Causal Deep Learning

Jeroen Berrevoets *jeroen.berrevoets@maths.cam.ac.uk*
DAMTP, University of Cambridge

Krzysztof Kacprzyk *kk751@cam.ac.uk*
DAMTP, University of Cambridge

Zhaozhi Qian *zhaozhi.qian@maths.cam.ac.uk*
DAMTP, University of Cambridge

Mihaela van der Schaar *mv472@cam.ac.uk*
DAMTP, University of Cambridge
The Alan Turing Institute

Abstract

Causality has the potential to truly transform the way we solve a large number of real-world problems. Yet, so far, its potential remains largely unlocked since most work so far requires strict assumptions which do not hold true in practice. To address this challenge and make progress in solving real-world problems, we propose a new way of thinking about causality – we call this *causal deep learning*. The framework which we propose for causal deep learning spans three dimensions: (1) a structural dimension, which allows incomplete causal knowledge rather than assuming either full or no causal knowledge; (2) a parametric dimension, which encompasses parametric forms which are typically ignored; and finally, (3) a temporal dimension, which explicitly allows for situations which capture exposure times or temporal structure. Together, these dimensions allow us to make progress on a variety of real-world problems by leveraging (sometimes incomplete) causal knowledge and/or combining diverse causal deep learning methods. This new framework also enables researchers to compare systematically across existing works as well as identify promising research areas which can lead to real-world impact.

Contents

1 Introduction

Across areas such as medicine, economics, marketing, and beyond we ask questions like: *"would sales increase if we do this?"* or *"would that drug have prevented the cancer from spreading?"*. While these questions are causal in nature [1], it is often not possible to use causality in these real-world settings. The reason is that causality often requires a large set of assumptions which cannot be met in practice. These assumptions vary from: knowing the ground-truth causal graph [2], or assuming all relevant variables are present in the data which cannot be guaranteed [3, 4]. As such, *the promise of causality remains unfulfilled* [5].

Causal deep learning in three dimensions. Instead, we propose *causal deep learning* (CDL) to tackle these real-world problems. With much less strict assumptions and based on the enormous flexibility of deep learning models [6–8], CDL is uniquely positioned to provide solutions to these challenging settings [9]. We characterise causal deep learning along three dimensions, which we formalise in detail in this paper. First, a **structural dimension**, which allows incomplete causal knowledge rather than assuming either full or no causal knowledge. By allowing incomplete structural causal information, we effectively introduce a *"rung 1.5"* on the now well-known ladder of causation introduced by Judea Pearl [10]. Second, a **parametric dimension**, formalising parametric forms which are typically ignored yet are incredibly important in causality. As an example, we not only care *if* a variable is causing another but also *how* it is causal. Third, a **temporal dimension**, which explicitly allows for situations which include exposure times or temporal structure. Not only is time an essential component in causality (as causes precede effects [11]), but time can also influence *whether or not* something is causal. For example, only when exposed *enough* does a drug affect certain diseases. As such, using these three dimensions, we can define CDL as follows:

Causal deep learning studies (deep) machine learning models that leverage *causal knowledge*. Causal knowledge, either learned or given, is expressed using structures (such as graphical models), functions (explicit relationships between variables), and time, allowing methods to complete certain tasks, such as prediction, forecasting, generation, optimisation, and others.

Note that our definition generalises to more than just *deep* models. In fact, as we shall see in section 2 (and beyond), being able to differentiate between all sorts of parametric forms is key to our parametric dimension.

We can illustrate our framework conceptually as a *topographic map*, which we did in fig. 1 on the right. This map is further refined in section 3, which allows us to use our framework for categorising and comparing various methods in CDL. Essentially, our map functions as an alternative to the ladder of causation [1, 10]. As we show in section 4, our refined visualisation can characterise methods, allowing careful comparison which further aids adoption. Furthermore, our visualisation also guides us to *combine* various CDL methods to solve real-world problems as we shall see in section 3.4.

Application	Challenge addressed using causality	Ref.
Imputation	Define interaction effects of missingness and covariates	[12–17]
Fairness	Identify of sources of unfairness	[18–22]
Generalisation	Constrain methods using a general multi-domain causal structure	[23–25]
Treatment effects	Learn models that respect identification assumptions	[26–39]
Drug repurposing	Relate multi-modal data using causal structure	[40–44]
Recommendation	Infer causal effect of recommendation	[45–49]

Table 1: **Real-world impact.** We list exemplary real-world applications of CDL.

Real-world impact. In table 1 we list a subset of CDL methods that have already impacted the real world. However, without our framework, CDL faces a significant hurdle in terms of further adoption and implementation, which could hamper its potential impact in the real world. A clear definition and an associated framework to think about the different requirements and needs that model real-world problems, help us understand and incorporate the latest advances correctly. Furthermore, the lack of clarity also

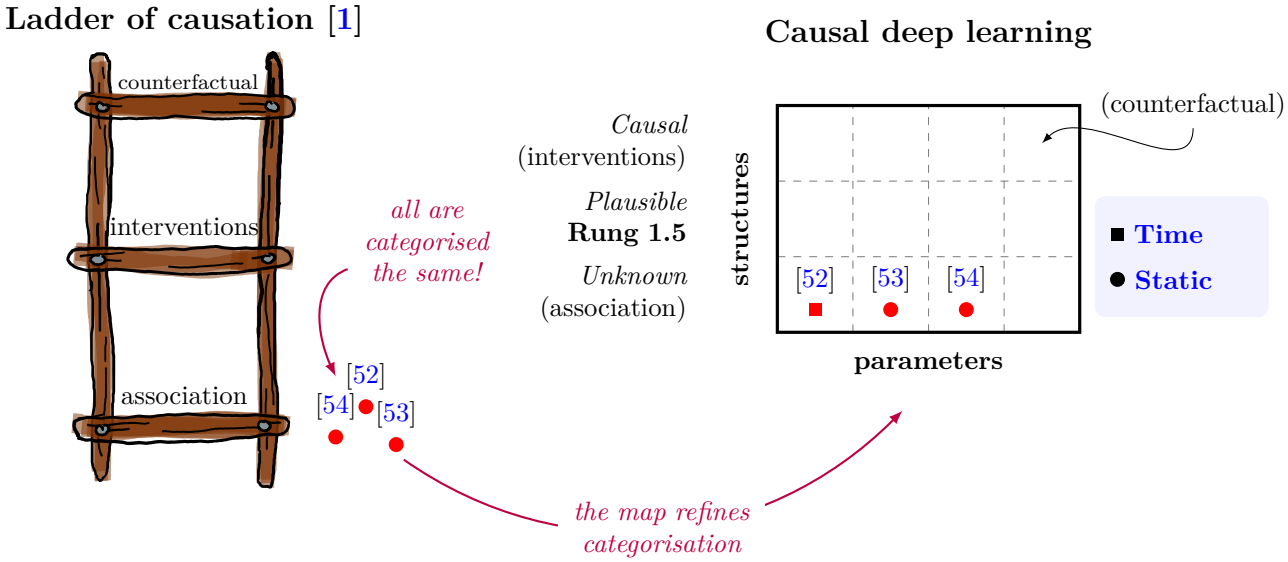


Figure 1: **Causal deep learning.** The higher one climbs on the ladder of causation (left), the more causal knowledge one needs. However, the ladder does not differentiate between (at the extreme) non-parametric models or highly parametric models. For this, we introduce the axis denoted with "**parameters**" and is discussed in section 3.2. Furthermore, we find there are more structures to be taken into account beyond causal structures, in particular the "rung 1.5" structures, discussed in section 3.1. The structures are categorised using the axis denoted as "**structures**". Finally, we also explicitly differentiate between temporal and static methods, discussed in section 3.3. Details on all the axes are provided in section 3.

makes it challenging to *create a common research agenda and careful characterisation*, which is essential for advancing the field. In section 5 we use our framework to discuss the application areas in table 1.

Defining a clear research agenda. To push this important field forward, we need a comprehensive and clear research agenda. Review papers (such as [50, 51]) are very useful for the community as they collect and frame an enormous amount of relevant papers, but they do not help us to define a clear research agenda as they never (a) provide a distinction between causality, deep learning, and causal deep learning as the selection criterion for inclusion into the field seems to be "machine learning with causal awareness"; and (b) differentiate the methods with proper (and shared) characterisation. Instead our framework enables such careful characterisation which:

(i) identifies which papers are already making progress towards more powerful methods that leverage (sometimes incomplete) causal knowledge

(ii) shows how to combine diverse causal deep learning methods to solve real-world problems, and

(iii) identifies promising areas ripe for further investigation.

These are important contributions which not only aid adoption but also drive the field forward.

2 Is the "ladder of causation" sufficient?

The lack of good ways to characterise the field has prevented us from properly comparing methods. Of course, we could characterise proposals by their deep learning architecture [8] (or used machine learning

model in general [55, 56])[1]. But that does not help us to determine in what respect a proposal leverages causal knowledge. For example, standard Gaussian processes [53] hold an equal amount of causal information as a very deep (but standard) transformer networks [52]: *none*. However, there exist Gaussian processes [26] and transformer networks [57] alike, which have been trained explicitly to leverage causal knowledge. In fact, they seem to exhibit *the same amount* of causal knowledge.

The above has not gone unnoticed by the causal community which has adopted a different strategy of characterisation: *the ladder of causation*. Originally introduced by Judea Pearl [1], the ladder is now considered central to the "Pearl causal hierarchy" (PCH) [10]. Depicted in fig. 1 (on the left), the ladder is composed of three rungs, from the bottom up a model is characterised by their ability to model: associative queries, interventional queries, and counterfactual queries. Using the ladder to characterise models would clearly differentiate standard Gaussian processes from those that answer interventional queries [26]. Yet, it doesn't properly define causal deep learning, as we explain below.

Structures do not tell the complete story. Pearl's ladder of causation categorises models on the basis of the type of query they can answer. From easy to hard, it helps us put into perspective the difficulties of performing *causal inference*. At its essence, the ladder tells us that increasing a model's potency requires increasing its causal knowledge.

The associative (bottom) rung requires essentially no additional information beyond what is given in observational data. To infer the effect of interventions we require structural causal knowledge. And to answer counterfactual queries we require structural causal knowledge *and* structural equations. While this helps bring across the point of how difficult causality is, it does not help us at all when trying to compare methods as the ladder obfuscates the difference between structures and functions (in particular in the counterfactual rung). Furthermore, it is a long road from no structural knowledge to causal structural knowledge (association to intervention).

To illustrate this point, consider the remainder of fig. 1. While Pearl's ladder is not able to differentiate between transformer networks and linear regression, we require a method that *can do this*. Naturally, we discuss in much more detail how we arrive at our map in section 3.3. However, fig. 1 shows (on the right) that our map *can* indeed differentiate between these wildly different examples.

Is knowledge learned or given? There is a big difference in methods that require the knowledge as given, and those that learn. This is not reflected in the ladder. Like parametric assumptions, the source of knowledge does not seem to matter much. However, depending on the source, a method can or cannot be used in practice!

In order to provide some clarity in causal deep learning, we have to be crystal clear about what a method requires and what it adds. We stress however, that there is definitely merit to methods that already require causal knowledge as we discuss in more detail in section 3.4.

Time. Despite its special position in causality, time is not considered explicitly in fig. 1. Perhaps the absence of a temporal indicator is the most glaring shortcoming as causality is in essence a temporal phenomenon. Furthermore, inferring causality over time requires very different assumptions and models and should not be compared directly with models that operate in a static setting [11, 58]. In order to compare accurately, we should not categorise temporal models in the same way we would categorise static models.

In the following section, we propose a method to map and later navigate causal deep learning. Our proposal should be considered a refinement of the now widely adopted ladder of causation as a way to think about and discuss methods in causal deep learning.

Central to our map is how each method handles assumptions which in turn helps us understand in what scenarios a method is actually applicable. For example, if a method assumes full causal knowledge, it cannot be used in settings where we lack such causal knowledge. However, with the help of our map, we can easily tie

[1] With deep learning architectures we mean network structures such as transformers, diffusion models, recurrent structures, etc.; With machine learning models, we mean well-known model classes such as Gaussian processes, support vector machines, random forests, etc.

multiple methods together which possibly *relax* some of a method's original assumptions. These relaxations are possible through *transitions*, either structural or parametric (discussed in detail in section 3.4).

3 Introducing the research agenda in causal deep learning

We propose a map to organise past and future research in CDL. Our map helps us to understand and categorise the many differences and similarities between models in CDL, but also to understand how to situate CDL (and these models) in the wider field of machine learning. Given the increasing amount of new ideas and proposals in CDL, we believe this to be a necessary exercise.

Causal deep learning refines machine learning by leveraging a different set of assumptions and inductive biases. As these are vital to CDL, we build our map around their classification. Specifically, a these assumptions and biases are defined by two ingredients: (i) the structure linking individual variables, and (ii) the functions that model these links.

As such, the map of causal deep learning operates under two main axes: (i) *the structural scale* concerning the links between variables, described in section 3.1; and (ii) *the parametric scale* concerning the shape of the functions that model links between variables, described in section 3.2.

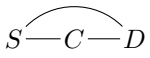
Our two scales form the basis of the map of causal deep learning, which we introduce in detail in section 3.3.

3.1 The structural scale and rung 1.5

Crucial to causal deep learning is the ability to leverage knowledge about the relationships between variables. Depicted as graphs, we can place these relations on a scale from less detailed to more detailed – this gives the structural scale, which is central to our map of causal deep learning.

fig. 2 depicts the structural scale, which ranges from **less** structure, to **lots of** structure. The structural scale governs information about the statistical or causal relationships in a system. Essentially, it quantifies how much we know, expressed in graphical structures. Below, we discuss each level using an illustrative example of three variables: smoking, denoted as S; cancer, denoted as C; and death, denoted as D.

Level 1 – no structure. At the lowest end of our scale we have structures that exhibit zero knowledge. These structures connect each variable with each other. An example of such a structure is denoted as,

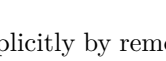


$$\textbf{(unknown).} \tag{1}$$

While technically a structure, we have no idea about the potential relationships these variables have with respect to each other. Having a fully connected structure such as in eq. (1) means that we assume *potential* dependence between each variable. Assuming a fully connected structure is not a strict assumption. In our example above, the structure in eq. (1) basically states that "smoking, cancer and death" are possibly related, which is a very general statement indeed.

That is because relatedness is a flexible term. While having an edge may appear to be a strict statement, the edge does not express the *level* at which two variables are related. In fact, the *absence* of an edge is a much stronger assumption which we discuss next.

Level 2 – plausibly causal structures. Let us assume now that we know a little more: cancer is the reason why smoking is related to death. Statistically, we can express this assumption as $S \perp\!\!\!\perp D|C$. In Level 2, we assume *statistical independence*. Independence is not directly the same as assuming some structural information, however, it can be expressed as such.

The easiest thing we can do to reflect $S \perp\!\!\!\perp D|C$ is to explicitly by remove the edge  from eq. (1) as we have below in eq. (2),

$$S - C - D\,. \tag{2}$$

Crucially, however, there are more structures that reflect $S \perp\!\!\!\perp D|C$. Consider the three structures below,

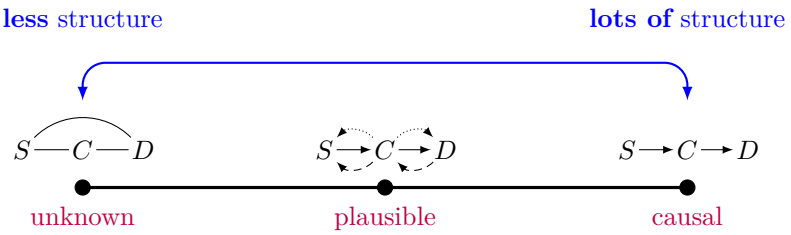


Figure 2: **The structural scale.** A first axis on the map of causal deep learning is the structural scale which categorises different structure types. The two extremes are non-informative structures, or a complete causal graph. Note that the structural scale makes no parametric assumptions on these structures (parametric assumptions are the focus of the parametric scale in fig. 3).

$$S \rightarrow C \rightarrow D, \qquad S \leftarrow C \rightarrow D, \qquad S \leftarrow C \leftarrow D,$$

which all model the same independencies.

For the sake of our discussion, let us assume $S \perp\!\!\!\perp D|C$ is correct[2]. As (in)dependence is a statement of statistical association, we cannot assume that the structures based on it are causal in nature. For example, $S \leftarrow C \leftarrow D$, implies that dying impacts whether or not the deceased is a smoker, which cannot be true.

The type of structure of interest in Level 2, are the probabilistic graphical models (PGMs) [60]. These include structures such as directed acyclic graphs (DAGs), undirected graphs, or even factor graphs. Depending on the set of independence statements one wishes to assume, one may express them as a certain type of PGM.

For example, if one wishes to assume $S \perp\!\!\!\perp D$ while also assuming $S \not\perp\!\!\!\perp D|C$, we best use a DAG:

$$S \rightarrow C \leftarrow D, \tag{3}$$

which is often called a "collider" structure. This collider structure (and more importantly the (in)dependence statements it implies) can only be modeled with a DAG.

Being able to model independence is a clear step up from Level 1, where we assumed nothing at all. While independence is a statement of statistical association (i.e. it is bidirectional and not causal), they do pose constraints for a later causal model. As we can check (in)dependence in data, a recovered causal model should respect (or even explain) these (in)dependence statements. Hence, the above structures are plausibly causal.

In level 2 we have no consensus on one graph in particular, only about dependence and independence (predominately statistical concepts). For the purpose of illustrating Level 2 on our structural scale in fig. 2 above, we denote these plausible structures as,

$$S \rightleftarrows C \rightleftarrows D \qquad \textbf{(plausible).} \tag{4}$$

where we combined all possible DAGs that respect, $S \perp\!\!\!\perp D|C$ [61]. Applying the rules of d-separation [62].

While Level 2 includes all types of PGMs, we chose to illustrate Level 2 using DAGs only, as DAGs are the structure of choice when modelling causality [2]. In eq. (4) all of the structures are DAGs, but only one of them is causal (likely to be the first one).

Level 3 – causal structure. That leaves assuming the exact causal structure as opposed to the many plausible causal structures in Level 2. With eq. (5) we say that smoking *causes* cancer which causes death. Such knowledge translates into the directed structure,

$$S \rightarrow C \rightarrow D \qquad \textbf{(causal).} \tag{5}$$

[2]If cancer renders smoking and death independent, we imply that cancer is the *only* path from smoking to death, which is known to be false [59]. For example, smoking may result in chronic obstructive pulmonary disease (COPD) or ischemic heart disease, which can also lead to premature death. However, for the purpose of illustration, we assume it is true.

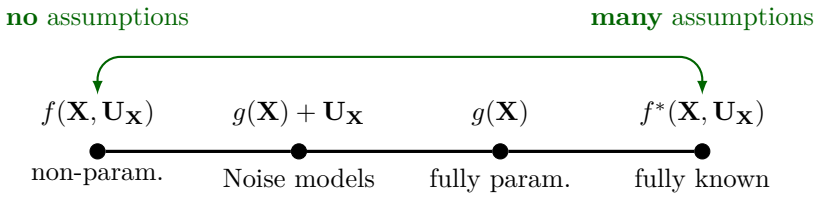


Figure 3: **The parametric scale.** A second axis in the map of causal deep learning. The parametric scale logs the type of assumptions made on the factors of the assumed distribution or model. The extremes are no assumptions at all (leaving completely non-parametric factors), or a fully known model. The parametric scale further discerns assumptions on the way noise interacts in the system (ϵ) and functional shape of the system's factorisation.

Which was indeed one of the plausible DAGs in Level 2 (eq. (4)).

Comparing eq. (2) with eq. (5) may lead to believe there is little difference between either end of the structural scale– *nothing could be further from the truth [1, 10]!* Rather, the step from association to causation is very large and is a central argument in causality. So large in fact that we recognise several intermediary structures.

An important realisation is that each level of information further specifies the environment, i.e. with information, we *restrict* possible interpretations and data-distributions. For example, from eqs. (1) and (2) we expect S to change if C is changed as both variables are associated with each other. However, eq. (5) tells us this is not necessarily the case: patients do not automatically start smoking when diagnosed with cancer. However, the reverse *is* true; when a cancer-patient *stops* smoking they have a higher chance of remission [63].

"Rung 1.5" Pearl's ladder of causation [1, 10] ranks structures in a similar way as we do, i.e., increasing a model's causal knowledge will yield a higher place upon his ladder. Like Pearl, we have three different levels in our scale. However, they do not correspond one-to-one.

In particular, we find that our Level 2 is not well represented in Pearl's ladder of causation[3]. Our reasoning is that our Level 2 *does* encode some prior knowledge into a model, which is more than encoding no prior knowledge at all. Yet, both would be categorised under Pearls first rung. Of course, we recognise that the models categorised under Level 2 have no embedded *causal* knowledge. While Level 2 does not reach the level of a rung 2 model (but Level 3 does), we consider a Level 2 model to correspond with a hypothetical "rung 1.5".

At the structural level, we find that Pearl's rung 3 (counterfactuals) is the same as rung 2 (interventional). Of course, rung 3 goes *beyond* structure, in that it also requires functional knowledge. Instead, we dedicate an entire new scale to functional knowledge as (like structure) we find there are different levels of functional knowledge one can assume (or learn). We introduce this additional scale in the following subsection.

3.2 The parametric scale

In section 3.1 we defined structures governing the (statistical or causal) relationships between variables. Now we turn to the parameterisation of these relationships. Intuitively, the structures depict which variables express an *affinity* with each other [60, Chapter 4]. In particular, each structure dictates how a distribution can be factorised as a series of functions. These functions are called *factors*, which are defined as,

Definition 1 (factor)**.** *Let* $\mathbf{X}$ *be a set of random variables. We define a factor* f *to be a function from* $Val(\mathbf{X}) \to \mathbb{R}$*. A factor is nonnegative if all its entries are nonnegative. The set of variables* $\mathbf{X}$ *is called the scope of the factor and denoted* $Scope[f]$ *[60, Definition 4.1].*

[3]Similarly, his counterfactual rung is not represented in our structural scale.

As such, the parametric scale— which we introduce in this section —models not *if* variables interact, but *how* they interact. For example, a statement such as a distribution $p \models (S \perp\!\!\!\perp D|C)$ holds iff we factorise the distribution as $p(S, C, D) = f_1(S, C)f_2(C, D)$ (lacking a factor $f_3(S, D)$). As such, a structure becomes parameterised by associating a set of factors with it.

Note that associating factors do not necessarily correspond to associating functions for each edge, at least in the undirected setting (such as eq. (2)). Rather, a factor can be generalised by allowing them over arbitrary subsets of variables. As such, in its most general form, we define a factorisation as,

$$p(\mathbf{X}) = \frac{1}{Z} \prod_{i \in [I]} f_i(\mathbf{X}^{(i)}), \tag{6}$$

where Z is a normalising constant, and one factor f_i takes a set of variables, $\mathbf{X}^{(i)} \subseteq \mathbf{X}$, as arguments [60].

A trivial example of such a factorisation is the decomposition of a joint distribution into conditional distributions: $p(A, B) = f_1(A, B)f_2(B)$ with $f_1(A, B) = p(A|B)$ and $f_2 = p(B)$. With eq. (6) one can generalise decomposition to accommodate parameterisation of symmetrical relationships as well as directional relationships (which are generally modelled as conditional probability densities).

Notice that in eq. (6) we iterate over I factors. However, the amount of factors is, at this stage, not yet fully determined. For example, one could have a factorisation which is comprised of the least amount of factors: $p(\mathbf{X}) = f(\mathbf{X})$, or one could have one where we hope to include a factor with only X^1 as an argument: $p(\mathbf{X}) = p(X^1)p(X^2|X^1)...p(X^n|X^{n-1}, ..., X^1)$.

Defining a factor, f_i as in eq. (6), is non-parametric (we assume nothing about the factors) and agnostic to the structure it models (because there is no implied correspondence between edges and factors). What *does* change is the arguments that need to go into a factor. Those *are* determined by the accompanying structure[4].

In our example above, having $A \perp\!\!\!\perp B$ would imply that $p(A|B) = p(A)$, which actively alters the variables that need to go into f_1. Recall, that Level 1 of the structural scale is not able to model this explicitly– we need at least a Level 2 assumption for this.

Level 1 – non-parametric. As with the structural scale, Level 1 of the parametric scale encompasses the least strict assumption: namely, there exists a factorisation such as eq. (6). Beyond that, level 1 assumes no specific functional form of its factorisations, f_i.

Of course, as we move up the structural scale, the structures themselves may dictate the purpose of the factorisation. For example, in a directed structure, the set of arguments changes from an arbitrary set $\mathbf{X}^{(i)}$ to the set containing a variable, its parents and noise: $f_X(X, \text{pa}(X), U_X)$. Nevertheless, despite the factor's arguments changing the function remains non-parametric.

Level 2 – noise models. The next level in our parametric scale is to first make an assumption on the composition of the set of random variables, $\mathbf{X}$. In particular, we assume that $\mathbf{X}$ can be separated in noise terms or exogenous variables $\mathbf{U}$ and variables $\mathbf{X} \setminus \mathbf{U}$ where each element in $\mathbf{U}$ corresponds with exactly one element in $\mathbf{X} \setminus \mathbf{U}$ and $|\mathbf{X}| = 2|\mathbf{U}|$.

The above assumption has implications for the factors. Specifically, if a factor takes $X \in \mathbf{X} \setminus \mathbf{U}$ as argument, it will always take $U_X \in \mathbf{U}$ (which is the corresponding noise variable of X) as argument also.

Note that the above is *only* an assumption on the factor arguments and is per our earlier discussion a *graphical* assumption. As such, currently, Level 2 is no different from level 1 in the parametric scale. However, in Level 2, we make an additional assumption which *is* a parametric assumption on the factors. For a factor $f_i(\mathbf{X}^{(i)}, \mathbf{U}_{\mathbf{X}^{(i)}})$ we make an assumption on how the noise variables $\mathbf{U}_{\mathbf{X}^{(i)}}$ are incorporated in f_i [68–70], for example, additive noise:

$$f(\mathbf{X}, \mathbf{U}_{\mathbf{X}}) = g(\mathbf{X}) + \mathbf{U}_{\mathbf{X}}. \tag{7}$$

[4]While not the topic of discussion here, it is this realisation that is often used as an argument in favour of using causal knowledge to increase efficiency by reducing the input required for correct inference (e.g. in [18, 64–67]).

In short, eq. (7) assumes we can decompose a factor f into the noise, $\mathbf{U_X}$, and some (non-parametric) function, $g(\mathbf{X})$. While the way in which $\mathbf{U_X}$ interacts in f is fixed, g is still non-parametric and therefore only level 2 of this scale. As discussed below, we find many papers in causal deep learning making assumptions of this level.

When performing causal discovery, making an assumption on the noise of a system may aid in causal identification [71]. Conceptually, making this type of assumption is another way to exert some outside expert knowledge into the system. In Hoyer et al. [70] g remains non-linear (through the use of neural networks) but $\mathbf{U_X}$ is assumed additively independent (as above) resulting in a clear level 2 assumption.

Level 3 – fully parametric. Naturally, the next level is to make a parametric assumption on g in eq. (7). These assumptions vary from making linear assumptions, to n times differentiable, to many more types of assumptions on the functional form of g and consequentially f. Note that because we make the assumption on g (and not f directly), we automatically assume that f can be decomposed in g and noise, i.e. Level 3 subsumes assumptions from Level 2.

Less parametric assumptions imply a larger model hypothesis class. For example, having linear assumptions allows much less possible factors than a non-linear class of assumptions. In fact, it has been shown that neural networks[5] have a connection to non-parametric regression, leading to high performance models [72].

As such, making a parametric assumption is typically done in favour of easier regression, but at the cost of less general model classes.

Level 4 – fully known factors. Rather than making assumptions on the functional class of the factors, some methods assume full access to a known set of factors which we term $f^*(\mathbf{X}, \mathbf{U_X})$. This is of course a very strict assumption, resulting in the final level of the parametric scale.

Having defined our two axes, we can now compose the complete *map of causal deep learning* in section 3.3 which governs both model input as well as model output.

3.3 The map of causal deep learning

Consider some data, $\mathcal{D} \coloneqq \{\mathbf{X}_i : \mathbf{X}_i \in \mathcal{X}, i \in [N]\}$, with $\mathcal{X}$ some space such as $\mathbb{R}^d$. Generally, the goal of deep learning is to represent these data using a model,

$$m : \mathcal{D} \to \mathcal{H}, \tag{8}$$

with m the model, and $\mathcal{H}$ a suitably defined space of learned representations which depends on the method to learn m. At a later stage, $\mathcal{H}$ is used to perform a task (such as prediction, or data-generation, or decision making). To perform this task accurately, one has to carefully consider: (i) the type of data ($\mathcal{D}$), and (ii) the type of representation ($\mathcal{H}$) upon which we base the model's output.

The same holds for models in causal deep learning. The difference, however, lays in the types of assumptions and restrictions we pose on (i) the data, and (ii) the representation. Naturally, despite being classified under the umbrella term causal deep learning, there are many differences in how these restrictions manifest. As such, the map of causal deep learning maps (i) the data, and (ii) the learnt representation.

Note that in many cases the representation is the same as the model output. An example of this are causal discovery methods where the aim is to structure the input data as a causal graph. This structure is of course a "representation". Specifically, a causal discovery method will transform a dataset into a graph: $m : \mathcal{D} \to \mathcal{G}$. In the latter case, we have that $\mathcal{G} = \mathcal{H}$, and $\mathcal{H}$ and the output are the same, corresponding with eq. (8).

Using our scales in figs. 2 and 3, we provide some categorisation of both objects, leading to *the map of causal deep learning*, illustrated in fig. 4. The map of causal deep learning has three components: Input, Time/static, and Representation. Input and Representation are built from the scales we introduced above. The Time/Static component is a simple binary structure which helps us discern methods that operate on temporal data or static data, respectively.

[5] In particular with a ReLU activation.

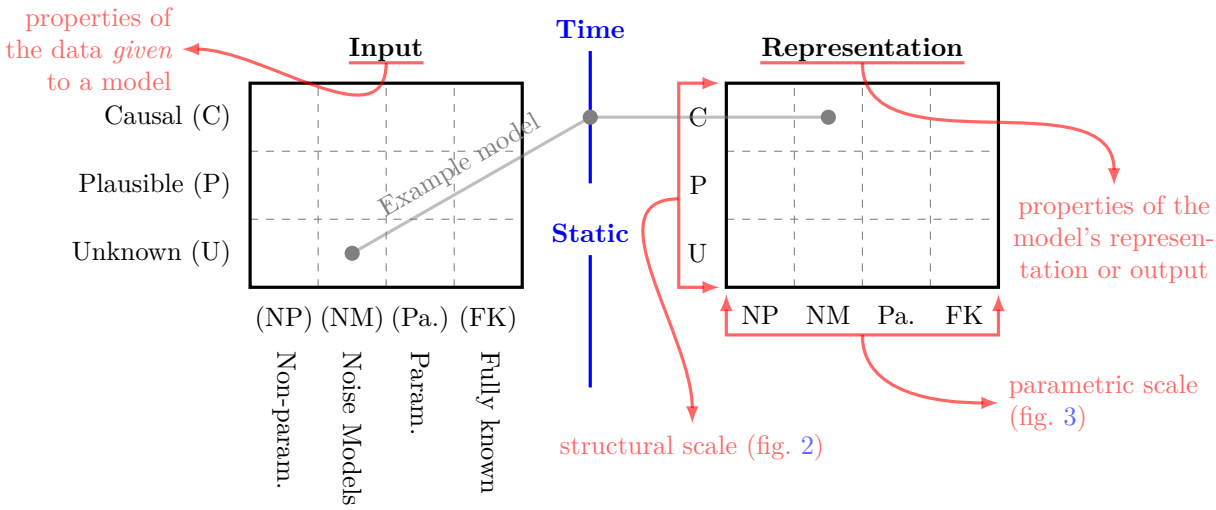


Figure 4: **The map of causal deep learning.** The map evaluates each model based on the required input and learnt representations. For each, we categorise using both the structural scale (fig. 2) and the parametric scale (fig. 3). Furthermore, we also discern between models that handle temporal or static data. As an example, we show a fictitious method which assumes no structure but *does* make an additive noise assumption; operates in the temporal domain; and provides a truly causal representation under the same additive noise assumption. Our map allows easy examination of a method on many properties at once.

Input. The input field (leftmost grid in fig. 4) categorises the data in terms of assumptions on the data-generating process (DGP). Here, the structural scale measures how much we already know about the DGP in terms of *which* variables are dependent; whereas the parametric scale measures *how* these variables are dependent.

Time/Static. We explicitly differentiate between temporal and static models. While evaluated similarly, time may lead to alternative considerations in the causal setting. For example, a typical assumption in the temporal domain is that causes precede effects [11]. While we do not express any opinion on whether or not this makes things harder or easier, we do believe it is important to differentiate in order to properly compare methods. Furthermore, when using our map to search for an appropriate method from a practical point of view, it makes sense to search across static methods for a static problem, and temporal methods for a temporal problem.

Representation. A model's representation is evaluated using the same scales as the **Input** of the model. This is a deliberate choice as we wish to document how much information a model adds to existing data.

In terms of representation, the structural scale measures what type of structure a model yields; whereas the parametric scale measures how the model is parameterised. Note that the latter is different from assumptions on the DGP as it concerns the maximum capacity of a particular model.

3.4 Transitions and cascading models

In section 3.3 we have alluded to methods that learn (or discovery) a causal graph ($\mathcal{G}$) from data ($\mathcal{D}$). In our map, a causal learning method shows a difference between input and representation on the structural level: the input would have no (or plausible) causal structure, while the output *would have* causal structure. Whenever such a difference is documented, we say that a method makes a *transition*.

Definition 2 (Transition)**.** *When a method has a different level input than representation, either structural (section 3.1) or parametric (section 3.2), we say that a model invokes a transition from the input to the representation.*

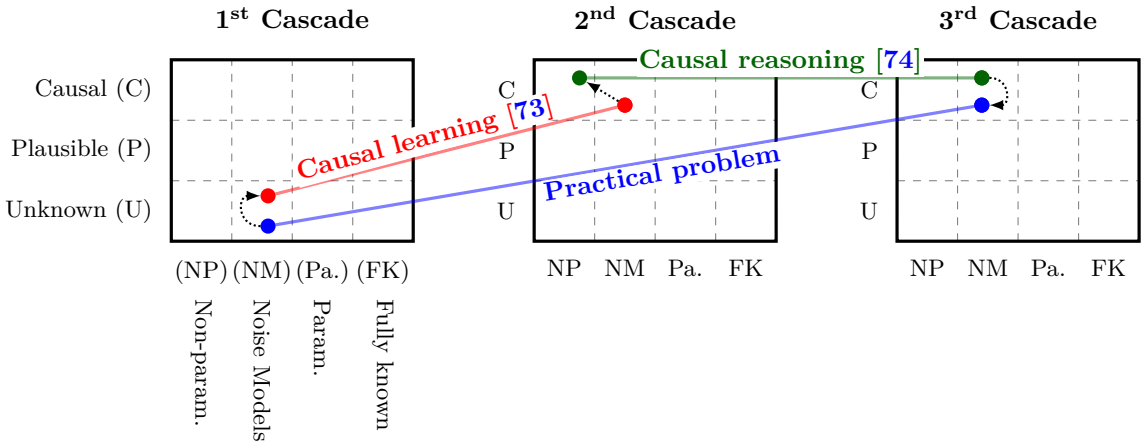


Figure 5: **Cascading CDL models.** As the input and representation in our map (fig. 4) is defined on the same domain– the structural scale × the parametric scale –we can use the representation of a model as input for a subsequent model. Documenting these cascading models is easy with our map as we can simply cascade multiple maps after one another as we have done above. Doing so shows us how we can use mostly theoretical work (such as causal reasoning) in practice.

A transition in definition 2 is a very general concept which allows us to appropriately discuss topics such as: causal learning, model cascading, or relaxing assumptions– using our map. To illustrate, please consider two algorithms:

Causal learning. First, a causal learning algorithm (called RESIT) where a causal structure is discovered from observational data, by Peters et al. [73].

Causal reasoning. and Second, a bandit algorithm which exploits a known causal graph to arrive at an optimal regret more efficiently than using no causal graph, by Lattimore et al. [74].

As discussed, causal learning is a prime example of a structural transition, where the input structure (typically unknown) is lower level as the output (a causal graph). Causal reasoning, however, constitutes *no* transition as causal reasoning algorithms typically assume access to a known causal structure. Examples of this include bandit algorithms such as Lattimore et al. [74], which we discuss further using our map of CDL in section 4.

While algorithms such as those proposed in Lattimore et al. [74] are useful, it is hard to apply them in practice as a causal graph is seldom known. Naturally, one can gather some expert knowledge and craft a causal structure manually, but that is typically cumbersome or doesn't yield a complete graph (which requires different reasoning algorithms altogether).

Consider fig. 5, where we have illustrated a practical scenario in blue: a practitioner wants to perform a causal reasoning task (such as applying Lattimore et al. [74]), but has no access to a ground truth causal DAG. From the map, it is clear that the practitioner needs to incorporate a transition model as Lattimore et al. [74] (illustrated in green) cannot function in the practitioner's scenario by itself. As such, we first learn a causal graph, using for example Peters et al. [73], and hand our discovery over to the reasoning algorithm. Cascading Peters et al. [73] before Lattimore et al. [74] matches the problem setup of our practitioner.

Naturally, our illustration is quite straightforward and one could likely arrive at the same solution without our map. However, often real-world scenarios can be much more complicated and require more than two cascading models to solve a particular problem [75, 76]. Imagine problems such as data imputation, transformation, or life-long learning. All of these intermediary steps have to take into account the assumptions made on each model present in the pipeline. Building on a set of non-cascading assumptions can have detrimental consequences as there may be unwanted bias creeping into your solution.

For example, in the our above scenario, the observational data used to learn a casual graph may have missing values. Depending on which imputation strategy we use, we may introduce additional parametric assumptions into our pipeline. Furthermore, depending on which mechanism governs the missingness patterns, we may have to make additional structural assumptions before imputation can even begin [14, 16, 77]. Making such a structural assumptions, *before* learning a graph using Peters et al. [73] would make our solution for the practitioner's problem invalid once again.

3.4.1 Causal learning on the map

There are many algorithms which promise a structural transition, ending up in causal structure. Ironically, *all of them are categorised the same in the ladder of causation (fig. 1)!* By introducing a parametric scale, the map allows to better differentiate between methods. For example, comparing Zheng et al. [78] with the later Zheng et al. [79] using the latter of causation, implies that Zheng et al. [79] wrongly claimed an improvement on Zheng et al. [78] as they both transition from having no structural knowledge to plausible structural knowledge. Of course, only when we introduce the parametric scale it becomes clear that the difference does not lie in structure, but in their parametric assumptions!

3.4.2 Non-matching assumptions

In our example in fig. 5, we have not addressed the fact that the used structure learner actually yields a graph on a different parametric level than what is assumed by the reasoning task. Namely, RESIT leans a causal structure assuming noise model structural equations, while Lattimore et al. [74] actually assumes a non-parametric causal graph.

Luckily, because of the composition of both the structural and parametric scale, each stricter assumption is subsumed by a lower level assumption. As such, from left to right, we can always relax the assumptions (such as allowing non-parametric reasoning based on a noise-model causal structure), but not the other way around. If, for example, our structure learner yielded a plausible but non-parametric causal structure, we are not guaranteed optimal regret from Lattimore et al. [80]. On our map, that would become clear as we would move from a less strict assumption (plausible causality) to a strict assumption (full causality).

4 Using our topographic map

Using the map of causal deep learning we can: (i) categorise and compare literature, as well as (ii) identify some areas that are not well explored. Using these two functionalities, we can perhaps attribute (i) to practitioners, and (ii) to researchers. A practitioner is faced with a problem they wish to solve. Using the map of CDL, the practitioner can map their available data onto the input field and search for a method in the representation field that would solve their needs. Perhaps guided by the required assumptions the practitioner is willing to make, the set of solutions is more digestible than scanning all the literature in CDL. The researcher may use the map differently. Rather than scanning the potential solutions, a researcher can use the map to scan which solutions are still lacking (or more likely, underrepresented).

Problems and goals. Learning a structured representation space, while retaining flexible assumptions, is difficult. Consider fig. 6 where we annotated learning a causally structured representation without any assumptions from arbitrary data with "*Goal*". This learning setup is the most ambitious setup included in our map. It is also long thought to be impossible [81–84].

However, fig. 6 also shows the relative complexity of other setups. In principle, the less strict assumptions we make, the harder it becomes to narrow down the structure in our representation space. For some problem setups it may be sufficient to learn a flexible structure, which in turn allows for more flexible assumptions on the input. Similarly, if a problem requires a completely identified causal structure, our map shows that one may have make to some strict assumptions on the input. As such, the map of CDL exposes a certain balance between input assumptions and achievable structure.

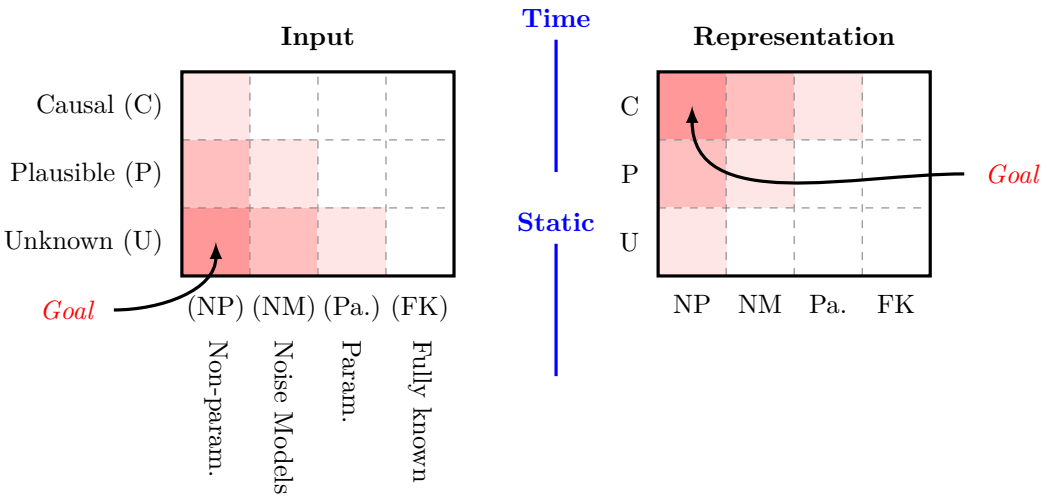


Figure 6: **Goals and direction of the field.** We can use the map of causal deep learning to specify the future of our field. Darker shades of red indicate a harder, but more desirable goal. In the case of input, we wish to minimise the required graphical input and made assumptions; with which we hope to achieve a more knowledgeable representation with minimal assumptions.

4.1 Comparing methods

To show how one can use the map to compare methodologies, we take *supervised learning* as an example, presented in fig. 7. Of course, CDL spans more than only supervised learning and so does our map.

To remind ourselves, supervised learning is a problem where we wish to map an input, to a label [55]:

$$m : \mathcal{D} \to \mathcal{Y},$$

where $\mathcal{D}$ is once again a dataset, but now including labels: $\{(X_i, Y_i) : i \in [N]\}$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. The labels $Y_i \in \mathcal{Y}$ can be anything from a real variable ($\mathcal{Y} = \mathbb{R}$), to a binary label ($\mathcal{Y} = \{0, 1\}$).

As we discussed in section 3.3, deep learning methods first map the data to a representation before it is mapped to the outcome-label [8]: $m : \mathcal{D} \to \mathcal{H} \to \mathcal{Y}$. Here, $\mathcal{H}$ corresponds to the representation as in eq. (8). As such, our discussion here does not concern $\mathcal{Y}$, but is instead focused on $\mathcal{H}$ and the structure it may respect.

Why use CDL for supervised learning? A fair question indeed. Typically, supervised learning models are evaluated only on the accuracy a model achieves on a hold-out test set. Yet, there is no immediate reason why a causal representation would help in this regard. Rather than maximising accuracy on a hold-out test set, one may resort to CDL to also achieve high accuracy on a subset that is not representative of the training data. Examples such as these include: domain adaptation [24, 85], transfer learning [86, 87], interpretability [88–91], or general robustness [92–95].

4.2 Developing methods

Let us discuss our map using a non-exhaustive list of some of the key tasks the map of CDL may help a researcher.

Identification. A big part in a researcher's workflow is to identify gaps in the literature. For researchers active in supervised learning, our map of CDL can be a useful tool to do exactly this. For example, from fig. 7 we learn that the models that take a fully known function as a parametric assumption are not well represented– they are not represented at all! Of course, this makes sense. If one assumes a fully known model, before commencing the learning task, there is no point learning it anymore.

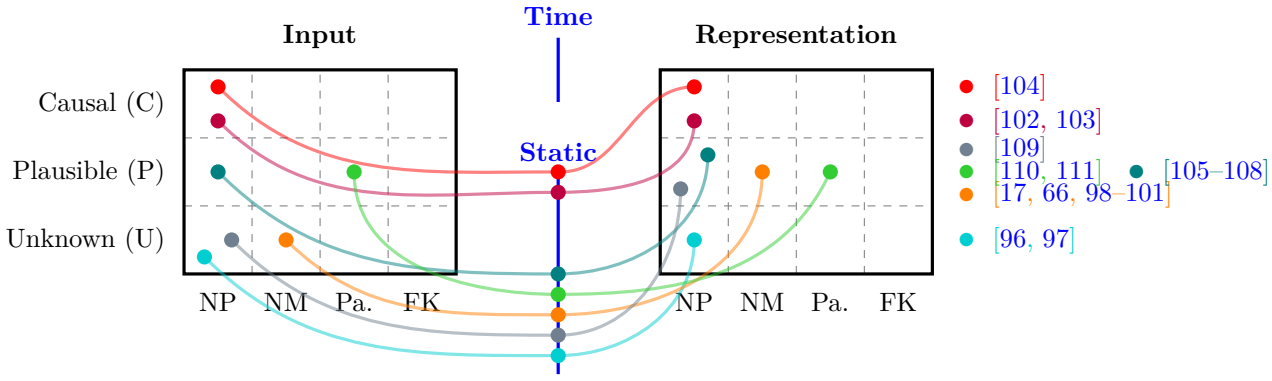


Figure 7: **Navigating (static) supervised learning.** Here we focus only on methods that share the task of mapping data to a label in a supervised way. We discern between methods based on what input they expect (or assume) and what type of representation they learn *before* mapping to a label. Our map presents a straightforward way to categorise methods in causal deep learning. Doing this is useful for practitioners (to identify a suitable method), and to researchers (to identify a potential research gap). Here too: Non-parametric (NP), Noise models (NM), Parametric (Pa.), and Fully known (FK).

Table 2: **Comparing different fields.** Section 4.4 compares methods in fields beyond supervised learning (cfr. section 4.1). Our table below provides a quick overview of the methods and fields we discuss. Beyond that, we also list the target aimed to learn, as well as the minimum data required to learn that target. We have excluded any additional information (beyond the data) such as parametric or structural assumptions, as this is exactly what the map of CDL can be used for.

Field	Data	Target	References	Map
Supervised learning (SL)	$\{(X_i, Y_i) : i \in [N]\}$	$\mathcal{X} \to \mathcal{Y}$	[17, 66, 96–111]	fig. 7
Forecasting/temporal SL	$\{(X_i, Y_i)_{1:T_i} : i \in [N]\}$	$\mathcal{T} \times \mathcal{X} \to \mathcal{Y}$	[52, 112–127]	fig. 8
Bandits & RL	$\{(X_i, A_i, Y_i)_{\pi^b} : i \in [N]\}$	$\mathcal{X} \times \mathcal{A} \times \mathcal{Y} \to \pi^*$	[74, 128–161]	fig. 9
Generative modelling	$\{(X_i) : i \in [N]\}$	$\mathcal{X} \to \mathcal{X}$	[18, 162–172]	fig. 10

Looking further, we find that mapping additive noise models with unknown structure, to additive noise model *with* structure actually *are* well represented. The reason for this is that recent contributions in differentiable structure learning are a great candidate to regularise models used for other tasks than structure learning. As such, a researcher would learn that there is indeed much competition in this field.

Navigation and related works. This brings us to a second task we can employ our map for: building a body of related works. The map will make it easier for researchers to find and learn about related works. Essentially enabling a platform for researchers to share their work with other interested researches.

Furthermore, the map allows researchers to think critically about the work they present. For example, in fig. 6 we clearly state that the "ultimate goal" to map data to a non-parametric causal structure without making any assumptions is thought to be impossible. Not only will this help researchers to discuss their proposal with a sense of realism, the same can be said about practitioners (and reviewers) alike.

Data. Beyond the above, the map also provides guidance for model evaluation. In particular, when evaluating a model, it is important to use data that actually matches the models input assumptions. Similarly for the benchmarks the proposal is compared against.

For example, if a method assumes a causal graph as input, there should be some assurance that the presented graph is indeed causal. One way to do this is by also provided some empirical evidence that this is the case, such as running clinical trials or assuming interventional data.

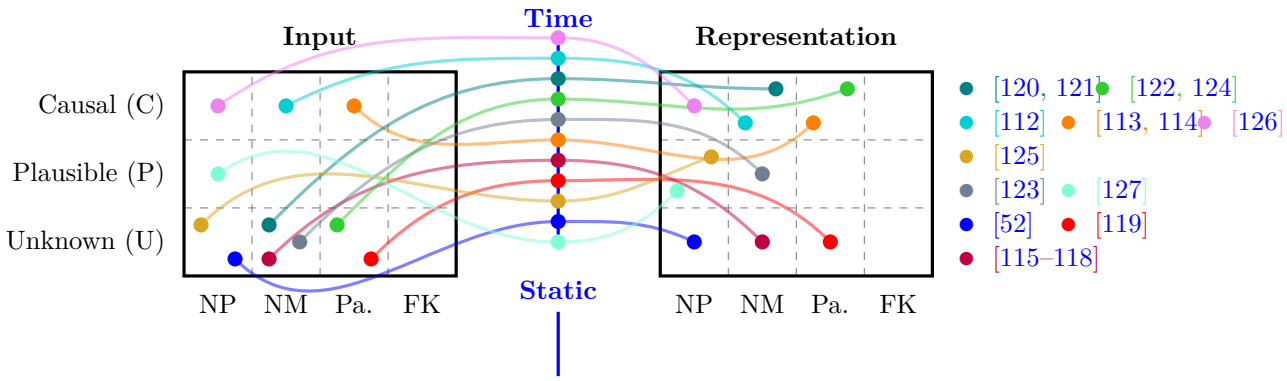


Figure 8: **Navigating (temporal) supervised learning and forecasting.** Here we focus on methods that solve a prediction task in the temporal domain. As in fig. 7 we discern between methods based on what input they expect (or assume) and what type of representation they learn *before* mapping to a label. Here too: Non-parametric (NP), Noise models (NM), Parametric (Pa.), and Fully known (FK).

4.3 Using methods

Navigate. While the researcher's perspective is mostly based around identifying methods that *don't* exist, the practitioner's is the opposite. When faced with a certain problem, the practitioner may use the map to identify which methods assume the input the practitioner has available, while also solving the problem they wish to solve.

Finding a suitable method can be extremely challenging for practitioners as there are new methods proposed almost daily. This latter point is in fact one of the prime reasons why we propose the map in the first place.

Validation and education. On the other hand, it may be that the practitioner wishes to solve a problem that is not yet solved before (stumbling upon a gap in the literature). Or, more strikingly, they are faced with a problem that is simply impossible to solve.

Especially the latter could be an interesting use of our map: education. We strongly believe that the map of CDL can form a bridge between researches and practitioners. Educating which problems *can* be solved (which is communicated from research to practice), and which problems *have to be* solved (which is in turn communicated from practice to research).

As we have observed in fig. 7, we find that there is a heavy focus on learning structure from nothing in the additive noise setting. But, perhaps this is not interesting from a practical point of view as it could be that practitioners are well aware of some conditional independence, or some parametric shape of their environment. Having clear communication between both parties may avoid researchers to waste valuable resources and time to solve problems that need not be solved.

4.4 Beyond supervised learning

Having explained how our map can categorise methods within the supervised learning problem (and what benefits this yields in sections 4.1 to 4.3), let us now consider other machine learning problems. We have listed these in table 2 where we– like we have for supervised learning in section 4.1 –present: the minimal data required to solve the problem, the problem's target, and provide a link to the map specific to the problem as well as the references they categorise. Naturally, more references and problems can be added, but we believe table 2 spans a broad range of the most researched ML problems. Thereby showing the versatility of the map of CDL.

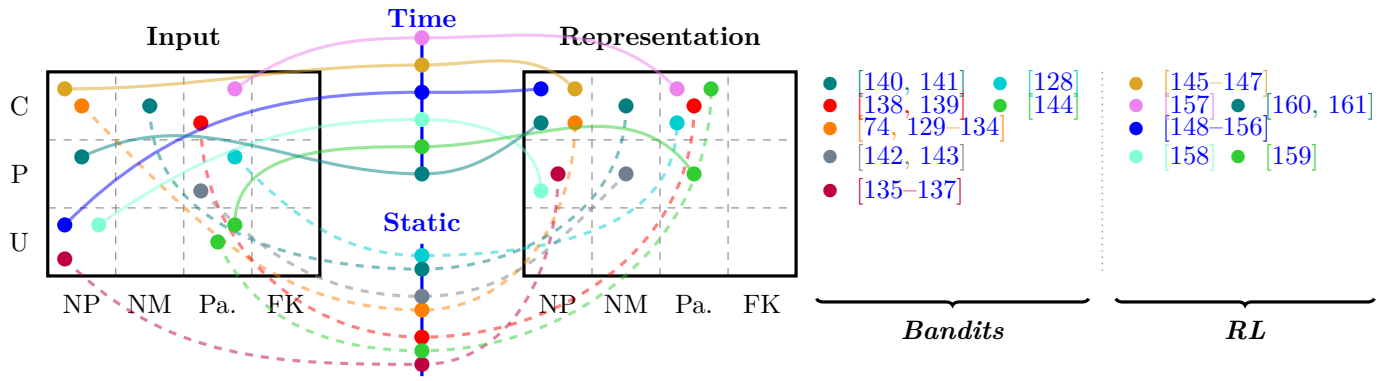


Figure 9: **Navigating bandit algorithms and reinforcement learning (RL).** Here we focus on methods that have interaction in the system. In particular, we consider bandit algorithms (where a sequence has no meaning), and RL (where sequence *has* meaning). As in figs. 7 and 8 we discern between methods based on what input they expect (or assume) and what type of representation they learn *before* mapping to a label. Here too: Non-parametric (NP), Noise models (NM), Parametric (Pa.), and Fully known (FK); and Unknown (U), Plausible (P), and Causal (C).

4.4.1 Forecasting and temporal supervised learning

Perhaps a logical *next step* beyond supervised learning is supervised learning in a dynamic setting. In particular, we will now categorise methods that forecast a prediction multiple timesteps ahead, both in discrete as in continuous time. As with static supervised learning, we assume each method has access to at least a dataset comprising features and labels. The difference is, though, that the features (and potentially the label) may be observed more than once. For example, for one patient, we may observe their blood-pressure at multiple timepoints (with a possibly changing blood-pressure). An overview of methods in this category can be found in fig. 8.

Strikingly, when we compare the static setting (fig. 7), with the temporal setting (fig. 8), we see that the static setting seems to be much more focused on non-parametric estimation than the temporal setting. Recall that, this difference is situated entirely on the parametric scale, which was not considered prior to the introduction of our map of CDL! If we were to only consider the structural scale (as one would when categorising according the ladder of causation in fig. 1), it would appear that the static setting and the temporal setting were quite similar in focus, i.e. papers would be similarly distributed over the unknown, plausible, and causal categories in fig. 2.

4.4.2 Bandit algorithms and reinforcement learning

In fig. 9 we have classified both bandit algorithms and reinforcement learning (RL) algorithms. The reason we categorise them together, is that they only differ in dependence across time (or sequential decisions). As such, they are easily differentiated using our Time/Static axis between input and representation. However, to clarify things further, we use dashed lines to represent bandit algorithms and full lines to represent RL algorithms.

Bandits. Let us first discuss the bandit side. From fig. 9 we can make a few observations. A first is that bandit algorithms seem to be very explicit when it comes to making structural assumptions. We believe a reason for this is that causal bandits tend to come from the PCH community. In contrast, it is not very common to make explicit the assumptions made on a parametric level. That is, most bandit papers are written in a non-parametric setting, or a linear setting, with the exception of a few algorithms situated in Bayesian optimisation (using Gaussian processes). Essentially, these papers provide regret analysis completely agnostic of the type of estimator used infer reward. Furthermore, only a few papers make *structural transitions*, where the algorithm aims to learn causal information, rather than simply assuming access to it.

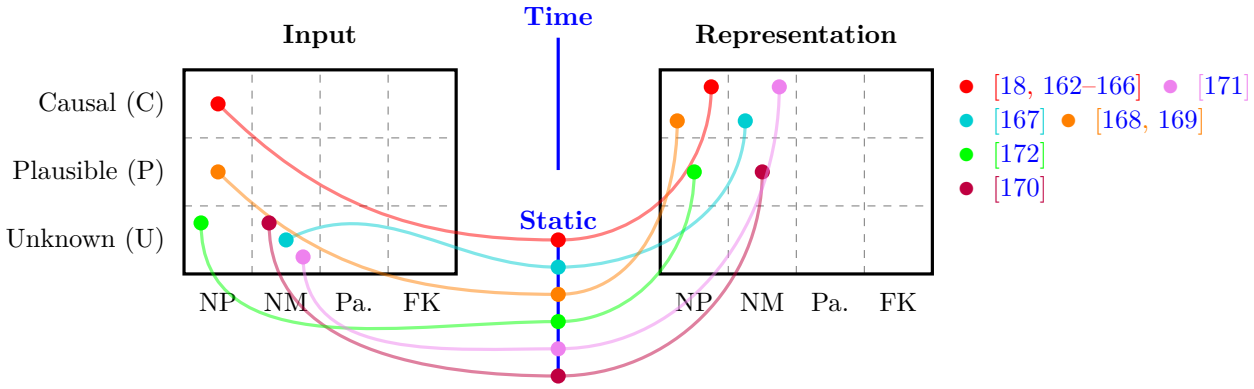


Figure 10: **Navigating generative modelling.** Here we consider methods that use causality to generate synthetic data. As in fig. 7 we discern between methods based on what input they expect (or assume) and what type of representation they learn *before* mapping to a label. Here too: Non-parametric (NP), Noise models (NM), Parametric (Pa.), and Fully known (FK).

RL. A second in fig. 9 are the reinforcement learning algorithms. Combining bandits and RL in one figure shows some interesting differences and equivalencies. First, it seems that– like in bandits –RL is also focused on non-parametric situations. The reason for this is a little different though. While research on bandits seem to provide algorithms which are agnostic to estimation, RL seems to stem more from a deep learning community; hence most papers in RL rely on neural networks to approximate the reward (based on, for example, Mnih et al. [173]).

Another important difference is the amount of RL papers that seem to succeed in a *structural transition* from unknown structure to causal structure, even in the non-parametric setting! Compare this with figs. 7 and 8 where *none* of the algorithms achief this. Why is this? Contrasting supervised learning, bandits and RL algorithms are allowed to perform *interventions* in the environment, whereas supervised learning algorithms can only rely on observational data. These experiments are of course hugely informative when inferring causality.

This brings us at an import point when considering fig. 9: we have only included papers that make an explicit point about causality in their investigation. However, recent work has shown that there may be an even deeper connection between causality and bandits and RL [174]. In fact, some results seem to indicate that some models may even exhibit causal knowledge without being trained for it explicitly [152]. We hope that these results encourage RL-researchers to consider causality in their papers.

4.4.3 Generative modelling

Let us now consider fig. 10 where we map generative models. Note that, like in fig. 9 we only consider research that explicitly discusses causality [175]. When comparing figs. 9 and 10, it should be apparent immediately that a *lot* more effort is spent on bandits and RL in CDL. We believe a reason for this is that bandits and RL may provide a more *natural* fit with causality given their ability to perform actual interventions in the environment. However, we wish to argue that generative modelling can *also* link more tightly with causality. In particular, generative models share an important taks with causal models: both model a distribution.

Most papers in fig. 10 exploit this parallel. Consider the setting where we are provided a causal model. From fig. 10 it seems that most researchers focus on this setting in particular. Effectively, given a causal structure, most research aims to incorporate the causal knowledge provided by this structure into the representation space of deep generative models. Any generated data thus respects a given causal graph. Naturally, there are methods that aim to combine structure discovery with generative modelling (providing a *structural transition*). However, these methods are equally bound by the same limitations as other causal discovery set-

tings. Specifically because– unlike bandits and RL –generative modelling typically assumes one observational dataset, rather than access to a live environment in which they can perform interventions.

5 Real-world applications

We discuss the impact of CDL on five different real-world applications: Imputation, Fairness, Generalisation, Treatment effects, and Recommendation. We will examine how CDL can help define interaction effects of missingness and covariates, identify sources of unfairness, constrain methods using a general multi-domain causal structure, learn models that respect identification assumptions, and infer the causal effect of recommendation.

Imputation Imputation is a common problem in many datasets, and missing data can lead to biased solutions. CDL can help define interaction effects of missingness and covariates to improve the accuracy of imputation methods. For example, CDL-based approaches allow deep learning to accurately impute missing values, respecting the causal interaction between missingness indicators and treatment selection [14–16]. CDL-based methods outperform existing imputation techniques in terms of both imputation accuracy *and* unbiased estimation from data with missing values. Another approach is to leverage causal knowledge to perform the imputation task directly [13, 93].

Fairness. CDL can also be used to identify sources of unfairness in machine learning models. In particular, CDL can help identify which features in the data are responsible for observed disparities in outcomes. CDL-based approaches can help mitigate unfairness in machine learning models by learning a representation of the data that is fair with respect to a given set of protected attributes and their causal neighbors [18–20, 22]. The authors showed that their method outperforms existing fairness methods in terms of both fairness and accuracy. Other CDL techniques evaluate models post training using causal evaluation metrics [21].

Generalisation. CDL can help constrain methods using a general multi-domain causal structure, which can improve generalisation performance in new domains. Specifically, causal structures can be used to select models in a new domain as (partial) causal knowledge is assumed invariant [23–25], yielding more robust model selections with increased performance as a result. While it would be beneficial to gather *complete* causal knowledge, we can already advance in this application using partial knowledge.

Treatment effects. Perhaps one of the most natural application domains for CDL is estimating causal treatment effects [26–29]. With a long history in statistics, it seems that treatment effect literature is truly benefiting from taking a causal deep learning approach. With applications in donor-organ transplantation [31, 32], cancer [33–35], and COVID-19 [36] and its impact [37].

Drug repurposing. At the intersection of generalisation and treatment effects is drug repurposing, an incredibly exciting area for CDL to have impact [43, 44]. With applications in COVID-19 [40] and dementia [41, 42], CDL is already proving to be a useful framework. However, as pointed out in table 3, these high-stakes application domains are a clear example of why we require a careful analysis of the made assumptions.

Recommendation. CDL can also be used to infer the causal effect of online store recommendations [45, 47, 48], which is essential for accurate evaluation of recommendation algorithms. Furthermore, Basilico and Raimond [46] illustrate the importance of causality *and* time to provide meaningful recommendations, again highlighting the importance of our temporal dimension! Beyond this, Ghazimatin et al. [49] show how causality can be used to *evaluate* a recommendation system by estimating counterfactual explanations.

We summarise the above in table 3, where we also indicate the range of structural, parametric and temporal assumptions, as per our topographic map. Clearly, from table 3 as well as the examples presented in section 4, it should be clear that we require our framework to truly have detailed discussions and clear adoption of methods going forward.

6 Conclusion

In conclusion, our framework for causal deep learning (CDL) represents a major step forward in the field of machine learning research. By incorporating a structural dimension to measure absent, present, or partially

Table 3: **Real-world impact.** Some methods in CDL have already impacted the real world. We list some of them in the table above, spanning a few important application areas. In our table, we use the following abbreviations: Unknown (U), Plausible (P), Causal (C), Non-parametric (NP), Noise models (NM), Parametric (Pa.), and Static (St.). These abbreviations correspond with our framework presented in section 3.

Application	**How they use causality**	**Ref.**	**Range of assumptions**		
			Struct.	Param.	Time
Imputation	Define missingness and covariates interaction	[12–17]	U → C	NP → Pa.	St.
Fairness	Identify of sources of unfairness	[18–22]	U → C	NP → Pa.	St.
Generalisation	Respect multi-domain causal structure	[23–25]	U → C	NP → NM	St.
Treatment effects	Respect causal identification assumptions	[26–39]	U → P	NP → Pa.	Both
Drug repurposing	Relate multi-modalities using causal structure	[40–44]	U → C	NP → Pa.	Both.
Recommendation	Infer causal effect of recommendation	[45–49]	U → P	NP → NM	Both

known structural causal knowledge in methods, a parametric dimension that maps the strengths of deep learning methods to causality, and a temporal dimension that accounts for the impact of time on a causal system, we have provided a comprehensive guide for using CDL to solve real-world problems.

Our framework extends the levels of structural knowledge in causality by introducing a new level, "plausible causal knowledge", that accommodates many existing papers and provides a clear pathway to powerful tools and methods. With our clear definition, we can now identify real problems and select the method that best fits the setting of interest, empowering us to better understand and address the shortcomings of current methods.

Furthermore, our framework identifies gaps in current research and proposes avenues for future work, including the need for research on recovering fully known structural equations and reflection on the relevance of current gaps to practical problems. We are confident that our framework will make a significant contribution to the advancement of CDL research and its practical applications, paving the way for a more efficient and effective approach to solving complex real-world problems.

Acknowledgements

We truly value the input of the CDL community. We wish to explicitly thank the following community members for their comments (alphabetised on last name):

Alicia Curth (University of Cambridge)
Isabelle Guyon (Google Brain)
Sam Holt (University of Cambridge)
Alihan Hüyük (University of Cambridge)
Trent Kyono (Meta)
Andrew Rashbass (University of Cambridge)
Paulius Rauba (University of Cambridge)
Guilermo Sapiro (Duke University & Apple)
Ricardo Silva (UCL)
Ruibo Tu (KTH)
Boris van Breugel (University of Cambridge)
David Watson (King's College London & Meta)
Simon Woodhead (Eedi)
Cheng Zhang (Microsoft Research)
Kun Zhang (CMU & MBZUAI)

References

[1] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[2] Judea Pearl. *Causality*. Cambridge university press, 2009.

[3] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[4] Donald B. Rubin. Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75(371):591, 1980. doi: 10.2307/2287653. URL https://doi.org/10.2307/2287653.

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

[6] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

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

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[9] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 857–869. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

[10] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. 2022.

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

[12] Pablo Morales-Alvarez, Wenbo Gong, Angus Lamb, Simon Woodhead, Simon Peyton Jones, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Simultaneous missing value imputation and structure learning with groups. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4rm6tzBjChe.

[13] Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions. In *Conference on Causal Learning and Reasoning*, pages 688–711. PMLR, 2022.

[14] Jeroen Berrevoets, Fergus Imrie, Trent Kyono, James Jordon, and Mihaela van der Schaar. To impute or not to impute?–missing data in treatment effect estimation. *arXiv preprint arXiv:2202.02096*, 2022.

[15] Daniel Westreich, Jessie K Edwards, Stephen R Cole, Robert W Platt, Sunni L Mumford, and Enrique F Schisterman. Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology*, 44(5):1731–1737, 2015.

[16] Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. 2020.

[17] Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34: 23806–23817, 2021.

[18] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[19] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[20] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in neural information processing systems*, 32, 2019.

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[22] Rik Helwegen, Christos Louizos, and Patrick Forré. Improving fair predictions using variational inference in causal models. *arXiv preprint arXiv:2008.10880*, 2020.

[23] Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pages 1546–1555. PMLR, 2021.

[24] Trent Kyono and Mihaela Van der Schaar. Exploiting causal structure for robust model selection in unsupervised domain adaptation. *IEEE Transactions on Artificial Intelligence*, 2(6):494–507, 2021.

[25] Trent Kyono, Ioana Bica, Zhaozhi Qian, and Mihaela van der Schaar. Selecting treatment effects models for domain adaptation using causal knowledge. *arXiv preprint arXiv:2102.06271*, 2021.

[26] Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

[27] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[28] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

[29] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

[30] Yoav Kan-Tor, Naama Srebnik, Matan Gavish, Uri Shalit, and Amnon Buxboim. Evaluating the heterogeneous effect of extended incubation to blastocyst transfer on the implantation outcome via causal inference. *bioRxiv*, pages 2021–11, 2021.

[31] Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.

[32] Jeroen Berrevoets, Ahmed Alaa, Zhaozhi Qian, James Jordon, Alexander ES Gimson, and Mihaela Van Der Schaar. Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In *International Conference on Machine Learning*, pages 792–802. PMLR, 2021.

[33] Sami Tabib and Denis Larocque. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics*, 36(2):629–636, 2020.

[34] Anirban Basu. Estimating person-centered treatment (pet) effects using instrumental variables: an application to evaluating prostate cancer treatments. *Journal of Applied Econometrics*, 29(4):671–691, 2014.

[35] Wouter AC van Amsterdam, Joost JC Verhoeff, Netanja I Harlianto, Gijs A Bartholomeus, Aahlad Manas Puli, Pim A de Jong, Tim Leiner, Anne SR van Lindert, Marinus JC Eijkemans, and Rajesh Ranganath. Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage iii non-small cell lung cancer. *Scientific reports*, 12(1):5848, 2022.

[36] William R Zame, Ioana Bica, Cong Shen, Alicia Curth, Hyun-Suk Lee, Stuart Bailey, James Weatherall, David Wright, Frank Bretz, and Mihaela van der Schaar. Machine learning for clinical trials in the era of covid-19. *Statistics in Biopharmaceutical Research*, 12(4):506–517, 2020.

[37] Florian Wozny. The impact of covid-19 on airfares—a machine learning counterfactual analysis. *Econometrics*, 10(1):8, 2022.

[38] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

[39] Achille Nazaret and Guillermo Sapiro. A large-scale observational study of the causal effects of a behavioral health nudge. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022. URL https://openreview.net/forum?id=4NocV3GQRG9.

[40] Anastasiya Belyaeva, Louis Cammarata, Adityanarayanan Radhakrishnan, Chandler Squires, Karren Dai Yang, GV Shivashankar, and Caroline Uhler. Causal network models of sars-cov-2 expression and aging to identify candidates for drug repurposing. *Nature communications*, 12(1):1024, 2021.

[41] Marie-Laure Charpignon, Bella Vakulenko-Lagun, Bang Zheng, Colin Magdamo, Bowen Su, Kyle Evans, Steve Rodriguez, Artem Sokolov, Sarah Boswell, Yi-Han Sheu, et al. Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia. *Nature communications*, 13(1):7652, 2022.

[42] Marie-Laure Charpignon, Bella Vakulenko-Lagun, Bang Zheng, Colin Magdamo, Bowen Su, Kyle Evans, Steve Rodriguez, Artem Sokolov, Sarah Boswell, Yi-Han Sheu, et al. Drug repurposing of metformin for alzheimer's disease: Combining causal inference in medical records data and systems pharmacology for biomarker identification. *medRxiv*, pages 2021–08, 2021.

[43] Fan Yang, Qi Zhang, Xiaokang Ji, Yanchun Zhang, Wentao Li, Shaoliang Peng, and Fuzhong Xue. Machine learning applications in drug repurposing. *Interdisciplinary Sciences: Computational Life Sciences*, 14(1):15–21, 2022.

[44] Ruoqi Liu, Lai Wei, and Ping Zhang. When deep learning meets causal inference: a computational framework for drug repurposing from real-world data. *arXiv preprint arXiv:2007.10152*, 2020.

[45] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 426–431, 2020.

[46] Justin Basilico and Yves Raimond. Déja vu: The importance of time and causality in recommender systems. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 342–342, 2017.

[47] Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 453–470, 2015.

[48] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, Austria*, pages 23–29, 2022.

[49] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. Prince: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 196–204, 2020.

[50] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

[51] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[54] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[55] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[56] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[57] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15329. PMLR, 2022.

[58] Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. 2022.

[59] Brian D. Carter, Christian C. Abnet, Diane Feskanich, Neal D. Freedman, Patricia Hartge, Cora E. Lewis, Judith K. Ockene, Ross L. Prentice, Frank E. Speizer, Michael J. Thun, and Eric J. Jacobs. Smoking and mortality — beyond established causes. *New England Journal of Medicine*, 372(7):631–640, 2015. doi: 10.1056/NEJMsa1407211. URL https://doi.org/10.1056/NEJMsa1407211. PMID: 25671255.

[60] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[61] Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible markov kernels. In *Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006)*, pages 1–11, 2006.

[62] Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, 1990.

[63] Mahdi Sheikh, Anush Mukeriya, Oxana Shangina, Paul Brennan, and David Zaridze. Postdiagnosis smoking cessation and reduced risk for lung cancer progression and mortality: a prospective cohort study. *Annals of internal medicine*, 174(9):1232–1239, 2021.

[64] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems*, 31, 2018.

[65] Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4164–4172, 2019.

[66] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

[67] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[68] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

[69] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

[70] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[71] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[72] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

[73] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

[74] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

[75] Jeroen Berrevoets, Ahmed Alaa, Zhaozhi Qian, James Jordon, Alexander E. S. Gimson, and Mihaela van der Schaar. Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In Marina Meila and Tong Zhang, editors, *Proceedings of*

the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 792–802. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/berrevoets21a.html.

[76] Andreas Kunft, Asterios Katsifodimos, Sebastian Schelter, Sebastian Breß, Tilmann Rabl, and Volker Markl. An intermediate representation for optimizing machine learning pipelines. *Proceedings of the VLDB Endowment*, 12(11):1553–1567, 2019.

[77] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

[78] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[80] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/b4288d9c0ec0a1841b3b3728321e7088-Paper.pdf.

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

[82] Dan Geiger and Judea Pearl. On the logic of causal models. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 3–14. Elsevier, 1990.

[83] Christopher Meek. Strong completeness and faithfulness in bayesian networks. *arXiv preprint arXiv:1302.4973*, 2013.

[84] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.

[85] Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, and Weidong Geng. Learning causal representation for training cross-domain pose estimator via generative interventions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11280, 2021.

[86] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[87] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Causal transfer learning. *arXiv preprint arXiv:1707.06422*, 2017.

[88] Dong Wang, Yuewei Yang, Chenyang Tao, Zhe Gan, Liqun Chen, Fanjie Kong, Ricardo Henao, and Lawrence Carin. Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models. *arXiv preprint arXiv:2012.03369*, 2020.

[89] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[90] Carolyn Kim and Osbert Bastani. Learning interpretable models with causal guarantees. *arXiv preprint arXiv:1901.08576*, 2019.

[91] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. Causality learning: a new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789*, 2020.

[92] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 378–379, 2020.

[93] Trent Kyono and Mihaela van der Schaar. Improving model robustness using causal knowledge. *arXiv preprint arXiv:1911.12441*, 2019.

[94] Jens Müller, Robert Schmier, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. Learning robust models using the principle of independent causal mechanisms. In *DAGM German Conference on Pattern Recognition*, pages 79–110. Springer, 2021.

[95] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[96] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[97] András Faragó and Gábor Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

[98] Liuyi Yao, Yaliang Li, Bolin Ding, Jingren Zhou, Jinduo Liu, Mengdi Huai, and Jing Gao. Path-specific causal fair prediction via auxiliary graph structure learning. 2021.

[99] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[100] Bahare Fatemi, Layla El Asri, and Seyed Mehran Kazemi. Slaps: Self-supervision improves structure learning for graph neural networks. *Advances in Neural Information Processing Systems*, 34:22667–22681, 2021.

[101] Pablo Morales-Alvarez, Angus Lamb, Simon Woodhead, Simon Peyton Jones, Miltiadis Allamanis, and Cheng Zhang. Vicause: Simultaneous missing value imputation and causal discovery with groups. *arXiv preprint arXiv:2110.08223*, 2021.

[102] Fabrizio Russo and Francesca Toni. Causal discovery and injection for feed-forward neural networks. *arXiv preprint arXiv:2205.09787*, 2022.

[103] Takeshi Teshima and Masashi Sugiyama. Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 86–96. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/teshima21a.html.

[104] Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, and Amit Sharma. Matching learned causal effects of neural networks with domain priors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10676–10696. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kancheti22a.html.

[105] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[106] Liangxiao Jiang, Harry Zhang, and Zhihua Cai. A novel bayes model: Hidden naive bayes. *IEEE Transactions on knowledge and data engineering*, 21(10):1361–1371, 2008.

[107] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, and Harry Zhang. Improving tree augmented naive bayes for class probability estimation. *Knowledge-Based Systems*, 26:239–245, 2012.

[108] Liangxiao Jiang, Chaoqun Li, Shasha Wang, and Lungan Zhang. Deep feature weighting for naive bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, 52: 26–39, 2016.

[109] Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19314–19326. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e05c7ba4e087beea9410929698dc41a6-Paper.pdf.

[110] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[111] Pat Langley, Wayne Iba, Kevin Thompson, et al. An analysis of bayesian classifiers. In *Aaai*, volume 90, pages 223–228, 1992.

[112] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[113] John R Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, pages 327–358, 1983.

[114] Guglielmo Maria Caporale and Nikitas Pittis. Causality and forecasting in incomplete systems. *Journal of Forecasting*, 16(6):425–437, 1997. doi: https://doi.org/10.1002/(SICI)1099-131X(199711)16:6<425::AID-FOR657>3.0.CO;2-9. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-131X%28199711%2916%3A6%3C425%3A%3AAID-FOR657%3E3.0.CO%3B2-9.

[115] Ben Calderhead, Mark Girolami, and Neil Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *Advances in neural information processing systems*, 21, 2008.

[116] Jim O Ramsay, Giles Hooker, David Campbell, and Jiguo Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

[117] Benn Macdonald and Dirk Husmeier. Gradient matching methods for computational inference in mechanistic models for systems biology: A review and comparative analysis. *Frontiers in Bioengineering and Biotechnology*, 3, 2015. ISSN 2296-4185. doi: 10.3389/fbioe.2015.00180. URL https://www.frontiersin.org/articles/10.3389/fbioe.2015.00180.

[118] Philippe Wenk, Alkis Gotovos, Stefan Bauer, Nico S. Gorbach, Andreas Krause, and Joachim M. Buhmann. Fast gaussian process based gradient matching for parameter identification in systems of nonlinear odes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1351–1360. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/wenk19a.html.

[119] Itai Dattner and Chris AJ Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9(2):1939–1973, 2015.

[120] Chris J. Oates, Frank Dondelinger, Nora Bayani, James Korkola, Joe W. Gray, and Sach Mukherjee. Causal network inference using biochemical kinetics. *Bioinformatics*, 30(17):i468–i474, 08 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu452. URL https://doi.org/10.1093/bioinformatics/btu452.

[121] Niklas Pfister, Stefan Bauer, and Jonas Peters. Identifying causal structure in large-scale kinetic systems. *arXiv*, pages 1810–11776, 2018.

[122] Catriona M. Queen and Casper J. Albers. Intervention and causality: Forecasting traffic flows using a dynamic bayesian network. *Journal of the American Statistical Association*, 104(486):669–681, 2009. doi: 10.1198/jasa.2009.0042. URL https://doi.org/10.1198/jasa.2009.0042.

[123] Frederik Vissing Mikkelsen and Niels Richard Hansen. Learning large scale ordinary differential equation systems. *arXiv preprint arXiv:1710.09308*, 2017.

[124] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[125] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17081–17092, June 2022.

[126] Kshitij Sharma, Yogesh K Dwivedi, and Bhimaraya Metri. Incorporating causality in energy consumption forecasting using deep neural networks. *Annals of Operations Research*, pages 1–36, 2022.

[127] Dashe Li, Jiajun Sun, Huanhai Yang, and Xueying Wang. An enhanced naive bayes model for dissolved oxygen forecasting in shellfish aquaculture. *IEEE Access*, 8:217917–217927, 2020. doi: 10.1109/ACCESS.2020.3042180.

[128] Jingwen Zhang, Yifang Chen, and Amandeep Singh. Causal bandits: Online decision-making in endogenous settings. *arXiv preprint arXiv:2211.08649*, 2022.

[129] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/c0a271bc0ecb776a094786474322cb82-Paper.pdf.

[130] Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4164–4172, Jul. 2019. doi: 10.1609/aaai.v33i01.33014164. URL https://ojs.aaai.org/index.php/AAAI/article/view/4320.

[131] Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8565–8576. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/61a10e6abb1149ad9d08f303267f9bc4-Paper.pdf.

[132] Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2017–2025. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/nair21a.html.

[133] Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[134] Chandrasekar Subramanian and Balaraman Ravindran. Causal contextual bandits with targeted interventions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F5Em8ASCosV.

[135] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24817–24828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d010396ca8abf6ead8cacc2c2f2f26c7-Paper.pdf.

[136] Blair Bilodeau, Linbo Wang, and Daniel M. Roy. Adaptively exploiting d-separators with causal bandits. 2022. doi: 10.48550/ARXIV.2202.05100. URL https://arxiv.org/abs/2202.05100.

[137] Jeroen Berrevoets, Sam Verboven, and Wouter Verbeke. Treatment effect optimisation in dynamic environments. *Journal of Causal Inference*, 10(1):106–122, 2022. doi: doi:10.1515/jci-2020-0009. URL https://doi.org/10.1515/jci-2020-0009.

[138] Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual Bandits with Latent Confounders: An NMF Approach. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 518–527. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/sen17a.html.

[139] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 141–150. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/lu20a.html.

[140] Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3155–3164. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/aglietti20a.html.

[141] Virginia Aglietti, Neil Dhir, Javier González, and Theodoros Damoulas. Dynamic causal bayesian optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10549–10560. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/577bcc914f9e55d5e4e4f82f9f00e7d4-Paper.pdf.

[142] Guy Tennenholtz, Uri Shalit, Shie Mannor, and Yonathan Efroni. Bandits with partially observable confounded data. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 430–439. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/tennenholtz21a.html.

[143] Nihal Sharma, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

[144] Arnoud De Kroon, Joris Mooij, and Danielle Belgrave. Causal bandits without prior knowledge using separating sets. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 407–427. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/kroon22a.html.

[145] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21164–21175. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/b0b79da57b95837f14be95aaa4d54cf8-Paper.pdf.

[146] Junzhe Zhang. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11012–11022. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20a.html.

[147] Andreas W.M. Sauter, Erman Acar, and Vincent Francois-Lavet. A meta-reinforcement learning algorithm for causal discovery. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. URL https://openreview.net/forum?id=2Y2HwBqORqz.

[148] Arquímides Méndez-Molina, Eduardo F.Morales, and L. Enrique Sucar. Causal discovery and reinforcement learning: A synergistic integration. In Antonio Salmerón and Rafael Rumi, editors, *Proceedings of The 11th International Conference on Probabilistic Graphical Models*, volume 186 of *Proceedings of Machine Learning Research*, pages 421–432. PMLR, 05–07 Oct 2022. URL https://proceedings.mlr.press/v186/mendez-molina22a.html.

[149] Clare Lyle, Amy Zhang, Minqi Jiang, Joelle Pineau, and Yarin Gal. Resolving causal confusion in reinforcement learning via robust exploration. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*, 2021.

[150] Tue Herlau and Rasmus Larsen. Reinforcement learning of causal variables using mediation analysis. In *36th AAAI Conference on Artificial Intelligence*, pages 6910–6917. Association for the Advancement of Artificial Intelligence, 2022.

[151] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.

[152] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

[153] Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22905–22918. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c1722a7941d61aad6e651a35b65a9c3e-Paper.pdf.

[154] Chao-Han Huck Yang, I Hung, Te Danny, Yi Ouyang, and Pin-Yu Chen. Causal inference q-network: Toward resilient reinforcement learning. *arXiv preprint arXiv:2102.09677*, 2021.

[155] X Wang, Y Du, S Zhu, L Ke, Z Chen, J Hao, and J Wang. Ordering-based causal discovery with reinforcement learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 3566–3573. IJCAI International Joint Conferences on Artificial Intelligence Organization, 2021.

[156] Tom He, Jasmina Gajcin, and Ivana Dusparic. Causal counterfactuals for improving the robustness of reinforcement learning. *arXiv preprint arXiv:2211.05551*, 2022.

[157] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient reinforcement learning with prior causal knowledge. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 526–541. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/lu22a.html.

[158] Wenxuan Zhu, Chao Yu, and Qiang Zhang. Causal deep reinforcement learning using observational data. *arXiv preprint arXiv:2211.15355*, 2022.

[159] Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, pages 1–13, 2022.

[160] Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*, 2021.

[161] Mirco Mutti, Riccardo De Santi, Emanuele Rossi, Juan Felipe Calderon, Michael M. Bronstein, and Marcello Restelli. Provably efficient causal model-based reinforcement learning for environment-agnostic generalization. In *A causal view on dynamical systems, NeurIPS 2022 workshop*, 2022. URL https://openreview.net/forum?id=VDunku3bDk.

[162] Bingyang Wen, Luis Oliveros Colon, KP Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Generating tabular data using causal generative adversarial networks. *arXiv preprint arXiv:2104.10680*, 2021.

[163] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

[164] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

[165] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. doi: 10.24963/ijcai.2019/201. URL https://par.nsf.gov/biblio/10126320.

[166] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23: 1–55, 2022.

[167] Martina Cinquini, Fosca Giannotti, and Riccardo Guidotti. Boosting synthetic data generation with effective nonlinear causal discovery. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pages 54–63, 2021. doi: 10.1109/CogMI52975.2021.00016.

[168] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.

[169] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3947–3956, June 2021.

[170] Wenbin Zhang, Jun Liao, Yi Zhang, and Li Liu. Cmgan: A generative adversarial network embedded with causal matrix. *Applied Intelligence*, pages 1–13, 2022.

[171] Wenqian Li, Yinchuan Li, Shengyu Zhu, Yunfeng Shao, Jianye Hao, and Yan Pang. Gflowcausal: Generative flow networks for causal discovery. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022. URL arXivpreprintarXiv:2210.08185.

[172] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE: Generative modelling for tabular data by learning relational structure. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fPVRcJqspu.

[173] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[174] James Bannon, Brad Windsor, Wenbo Song, and Tao Li. Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains, 2020. URL https://arxiv.org/abs/2006.02579.

[175] Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, Liming Zhu, and Kun Zhang. On the opportunity of causal deep generative models: A survey and future directions. *arXiv preprint arXiv:2301.12351*, 2023.