performance of WL relative to the various LiNGAM models shown in Table 5. By tuning K , we compare the complexities of the models in a complexity hierarchy (for visualization, WL is in boldface):

There are other hyperparameters for WL, but K is what contributes directly to big-O complexity.

1. For K 4
   F_4 $C_4 > B > A > E > D$ C_2
2. For K = 3
   $C_4 > B > A > E >$ $F_3 > D$ C_2
3. For K = 2
   $C_4 > B > A > E > D$ F_2 C_2

In this way, we learn why K = 4 is suggested as the upper limit for WL, and why it was mentioned by the industry mentor that no more than 4 should be seriously considered: the computational cost of WL for K 4 is competitive with the most expensive LiNGAM model for computational complexity. Additionally, if a value K 4 is considered, then are we really gaining anything meaningful for CVD? That is, this then induces a technical and philosophical dilemma between Occam's razor and Hickam's dictum [13, 25, 59]: more features in a condition tend to break Occam's razor, but Hickam's dictum counters this in that the world is complicated and as many features as can happen very well might happen and are responsible for a result.

As a heuristic, we put a limiter at K = 4 for practical computational reasons, but the extent to which such a limiter impacts CVD as a result of the dilemma is a very real problem. Do we discover causal relationships we might have otherwise missed with our hard limit? Possibly, but given computational resources at hand, we will set that aside for now.

6.2 Complexities of LiNGAM models and WL

Let us understand where the model complexities come from for our causality project, as well as the nature of datasets that each model is better prepared with which to work (that is, how sample size and dimensionality relate to model performance).

In general: root-based frameworks are sensitive to sample size, especially with high dimensional data (due to the iterative data updating process); computational complexity and accuracy are usually unsatis ed when the data is high-dimensional or the sample

Experimental results validating model performance may be found in the appropriate literature; here we are evaluating based on complexity.

size is too small. If either the sample size or the magnitude of nongaussianity is small, LiNGAM analyses tend to provide signi cantly di erent results for di erent bootstrap samples; smaller nongaussianity causes the model to become closer to not being identi - able. Essentially, the data updating process needs enough samples to evaluate, and if the dimensionality is high, then more data updating processes must iterate.

The following is not meant to be in-depth or in-detail; rather, it is meant to see how the nature of data sample size and dimensionality contribute to the complexity of each model, such that if complexity is a heuristic for model choice, then we have some principled representation metric with which to work. It may be important to choose a suitable algorithm depending on data dimension, sample size, noise level, the distributions of the external in uences, and other relevant factors [37].

6.2.1 ICA-LiNGAM

- Independent Component Analysis (ICA)-LiNGAM [76, 77]
- Complexity: $O\ np^3 + p^4$

  Iterations in FastICA (used in ICA-LiNGAM) are known to be $O\ np^2$. If we assume some number N of iterations, then ICA-LiNGAM has complexity $O\ Np^2 + p^4$. N is conjectured to grow linearly with p (conjecture only; it is practically di cult to validate this since, in general, the required number of iterations is not known). Therefore, ICA-LiNGAM total budget is $O\ np^3 + p^4$.
- With low samples n, there is not enough data for the data updating process, which, in turn, induces errors that may cause the wrong identi cation of the next root node. CPU time of ICA-LiNGAM has been shown to decrease with higher n.

  When high-dimensional p, we induce a high complexity in addition to cascading errors; the higher the dimensionality p, the more the data updating processes should be conducted.
- ICA-LiNGAM, therefore, wants: high n and low p

6.2.2 Direct LiNGAM

- Direct LiNGAM [77]
- Complexity: $O\ np^3M^2 + p^4M^3$

  DirectLiNGAM has two parts of its algorithm design that dominate the computation process.

  1. Compute pairwise independence (i.e., a kernel-based independence) for each variable. This process requires $O\ np^2M^2 + p^3M^3$ operations across p 1 iterations. Therefore, $O\ np^3M^2 + p^4M^3$
  2. Compute regression to estimate the weight matrix of the linear f_i to determine x_i. Complexity follows from many representative regressions (including a least squares algorithm) such that $O\ np^3$.

  Therefore, DirectLiNGAM total budget is $O\ np^3M^2 + p^4M^3$

- DirectLiNGAM is assumed to work with in nite samples n

  While samples are assumed to be in nite, there must be a nite number of variables p (else the complexity explodes); DirectLiNGAM guarantees convergence within xed number of steps equal to number of variables if all assumptions met and sample size is in nite

- Computation of DirectLiNGAM is larger than ICA-LiNGAM when the sample size n increases, in accordance to the sample-size assumption of DirectLiNGAM, but DirectLiNGAM guarentees convergence. DirectLiNGAM, therefore, wants: high (in- nite) n and low p.

6.2.3 ALiNGAM

- Altered-LiNGAM [62]

- Complexity: max: $O(np^4)$; min: $O(np^2)$

  ALiNGAM is exactly sensitive to the number of nodes (i.e., the number of features). ALiNGAM computes directions for two nodes at a time by checking their opposite directions, so there exists $C(p;r) = C(p;2)$ number of combinations of nodes for p. This implies there are $2\ C(p;2)$ equations to be solved to nd all probable causal directions. If a pair of nodes are visited only once, then complexity is $O(np^2)$; with more visits, however, complexity tends towards $O(np^4)$. Therefore, ALiNGAM total budget is minimally $O(np^2)$ and maximally $O(np^4)$.

- Dimensionality for ALiNGAM directly impacts the combinatorial calculation for the number of equations to solve in addition to the choice of number of visitations to each node. A high number of samples, however, are needed for convergence. ALiNGAM, therefore, wants: high n and low p.

6.2.4 GPL LiNGAM

- Gives Priority to Leaf Nodes (GPL) LiNGAM [97]

- Complexity: $O(np^2)$

  GPL LiNGAM has two parts of its algorithm design that dominate the computation process.

  1. Iterating independence tests for each feature p $\frac{(p-1)}{2}$ times requires $O(n)$ operations for each iteration. Therefore, $O(np^2)$.
  2. Process of maximum entropy approximation to extimate the
  3. Estimate the likelihood ratio matrix via approximation of maximum entropy requires $O(n)$ in $p(p-1)$ iterations. Therefore, $O(np^2)$.

  Therefore, GPL LiNGAM total budget is $O(np^2)$.

- Relative to root-based LiNGAM models, GPL LiNGAM performs better for higher dimensional data. Additionally, compared to DirectLiNGAM, GPL LiNGAM tends to favorably perform when given a small sample size relative to high dimensional data. GPL LiNGAM, therefore, wants: low n and high p.

6.2.5 Pairwise LiNGAM

- Pairwise LiNGAM [37, 97]
- Complexity: $O\ np^2 + np^3$

  Pairwise LiNGAM has no calculation for independence tests (contrary to the other LiNGAM models), but it does iterate the data updating process and compute the likelihood ratio matrix each time after choosing the next root.

  1. For the rst iteration (i.e., the rst root node): $O\ np^2$
  2. For the second iteration (i.e., the second root node): $O\ np + np^2$

  Repeating the updating process for $(p \quad 1)$ iterations results in Pairwise LiNGAM total budget $O\ np^2 + np^3$.

- Similar to other root-based LiNGAM models, with a limited number of samples, errors may occur during the data updating process, which may adversely in uence successive nodes or cause wrong identi cation of nodes. Likewise for dimensionality, high dimensionality simply scales the issues with the number of samples. Performance of Pairwise LiNGAM, however, achieves high accuracies for when number of data points is small compared to the dimension of the data, or the data is noisy. Pairwise LiNGAM, therefore, wants: low n and high p, or noisy data.

6.2.6 WL

- Wide Learning™ [57, 61]
- Complexity: $O\ np^K$

  For each variable p, combinations up to length K are checked for relevance, scaled by the sample size. Therefore, WL total budget is $O\ np^K$.

- Immediately, one notices a risk for explosion in WL, due to the combinatorics of the algorithm (as combinations scale as p^K and overall complexity follows as this scaled by n); [57] mitigate explosion via a method derived from contrast pattern search for sparse and dense data [41], a kind of `dynamic pruning` algorithm based on depth- rst search. WL, therefore, wants: low K (to be better than LiNGAM, maximum value 4), high n and low p.

The details of [41] are not relevant here, other than we note that this algorithm mitigates explosion.