Relation-Oriented: Toward Causal Knowledge-Aligned AI

Anonymous authors
Paper under double-blind review

Abstract

This study examines the inherent limitations of the prevailing *Observation-Oriented* modeling paradigm by approaching relationship learning from a unique dimensionality perspective. This paradigm necessitates the identification of modeling objects prior to defining relations, confining models to observational space, and limiting their access to temporal features. Relying on a singular, absolute timeline often leads to an oversight of the multi-dimensional nature of the temporal feature space. This oversight compromises model robustness and generalizability, contributing significantly to the AI misalignment issue.

Drawing from the relation-centric essence of human cognition, this study presents a new *Relation-Oriented* paradigm, complemented by its methodological counterpart, the *relation-defined representation* learning, supported by extensive efficacy experiments.

1 Introduction

The prevailing modeling paradigm rules that observed variables (and outcomes) are the premise of building relationships. Model variables are often estimated by their observational values with an independent and identical distribution (i.i.d.) setting. Back in the 1890s, Picard-Lindelof theorem introduced a *logical timeline* t to record observational timestamps, establishing the paradigm $x_{t+1} = f(x_t)$ to depict variable X's time evolution. Since then, this ***Observation-Oriented*** principle has become our learning convention, where the temporal dimensionality is equated to the counts of $\{t, t+1\}$ unit, a predetermined constant time lag.

For a relationship $X \to Y$, the model can be in form $y_{t+m} = f(x_t)$, or $y_{t+m} = f(\{x_t\})$, where $\{x_t\} = \{x_1, \ldots, x_t, x_{t+1}, \ldots, x_T\}$ represents a time sequence of X within a certain length T, and a predetermined time progress m from X to Y. No matter in which form, the outcome Y is strictly observational only, leaving all potentially significant temporal changes of Y completely managed by $f(\cdot)$. However, although function $f(\cdot)$ can be selected as *linear* or *nonlinear*, the time evolution from t to $t+m$ is always left as *linear*.

Such a conventional linearity on the temporal dimension may be sufficient in the past, but not present, given the current technological advancements in data collection and Artificial Intelligence (AI) learning. Exploring nonlinear temporal distributions is gradually becoming essential. From a broader viewpoint, this is calling for a new modeling paradigm Scholkopf et al. (2021), which does not rest on the conventional i.i.d. assumed observations, but can treat t as a distinct computational dimension.

This study aims to fundamentally reveal the inherent deficiency of the current *Observation-Oriented* modeling paradigm (Chapter I: Sections 2-4), and accordingly propose the new ***Relation-Oriented*** one as desired, along with feasibility assessments (Chapter II: Sections 5-7). Particularly, the single absolute timeline t that we conventionally use, inherently cannot capture the multifaceted nature of temporal dimensionality, leading to widespread biases and resulting in AI models misaligned with our cognitive understanding, contributing significantly to the AI misalignment issue Christian (2020).

In this paper, we approach the concept of relationships in modeling through a novel *dimensionality framework*, offering a unique perspective. The remainder of this section aims to lay the groundwork. Then, in Chapter I, we will inspect causal learning from the view of temporal dimensionality, highlighting the key role of relations in modeling. Subsequently, Chapter II will concentrate on the proposed ***relation-defined representation*** learning method, which embodies the advocated *Relation-Oriented* modeling paradigm.

1.1 Manifestation of AI Misalignment

Today, AI has displayed capabilities surpassing humans in solely observational learning tasks, such as generating images, Go gaming, and so on. However, AI may appear "unintelligent" in comprehending certain relations that humans find intuitive. For instance, AI-created personas on social media can have realistic faces but barely with the presence of hands, due to AI struggling with the complex structure, instead treating hands as arbitrary assortments of finger-like items.

Moreover, when it comes to time evolution, causal reasoning presents a substantial challenge for AI, although it is innate for humans. Traditional causal learning methods, while having made valuable contributions to various fields of knowledge over the years Wood (2015); Vuković (2022); Ombadi et al. (2020), often suffer from a limitation in their generalizability Scholkopf et al. (2021). Unsuccessful neural network applications are particularly evident when addressing large-scale causal questions Luo et al. (2020). As a result, these methods are often confined to context-specific applications and encounter difficulties in extending to diverse scenarios. Thus, it is not strange that AI's capability on the temporal dimension remains notably constrained.

The questions "How to leverage AI's capability in causality" and "How to simulate hands with reasonable fingers" may seemly pertain to specific domains such as causal inference and computer vision. However, they fundamentally converge toward the broader challenge of AI Alignment, encapsulated by the essential question: "*Why are these relations unseen to AI?*" Reflecting on Dr. Geoffrey Hinton's warning, the misalignment of AI capabilities with human values can result in unintended and potentially harmful consequences. It is becoming increasingly critical to address this essential question.

1.2 Relations in Hyper-Dimension

Consider a pairwise relationship comprised of three elements: two *observable* objects, and a relation connecting them, which comes from our knowledge. The two objects can be solely observational (e.g., images, spatial coordinates of a quadrotor, etc.), or either observational-temporal (e.g., trends of stocks, persistent rain for five hours, etc.). Interestingly, the "relation" has to be ***unobservable*** to make this relationship meaningful for machine learning, distinguished from mere statistical dependencies.

This principle was initially introduced in the form of Common Cause Dawid (1979); Scholkopf et al. (2021), suggesting that any nontrivial conditional independence between two observables requires a third, mutual cause (i.e., our *unobservable* "relation"). Take the relationship "Bob has a son named Jim" as an example. The father-son relation is unobservable information that exists in our knowledge, which can also be seen as the common cause that makes their connection unique rather than any random pairing of "Bob" and "Jim". Given sufficient observed social activities, AI may deduce this pair of "Bob" and "Jim" have some special connection, but that does not equate to discerning their genuine father-son relation.

Put simply, the existence of unobservable element(s) makes the relationship model informative. In other words, the information contained by the model stems from our knowledge, rather than direct observations. Let's denote the model as $Y = f(X; \theta)$ with θ indicating the function parameter in demand. Then, in the context of modeling, the term "relation" can be represented by θ.

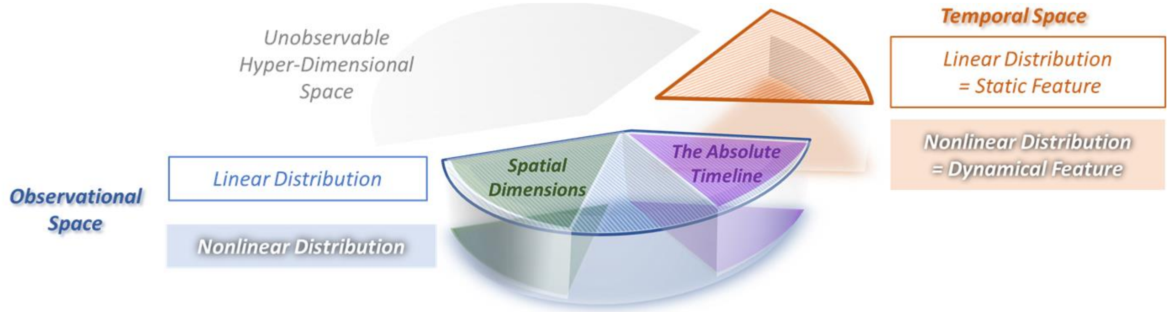

Unobservable Hyper-Dimensional Space
Temporal Space
Linear Distribution = Static Feature
Nonlinear Distribution = Dynamical Feature
Spatial Dimensions
The Absolute Timeline
Observational Space
Linear Distribution
Nonlinear Distribution

Figure 1: *Observational*, *Temporal*, and *Hyper-Dimensional* spaces, with the former two *Observable*.

From a dimensionality standpoint, a relationship can be viewed as a joint distribution across multiple dimensions: The observable objects feature the distribution on observational-temporal dimensions, while the unobservable relation manifests as some unseen distribution on a hyper-dimension. As illustrated in Figure 1, our cognitive space storing the knowledge relationships can be divided into three categories accordingly, where the ***Hyper-Dimensional*** space symbolizes the collective of all unobservable relations within our knowledge. Chapter I of this study aims to examine why AI cannot autonomously model certain relations in this space and understand the implications for its learning results.

1.3 Observational and Temporal Spaces

Under the *Observation-Oriented* principle, current models largely operate within the observational space. For example, CNNs (Convolutional Neural Networks) can learn observational associations among two-dimensional pixels; a quadrotor's movement can be estimated in three spatial dimensions; LLMs (Large Language Models) work in a semantic space along a logical timeline representing the order of words. Some applications (e.g., the last two examples) are aligned with the Picard-Lindelof theorem, using a single logical timeline to depict the absolute time evolution, thus often referred to as spatial-temporal analysis Alkon (1988); Turner (1990); Andrienko (2003). However, in a modeling context, an attribute of timestamps is not distinguishable from other observational attributes, unnecessarily to be temporally significant. Thus, we classify this single absolute timeline scenario as within the observational space.

According to our discussions at the beginning, the form of timestamps can only capture *linear* relationships on the temporal dimension, thus fundamentally impeding AI's ability to handle the temporal *nonlinearity*. This inherent disparity between our knowledge understanding and established models results in misalignment (see Section 3.3 for further discussions), accentuated by the rise of highly efficient AI applications.

Moreover, in our cognition (not the modeling context), ***multiple*** *logical timelines* may exist to form the cognitive temporal space in Figure 1 (see Section 4 for further insights). However, the current modeling paradigm has determined they cannot be distinguished as different dimensions in computation but crudely represented by a column of timestamps, i.e., consolidated as a single timeline.

In the observational-temporal joint space, as shown in Figure 1, observable distributions can be categorized as either *linear* or *nonlinear*. The temporal-significant ones can manifest as "static" or "dynamical" temporal features within a modeling context. For example, in the relationship "rain leads to wet floors", the events "rain" and "wet floors" are snapshots at specific timestamps and are thus viewed as ***static*** *temporal* objects. In contrast, events such as "persistent rain for five hours" and "floors becoming progressively wetter" are considered ***dynamical*** *temporal* due to their indispensable sequential patterns on the temporal dimension.

In this paper, we use the term "feature" to indicate the potential variable that fully represents the distribution of interest in any dimension. Additionally, the observational-temporal joint space may also be referred to as "observable data space", in contrast to the "latent feature space".

1.4 Hyper-Dimensional Space

Unobservable relations that fall outside the primary modeling objective can profoundly affect relationship models. This can be traced back to an undetected joint distribution within the hyper-dimensional space.

For example, when evaluating the impact of spicy foods on health, the direct link between spiciness and health is our primary modeling focus. However, there are underlying relations at play - such as how personal traits (individual-level features) are influenced by their cultural context (population-level features). Even if cultural differences are out of our modeling concern, overlooking these hierarchical distributions may introduce biases into our relationship model. For clarity, we term these hidden hyper-dimensional distributions as *unobservable* ***hierarchies***, sidestepping their relational aspects that fall outside the modeling objective.

These unobservable hierarchies often signify different granularity levels within the population. Achieving model ***generalizability*** across these levels is a common concern, dependent on the model's ability to reuse learned lower-level relationships for higher-level learnings Scholkopf et al. (2021). We argue that a shift from *Observation-Oriented* to *Relation-Oriented* is essential to realize this goal, in light of the *relation-centric* nature of human intelligence. In human understanding, relations function as ***indices*** that point to our mental representations Pitt (2022), crafting interconnected knowledge systems in memory, inclusive of their hierarchical structures. In line with this perspective, our proposed *relation-defined representation* learning is conceived as an attempt to "simulate" the process of human knowledge construction.

Chapter I: Deficiency of Current Observation-Oriented Paradigm

This chapter begins by examining the impact of unobservable hierarchies on models in Section 2, to highlight how these hierarchies can result in significant information loss on the temporal dimension, and its challenges for conventional causal inference. In Section 3, we offer a comprehensive critique of the prevailing *Observation-Oriented* causal learning paradigm. Finally, Section 4 delves into the temporal space untouched by the current paradigm, spotlighting its multi-dimensional nature that leads to inherent modeling issues.

2 Impact of Unobservable Hierarchies

Unobservable hierarchies in knowledge suggest unknown distributions in the hyper-dimensional space, which are related to but distinct from the modeling objective. For solely observational learning tasks, such unknowns may lead to troubles, but still have the potential to be uncovered through methods like reinforcement learning. However, when it comes to observational-temporal causal learning, the *Observation-Oriented* paradigm inherently falls short in capturing dynamical temporal features across all hierarchical levels. This section will illustrate these phenomena via two examples: one from computer vision and another from health informatics. For the latter, we will further dissect the issue from a traditional causal inference perspective.

2.1 Observational Hierarchy

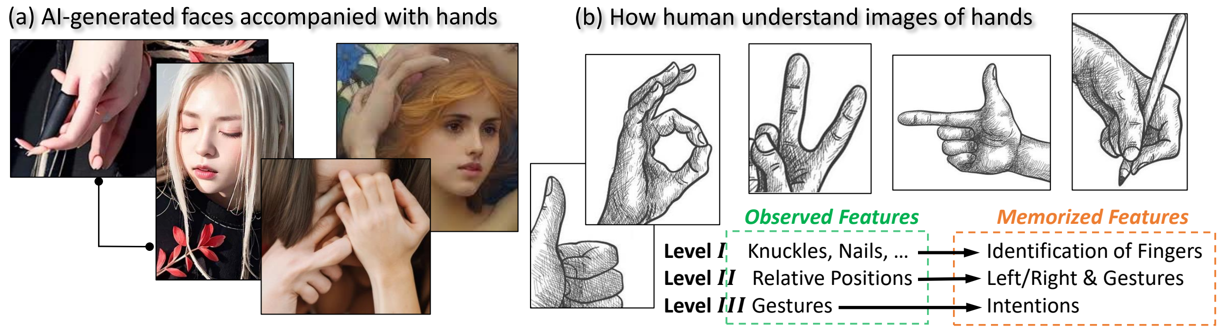

(a) AI-generated faces accompanied with hands
(b) How human understand images of hands
Observed Features
Memorized Features
Level *I* Knuckles, Nails, ...
Identification of Fingers
Level *II* Relative Positions
Left/Right & Gestures
Level *III* Gestures
Intentions

Figure 2: A comparison of AI-generated and human-sketched hand images. AI processes observed features simultaneously, thus treating hands as arbitrary mixtures of finger-like items. The process is hierarchical for humans, indexed using relations, where higher-level recognition relies on lower-level conclusions.

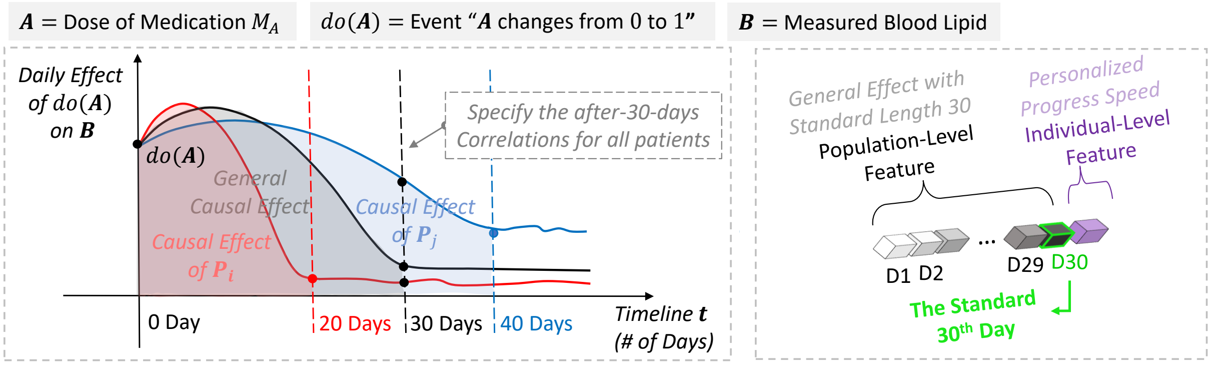
Figure 2(a) showcases AI-created hands with faithful color but unrealistic shapes, while humans can easily recognize a plausible hand from simple grayscale sketches in (b). Indeed, we can rapidly decompose our observations hierarchically according to different relations in our knowledge, and process sequentially from lower to higher levels: **I** identifies fingers through knuckles, nails, and relative lengths; **II** denotes hand gestures through positions; **III** retrieves the gesture's meaning from memory. However, such an intuitive hierarchy exists in our cognitions only. To AI, or similarly, to an extraterrestrial without our knowledge, the hands in Figure 2(a) may seem as reasonable as the actual hands.

Such observational hierarchy may not always create major problems. If features at different levels do not significantly overlap, AI may successfully "distinguish" them. For instance, AI can generate convincing faces because the appearance of eyes is strongly indicative of the facial angle, eliminating the need for AI to recognize "eyes" from "faces". But various hand gestures may have similar appearances, leading to chaos.

Even with problems, AI may learn the hidden knowledge via reinforcement learning Sutton & Barto (2018), under the guidance of human feedback. For example, human approval of five-fingered hands could lead AI to start identifying fingers autonomously. It works because of *completely* captured observational features at each level, while may not function when involving distributions across temporal dimensions.

2.2 Observational-Temporal Hierarchy

Figure 3(a) depicts patients' daily effects on B following $do(A)$, with t indicating the elapsed days. For simplicity, let's assume the patient's (unobserved) personal characteristics linearly influence M_A's release, i.e., uniformly accelerate or decelerate its effective progress. The individualized causal effects (i.e., the red and blue curves in (a)) are shaped by two levels of *dynamical temporal* features: 1) the population-level effect sequence with a standard length of 30, and 2) the individual-level progress speed. An accurate estimation of the level 1) dynamical feature provides the desired clinical effectiveness evaluation of M_A.

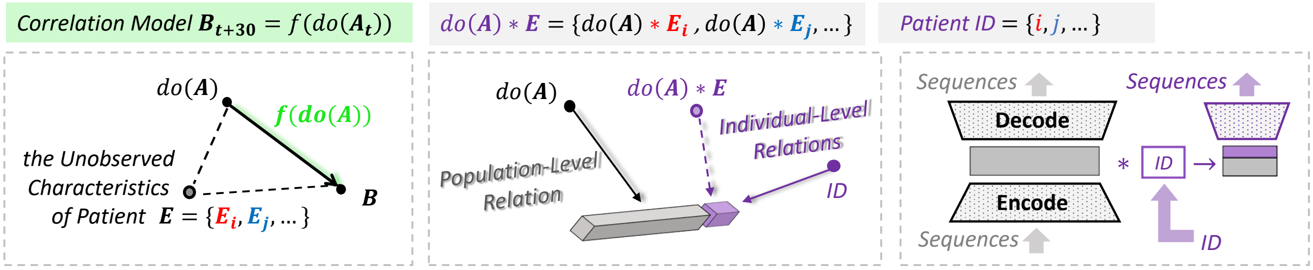


(a) Observational Time Sequences (b) Complete Dynamical Features

Figure 3: Medication M_A treats high blood lipid, with $do(A)$ denoting its initial use. The population-level effect takes about 30 days to fully manifest ($t = 30$ at the elbow), depicted by the black curve in (a). Patient P_i achieves this effect curve elbow in 20 days, while P_j takes 40 days.

Figure 3(b) represents patients' effects in a 31-length feature vector, disentangled by two hierarchical levels. Traditional medical effect estimation is often obtained by averaging the patients' after-30-day performances. This essentially builds a correlation model $B_{t+30} = f(do(A_t))$, which only captures a *static temporal* feature B_{t+30}, the last step of the level 1) dynamic, disregarding the preceding 29 steps. Moreover, even if the estimation method employs a sequence of length 30 (e.g., Granger causality), it can capture the level 1) dynamic at most and is exclusive of further levels. Causal effects with multiple levels of dynamics are prevalent in various causal learning applications, such as epidemic progression, economic fluctuations, strategic decision-making, etc. The *Observation-Oriented* paradigm necessitates identifying objects before establishing relations, making it often difficult to comprehensively encompass all levels of dynamics.

2.3 Strange Hidden-Confounder in Causal Inference


Correlation Model $\boldsymbol{B}_{t+30} = f(do(\boldsymbol{A}_t))$
$do(\boldsymbol{A}) * \boldsymbol{E} = \{do(\boldsymbol{A}) * \boldsymbol{E}_i, do(\boldsymbol{A}) * \boldsymbol{E}_j, \ldots\}$
Patient ID = $\{i, j, \ldots\}$
$do(\boldsymbol{A})$
$f(do(\boldsymbol{A}))$
the Unobserved Characteristics of Patient $\boldsymbol{E} = \{\boldsymbol{E}_i, \boldsymbol{E}_j, \ldots\}$
$\boldsymbol{B}$
$do(\boldsymbol{A})$
$do(\boldsymbol{A}) * \boldsymbol{E}$
Individual-Level Relations
Population-Level Relation
ID
Sequences
Decode
Encode
Sequences
Sequences
*
ID
ID

(a) DAG with Hidden Confounder (b) Relation-Oriented Disentanglement (c) Latent Space Representation of (b)

Figure 4: (a) Traditional causal inference DAG. (b) Hierarchical disentanglement of dynamics using relations as indices. (c) Autoencoder-based generalized and individualized reconstructions of the sequential data.

For patients P_i and P_j, the estimated last-day effect B_{t+30} is biased, as P_i exceeds 100% full effect, while P_j only achieves about 75%. To account for such individual-level biases, causal inference usually introduces a *hidden confounder* into DAG (Directed Acyclic Graph), to represent the unobserved personalized characteristics, depicted as the node E in Figure 4 (a), a strangely involved outer variable. It implies an illogical assertion: "Our model is biased due to some unknown aspects we have no intention to know."

It is because, while E is unknown, its effect, the individual-level dynamical feature, is observable, but excluded by the correlation model f. Although hidden, E is solely observational and thus could be incorporated by f if revealed. Thus, introducing a hidden confounder transforms *observed dynamical* variables into *unobserved observational* ones, which enhances human understanding but unnecessarily benefits the model.

As depicted in Figure 4(b), traditional causal inference views the individual-level effect as caused by the unobserved composite cause $do(A) * E$, not a directly modelable relationship. Conversely, a *Relation-Oriented* approach just treats relation as an index, to extract representations of *observational-temporal* effects from sequential data, so we can employ any observed identifier, e.g., patient ID. Figure 4(c) illustrates its implementation architecture, to realize a relation-defined hierarchical disentanglement.

3 Causality on Temporal Dimension

Causality research acts as a gateway into the temporal dimension, going beyond the observational space. However, the current causal learning models, represented as $y_{t+m} = f(x_t)$ for causality $X \to Y$, do not fully integrate t as a computational dimension.

Under the prevailing *Observation-Oriented* paradigm, the objects - cause on X and effect on Y - must be pre-identified prior to formulating the relation function f. While it remains feasible to assign a sequence of X to encompass key *dynamical temporal* features for the cause, identifying the exact start and end timestamps for the effect becomes problematic. Consequently, traditional causal inference typically treats effects as *observational* only. Even when trying to represent *static temporal* features, determining the appropriate value of m to capture a relevant snapshot can be challenging, especially if the effect carries dynamic significance.

Indeed, integrating the concept of *temporal distribution* could greatly streamline causal inference theories, making associated ideas more intuitive. For instance, when we acquire *Counterfactuals* Pearl (2009), we are essentially capturing temporal distributions in response to conditional queries. Also, as demonstrated in the prior section, fully capturing the observed dynamical temporal features within the model could potentially eliminate the need for hidden confounders.

Next, we begin by redefining the notion of causal models concerning the temporal dimension in Section 3.1, then delve into existing methodologies in Section 3.2, focusing on their capacity to capture temporal distributions, with a particular exploration of the essence of do-calculus. Section 3.3 discusses inherent limitations of the dominant *Observation-Oriented* causal modeling paradigm.

3.1 Redefined Causality vs. Correlation

Traditional causal inference places a strong emphasis on interpreting causal models; for instance, discerning causal directions is crucial, especially when differentiating them from mere correlations. From a modeling standpoint, once the domain is established, it functions regardless of the temporal logic behind the dimensions. Hence, it is reasonable that temporal-evolving causal aspects are primarily evident in interpretations, not directly within the modeling framework. Given this, we distinguish causality from correlation in the context of modeling, by incorporating distributions along the temporal dimension.

Theorem 1. Causality vs. Correlation in the modeling context.
- Causality is the relationship between *observational-temporal* features, which can be ***dynamical***.
- Correlation is the relationship between features ***not dynamical***.

A causality $X \to Y$ can be divided into two parts: 1) the informative relation connecting X and Y, which is crucial for modeling, and 2) the causal direction, distinguishing between cause and effect, which holds significance mainly in interpretation. Specifically, in a modeling context, we can employ $Y = f(X; \theta)$ to predict the effect on Y, and, conversely, utilize $X = g(Y; \psi)$ to deduce the cause X given Y. Both parameters, θ and ψ, are derived from the joint probability $\mathbf{P}(X, Y)$ without imposing modeling restrictions.

However, practical scenarios often emphasize determining the causal direction prior to modeling, suggesting underlying modeling concerns. One reason stems from the importance of aligning the modeling direction with our intuitive understanding of temporal progression. Moreover, the prevailing causal modeling paradigm displays an imbalanced capacity for capturing dynamical temporal features between the cause X and the effect Y. For example, in Figure 4, an inverse modeling of $do(A) = f(\{B_t\})$ through RNNs, given a sufficiently long sequence $\{B_t\} = \{B_{t+1}, \ldots, B_{t+40}\}$, might negate the need for a hidden confounder.

Within the suggested *Relation-Oriented* approach, we can utilize relations to accurately identify the effect's observational-temporal features and fully extract their representations. As a result, the modeling function f is relieved from encapsulating temporal facets. The differentiation between causality and correlation becomes a matter of connected features, rather than the nature of the relational model.

3.2 Learning Temporal Distributions

Numerous methods are dedicated to capturing the dynamical features of the ***cause*** alone, such as autoregressive models Hyvärinen et al. (2010) and RNNs Xu et al. (2020), both employing the modeling formate $y_{t+m} = f(\{x_t\})$ with $\{x_t\} = \{x_1, \ldots, x_t, x_{t+1}, \ldots, x_T\}$. Meanwhile, Granger causality Maziarz (2015), a method widely recognized in economics, employs a sequence for the ***effect*** that exhibits significant temporal patterns, in the formate $\{y_\tau\} = f(\{x_t\})$, where t and τ signifying two separate timelines.

Yet, using a sequence does not equate to capturing dynamics. The distinction between "a sequence of static variables" and "a dynamical variable" hinges on whether the *nonlinear* mutual relationships among these variables can be identified. For autoregressive, if the selected model is linear, then $\{x_t\}$ remains a static sequence, which can capture dynamics at a single level at most, akin to the level 1) sequence in Figure 3(b). Conversely, RNNs can harness the nonlinearity of $\{x_t\}$, enabling them to encapsulate multiple levels of dynamics. However, for the effect sequence $\{y_\tau\}$ in Granger causality, due to its *Observation-Oriented* nature, it can capture only a single-level dynamic at most, referring to Figure 3(b)'s discussions.

A more universal approach to represent temporal distributions is do-calculus Pearl (2012); Huang & Valtorta (2012). Instead of specifying time sequences, it takes the *identifiable* temporal events as modeling objects to conduct elementary calculus. The $do(\cdot)$ format flexibly modulates the temporal features for the ***cause*** alone. However, such a *differential*-calculus essence also introduces elevated complexity. Here, we reinterpret its three core rules from an *integral*-calculus perspective, aiming for a more intuitive comprehension.

For the time sequence $\{x_t\} = \{x_1, \ldots, x_T\}$, let $do(x_t) = \{x_t, x_{t+1}\}$ indicate the occurrence of an instantaneous event $do(x)$ at time t. Time lag Δt between $\{t, t+1\}$ is sufficiently small to make this event identifiable, such that $do(x_t)$'s *interventional* effect can be depicted as a function of the resultant distribution at $t+1$. Conversely, the effect provoked by static x_t snapshot is called *observational* effect. Then, the observational-temporal distribution of the cause $X \in \mathbb{R}^d$ can be formulated as below:

$$\text{Given } \mathcal{X} \to Y \mid Z, \text{ where } \mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1} \text{ encompass the temporal dimension, we have}$$

$$\mathcal{X} = \int_0^T do(x_t)\cdot x_t \; dt \quad \text{with} \begin{cases} (do(x_t)=1) \mid do(z_t), & \textit{Observational} \text{ only (Rule 1)} \\ (x_t = 1) \mid do(z_t), & \textit{Interventional} \text{ only (Rule 2)} \\ (do(x_t)=0) \mid do(z_t), & \text{No } \textit{interventional} \text{ (Rule 3)} \\ \text{otherwise} & \text{Associated } \textit{observational} \text{ and } \textit{interventional} \end{cases}$$

$$\text{The effect of } \mathcal{X} \text{ can be derived as } f(\mathcal{X}) = \int_0^T f_t\big(do(x_t)\cdot x_t\big)\; dt = \sum_{t=0}^{T-1} (y_{t+1} - y_t) = y_T - y_0$$

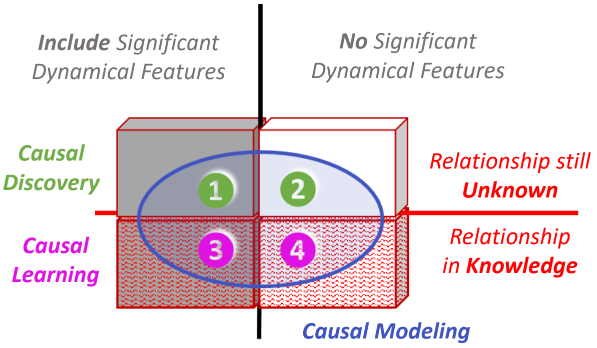
Within the graphical system $\{X, Y, Z\}$, the rules of do-calculus tackle three specific scenarios (notably, a specifiable $do(x_t)\cdot x_t$ pertains to Rule 2), where conditional independence is maintained between the *observational* and *interventional* effects. However, these rules bypass more generalized cases.

Utilizing the $do(\cdot)$ format, we can also represent observational-temporal distributions of Y as $\mathcal{Y} = \langle Y, \tau\rangle$, by incorporating an additional timeline τ. However, under the *Observation-Oriented* paradigm, identifiable events for both $\mathcal{X}$ and $\mathcal{Y}$ still require our prior identifications, as opposed to the automatic construction of $\mathcal{Y}$ automatically within the suggested *Relation-Oriented* in the proposed Relation-Oriented methodology.

3.3 Limitation of Current Causal Modeling Paradigm

Our inherent understanding of causality complies with Theorem 1. But *Observation-Oriented* models are mainly confined to the observational space, resulting in potential misalignments between established causal models and our intuitive knowledge. We have categorized causal modeling into four scenarios shown in

Figure 5. Depending on whether the relationship is already in knowledge, queries can be divided into causal discovery, which seeks new insights, and causal learning, which leverages knowledge to model causality. Further, these applications can be sorted based on the dynamical significance of the effects. For instance, the causality "raining → wet floor" includes only static temporal features; it is logically a causality but not distinguishable from correlation when modeled. We explore these scenarios from two perspectives: the *relation* connecting features, critical for modeling, and the *causal direction*, essential for interpretation.



	Modeled Relation	Modeled Causal Direction
①	Observational Only. Undiscovered Dynamics covered by *Faithfulness Assumption.*	Observational Information Determined. Not Logical Causal Significant.
②	Observational Only. Aligned with Knowledge.	Observational Information Determined. Maybe Logical Causal Significant.
③	Knowledge Determined. Unmodeled Dynamics covered by *Hidden Confounders* or *Sufficiency Assumption.*	Knowledge Determined.
④	Knowledge Determined.	Knowledge Determined.

Figure 5: An overview of the current *Observation-Oriented* causal modeling paradigm's limitations. On the left, the rectangle means all logical causal relationships, while its potentially modelable scope is blue-circled.

(1) Modeled Relation Connecting Features

Traditional causal inference has made notable advancements in "downgrading" dynamical temporal features to be observationally accessible. For instance, do-calculus Pearl (2012) explores independence conditions on the temporal dimension. For overlooked dynamical features of the effect, if existing knowledge can suggest its potential cause, creating a hidden confounder can enhance comprehension; if not, these dynamics may be dismissed based on the *causal Sufficiency assumption*, potentially leading to subsequent challenges.

On the other hand, causal discovery mainly scans the observational space to explore dependencies. As a result, if the underlying causality does not encompass dynamical features, causal discovery can be effective. However, if such dynamics exist, they largely go undetected. This potential gap may be negated under the *causal Faithfulness assumption* suggesting that observed variables fully represent the causal reality.

(2) Modeled Causal Direction

Consider observed variables X and Y in a graphical system, with specified models $Y = f(X;\theta)$ and $X = g(Y;\psi)$. Based on observations, the discovered causal direction between X and Y is determined by the likelihoods of estimated parameters $\hat{\theta}$ and $\hat{\psi}$. Given the joint distribution $\mathbf{P}(X,Y)$, one would prefer $X \to Y$ if $\mathcal{L}(\hat{\theta}) > \mathcal{L}(\hat{\psi})$. Now, let $\mathcal{I}(\theta)$ be a simplified form of $\mathcal{I}_{X,Y}(\theta)$, the Fisher information, representing the amount of information contained by $\mathbf{P}(X,Y)$ about unknown θ. Assume $p(\cdot)$ to be the probability density function; then, in this context, $\int_X p(x;\theta)dx$ remains constant. So, we have

$$\mathcal{I}(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}\log p(X,Y;\theta))^2 \mid \theta] = \int_Y\int_X (\frac{\partial}{\partial\theta}\log p(x,y;\theta))^2 p(x,y;\theta)dxdy$$

$$= \alpha\int_Y (\frac{\partial}{\partial\theta}\log p(y;x,\theta))^2 p(y;x,\theta)dy + \beta = \alpha\mathcal{I}_{Y|X}(\theta) + \beta, \text{with } \alpha,\beta \text{ constants.}$$

$$\text{Thus, } \hat{\theta} = \arg\max_{\theta} \mathbf{P}(Y \mid X,\theta) = \arg\min_{\theta} \mathcal{I}_{Y|X}(\theta) = \arg\min_{\theta} \mathcal{I}(\theta), \text{ and } \mathcal{L}(\hat{\theta}) \propto 1/\mathcal{I}(\hat{\theta}).$$

Subsequently, the likelihoods of the estimated parameters $\hat{\theta}$ and $\hat{\psi}$ depend on the amount of information, $\mathcal{I}(\hat{\theta})$ and $\mathcal{I}(\hat{\psi})$. That means, the learned directionality between X and Y essentially indicates how much their specified distributions are reflected in the data, with the more dominant one deemed the "cause". It presumes that the cause is more comprehensively captured in the observations than the effect by default. Due to restricted data collection techniques, such a presumption was justifiable in past decades. But in the present era, assuming such discovered directions to have logical causal meaning is no longer appropriate.

4 The Overlooked Temporal Space

Data is commonly stored in matrices, with time series data incorporating an extra attribute for the timestamps, which forms a logical timeline to reflect the absolute time evolution in reality. Traditionally, modeling has relied on this timeline to determine the chronological order of all potential events. However, our intuitive understanding of time is far more complex than this singular, simplified absolute timeline.

Consider an analogy where ants dwell on a two-dimensional plane of a floor. If these ants were to construct models, they might use the nearest tree as a reference to specify the elevation in their two-dimensional models. By modeling, they observe an increased disruption at the tree's mid-level, which indicates a higher chance of encountering children. However, since they fail to comprehend humans as three-dimensional beings, instead of interpreting this phenomenon in a new dimension "height", they solely relate it to the tree's mid-level. If they migrate to a different tree with a varying height, where mid-level no longer presents a risk, they might conclude that human behavior is too complex to model effectively. Similarly, when modeling time series, we usually discount the dimension "time" as the single absolute timeline, which has become our "tree".

Our understanding allows for the simultaneous existence of multiple logical timelines. If one is designated as the absolute timeline, the remaining ones can be viewed as relative timelines, each representing distinctive temporal events, which can be interconnected via specific relationships. In such *Relation-Oriented* perspective, like, during a causal inference analysis, the temporal dimension contains numerous possible logical timelines that we could choose to construct any necessary scenarios. However, once we enter a modeling context, like, using AI to model the time series along a single timeline, the temporal significance no longer exists, but only a regular dimension containing timestamp values, indistinguishable from other observational attributes. Metaphorically, if we consider the observational space for AI modeling as Schrödinger's box and our interest is the "cat" within, our task is to accurately construct the box, giving adequate consideration to all potential logical timelines, to ensure the "cat" remains *reasonable* upon unveiling.

Theorem 2. The term *Temporal Dimension* encompasses all potential logical timelines, not just a singular one. Consequently, a *Temporal Space* is defined as the space built by chosen timeline axes.

Fundamentally, as three-dimensional beings, we are limited from truly understanding temporal dimensionality. As the term "space" typically evokes a three-dimensional conception, the notion of "temporal space" might seem odd for a four-dimensional creature. Like ants can use trees as references without the need to fully comprehend the third dimension, we rely on logical timelines to interpret the fourth. At this juncture, our mission is to recognize the potential "forest" beyond the present single "tree".

This section will demonstrate how the single-timeline-based timestamp specification operation, rooted in the *Observation-Oriented* paradigm, inherently biases modeling and hinders model generalizability. Then we will summarize advancements and challenges on our journey towards realizing causal knowledge-aligned AI.

4.1 Inherent Temporal Bias Scheme

Model specification typically requires event timestamps to be set up, often based on a single absolute timeline present in the data. Yet, for structural causal models (SCMs), this can introduce *inherent temporal biases*, considerably constraining our ability to leverage AI's potential in the temporal dimension. This challenge can be more pronounced in large-scale causal relationships, which may hide more logical timelines.

To better ascribe this issue, we *redefine* the causal Directed Acyclic Graph (DAG) Pearl (2009) as follows: **1)** incorporating (potentially multiple) logical timelines as axes into the DAG space, and **2)** defining edges along timeline axes to be vectors with meaningful lengths indicating the timespans of causal effects. For example, the single-timeline scenario in Figure 3 has the redefined DAG depicted in Figure 6(b), with (a) showing the traditional one as a comparison. The edge $do(A) \to B$ in Figure 6(a) can only (partially) represent population-level effect, thus necessities a hidden confounder to explain the individual-level diversities, while in Figure 6(b), they can be explicitly represented by varying lengths of $\overrightarrow{do(A)\ B}$.

Consider an expanded two-timeline scenario in Figure 7(a), where A shorthandly represents $do(A)$. Apart from its primary effect on B, A also indirectly influences B through its side effect on another vital sign, C, depicted as edges $\overrightarrow{AC}$ and $\overrightarrow{CB}$. For simplicity, assume the timespan for $\overrightarrow{AC}$ is 10 days for all patients, with the individual-level diversity solely confined to timeline T_X. In conventional single-timeline causal modeling, the SCM function would be $B_{t+30} = f(A_t, C_{t+10})$. Let's assume $f(A_t, C_{t+10})$ is implemented using RNNs, which could accurately depict the individual-level final effects of A on B for any patient.

The confounding relationship over nodes $\{A, B, C\}$ forms a triangle across timelines T_X and T_Y - such shape geometrically holds for any hierarchical level relationship. For patients P_i and P_j, the *individualization* process is to "stretch" this triangle along T_X by different ratios, which is a homographic *linear transformation* in this space. However, as illustrated in Figure 7 (b) and (c), for either P_i or P_j, equating the outcome of f to be B_{t+30} violates the *causal Markov condition* necessary for reasonable SCMs.

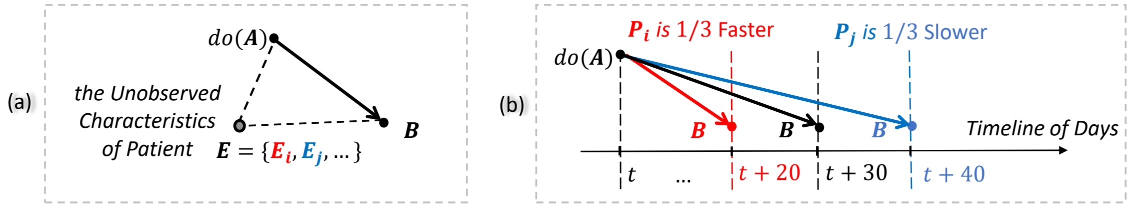


Figure 6: (a) Traditional Causal DAG introducing hidden E. (b) Redefined DAG: the standard black vector signifies the population-level effect, while the individual-level ones are represented by its different scaling.

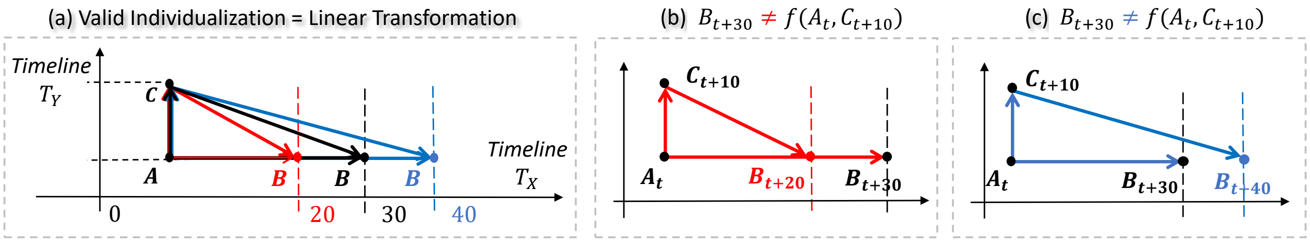

(a) Valid Individualization = Linear Transformation
Timeline
T_Y
C
A
B
B
B
Timeline
T_X
0
20
30
40
(b) $B_{t+30} \neq f(A_t, C_{t+10})$
C_{t+10}
A_t
B_{t+20}
B_{t+30}
(c) $B_{t+30} \neq f(A_t, C_{t+10})$
C_{t+10}
A_t
B_{t+30}
B_{t+40}

Figure 7: (a) A two-timeline (redefined) DAG space, where a valid individualization presents a linear transformation. (b)(c) Violations of the Markov condition for the prevailing SCM with confounding across timelines.

Notably, the violation may not cause significant issues for AI models like RNNs in this specific case. Given the *independence* of dynamical features on T_X and T_Y, the SCM can be formulated as $B_{t+30} = f_1(A_t) + f_2(C_{t+10})$, which suggests that the cross-timeline confounding can be broken down into two single-timeline issues. However, making assumptions such as independence or non-confounding is unrealistic. Given that each cause-and-effect pair might exist on its own unique logical timeline, such biases could accumulate exponentially, profoundly affecting causal applications regardless of our model choices.

Theorem 3. The ***inherent temporal bias*** may occur in SCM if it contains: **1)** *Confounding* dynamical features across *Multiple* logical timelines, and **2)** Unobservable hierarchy.

It is interesting to notice that most of the successful applications instinctively avoid one of the two factors: *confounding* or *multi-timeline*. Statistical causal models can be particularly adjusted to facilitate de-confounding, e.g., the backdoor adjustment Pearl (2009). For AI models, most of the sweeping achievements do not potentially involve relative timelines, e.g., the large language model (LLM) in a semantic space, where the phrases are ordered consistently along a single logical timeline.

Unlike AI's black-box nature, causal inference inherently takes a *Relation-Oriented* view. But in its context, the inherent temporal biases are difficult to recognize, as they often intermingle with the biases resulting from unsolvable nonlinearity - They have similar manifestations, and both can be addressed by de-confounding. Consider Figure 6(a), a linear causal model overlooking the individual-level dynamical feature can mismatch with individuals P_i and P_j, which may not be distinguishable from the model mismatching in Figure 7(b)(c), caused by dynamics crossing two timelines.

4.2 Inherent Impact on SCM Generalizability

The traditional SCM possesses an *Observation-Oriented* nature, typically necessitating timestamp specification for object events prior to formulating relations. Due to the inherent temporal bias, such specifications can influence the precision of ongoing models. However, more significantly, they might render the established SCM *non-generalizable* for different scenarios, such as a new population that maintains the same core causal relationship but manifests differently from the current training group.

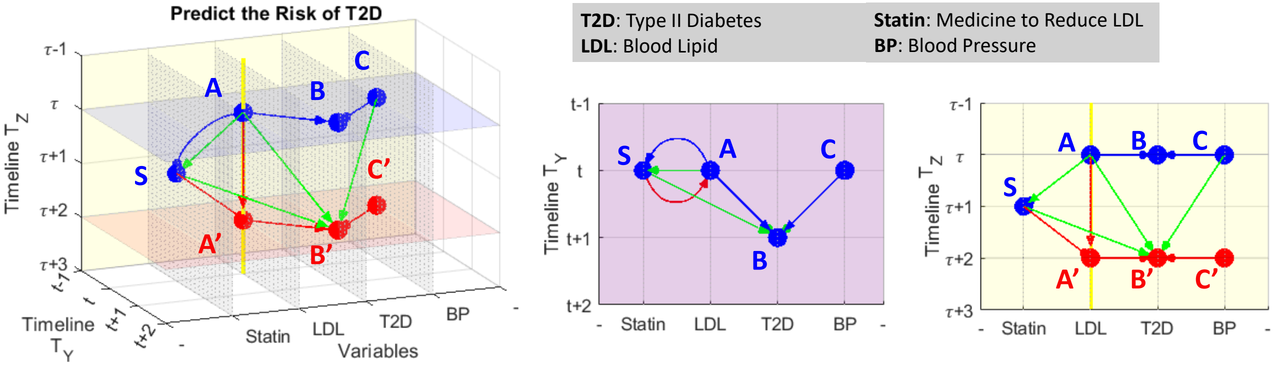


Figure 8: An example DAG in 3D observational-temporal space, where the SCM function $B' = f(A, C, S)$ aims to evaluate Statin's medical effect on reducing the risk of T2D, with two logical timelines $\mathcal{T}_Y$ and $\mathcal{T}_Z$. On $\mathcal{T}_Y$, the step Δt from t to $(t+1)$ allows A and C to fully influence B, while the step $\Delta\tau$ on $\mathcal{T}_Z$, from $(\tau+1)$ to $(\tau+2)$, let medicine S completely release its effect on LDL, which is, switching from A to A'.

Let's consider the practical scenario depicted in Figure 8. Here, Δt and $\Delta\tau$ represent actual time spans. The main point is not determining their exact values, but on their intended causal relationship: As each unit of Statin's effect is delivered on LDL via $\overrightarrow{SA'}$, it immediately impacts T2D through $\overrightarrow{A'B'}$. Simultaneously, the next unit effect begins generation. This dual action runs concurrently until S is fully administered. The ultimate aim of this process is to evaluate the total cumulative influence stemming from S at B'.

Given the relationship $\overrightarrow{SB'} = \overrightarrow{SA'} + \overrightarrow{A'B'}$, specifying the $\overrightarrow{SB'}$ time span inherently sets the $\Delta t : \Delta\tau$ ratio, defining the ASB' triangle's shape in the model. While the mean effect at B' might be precise for the present population, the preset $\Delta t : \Delta\tau$ ratio's universality is questionable, potentially constraining the established SCM's generalizability.

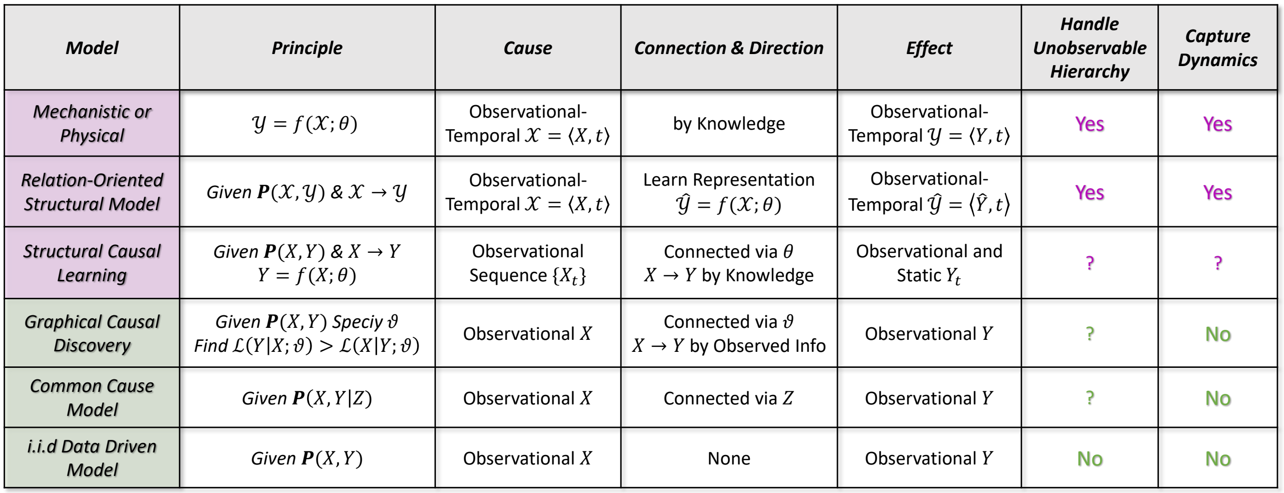
4.3 Toward Causal Knowledge-Aligned AI

Our quest for causal reasoning AI involves broadening our modeling techniques from purely observational to include temporal dimensions, as summarized in Figure 9. The present challenge lies in enabling structural causal models in the temporal space. Recognizing the underlying logical timelines is critical to avoid the inherent biases and enhance model generalizability. However, since manual identification is unrealistic, it may have been time for us to consider the new paradigm.

The initial models under i.i.d. assumption only approximate observational associations, proved unreliable for causal reasoning Pearl et al. (2000); Peters et al. (2017). Correspondingly, the common cause principle highlights the significance of the nontrivial conditional properties, to distinguish structural relationships from statistical dependencies Dawid (1979); Geiger & Pearl (1993), providing a basis for effectively uncovering the underlying structures in graphical models Peters et al. (2014).

Graphical models, employing conditional dependencies to construct Bayesian networks (BNs), often operate in observational space and neglect temporal aspects, reducing their causal relevance Scheines (1997). Notably, causally significant models, such as Structural Equation Models (SEMs) and Functional Causal Models (FCMs) Glymour et al. (2019); Elwert (2013), are able to address *counterfactual* queries Scholkopf et al. (2021). Typically, these models leverage prior knowledge to construct causal DAGs.

State-of-the-art deep learning applications on causality, which encode the DAG structural constraint into continuous optimization functions Zheng et al. (2018; 2020); Lachapelle et al. (2019), undoubtedly enable highly efficient solutions, especially for large-scale problems. However, larger question scales indicate more underlying logical timelines, which may lead to snowballing temporal biases. It can be evident from the limited successful applications of incorporating DAG structure into network architectures Luo et al. (2020); Ma (2018), e.g., neural architecture search (NAS).



Model	*Principle*	*Cause*	*Connection & Direction*	*Effect*	*Handle Unobservable Hierarchy*	*Capture Dynamics*
Mechanistic or Physical	$\mathcal{Y} = f(\mathcal{X};\theta)$	Observational-Temporal $\mathcal{X} = \langle X, t\rangle$	by Knowledge	Observational-Temporal $\mathcal{Y} = \langle Y, t\rangle$	Yes	Yes
Relation-Oriented Structural Model	*Given* $\mathbf{P}(\mathcal{X},\mathcal{Y})$ & $\mathcal{X} \to \mathcal{Y}$	Observational-Temporal $\mathcal{X} = \langle X, t\rangle$	Learn Representation $\hat{\mathcal{Y}} = f(\mathcal{X};\theta)$	Observational-Temporal $\hat{\mathcal{Y}} = \langle \hat{Y}, t\rangle$	Yes	Yes
Structural Causal Learning	*Given* $\mathbf{P}(X,Y)$ & $X \to Y$ $Y = f(X;\theta)$	Observational Sequence $\{X_t\}$	Connected via θ $X \to Y$ by Knowledge	Observational and Static Y_t	?	?
Graphical Causal Discovery	*Given* $\mathbf{P}(X,Y)$ *Speciy* ϑ *Find* $\mathcal{L}(Y\|X;\vartheta) > \mathcal{L}(X\|Y;\vartheta)$	Observational X	Connected via ϑ $X \to Y$ by Observed Info	Observational Y	?	No
Common Cause Model	*Given* $\mathbf{P}(X,Y\|Z)$	Observational X	Connected via Z	Observational Y	?	No
i.i.d Data Driven Model	*Given* $\mathbf{P}(X,Y)$	Observational X	None	Observational Y	No	No

Figure 9: Simple Taxonomy of Models (Adapted in part of Table 1 in Scholkopf et al. (2021)), from more knowledge-driven (top in purple) to more data-driven (bottom in green). Notations: ϑ or θ = parameters derived from joint or conditional distribution, $\langle X, t\rangle$ = augment t-dimension, "?" = depending on practice.

Schölkopf Scholkopf et al. (2021) summarized three key challenges impeding causal AI applications to achieving generalizable success: 1) limited model robustness, 2) insufficient model reusability, and 3) inability to handle data heterogeneity (caused by unobservable hierarchies in knowledge). Notably, all these challenges can be attributed to the timestamp specification required by *Observation-Oriented* structural models.

On the other side, physical models, which explicitly integrate temporal dimensions in computation, and are able to establish abstract concepts through relations, may provide insights into these challenges. We believe that the *Relation-Oriented* approach can help bridge the gap between observational and temporal spaces.

Chapter II: Realization of Proposed Relation-Oriented Paradigm

This chapter begins by formulating the factorizations to achieve hierarchical disentanglement in the latent space. Then, we explore the proposed *relation-defined representation* methodology as an embodiment of the *Relation-Oriented* paradigm. Lastly, we validate its feasibility through comprehensive experiments.

5 Hierarchical Disentanglement in Latent Space

For observational variable $X \in \mathbb{R}^d$ with time sequence $\{x_t\} = \{x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T\}$, We aim to devise a latent feature space $\mathbb{R}^L$ for two purposes: 1) Fully represent the observational-temporal features of $\mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1}$. 2) Hierarchically disentangle $\mathcal{X}$'s representation according to knowledge. Consequently, the established system can facilitate the reusability of models at any level, indexed by relations in knowledge.

For $\mathcal{Y} = \langle Y, \tau\rangle \in \mathbb{R}^{b+1}$, if the relationship $\mathcal{X} \to \mathcal{Y}$ identifies a level in the unobservable hierarchy for $\mathcal{Y}$, the proposed *relation-defined representation* learning aims to extract the representation $\hat{\mathcal{Y}}$ as defined by the relation with $\mathcal{X}$. Moreover, the resulting $\hat{\mathcal{Y}}$ should be reusable in the development of further levels of representations based on it, facilitating the model's generalizability. For example, in the graphical system $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$ with relationship $\mathcal{X} \to \mathcal{Y} \leftarrow \mathcal{Z}$, $\mathcal{Y}$ can be viewed within a two-level hierarchy. The first level is defined by $\mathcal{X} \to \mathcal{Y}$ and the second by $(\mathcal{X}, \mathcal{Z}) \to \mathcal{Y}$, where the second level enhances the first by incorporating an additional data stream from $\mathcal{Z}$.

5.1 Factorize Observational-Temporal Hierarchy

Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$, and assume $\mathcal{X} = \langle X, t \rangle \in \mathbb{R}^{d+1}$ has an n-level hierarchy. Define Θ_i as the i-th level component of $\mathcal{X}$ in the *observable data space*, and its counterpart in the *latent feature space* $\mathbb{R}^L$ as θ_i. The representation function f_i facilitates the transformation from $\mathbb{R}^{d+1}$ to $\mathbb{R}^{L_i}$ for the i-th level, considering all prior lower-level features as attributes. θ_i is a vector in $\mathbb{R}^L$, with its significant value residing in a subset of the L dimensions, denoted as $\mathbb{R}^{L_i}$, forming the disentanglement $\{\mathbb{R}^{L_1}, \dots, \mathbb{R}^{L_i}, \dots, \mathbb{R}^{L_n}\}$. Then, we obtain:

$$\mathcal{X} = \sum_{i=1}^{n} \Theta_i, \text{ where } \Theta_i = f_i\big(\theta_i;\ \Theta_1, \dots, \Theta_{i-1}\big) \text{ with } \Theta_i \in \mathbb{R}^{d+1} \text{ and } \theta_i \in \mathbb{R}^{L_i} \subseteq \mathbb{R}^L \tag{1}$$

To illustrate an observational hierarchy, refer to Figure 2 (b). Let $\theta_1 \in \mathbb{R}^{L_1}$, $\theta_2 \in \mathbb{R}^{L_2}$, and $\theta_3 \in \mathbb{R}^{L_3}$ represent the three levels of features, with each subspace being mutually exclusive. That is, $L = L_1 + L_2 + L_3$. The combined vector $\langle \theta_1, \theta_2, \theta_3 \rangle \in \mathbb{R}^L$ represent the whole image. In correspondence, Θ_1, Θ_2, and Θ_3 are full-scale images, each presenting unique content. For instance, Θ_1 highlights the details of the fingers, whereas $\Theta_1 + \Theta_2$ expands to showcase the entire hand.

In the context of an observational-temporal hierarchy, the component $\Theta_i \in \mathbb{R}^{d+1}$ can be expressed as the original time sequence $\{\Theta_t\}_i = \{\Theta_{t_i} \in \mathbb{R}^d \mid t_i = 1, \dots, T\}$. Consequently, we obtain a set of relative logical timelines $\{t_1, \dots, t_i, \dots, t_n\}$ which, in contrast to the absolute timeline t, are each uniquely determined by the relationship at their respective levels. In the *observable data space*, the observation at the i-th level, represented as the sum $\Theta_1 + \ldots + \Theta_i$, maintains its sequence along t.

5.2 Factorize Hierarchy of Relationship

Given a set of n-level hierarchical representation functions for $\mathcal{X}$, denoted by $\mathcal{F}(\vartheta) = \big\{f_i\big(\theta_i\big) \mid i = 1, \dots, n\big\}$, our goal is to define n relationship functions, collectively termed $\mathcal{G}$, such that $\mathcal{Y} = \mathcal{G}(\mathcal{X})$ exhibits an n-level hierarchy. Each i-th level relationship function is $g_i(\mathcal{X}; \varphi_i)$, where φ_i is its parameter. Then, we have:

$$\mathcal{G}(\mathcal{X}) = \sum_{i=1}^{n} g_i(\mathcal{X}; \varphi_i) = \sum_{i=1}^{n} g_i(\Theta_i; \varphi_i) = \sum_{i=1}^{n} g_i\big(\theta_i;\ \Theta_1, \dots, \Theta_{i-1}, \varphi_i\big) = \mathcal{Y} \tag{2}$$

The i-th level relation-defined representation for $\mathcal{Y}$ is $g_i(\theta_i; \varphi_i)$ considering the features of the preceding $(i-1)$ levels of $\mathcal{X}$. This relationship can be portrayed as the augmented feature vector $\langle \theta_i, \varphi_i \rangle$ in latent space $\mathbb{R}^L$. Using ϑ_X and ϑ_Y to distinguish the collective hierarchical representations for $\mathcal{X}$ and $\mathcal{Y}$ respectively, the overall relationship from $\mathcal{X}$ to $\mathcal{Y}$ becomes $\vartheta_Y = \langle \vartheta_X, \varphi \rangle$, where $\varphi = \{\varphi_1, \dots, \varphi_n\}$. The term $\langle \vartheta_X, \varphi \rangle$ represents the pairwise augmentations between collections ϑ_X and φ.

6 Relation-Defined Representation Methodology

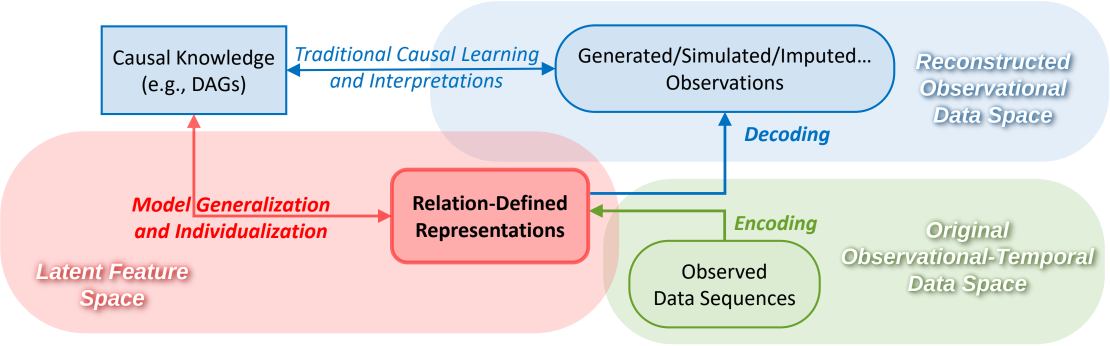

Causal Knowledge (e.g., DAGs)
Traditional Causal Learning and Interpretations
Generated/Simulated/Imputed… Observations
Reconstructed Observational Data Space
Decoding
Model Generalization and Individualization
Relation-Defined Representations
Encoding
Original Observational-Temporal Data Space
Latent Feature Space
Observed Data Sequences

Figure 10: Framework of utilizing *relation-defined representations* to benefit conventional models.

While the existing *Observation-Oriented* modeling paradigm has its limitations, it still forms the basis of many existing knowledge infrastructures. As showcased in Figure 10, relation-defined representations enable

AI to develop generalizable models within a latent feature space abundant with human-indecipherable data. Concurrently, this framework amplifies AI's potential to optimize observations, strengthening traditional models by enabling necessary counterfactual effects, simulating de-confounded observations, and more.

This section introduces a special autoencoder architecture required for implementing relation-defined representation. Building on this, we detail the process of stacking hierarchical levels of representations to construct graphical models. Finally, we present a causal discovery algorithm within the latent feature space.

6.1 Autoencoder Design for Higher-Dimensional Representation

Autoencoders are primarily used for dimensionality reduction Wang et al. (2016). In structural modeling, one often treats all variables (i.e., nodes in a DAG) as aligned observations to reduce data dimensionality. In contrast, our aim is to model individual relations and "stack" them to construct a DAG within the latent space $\mathbb{R}^L$. This requires a large dimensionality for $\mathbb{R}^L$ to capture all potential hierarchical features. As a result, we face the significant technical challenge of achieving a *higher-dimensional* representation extraction.

> **Corollary 1.** For a given graph G and a data matrix $\mathbf{X}$ that is column-augmented with all observational attributes of variables in G as well as timestamps, the dimensionality of the latent space L must be at least the rank of $rank(\mathbf{X})$ to adequately represent G.

Corollary 1 stems from the notion that the autoencoder-learned $\mathbb{R}^L$ is spanned by $\mathbf{X}$'s top principal components, often referred to in Principal Component Analysis (PCA) Baldi & Hornik (1989); Plaut (2018); Wang et al. (2016). Hypothetically, reducing L below $rank(\mathbf{X})$ may yield a less comprehensive but causally more significant latent space through better alignment Jain et al. (2021), although further exploration is needed. In this study, we will set aside discussions on the boundaries of dimensionality. Our experiments feature 10 variables with dimensions 1 to 5 (Table 1), and we empirically fine-tune and reduce L from 64 to 16.

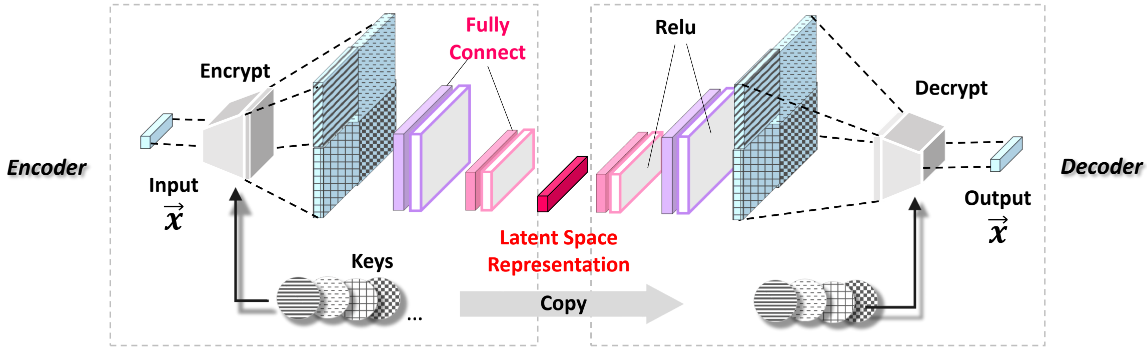


Figure 11: Invertible autoencoder architecture for extracting *higher-dimensional* representations.

Figure 11 depicts the proposed autoencoder architecture, which creates symmetrical *Encrypt* and *Decrypt* layers at the input and output. The Encrypt layer amplifies the input vector $\vec{x}$ by extracting its higher-order intrinsic features. Conversely, the Decrypt layer symmetrically reduces the input and restores $\vec{x}$ to its original form. To ensure reconstruction accuracy, the invertibility of these operations is naturally required.

Figure 11 illustrates a *double-wise* feature expansion. In this method, each pair of *two* digits from $\vec{x}$ is encoded into a new digit, thus capturing their association. This is accomplished using a *Key*, a set of random constants created by the encoder and mirrored by the decoder for reverse decryption. The double-wise expansion *Key* on $\vec{x} \in \mathbb{R}^d$ generates a $(d-1)(d-1)$ length vector. By augmenting these vectors using multiple *Keys*, $\vec{x}$ can significantly extend beyond its original length d. The four differently patterned squares in Figure 11 represent the results of four distinct *Keys*. Each square visualizes a $(d-1)(d-1)$ length vector (not suggesting 2-dimensionality), with the patterned grids indicating each *Key*'s unique "signature". As an analogy, higher-order extensions such as *triple-wise* ones across every three digits can also be employed, by appropriately adapting the *Key* to encapsulate more intricate associations within the data.

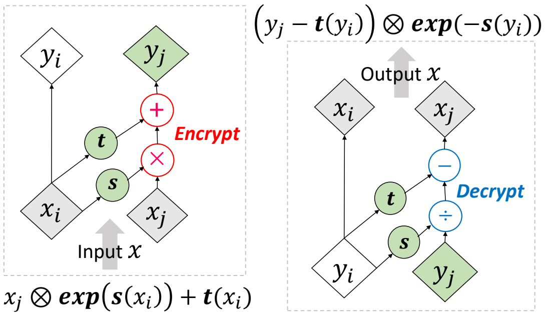


Figure 12: Encrypt (left) and Decrypt (right).

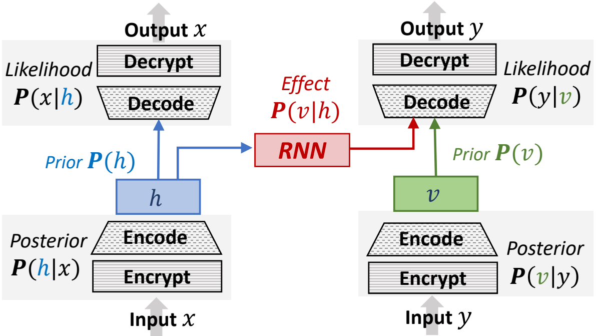


Figure 13: Architecture of Latent Effect.

Figure 12 illustrates the Encrypt and Decrypt functions executing a double-wise expansion. These processes transform a digit pair (x_i, x_j), $i \neq j \in 1, \ldots, d$, via encryption $f_\theta(x_i, x_j)$, with $\theta = (w_s, w_t)$ as the *Key* comprising two weights defining elementary functions $s(\cdot)$ and $t(\cdot)$. Specifically, $f_\theta(x_i, x_j) = x_j \otimes exp(s(x_i)) + t(x_i)$ is applied to each digit pair, transforming x_j into a new digit y_j using x_i as a parameter. The Decrypt layer uses the symmetric inverse function f_θ^{-1}, defined as $(y_j - t(y_i)) \otimes exp(-s(y_i))$. Importantly, this calculation sidesteps the need for s^{-1} or t^{-1}, permitting both linear and non-linear transformations. With the set of all f_θ functions denoted as $\mathcal{F}(X; \vartheta)$ - where X is the input variable and ϑ comprises all parameters - the Encrypt and Decrypt layers can be represented as $Y = \mathcal{F}(X; \vartheta)$ and $X = \mathcal{F}^{-1}(Y; \vartheta)$, respectively. Drawing from the seminal work of Dinh et al. (2016) on invertible neural network layers, we employ bijective functions to design our autoencoder. We specifically use the double-wise extension function $f_\theta(x_i, x_j)$, operating on digit pairs, thus preserving reconstruction accuracy. This bijective foundation ensures our architecture's robustness and adaptability, tailoring extension levels to application requirements. The source code for Encrypt and Decrypt is provided [1], along with a comprehensive experimental demo.

6.2 Structural Model with Hierarchical Representations

Consider a causal system comprising three variables: $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$. For each, a corresponding representation $\{\mathcal{H}, \mathcal{V}, \mathcal{K}\} \in \mathbb{R}^L$ is generated via independent autoencoders with the aforementioned architecture. Figure 13 portrays the process of connecting $\mathcal{H}$ and $\mathcal{V}$ to represent the relation $\mathcal{X} \to \mathcal{Y}$, while Figure 14 illustrates stacking these relations to represent the entire causal system, thereby enabling a hierarchical representation.

Assume x and y as instances of the relation $\mathcal{X} \to \mathcal{Y}$, with corresponding latent representations h and v. We utilize an RNN model to estimate the latent dependency $\mathbf{P}(v|h)$ as displayed in Figure 13. The training process involves three simultaneous optimizations per iteration:

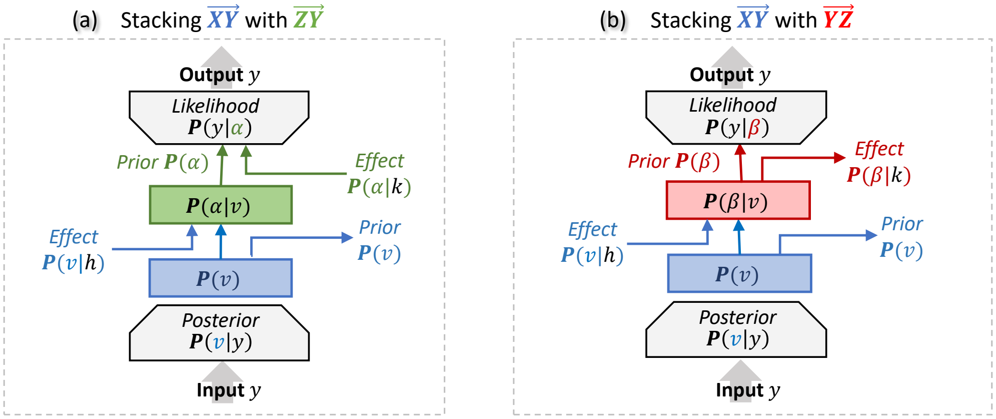
1. Optimizing encoder $\mathbf{P}(h|x)$, RNN model $\mathbf{P}(v|h)$, and decoder $\mathbf{P}(y|v)$ to reconstruct the effect $x \to y$.
2. Fine-tuning encoder $\mathbf{P}(v|y)$ and decoder $\mathbf{P}(y|v)$ to accurately represent y.
3. Fine-tuning encoder $\mathbf{P}(h|x)$ and decoder $\mathbf{P}(x|h)$ to accurately represent x.

Throughout the learning, h and v values are iteratively refined to minimize their latent space distance, and the RNN functions act as a bridge to traverse this distance, thereby estimating the causal effect $x \to y$.

Figure 14 presents two stacking scenarios for $\mathcal{Y}$ in the three-variable causal system comprising $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, based on different causal direction settings. For the established latent edge $\overrightarrow{XY}$, the left-side architecture completes the $X \to Y \leftarrow Z$ relationship, while the right-side caters to $X \to Y \to Z$. Stacking is achieved by adding an extra representation layer, thereby forming a hierarchical structure, enabling diverse input-output combinations (denoted as $\mapsto$). For example, in the left setup, $\mathbf{P}(v|h) \mapsto \mathbf{P}(\alpha)$ signifies the $X \to Y$ relationship, while $\mathbf{P}(\alpha|k)$ implies $Z \to Y$. Conversely, the right setup has $\mathbf{P}(v) \mapsto P(\beta|k)$ representing $Y \to Z$ with Y as input and $\mathbf{P}(v|h) \mapsto P(\beta|k)$ denoting the $X \to Y \to Z$ relationship.

---

[1] https://github.com/kflijia/bijective_crossing_functions/blob/main/code_bicross_extracter.py


(a) Stacking $\overrightarrow{XY}$ with $\overrightarrow{ZY}$

(b) Stacking $\overrightarrow{XY}$ with $\overrightarrow{YZ}$

Output y
Likelihood $\mathbf{P}(y|\alpha)$
Prior $\mathbf{P}(\alpha)$
Effect $\mathbf{P}(\alpha|k)$
$\mathbf{P}(\alpha|v)$
Effect $\mathbf{P}(v|h)$
Prior $\mathbf{P}(v)$
$\mathbf{P}(v)$
Posterior $\mathbf{P}(v|y)$
Input y
Output y
Likelihood $\mathbf{P}(y|\beta)$
Prior $\mathbf{P}(\beta)$
Effect $\mathbf{P}(\beta|k)$
$\mathbf{P}(\beta|v)$
Effect $\mathbf{P}(v|h)$
Prior $\mathbf{P}(v)$
$\mathbf{P}(v)$
Posterior $\mathbf{P}(v|y)$
Input y

Figure 14: Architecutres of the Relation-Defined Hierarchical Representation.

Causal effects of edges can be sequentially stacked based on known causal Directed DAGs by leveraging domain knowledge. Additionally, this method can facilitate causal structure discovery in the latent space, identifying potential edges among the initial representations of the variables.

6.3 Causal Discovery in Latent Space

Algorithm 1 details the heuristic process of discovering causal edges among the initially established variable representations. It employs the Kullback-Leibler Divergence (KLD) as a metric to assess causal relationship strength. Specifically, KLD measures the similarity between the RNN's output, $\mathbf{P}(v|h)$, and the prior $\mathbf{P}(v)$, as depicted in Figure 13. A lower KLD signifies a stronger causal relationship, given its closer alignment with the ground truth. Though Mean Squared Error (MSE) is a conventional evaluation metric, considering it may be influenced by data variances Reisach et al. (2021); Kaiser & Sipos (2021), we primarily employ KLD as the criterion and use MSE as a supplementary metric. For clarity, in the graphical context, for edge $A \to B$, we refer to variables A and B as the *cause node* and *result node*, respectively.

Algorithm 1: Latent Space Causal Discovery

```
Result: ordered edges set $\mathbf{E} = \{e_1, \ldots, e_n\}$
$\mathbf{E} = \{\}$ ; $N_R = \{n_0 \mid n_0 \in N, Parent(n_0) = \varnothing\}$ ;
while $N_R \subset N$ do
    $\Delta = \{\}$ ;
    for $n \in N$ do
        for $p \in Parent(n)$ do
            if $n \notin N_R$ and $p \in N_R$ then
                $e = (p, n)$;  $\beta = \{\}$;
                for $r \in N_R$ do
                    if $r \in Parent(n)$ and $r \neq p$ then
                        $\beta = \beta \cup r$
                    end
                end
                $\delta_e = K(\beta \cup p, n) - K(\beta, n)$;
                $\Delta = \Delta \cup \delta_e$;
            end
        end
    end
    $\sigma = argmin_e(\delta_e \mid \delta_e \in \Delta)$;
    $\mathbf{E} = \mathbf{E} \cup \sigma$;  $N_R = N_R \cup n_\sigma$;
end
```

$G = (N, E)$	graph G consists of N and E
N	the set of nodes
E	the set of edges
N_R	the set of reachable nodes
$\mathbf{E}$	the list of discovered edges
$K(\beta, n)$	KLD metric of effect $\beta \to n$
β	the cause nodes
n	the result node
δ_e	KLD Gain of candidate edge e
$\Delta = \{\delta_e\}$	the set $\{\delta_e\}$ for e
n,p,r	notations of nodes
e,σ	notations of edges

Figure 15 presents an exemplification of the causal structure discovery process within the latent space. Across four steps, two edges (e_1 and e_3) are successively selected. The selection of e_1 establishes node B as the starting point for e_3. In step 3, the causal effect of e_2 from A to C is deselected from the potential edges and re-evaluated. This is due to the introduction of edge e_3 to C, modifying C's existing causal conditions. As the procedure unfolds, the ultimately discovered causal structure is represented by the final DAG.

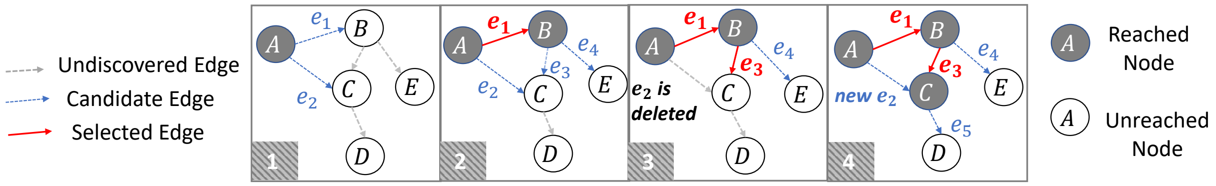


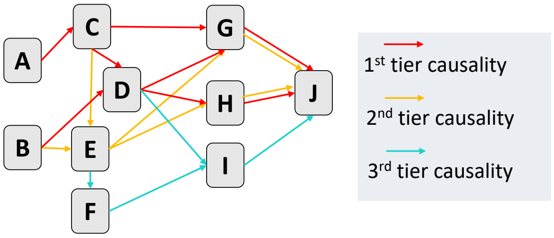
Figure 15: Example of the Latent Space Causal Discovery.

7 Feasibility Validation Experiments

The experiments aim to validate the efficacy of the proposed *relation-defined representation* learning method in 1) creating high-dimensional feature representations using our autoencoder architecture, 2) constructing latent effects and stacking them for hierarchical representation, and 3) latent space causal structure discovery.

We employ a synthetic hydrology dataset for the experiments, a prevalent resource in hydrology. The task involves predicting streamflow based on observed environmental conditions like temperature and precipitation. By using relation-defined representation learning on this hydrology data, we aim to construct generalizable causal models across diverse watersheds. Despite similarities in hydrological schemes, differences in unmeasurable conditions such as economic developments and land use complicate direct model application. Current physical knowledge-based models, however, are often constrained by limited parameters, which restricts their flexibility in capturing the common knowledge within heterogeneous data.

To assess models' robustness and generalizability, Electronic Health Record (EHR) data would be an ideal choice, given their rich confounding relationships across multi-timelines. However, due to empirical restrictions, we lost access to EHRs during this study. To confirm the existence of inherent temporal bias, we direct readers to the previous work Li et al. (2020). A complete demo of experiments in this study is provided [2].



ID	Variable Name	Explanation
A	Environmental set I	Wind Speed, Humidity, Temperature
B	Environmental set II	Temperature, Solar Radiation, Precipitation
C	Evapotranspiration	Evaporation and transpiration
D	Snowpack	The winter frozen water in the ice form
E	Soil Water	Soil moisture in vadose zone
F	Aquifer	Groundwater storage
G	Surface Runoff	Flowing water over the land surface
H	Lateral	Vadose zone flow
I	Baseflow	Groundwater discharge
J	Streamflow	Sensors recorded outputs

Figure 16: DAG of structured hydrology data, with tiers of routines ordered by decreasing causal strengths.

7.1 Hydrology Dataset

Our experiments leverage the Soil and Water Assessment Tool (SWAT), a comprehensive hydrology data simulation system rooted in physical modules. We use SWAT's simulation of the Root River Headwater watershed in Southeast Minnesota, selecting 60 consecutive virtual years with daily updates. The performance evaluations predominantly focus on the accuracy of the autoencoder reconstructions.

In hydrology, deep learning methodologies are frequently employed Goodwell et al. (2020) to distill effective representations from time series data, with RNN models emerging as a favored choice for streamflow prediction Kratzert (2018). Figure 16 illustrates the causal DAG used by SWAT, with accompanying node descriptions. The nodes signify different hydrological routines, with the intensity of causality between them determined by their contribution to the output streamflow, denoted by various colors. The surface runoff routine (1st tier causality) plays a significant role in causing swift streamflow peaks, followed by the lateral flow routine (2nd tier causality). The baseflow dynamics (3rd tier causality) exert a more subtle influence. In our causal discovery experiments, we aim to uncover these ground truths from the observed data.

---

[2] https://github.com/kflijia/bijective_crossing_functions.git

Table 1: Statistics of Attributes and the Reconstruction Performances.

Variable	Dim	Mean	Std	Min	Max	Non-Zero Rate%	RMSE on Scaled	RMSE on Unscaled	BCE of Mask
A	5	1.8513	1.5496	-3.3557	7.6809	87.54	0.093	0.871	0.095
B	4	0.7687	1.1353	-3.3557	5.9710	64.52	0.076	0.678	1.132
C	2	1.0342	1.0025	0.0	6.2145	94.42	0.037	0.089	0.428
D	3	0.0458	0.2005	0.0	5.2434	11.40	0.015	0.679	0.445
E	2	3.1449	1.0000	0.0285	5.0916	100	0.058	3.343	0.643
F	4	0.3922	0.8962	0.0	8.6122	59.08	0.326	7.178	2.045
G	4	0.7180	1.1064	0.0	8.2551	47.87	0.045	0.81	1.327
H	4	0.7344	1.0193	0.0	7.6350	49.93	0.045	0.009	1.345
I	3	0.1432	0.6137	0.0	8.3880	21.66	0.035	0.009	1.672
J	1	0.0410	0.2000	0.0	7.8903	21.75	0.007	0.098	1.088

Table 2: Brief Summary of the Latent Causal Discovery Results.

Edge	A→C	B→D	C→D	C→G	D→G	G→J	D→H	H→J	B→E	E→G	E→H	C→E	E→F	F→I	I→J	D→I
KLD	7.63	8.51	10.14	11.60	27.87	5.29	25.19	15.93	37.07	39.13	39.88	46.58	53.68	45.64	17.41	75.57
Gain	7.63	8.51	1.135	11.60	2.454	5.29	25.19	0.209	37.07	-5.91	-3.29	2.677	53.68	45.64	0.028	3.384

7.2 Higher-Dimensional Representation Reconstruction Test

As depicted in Figure 16, there are 10 nodes needing initial representation establishment. Table 1 displays the statistics of their attributes (post-normalization), and reconstruction performance using the proposed high-dimensional feature representation autoencoders. Accuracy is evaluated via root mean square error (RMSE); lower RMSE equates to higher accuracy, on both scaled (i.e., normalized) and unscaled data.

The task poses challenges due to the exceedingly low dimensionality of the 10 variables, with a maximum of just 5 and the target node, J, possessing a single attribute. To counter this, we duplicate their columns to achieve a uniform 12-dimensionality, supplemented by the dummy variables of the 12 months, yielding a 24-dimensional autoencoder input. Through a double-wise feature extension, we generate a 576-dimensional amplified input, from which we extract a 16-dimensional representation via the encoder and decoder.

Significant challenges also arise from considerable meaningful-zero values. For example, node D (Snowpack in winter) includes numerous zeros in other seasons, closely related to node E (Soil Water) values. We address this by concurrently reconstructing non-zero indicator variables, named masks, within the autoencoder, evaluated using binary cross entropy (BCE).

Despite these challenges, the shallow RMSE values within $[0.01, 0.09]$ suggest success, barring node F (the Aquifer). Considering that research into the physical schemes under the aquifer system is still in its infancy, it is plausible that in this synthetic dataset, node F is more representative of random noise than other nodes.

7.3 Latent Causal Effects Learning Test

Table 3 shows the results of the latent effect learning, organized by each result node. For convenience, the pairwise relationship performances are referred to as "pair-effect", and the hierarchical multi-level performances as "stacking-effect". To facilitate comparison, the baseline performances from the initial variable representation (Table 1) are also included. During latent effect estimation, each result node fulfills two roles: preserving an accurate self-representation (optimization 2), and reconstructing the effect (optimization 1). These dual roles are respectively depicted in the middle and right-hand side of Table 1.

The KLD metrics in Table 3 indicate the strength of learned causality, with a lower value signifying a stronger causal relationship. For instance, node J's minimal KLD values suggest a significant causal effect from nodes G (Surface Runoff), H (Lateral), and I (Baseflow). In contrast, the high KLD values imply that predicting variable I using D and F is challenging.

For nodes D, E, and J, the stacking-effect causal strengths hover at a middle range compared to their pair-effects, suggesting a potential associative uninformative among their cause nodes. In contrast, for nodes G and H, lower stacking-effect KLDs indicate effective capture of associations by hierarchical representations. The KLD metric also unveils the most contributive cause node to the causal effect. For instance, the $C \to G$ strength being closer to $CDE \to G$ indicates C as the primary source of this causal effect.

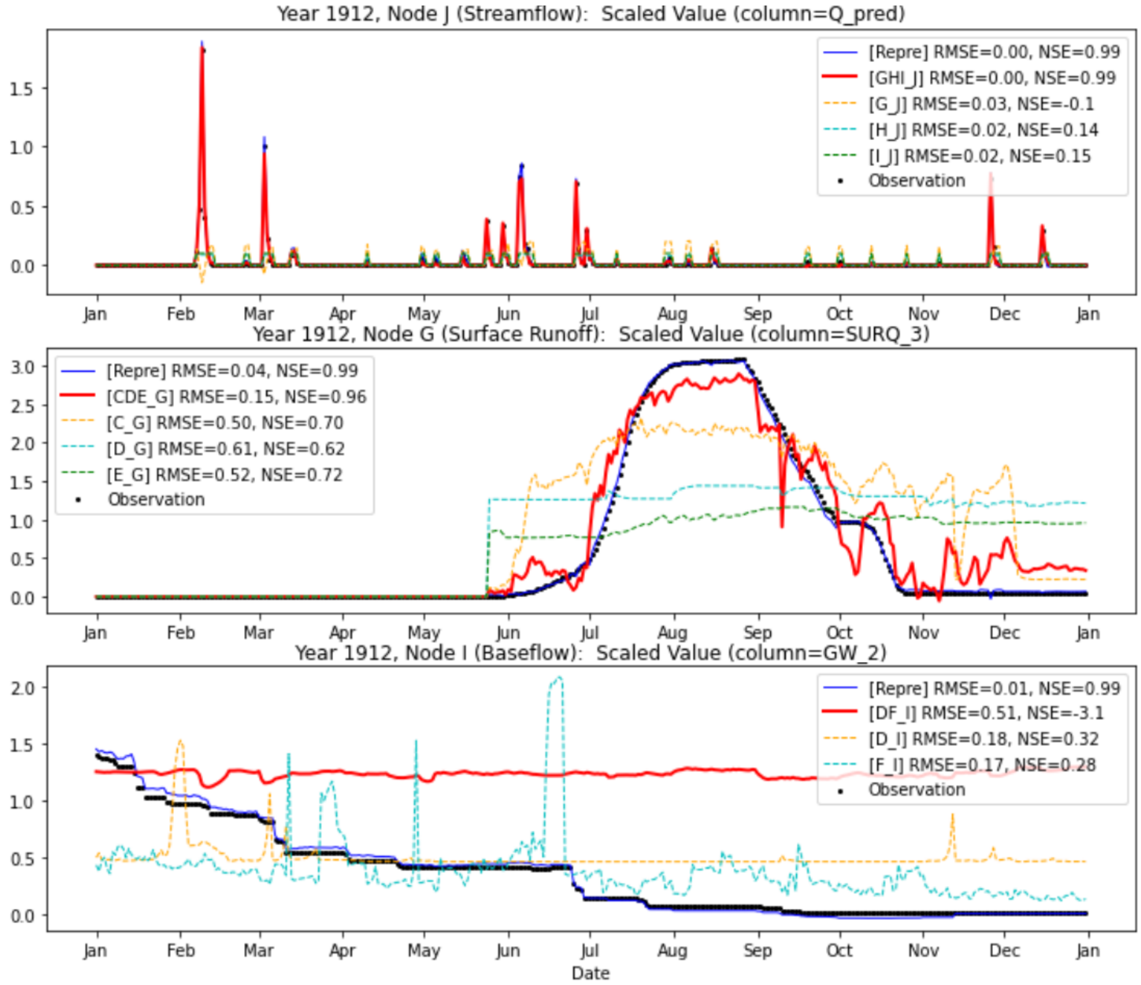


Figure 17: Time series simulation examples for the reconstruction performances' comparison.

Figure 17 showcases time series simulations of nodes J, G, and I, in the same synthetic year, to provide a straightforward overview of the hierarchical representation performances. Here, blue lines represent reconstructed data, black dots represent the ground truth, and red lines are hierarchical representations. Except for RMSE, we also employ the Nash–Sutcliffe model efficiency coefficient (NSE) for accuracy evaluation, which ranges from -∞ to 1.

The reconstructions closely mirror the ground truth, and as anticipated, the stacking-effect outperforms the pair-effect in Figure 17. Although node J has the best prediction, node I proves challenging. For node G, which is predicted from causes CDE, C offers the most potent causality.

One might observe via the demo that our experiments do not show smooth information flows along successive long causal chains. Given that RNNs are designed primarily for capturing the dynamics of causes rather than the effects, relying on them to spontaneously organize the effects' dynamical representations might prove unreliable. It underscores a significant opportunity for enhancing effectiveness by improving the architecture.

7.4 Latent Space Causal Discovery Test

Table 6 shows the order of discovered edges, with the KLD values after each edge's inclusion, and respective KLD gains. Cells follow the color-coding scheme from Figure 16, representing different tiers of causal routines.

For a detailed look at the causal discovery process, see 4, which presents sorted detection rounds. For comparison, we conducted a 10-fold cross-validation using the conventional FGES method; results can be found in Appendix A Table 5. The proposed method markedly outperforms the traditional FGES approach.

Table 3: Performances of Latent Causal Effect Learning via Reconstructions.

Result Node	Variable Representation (Initial)			Cause Node	Variable Representation (in Effect Learning)			Latent Causal Effect Reconstruction			
	RMSE		BCE		RMSE		BCE	RMSE		BCE	KLD
	on Scaled Values	on Unscaled Values	Mask		on Scaled Values	on Unscaled Values	Mask	on Scaled Values	on Unscaled Values	Mask	(in latent space)
C	0.037	0.089	0.428	A	0.0295	0.0616	0.4278	0.1747	0.3334	0.4278	7.6353
D	0.015	0.679	0.445	BC	0.0350	1.0179	0.1355	0.0509	1.7059	0.1285	9.6502
				B	0.0341	1.0361	0.1693	0.0516	1.7737	0.1925	8.5147
				C	0.0331	0.9818	0.3404	0.0512	1.7265	0.3667	10.149
E	0.058	3.343	0.643	BC	0.4612	26.605	0.6427	0.7827	45.149	0.6427	39.750
				B	0.6428	37.076	0.6427	0.8209	47.353	0.6427	37.072
				C	0.5212	30.065	1.2854	0.7939	45.791	1.2854	46.587
F	0.326	7.178	2.045	E	0.4334	8.3807	3.0895	0.4509	5.9553	3.0895	53.680
G	0.045	0.81	1.327	CDE	0.0538	0.9598	0.0878	0.1719	3.5736	0.1340	8.1360
				C	0.1057	1.4219	0.1078	0.2996	4.6278	0.1362	11.601
				D	0.1773	3.6083	0.1842	0.4112	8.0841	0.2228	27.879
				E	0.1949	4.7124	0.1482	0.5564	10.852	0.1877	39.133
H	0.045	0.009	1.345	DE	0.0889	0.0099	2.5980	0.3564	0.0096	2.5980	21.905
				D	0.0878	0.0104	0.0911	0.4301	0.0095	0.0911	25.198
				E	0.1162	0.0105	0.1482	0.5168	0.0097	3.8514	39.886
I	0.035	0.009	1.672	DF	0.0600	0.0103	3.4493	0.1158	0.0099	3.4493	49.033
				D	0.1212	0.0108	3.0048	0.2073	0.0108	3.0048	75.577
				F	0.0540	0.0102	3.4493	0.0948	0.0098	3.4493	45.648
J	0.007	0.098	1.088	GHI	0.0052	0.0742	0.2593	0.0090	0.1269	0.2937	5.5300
				G	0.0077	0.1085	0.4009	0.0099	0.1390	0.4375	5.2924
				H	0.0159	0.2239	0.4584	0.0393	0.5520	0.4938	15.930
				I	0.0308	0.4328	0.3818	0.0397	0.5564	0.3954	17.410

8 Conclusions

Prompted by the challenges of AI misalignment, this research delves into the inherent constraints of the predominant *Observation-Oriented* modeling paradigm. Accordingly, we introduce a new *Relation-Oriented* paradigm, complemented by its practical methodology, the *relation-defined representation* learning, and experimentally validate the efficacy of this approach.

Our perspective offers a new lens - viewing relationship learning through a dimensionality framework, where the relationships in our knowledge can be seen as distributions spanning various dimensions. It brings fresh insights into causal inference theories, presenting them in a more intuitively accessible manner.

The prevailing *Observation-Oriented* paradigm necessitates the identification of modeling objects prior to defining relations. This confines models to the *observational space* and limits their access to temporal features. By depending on a singular, absolute timeline, the paradigm overlooks the multi-dimensional nature of the *temporal space*, thereby compromising model robustness and generalizability, which is a major factor in the AI misalignment issue.

Human cognition, in essence, prioritizes relations, leading to our vast relation-centric knowledge systems. We can identify dynamical temporal features by navigating the intricate network of relations in the *hyper-dimensional space*. This insight inspired the *Relation-Oriented* paradigm.

While implementing the *relation-defined representation* learning method, we encountered formidable technical challenges, such as crafting an invertible autoencoder for higher-dimensional representation. Nevertheless, thorough experiments have affirmed the feasibility of our proposed method. AI alignment is never a question with a simple answer but calls for interdisciplinary efforts Christian (2020). Through this work, we aim to contribute to the development of more genuine AI and provide a foundation for future advancements.

References

Howard Alkon, Daniel L & Rasmussen. A spatial-temporal model of cell activation. *Science*, 239(4843): 998–1005, 1988.

Gennady & Gatalsky Peter Andrienko, Natalia & Andrienko. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Brian Christian. *The alignment problem: Machine learning and human values.* 2020.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Laurent Dinh, Jascha Sohl, and Samy Bengio. Density estimation using real nvp. *arXiv:1605.08803*, 2016.

Felix Elwert. Graphical causal models. *Handbook of causal analysis for social research*, pp. 245–273, 2013.

Dan Geiger and Judea Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The annals of statistics*, 21(4):2001–2021, 1993.

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

Allison E Goodwell, Peishi Jiang, Benjamin L Ruddell, and Praveen Kumar. Debates—does information theory provide a new paradigm for earth science? causality, interaction, and feedback. *Water Resources Research*, 56(2):e2019WR024940, 2020.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Saachi Jain, Adityanarayanan Radhakrishnan, and Caroline Uhler. A mechanism for producing aligned latent spaces with autoencoders. *arXiv preprint arXiv:2106.15456*, 2021.

Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery. *arXiv:2104.05441*, 2021.

Frederik et. al Kratzert. Rainfall–runoff modelling using lstm networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Jia Li, Xiaowei Jia, Haoyu Yang, Vipin Kumar, Michael Steinbach, and Gyorgy Simon. Teaching deep learning causal effects improves predictive performance. *arXiv preprint arXiv:2011.05466*, 2020.

Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427, 2020.

Jianzhu et. al Ma. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4):290–298, 2018.

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

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

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

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

David Pitt. Mental Representation. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv:1804.10253*, 2018.

Alexander G Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

Richard Scheines. An introduction to causal inference. 1997.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Monica G Turner. Spatial and temporal analysis of landscape patterns. *Landscape ecology*, 4:21–30, 1990.

Stefan Vuković, Matej & Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. 184:232–242, 2016.

Robert W Wood, Christopher J & Spekkens. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of bell-inequality violations require fine-tuning. *New Journal of Physics*, 17 (3):033002, 2015.

Haoyan Xu, Yida Huang, Ziheng Duan, Jie Feng, and Pengyu Song. Multivariate time series forecasting based on causal inference with transfer entropy and graph neural network. *arXiv:2005.01185*, 2020.

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.

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

A Appendix: Complete Experimental Results of Causal Discovery

Table 4: The Complete Results of Heuristic Causal Discovery in latent space. Each row stands for a round of detection, with '#" identifying the round number, and all candidate edges are listed with their KLD gains as below. 1) Green cells: the newly detected edges. 2) Red cells: the selected edge. 3) Blue cells: the trimmed edges accordingly.

# 1	**A→ C**	A→ D	A→ E	A→ F	**B→ C**	B→ D	B→ E	B→ F								
	7.6354	19.7407	60.1876	119.7730	**8.4753**	8.5147	65.9335	132.7717								
# 2	A→ D	A→ E	A→ F	B→ D	B→ E	B→ F	C→ D	C→ E	C→ F	C→ G	C→ H	C→ I				
	19.7407	60.1876	119.7730	**8.5147**	65.9335	132.7717	10.1490	46.5876	111.2978	11.6012	39.2361	95.1564				
# 3	**A→ D**	A→ E	A→ F	B→ E	B→ F	C→ D	C→ E	C→ F	C→ G	C→ H	C→ I	D→ E	D→ F	D→ G	D→ H	D→ I
	9.7357	60.1876	119.7730	65.9335	132.7717	**1.1355**	46.5876	111.2978	11.6012	39.2361	95.1564	63.7348	123.3203	27.8798	25.1988	75.5775
# 4	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	**C→ G**	C→ H	C→ I	D→ E	D→ F	D→ G	D→ H	D→ I		
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	**11.6012**	39.2361	95.1564	63.7348	123.3203	27.8798	25.1988	75.5775		
# 5	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ H	C→ I	D→ E	D→ F	D→ G	D→ H	D→ I	G→ J		
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	39.2361	95.1564	63.7348	123.3203	**2.4540**	25.1988	75.5775	5.2924		
# 6	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ H	C→ I	D→ E	D→ F	D→ H	D→ I	G→ J			
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	39.2361	95.1564	63.7348	123.3203	25.1988	75.5775	**5.2924**			
# 7	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	**C→ H**	C→ I	D→ E	D→ F	D→ H	D→ I				
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	**39.2361**	95.1564	63.7348	123.3203	**25.1988**	75.5775				
# 8	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ I	D→ E	D→ F	D→ I	H→ J					
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	95.1564	63.7348	123.3203	75.5775	**0.2092**					
# 9	**A→ E**	A→ F	B→ E	B→ F	C→ E	C→ F	C→ I	D→ E	D→ F	D→ I						
	60.1876	119.7730	65.9335	132.7717	**46.5876**	111.2978	95.1564	63.7348	123.3203	75.5775						
# 10	A→ F	B→ E	B→ F	C→ F	C→ I	**D→ E**	D→ F	D→ I	E→ F	E→ G	E→ H	E→ I				
	119.7730	**-6.8372**	132.7717	111.2978	95.1564	**17.0407**	123.3203	75.5775	53.6806	-5.9191	-3.2931	110.2558				
# 11	A→ F	B→ F	C→ F	C→ I	D→ F	D→ I	E→ F	**E→ G**	E→ H	E→ I						
	119.7730	132.7717	111.2978	95.1564	123.3203	75.5775	53.6806	**-5.9191**	-3.2931	110.2558						
# 12	A→ F	B→ F	C→ F	C→ I	D→ F	D→ I	E→ F	**E→ H**	E→ I							
	119.7730	132.7717	111.2978	95.1564	123.3203	75.5775	53.6806	**-3.2931**	110.2558							
# 13	**A→ F**	**B→ F**	**C→ F**	C→ I	**D→ F**	D→ I	**E→ F**	E→ I								
	119.7730	**132.7717**	**111.2978**	95.1564	**123.3203**	75.5775	**53.6806**	110.2558								
# 14	C→ I	D→ I	**E→ I**	**F→ I**												
	95.1564	75.5775	**110.2558**	**45.6490**												
# 15	C→ I	D→ I	**I→ J**													
	15.0222	3.3845	**0.0284**													
# 16	**C→ I**	**D→ I**														
	15.0222	**3.3845**														

Table 5: Average performance of 10-Fold FGES (Fast Greedy Equivalence Search) causal discovery, with the prior knowledge that each node can only cause the other nodes with the same or greater depth with it. An edge means connecting two attributes from two different nodes, respectively. Thus, the number of possible edges between two nodes is the multiplication of the numbers of their attributes, i.e., the lengths of their data vectors.
(All experiments are performed with 6 different Independent-Test kernels, including chi-square-test, d-sep-test, prob-test, disc-bic-test, fisher-z-test, mvplr-test. But their results turn out to be identical.)

Cause Node	A	B		C			D			E			F	G		H	I
True Causation	A→ C	B→ D	B→ E	C→ D	C→ E	C→ G	D→ G	D→ H	D→ I	E→ F	E→ G	E→ H	F→ I	G→ J		H→ J	I→ J
Number of Edges	16	24	16	6	4	8	12	12	9	8	8	8	12	4		4	3
Probability of Missing	0.038889	0.125	0.125	0.062	0.06875	0.039286	0.069048	0.2	0.142857	0.3	0.003571	0.2	0.142857	0.0		0.072727	0.030303
Wrong Causation Times of Wrongly Discovered				C→ F				D→ E	D→ F				F→ G	G→ H	G→ I	H→ I	
				5.6				1.2	0.8				5.0	8.2	3.0	2.8	

Table 6: Brief Results of the Heuristic Causal Discovery in latent space, identical with Table 3 in the paper body, for better comparison to the traditional FGES methods results on this page.
The edges are arranged in detected order (from left to right) and their measured causal strengths in each step are shown below correspondingly.
Causal strength is measured by KLD values (less is stronger). Each round of detection is pursuing the least KLD gain globally. All evaluations are in 4-Fold validation average values. Different colors represent the ground truth causality strength tiers (referred to the Figure 10 in the paper body).

Causation	A→ C	B→ D	C→ D	C→ G	D→ G	G→ J	D→ H	H→ J	C→ E	B→ E	E→ G	E→ H	E→ F	F→ I	I→ J	D→ I
KLD	7.63	8.51	10.14	11.60	27.87	5.29	25.19	15.93	46.58	65.93	39.13	39.88	53.68	45.64	17.41	75.57
Gain	7.63	8.51	1.135	11.60	2.454	5.29	25.19	0.209	46.58	-6.84	-5.91	-3.29	53.68	45.64	0.028	3.384