Relation-Oriented: Toward Causal Knowledge-Aligned AGI

Anonymous authors
Paper under double-blind review

Abstract

The prevalent *Observation-Oriented* modeling paradigm in machine learning, including AI, inherently views "time" as a *singular*, *linear* timeline, rather than as computational dimensions. Specifically, it requires identifying observational variables before modeling relations, limiting access to dynamical temporal features, and overlooking the multi-dimensional temporal feature space. These limitations introduce inherent bias, affecting the robustness and generalizability of structural causal AI models and contributing to AI Alignment issues.

This study examines these limitations uniquely through a *dimensionality* lens and presents a new *Relation-Oriented* paradigm. Inspired by the relation-centric nature of human cognition, this paradigm aims to enable interpretable Artificial General Intelligence (AGI) development grounded in human knowledge. As its methodological counterpart, the proposed *relation-defined representation* learning is substantiated 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 temporal dimensional computing is equated to counting $\{t, t+1\}$ 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. Regardless of its form, the outcome Y is strictly observational with a specified timestamp, leaving all potentially significant dynamics of the Y object entirely managed by $f(\cdot)$. However, whether the selected function $f(\cdot)$ is *linear* or *nonlinear* only influences the dimensionality of $\mathbb{R}^d$, where $X \in \mathbb{R}^d$, while the time evolution from t to $t+m$ for the Y object remains invariably ***linear***.

Such a linearity on the temporal dimension may be sufficient in the past, but not in the present, given the current technological advancements in data collection and Artificial Intelligence (AI) methods. Exploring nonlinear distributions on temporal dimension(s) is gradually becoming essential, calling for a new modeling paradigm Scholkopf (2021), which does not rest on the conventional i.i.d.-assumed observations, but can treat relative timelines, i.e., potentially multiple t-axes, as distinct computational dimensions.

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 validations (Chapter II: Sections 5-7). Particularly, the linear absolute timeline t that we conventionally use inherently fails to capture dynamics of causal effects within the multi-dimensional temporal feature space (see subsection 1.3). This limitation leads to biases, resulting in AI models misaligned with our cognitive understanding Christian (2020) and challenging to generalize Scholkopf (2021).

In this paper, we interpret the "relationship" modeling through a novel *dimensionality* framework (in Figure 1), offering a unique perspective. The remainder of this section aims to lay the groundwork. In Chapter I, we will inspect causal learnings with respect to the temporal dimensional distributions, highlighting the key role of relations. 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

AI has displayed capabilities surpassing humans in observational learning tasks, such as generating images, Go gaming (in a single absolute timeline), etc, but may appear "unintelligent" in comprehending some knowledge 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 treating them as arbitrary assortments of finger-like items.

Moreover, when it comes to time evolution, causal reasoning presents a substantial challenge for AI. Despite valuable contributions from traditional causal learning methods Wood (2015); Vuković (2022); Ombadi (2020), and the rise of neural network applications tackling large-scale causal questions Luo et al. (2020), limitations in model generalizability persist Scholkopf (2021). Accordingly, our causal model applications are often context-specific, and AI's nonlinear learning capability remains constrained on the temporal dimension.

The questions "How to utilize AI in causality" and "How to simulate reasonable hands" 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 some relations in knowledge are unseen to AI?", which is increasingly critical to address for today Christian (2020).

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, a quadrotor's movement in one hour, 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 (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 a special connection, but that does not equate to discerning the father-son relation between them.

Put simply, the existence of *unobservables* makes our relationship modeling 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.

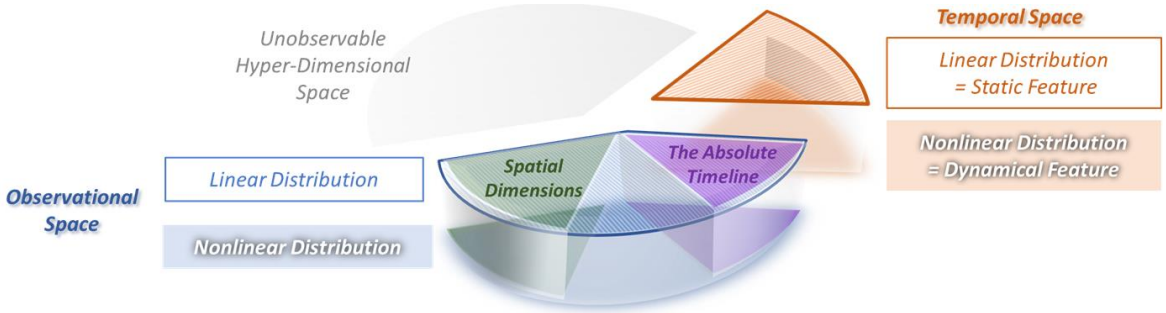


Figure 1: The knowledge space split by dimensionality types: *Observational*, *Temporal*, *Hyper-Dimension*.

Hence, from a dimensionality perspective, a relationship in modeling is interpretable as a joint distribution across multiple dimensions, with modeling objects existing on observational-temporal dimensions, and their relation, the modeling target, manifesting as an unseen distribution in a ***hyper-dimension***. Figure 1 categorizes our knowledge-storing cognitive space into three sections accordingly, with the *hyper-dimensional* space representing the aggregate of all unobservable relations in our knowledge. Current narrow AI, limited by the *Observation-Oriented* principle, can overlook crucial temporal features and hinder its ability to autonomously learn causal relations within this space, potentially impeding progress towards realizing AGI.

1.3 Observational and Temporal Spaces

Presently, most machine learning models primarily work within the observational space, maybe adhering to a single absolute timeline. For example, Convolutional Neural Networks (CNNs) recognize pixel associations; a quadrotor's movements are identifiable within three spatial dimensions; and Large Language Models (LLMs) operate in semantic space along a logical timeline representing word order. Applications similar to the latter two, in alignment with the Picard-Lindelof theorem, utilize absolute timestamp t to depict time evolution, commonly referred to as "spatial-temporal" analysis Alkon (1988); Turner (1990); Andrienko (2003).

Equating the "temporal dimension" to a singular t-timeline has become conventional Wes (2023), however, it seems to be a common misunderstanding. From a modeling viewpoint, timestamp values are not distinct from other observational attributes. In our comprehension of causal knowledge, ***multiple*** *relative* timelines can coexist within a complex causal structure, each representing different causal effects (see Section 4 for additional insights). Thus, we classify the absolute t-timeline as a dimension within the observational space, addressing the knowledge-aligned temporal distributions in a separate space as shown in Figure 1.

As initially discussed, solely relying on t-timeline allows for capturing only temporally linear relationships, yielding outcomes with ***static*** causal effects while excluding their ***dynamical*** aspects found in temporal nonlinearity (see subsection 3.3 for further discussions). Consider the example "rain leads to wet floors", both objects, the cause "rain" and the effect "wet floors", are status snapshotted at specific timestamps, thus viewed as *static*. In contrast, the effects such as "floors becoming progressively wetter" are considered *dynamical* due to the temporally significant patterns. More importantly, the presence of relative timelines in our knowledge can collectively form a *multi-dimensional* space, representing temporal features. Neglected nonlinear dynamics, combined with overlooked multiple timelines, can result in inherent bias (see subsection 4.1 for details) and compromise causal models' generalizability (see subsection 4.2). Presently, the advent of large AI-based model applications has exacerbated this inherent misalignment.

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

Like regular dimensions, joint distributions exist on hyper-dimensions but remain undetected due to their unobservable nature. Some unseen relations, although not targeted in the modeling process, can significantly influence our models. Consider joint relations $\langle \theta, \psi \rangle \in \mathbb{R}^h$ connecting observables X and Y, where $\mathbb{R}^h$ denotes the hyper-dimensional space. While the relationship model $Y = f(X; \theta)$ aims to model θ using given X and Y, the unseen ψ plays a crucial role - It can necessitate the model's *generalizability* across various scenarios, and also may contribute to a *misalignment* between the model and our knowledge comprehension.

For example, when examining how family income levels (i.e., X) influence (i.e., θ) grocery shopping frequencies (i.e., Y), underlying cultural factors (i.e., ψ) may play a role, such that established model $Y = f(X; \theta)$ proves practically useful only when conditioned on a specific country (i.e., a particular ψ value). Accordingly, there are two levels for the objective relation: the global-level θ without a ψ value, and the local-level θ with a specified ψ value. Model *generalizability* can be viewed as the ability to cross these levels, allowing the learned lower-level relationships to inform higher-level learnings Scholkopf (2021). Broadly, this also signifies the ability to *individualize* established models from higher to lower levels for different ψ values.

For clarity, we term such unseen relation ψ as *unobservable* ***hierarchy***, which, while external to the modeling target θ, remains vital for $f(\cdot)$. Such hierarchies are common in learning tasks and hold various meanings in different applications. For instance, they may signify levels of granularity (e.g., population vs. individual), as illustrated in subsection 2.2, or denote decision-making dependencies, as seen in subsection 2.1.

These hidden relations, while unobservable to AI, exist within our knowledge. Accordingly, their absence may lead to our current *Observation-Oriented* models *misaligned* with our anticipated understanding. While this absence can often be resolved when the modeling objects are purely observational (refer to subsection 2.1), it becomes a noticeable inherent deficiency under the current paradigm when critical temporal dynamics are involved (refer to subsections 2.2 and 2.3).

Chapter I: Deficiency of Current Observation-Oriented Paradigm

Human intelligence is inherently *relation-centric*, with relations serving as indices pointing to mental representations Pitt (2022), facilitating understanding of observational and temporal objects. This nature creates a fundamental misalignment with the prevailing modeling paradigm, which prioritizes observational objects as variables and outcomes. Depending on the application, this discrepancy can result in noticeable AI Alignment issues Christian (2020), ineffective use of causal knowledge in large AI-based models Luo et al. (2020), or challenges with model generalizability in traditional causal learning Scholkopf (2021).

This chapter explores the influences of hidden relations under the current paradigm (Section 2), re-evaluates causal learning in light of often-overlooked critical temporal features (Section 3), and highlights the multi-dimensionality of the temporal feature space, along with the inherent biases it introduces (Section 4).

2 Impact of Unobservable Hierarchy

Unobservable hierarchies indicate hidden relations, vital but separate from the modeling objective. For tasks solely involving observational learning, such information absence might be resolved by leveraging knowledge to enhance modeling (subsection 2.1). However, when it comes to temporally significant causal learning, these hierarchies may lead to a fundamental loss of dynamical features in the temporal dimension (subsection 2.2), presenting a substantial challenge to conventional causal inference methods (subsection 2.3).

2.1 On Solely Observational Learning

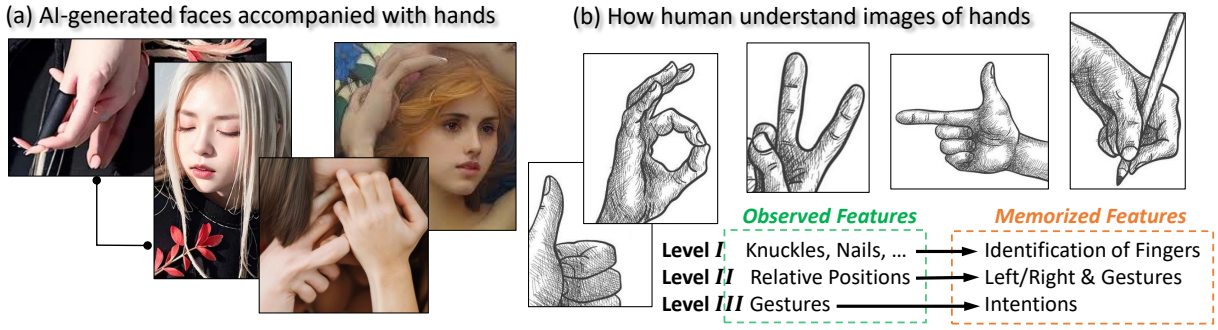


Figure 2: A comparison of AI-generated and human-sketched hand images. AI processes observable features simultaneously, thus treating hands as arbitrary mixtures of finger-like items. The process is hierarchical for humans, indexed through relations, where higher-level recognition relies on lower-level 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 rapidly make decisions according to different relations in our knowledge, hierarchically from lower to higher levels: **I** identifies fingers through knuckles and nails; **II** determines hand gestures through finger positions; **III** retrieves the gesture's meaning from memory. This intuitive hierarchy is unobservable, existing in our cognitions only. To AI, or similarly, to some extraterrestrial intelligent life without our knowledge, the hands in Figure 2(a) may seem reasonable.

In purely observational learning tasks, such hierarchies might not always pose significant issues. If features at different levels do not significantly overlap, AI may successfully "distinguish" them. For example, AI can generate convincing faces as the appearance of eyes strongly indicates facial angle, negating the need to recognize "eyes" from "faces", while various similar-looking hand gestures can create confusion. However, based on completely captured observational features at each level, AI may reveal hidden knowledge through methods like reinforcement learning Sutton & Barto (2018), guided by human feedback. For instance, human approval of five-fingered hands may lead AI to autonomously identify fingers.

2.2 On Temporally Significant Learning

Figure 3(a) illustrates an example from health informatics, depicting the causal effects of action $do(A)$ on B, with t indicating the elapsed days. For simplicity, assume the patient's hidden personal characteristics

linearly influence M_A's release, i.e., uniformly accelerate or decelerate its effective progress. The red and blue curves in (a) are individualized, shaped by two levels of *dynamical* features on the temporal dimension: 1) the standard population-level effect sequence with a length of 30, and 2) the individual-level progress speed. The modeling objective is to estimate level 1) dynamic, as the clinical effectiveness evaluation of M_A.

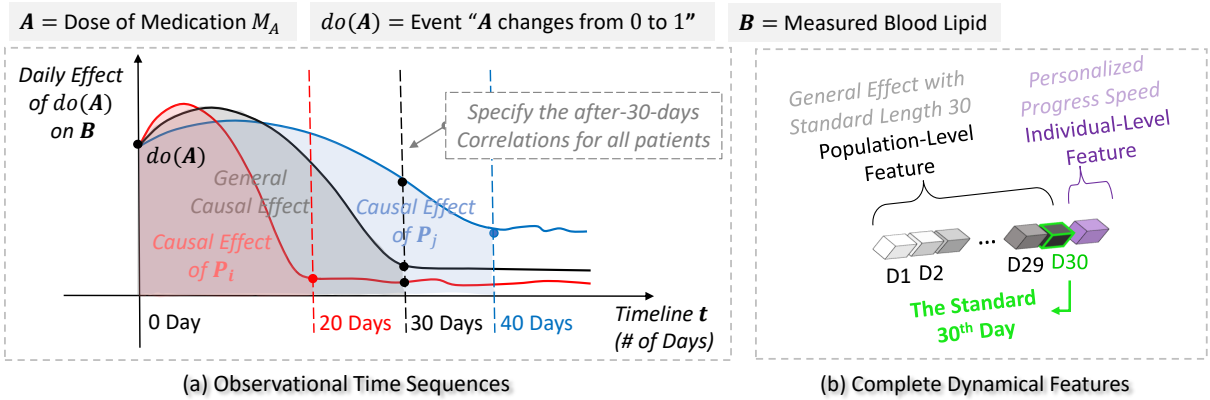


Figure 3: Medication M_A treats high blood lipid, with $do(A)$ denoting its initial use. It is given that the population-level effect takes about 30 days to fully release ($t = 30$ at the elbow), depicted by the black curve in (a). Patient P_i achieves this effect curve elbow in 20 days, while P_j takes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

Conventionally, the medical effect of M_A's is estimated by averaging the performances of all patients after 30 days, resulting in a correlation model $B_{t+30} = f(do(A_t))$. This model captures only the *static* feature B_{t+30}, the final step of level 1) dynamics, neglecting the preceding 29 steps. These steps are highlighted in Figure 3's complete feature vector, which is disentangled by hierarchical levels. Due to the lack of nonlinear modeling ability for effect objects, employing a sequence of length 30 to represent effects (e.g., in a Granger causality model) can, at best, capture level 1) sequence, but exclude further dynamical feature levels.

Dynamical causal effects, prevalent in applications like epidemic progression, economic fluctuations, and strategic decision-making, often occur within different granularity levels. Group-specific learning Fuller et al. (2007) methodologies are commonly used to address this, essentially equivalent to a manual specification of the value of ψ (see subsection 1.4 for reference of ψ).

2.3 The Elusive Hidden-Confounder

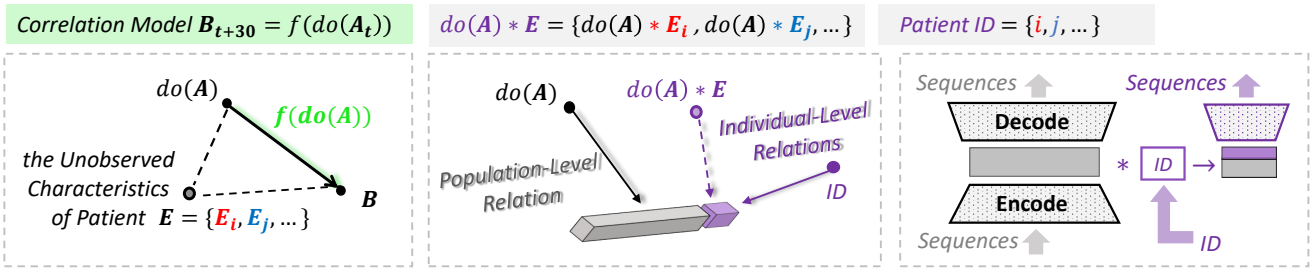


Figure 4: (a) Traditional causal inference DAG. (b) Hierarchical disentanglement of dynamics using relations as indices. (c) Autoencoder-based generalized and individualized reconstructions of the sequential data.

For patients P_i and P_j, the population-level estimated effect B_{t+30} is biased. To counter this individual-level bias and improve model interpretability, statistical causal inference incorporates the "hidden confounder" concept into Directed Acyclic Graphs (DAG), representing the concealed ψ as node E in Figure 4 (a). However, it does not necessitate collecting additional data for E, leading to an illogical assertion: "The model bias stems from unknown factors we don't intend to explore." This strategy compensates for the overlooked level 2) dynamics in causal models. Due to the observational learning essence, such *observable dynamical effect* features are explained through a *hidden observational* variable, E, associated with the *cause*.

As illustrated in Figure 4(b), the hidden composite cause $do(A) * E$ does not offer a modelable relationship - While introducing E may enhance human understanding, it unnecessarily improves the model. Conversely, the *Relation-Oriented* approach treats relations as indices, empowering AI to autonomously extract observational and temporal representations using any observed identifier, like patient ID. These representations, disentangled per the desired knowledge-derived hierarchy, enhance model generalizability.

3 Causality on Temporal Dimension

Causal learning acts as a portal to reach the temporal dimensional distributions beyond the observational space. However, the current causal learning paradigm fails to grasp temporally nonlinear relationships - It necessitates predefined timestamps for causal effects and focuses solely on their static, snapshotted observational features, thereby limiting their dynamism in the temporal dimension. Overlooking key dynamics can misalign the modeled relations with our anticipated knowledge (stored in the hyper-dimensional space).

The inherent nature of observational learning significantly complicates our modeling methods and the interpretation of inferences on causality. This section revisits causal learning tasks with a focus on temporal dimensional features, aiming to provide more intuitive insights into related theories and concepts. We start by redefining the concept of causal modeling (subsection 3.1), then assess the effectiveness of existing methods in learning dynamics (subsection 3.2), and finally highlight the inherent limitations of the prevailing *Observation-Oriented* causal model paradigm (subsection 3.3).

3.1 Redefine Causal Modeling

From a modeling perspective, specified timestamps are associated with observations rather than constituting a separate computational dimension. Consequently, in traditional causal inference, the temporal-evolving aspects that distinguish causality from correlation are not directly built into the current modeling framework. Instead, they are mainly evident in model interpretations, guiding potential improvements to the model. With this in mind, we identify causality in modeling by integrating temporal dimensional features.

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***.

In particular, causal modeling is crucial as it enables the answering of *counterfactual* questions Scholkopf (2021), such as "What effect would result if the cause were changed?" Simply put, causal models are valuable if they can effectively capture *temporal dimensional distributions*, thereby responding accurately to conditional queries regarding the temporal dimension.

The current causal modeling framework, which only captures linear features on a specified absolute timeline, may have sufficed in past decades but falls short of current needs. While AI-based models like RNNs facilitate nonlinear modeling, this capability depends on the accurate specification of time sequences (primarily as the cause only) and a correct timeline. Violating the former condition results in a failure to capture dynamics (see subsection 3.2), and violating the latter introduces inherent bias, fundamentally reducing model effectiveness (refer to Section 4).

The significance of model interpretation in traditional causal models is underscored by their emphasis on determining the "causal direction", i.e., the roles of cause and effect, which may not be crucial in the modeling context. Specifically, when selecting a model for the causal relationship $X \to Y$, one could use $Y = f(X; \theta)$ to predict the effect on Y, or inversely use $X = g(Y; \phi)$ to infer the cause X given Y. Both parameters, θ and ϕ, are obtained from the joint probability $\mathbf{P}(X, Y)$ without the need for imposing modeling constraints.

Empirically, concern for causal direction arises for a couple of reasons: firstly, there is a need for alignment with our intuitive understanding of temporal progression; secondly, the current paradigm exhibits an *imbalance* in capturing dynamics between the cause and the effect (see subsection 3.2 for details). For

instance, the hierarchical dynamics overlooked in Figure 4 can be fully captured by an inverse model of $do(A) = f(\{B_t, \ldots, B_{t+40}\})$ using RNNs, eliminating the need for a hidden confounder.

3.2 Learning Temporal Dynamics

Importantly, using a sequence as a variable does not necessarily entail the ability to capture dynamics. The difference between "a sequence of *static* variables" and "a *dynamical* variable" depends on the capability to capture the *nonlinearity* among them. The advent of RNNs Xu et al. (2020) addresses this dynamical learning concern by transforming data sequences into a latent feature space. In this space, temporal distributions are converted into representations that are indistinguishable from other observational representations, allowing the input timeline to function as a computational dimension.

Before the emergence of RNNs, autoregressive models Hyvärinen et al. (2010) usually appeared to embody the "sequential static learning" for the cause only. While both autoregressive models and RNNs utilize the format $y_{t+m} = f(\{x_t\})$ with $\{x_t\} = \{x_1, \ldots, x_t, x_{t+1}, \ldots, x_T\}$, the function $f(\cdot)$ conventionally selected for autoregressive models is often linear, treating $\{x_t\}$ as a static sequence, unlike the deep learning approach inherent to RNNs. With precise parameter settings (e.g., "30 days" in Figure 3(a)), at best, single-level dynamical features of the cause may be learned (e.g., population-level in Figure 3(b)).

Additionally, Granger causality Maziarz (2015), a methodology well-acknowledged in economics, utilizes another sequence for causal effects, expressed as $\{y_\tau\} = f(\{x_t\})$, with t and τ representing distinct timelines for cause and effect, respectively. This format aligns well with the causal modeling defined in Theorem 1, allowing for learning dynamics in both cause and effect if incorporating latent space representations like those in RNNs. Nonetheless, the fundamental challenge stems from ***identifiability difficulty*** Zhang (2012). Specifically, organizing sequential data according to a significant causal event (e.g., days of heavy rain) is feasible, but pinpointing the exact start of subsequent effects (e.g., the flood's start day caused by this rain) proves problematic. Furthermore, identifying appropriate timelines presents an additional challenge.

Remark 1. Identifiability difficulty is inevitable under the *Observation-Oriented* paradigm. In contrast, the proposed *Relation-Oriented* principle allows for autonomous effect identification in AI by indexing relations, mirroring the human understanding process.

Another method, do-calculus Pearl (2012); Huang (2012), sidesteps time sequence specifications, focusing instead on *identifiable* temporal events as modeling objects to perform an elementary calculus. Its *differential* nature, however, adds complexity, while we offer a simplified reinterpretation of its three core rules from an *integral* perspective. Let $do(x_t) = \{x_t, x_{t+1}\}$ indicate the occurrence of an instantaneous event $do(x)$ at time t, with the time step Δt sufficiently small to make this event's *interventional* effect identifiable as a function of the resultant distribution at $t+1$. Meanwhile, a separate *observational* effect is provoked by the static x_t. Then, for variable $X \in \mathbb{R}^d$, its dynamics $\mathcal{X}$ as the cause object can be expressed as follows:

Given $\mathcal{X} \to Y \mid \psi$, where $\mathcal{X} = \langle X, t \rangle \in \mathbb{R}^{d+1}$ with augmented t-dimension residing a T-length sequence,

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

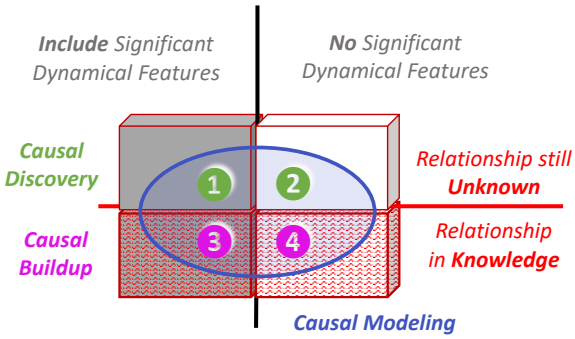
$$\text{The effect of } \mathcal{X} \text{ can be derived as } f(\mathcal{X}) = \int_0^T f_t\big(do(x_t) \cdot x_t\big) \; dt = \sum_{t=0}^{T-1} (y_{t+1} - y_t) = y_T - y_0$$

For the causal relationship $X \to Y$, given conditional knowledge ψ, the do-calculus rules tackle three specific scenarios, where conditional independence is maintained between the *observational* and the *interventional* effects, but bypassing more generalized cases (an identifiable $do(x_t) \cdot x_t$ pertains to Rule 2). Utilizing the $do(\cdot)$ format, we can also represent a dynamical effect as $\mathcal{Y} = \langle Y, \tau \rangle$. However, specification of the events for $\mathcal{Y}$ is still required under the current paradigm, while the proposed one is designed to construct $\mathcal{Y}$ autonomously.

3.3 Limitations of Current Causal Model Paradigm

In essence, traditional causal models employ an observational-based paradigm to learn relationships that are both observational and dynamical significant, thus often relying on fundamental hypotheses (such as causal *Sufficiency* and *Faithfulness* assumptions) or specific conditions derived from knowledge (as in do-calculus).

Figure 5 classifies causal modeling tasks into four scenarios. Modeling queries can be divided into Discovery and Buildup, depending on whether the objective relation θ is known; they can also be further categorized by the dynamic significance of the effect. For example, the causality "raining $\rightarrow$ wet floor" falls into area ④, while "raining $\rightarrow$ floor becoming wetter" is in area ③. They will be examined from two perspectives: the modeling critical objective Relation (i.e., θ), and the interpretationally essential causal Direction.



	Modeled Relation	Modeled Direction
1	Observational Only. Undiscovered Dynamics covered by *Faithfulness Assumption.*	Observation Determined. No Logical Meaning.
2	Observational Only. Aligned with Knowledge.	Observation Determined. Maybe Logically Meaningful.
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 model paradigm. The left rectangle cube represents all logical causal relationships, with the potentially modelable scope circled in blue.

(1) Modeled Relation

Causal inference has notably advanced by making particular dynamics observationally accessible and linearly modelable, (e.g., the independence explored in do-calculus). For overlooked dynamics in relationship buildup, if existing knowledge suggests a potential cause, introducing a hidden confounder can enhance comprehension. If not, they may be dismissed due to the assumed causal *Sufficiency*, potentially leading to later challenges.

Causal discovery mainly examines the observational space to uncover statistical dependencies. If the true causality of interest does not have significant dynamics, the discovered associations can be insightful. However, if such dynamics are present, especially along unknown hierarchy ψ, the potential undetected gap may be dismissed by the *Faithfulness* assumption, positing that observables can fully represent causal reality.

(2) Modeled Causal Direction

Consider causally related variables X and Y with potential directional models $Y = f(X;\theta)$ and $X = g(Y;\phi)$. In observational space, the discovered direction depends on the likelihoods of estimated $\hat{\theta}$ and $\hat{\phi}$. The preference is for $X \rightarrow Y$ if $\mathcal{L}(\hat{\theta}) > \mathcal{L}(\hat{\phi})$. Now, let $\mathcal{I}(\theta)$ be a simplified form of $\mathcal{I}_{X,Y}(\theta)$ (the Fisher information), representing the information in $\mathbf{P}(X,Y)$ about the relevant θ. If $p(\cdot)$ is the density function, then $\int_X p(x;\theta)dx$ is constant in this context. Thus, we have:

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

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

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

The likelihoods of the estimated $\hat{\theta}$ and $\hat{\psi}$ rely on the information $\mathcal{I}(\hat{\theta})$ and $\mathcal{I}(\hat{\psi})$. Consequently, the inferred directionality between X and Y reflects the extent to which their designated distributions appear in the data, with the predominant one deemed the "cause" - It assumes, by default, that observations capture the cause more thoroughly than the effect. While limited data collection techniques made it reasonable in the past, it is no longer safe to assume such observationally inferred directions to hold logical meaning for causality.

4 The Overlooked Temporal Space

To comprehend the issue of model generalizability in causal learning, consider the following analogy:

Imagine ants dwelling on a floor's two-dimensional plane. The ant scientists among them, aiming to predict risks, instinctively use the nearest tree's height as a reference for their two-dimensional models. During their modeling, they noticed an increase in disruptions at the tree's mid-level. This increase correlates to a higher likelihood of encountering children. However, without understanding humans as three-dimensional beings, the ants' interpretations are limited to observations at the tree's mid-level.

If these ants were to relocate to a tree of different heights, the mid-level would no longer correlate with risk, making their model ineffective. They may conclude that human behavior is too complex to model accurately. Similarly, when we specify a single, absolute timeline for all potential events, this timeline becomes our "tree".

Our logical understanding allows for the co-existence of multiple timelines Coulson (2009). With one absolute timeline, the others are relative, each embodying distinct causal effects. These effects, while possibly stemming from a single cause, can possess unique dynamical features, and interrelate with each other. Therefore, when identifying temporal events based on structural causal knowledge, it is crucial to consider all potential relative timelines comprehensively to avoid bias in the temporal dimension.

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 multiple timeline axes.

Inherently, causal inference adopts a *Relation-Oriented* perspective, acknowledging distinct effects and their respective dynamics. However, given that *Observation-Oriented* models typically rely on a single absolute timeline as a regular observational dimension, various de-confounding methods, including propensity score matching Benedetto (2018) and backdoor adjustment Pearl (2009), are utilized to eliminate inter-relations within the knowledge structure for more effective modeling.

In contrast, employing AI methods for causal questions is challenging due to their black-box nature and large scale, which makes manual inspections impractical. Crudely consolidating all dynamics into one timeline exacerbates the inherent biases, yielding uninterpretable results Luo et al. (2020). Think of traditional causal learning as manually building Schrödinger's box so the "cat" appears reasonable upon revelation; the proposed relation-indexing approach aims to have AI autonomously craft this box.

In this section, we'll first introduce the inherent bias resulting from neglecting underlying multi-timelines through an intuitive example (subsection 4.1). Next, we'll demonstrate how this bias affects the generalizability of structural causal models (subsection 4.2). Finally, we'll summarize the advancements and challenges encountered on our path toward causal knowledge-aligned AGI (subsection 4.3).

4.1 Scheme of the Inherent Bias

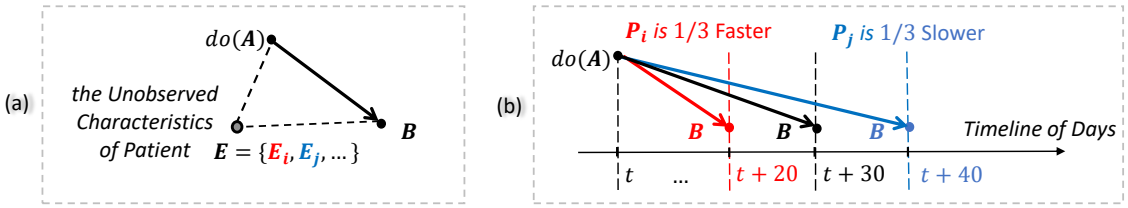


Figure 6: (a) Traditional Causal DAG introducing hidden E. (b) Enhanced DAG.

To more effectively address this issue, the causal DAG (directed acyclic graph) Pearl (2009) is enhanced in two ways: 1) by incorporating desired logical timelines as axes into the DAG space, and 2) by assuming causal effects are dynamically significant, with varying edge lengths indicating different timespans required to achieve identical effects. For instance, Figure 6(a) revisits the hidden-confounder example, which aims to interpret different individualized effects. Alternatively, the enhanced DAG shown in (b) provides a convenient representation of these effects.

Figure 7(a) expands on Figure 6, using A as shorthand for $do(A)$ for simplicity. In this figure, A has two distinct effects: it not only has a primary effect on B but also indirectly influences B through a side effect on another vital sign, C, represented by the edges $\overrightarrow{AC}$ and $\overrightarrow{CB}$. The confounding relationship among nodes $\{A, B, C\}$ forms a triangle across timelines T_X and T_Y According to the causal *Markov* condition, this shape should consistently hold for all individuals or populations. The process of individualization for patients P_i and P_j involves "stretching" this triangle along T_X at different ratios, conducting a homographic *linear transformation* within this DAG space, as depicted in Figure 7 (a).

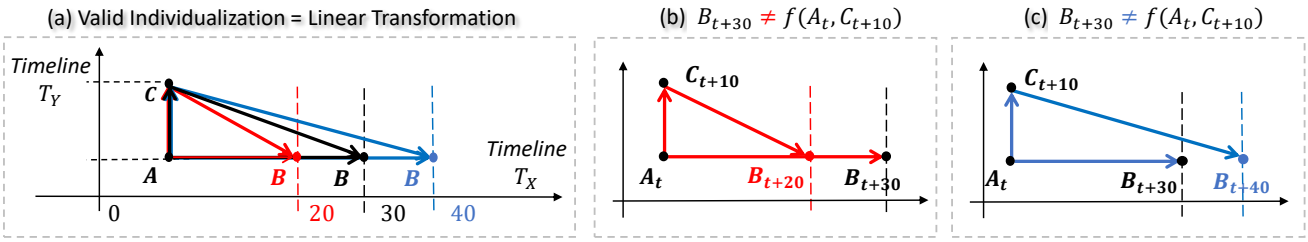


Figure 7: (a) A two-timeline DAG space, where a valid individualization presents a linear transformation. (b)(c) Violations of the Markov condition for the prevailing SCM with confounding dynamics across timelines.

Simplifying for discussion, assume individual-level diversity is confined to the primary effect $\overrightarrow{AB}$ only, with the timespan for $\overrightarrow{AC}$ being 10 days for all patients. Traditional Structural Causal Models (SCMs) typically assign effects a predefined timestamp, such as 30 in this instance, representing the population-level average. Consequently, the SCM function becomes $B_{t+30} = f(A_t, C_{t+10})$. However, as depicted in (b) and (c), setting the outcome of f to B_{t+30} for either P_i or P_j results in violations of the Markov condition. This discrepancy introduces individual-level biases in the SCM due to specifying a static timestamp for the effect.

Importantly, in this simplified example, violations may not pose significant issues for models able to address nonlinearity, due to the independence of dynamics on T_X and T_Y. The SCM can become $B_{t+30} = f_1(A_t) + f_2(C_{t+10})$, implying that cross-timeline confounding can be dissected into two single-timeline problems. However, invariably assuming independence or absence of confounding is impractical. In extensive causal AI applications, such inherent biases may emerge between any pair of distinct dynamical effects, accruing exponentially and substantially affecting model robustness, regardless of the selected modeling approach.

Theorem 3. The ***inherent*** *bias* may occur in SCM if it contains: 1) *Confounded* dynamical causal effects across *Multiple* logical timelines, and 2) Hierarchical diversity due to unobservable ψ.

Interestingly, most successful causal applications instinctively avoid either *confounding* or *multi-timeline*. Causal inference typically employs de-confounding, to mitigate inherent bias and other confounding biases resulting from unaddressed nonlinearity. Meanwhile, many AI accomplishments, including Large Language Models (LLMs), which operate in a semantic space, do not inherently deal with relative timelines, maintaining words consistently ordered along a singular timeline.

4.2 Inherent Impact on SCM Generalizability

Unobservable hierarchies imply different scenarios with identical core relationships. Traditional SCMs typically require specifying the event timestamps for modeling, according to a single absolute timeline. It affects not only model robustness but also hinders the generalizability of established SCMs across varied scenarios.

Consider the practical scenario depicted in Figure 8. Here, Δt and $\Delta \tau$ represent actual time spans. Yet, the crux is not on determining their exact values, but on realizing 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. At B', the ultimate aim of this process is to evaluate the total cumulative influence stemming from S.

Given the relationship $\overrightarrow{SB'} = \overrightarrow{SA'} + \overrightarrow{A'B'}$, specifying the $\overrightarrow{SB'}$ time span (= half of the $\overrightarrow{AB'}$ time span) inherently sets the $\Delta t : \Delta\tau$ ratio, defining the ASB' triangle's shape in the DAG space. While the estimated 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.

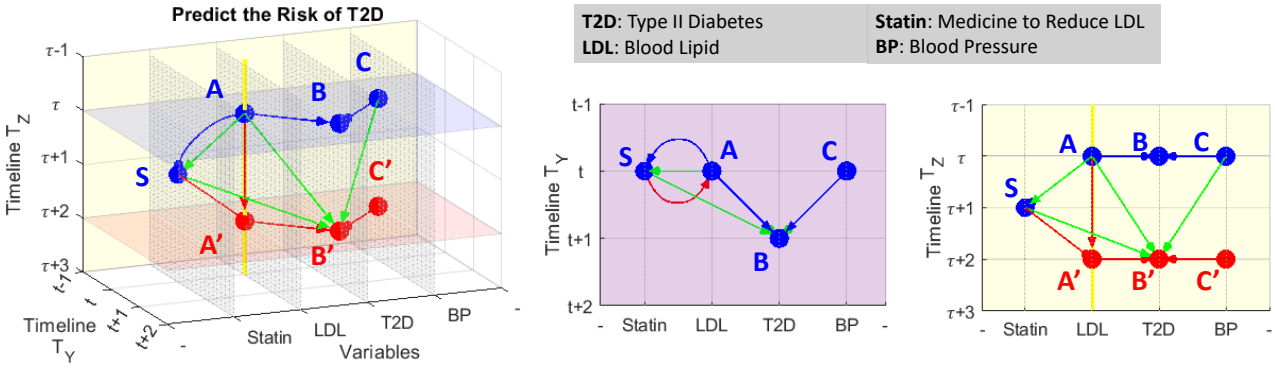


Figure 8: An exemplified 3D observational-temporal DAG space, with specified SCM, $B' = f(A, C, S)$, to evaluate Statin's medical effect on reducing the risk of T2D, including two logical timelines $\mathcal{T}_Y$ and $\mathcal{T}_Z$. On $\mathcal{T}_Y$, the step Δt from t to $(t+1)$ allows A and C to 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 to progress from A to A'.

4.3 Toward Causal Knowledge-Aligned AGI

In pursuit of causally interpretable AI, our modeling techniques expand beyond the purely observational to encompass temporal dimensions, as summarized in Figure 9. The present challenge is ensuring the generalizability of structural models across the temporal space. Acknowledging its multi-dimensional nature is critical to preventing inherent biases that render AI systems uninterpretable. Manually discerning underlying logical timelines for observables is impractical. Thus, it may have been time for us to consider the new paradigm.

Model	*Principle*	*Cause*	*Relation & 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-Indexing Methodology	*Given* $\boldsymbol{P}(\mathcal{X}, \mathcal{Y})$ & $\mathcal{X} \to \mathcal{Y}$	Observational-Temporal $\mathcal{X} = \langle X, t \rangle$	by Representation $\hat{\mathcal{Y}} = f(\mathcal{X}; \theta)$	Observational-Temporal $\hat{\mathcal{Y}} = \langle \hat{Y}, t \rangle$	Yes	Yes
Structural Causal Learning	*Given* $\boldsymbol{P}(X, Y)$ & $X \to Y$ $Y = f(X; \theta)$	Observational Sequence $\{X_t\}$	$X \to Y$ via Relation θ by Knowledge	Observational and Static Y_t	?	?
Graphical Causal Discovery	*Given* $\boldsymbol{P}(X, Y)$ *Find* $\mathcal{L}(Y\|X; \theta) > \mathcal{L}(X\|Y; \theta)$	Observational X	Observationally Associated X and Y	Observational Y	?	No
Common Cause Model	*Given* $\boldsymbol{P}(X, Y\|Z)$	Observational X	Related via Z	Observational Y	?	No
i.i.d Data Driven Model	*Given* $\boldsymbol{P}(X, Y)$	Observational X	None	Observational Y	No	No

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

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 causal models relying on conditional dependencies to construct Bayesian networks (BNs) often operate in observational space and neglect temporal aspects, reducing their causal relevance Scheines (1997). Causally significant models, such as Structural Equation Models (SEMs) and Functional Causal Models (FCMs) Glymour et al. (2019); Elwert (2013), can address counterfactual queries Scholkopf (2021), with respect to temporal distributions by leveraging prior knowledge, to construct causal DAGs accordingly.

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).

Schölkopf Scholkopf (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* SCMs.

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 the relation-indexing 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 efficacy through comprehensive experiments.

5 Hierarchical Disentanglement in Latent Space

Given an observational variable $X \in \mathbb{R}^d$, we denote its time sequence of length T as $\{x_t\} = \{x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T\}$. Our goal is to construct a latent feature space $\mathbb{R}^L$ for two specific 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 relations in knowledge. Consequently, the established system realizes the reusability of models at any hierarchical level by indexing through the corresponding relations.

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

5.1 Factorize Observational-Temporal Hierarchy

Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, 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 $\{\mathbb{R}^{L_1}, \ldots, \mathbb{R}^{L_i}, \ldots, \mathbb{R}^{L_n}\}$. Then, we obtain:

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

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 $\{\Theta_t\}_i = \{\Theta_{t_i} \in \mathbb{R}^d \mid t_i = 1, \ldots, T\}$. Consequently, we obtain a set of relative logical timelines $\{t_1, \ldots, t_i, \ldots, t_n\}$ which, in contrast to the absolute timeline t, are each uniquely determined by the relationship at their respective levels. However, in the *observable data space*, the i-th level observational-temporal feature, represented as the sum $\Theta_1 + \ldots + \Theta_i$, still maintains its timestamp attribute along t.

5.2 Factorize Hierarchy of Relationship

Given a set of n-level hierarchical representation functions for $\mathcal{X}$, denoted by $\mathcal{F}(\vartheta) = \{f_i(\theta_i) \mid i = 1, \ldots, n\}$, 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:

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

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 $\varphi = \{\varphi_1, \ldots, \varphi_n\}$. The term $\langle \vartheta_X, \varphi \rangle$ represents the pairwise augmentations between collections ϑ_X and φ.

6 Relation-Defined Representation Methodology

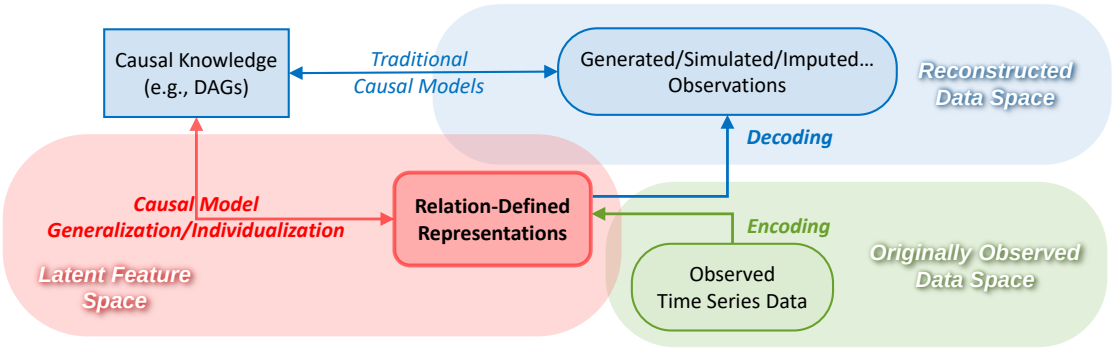


Figure 10: Framework of using *relation-defined representations* to enhance traditional models.

By extracting relation-defined representations, we facilitate the construction of causally interpretable AI systems in the latent feature space, adaptable to various scenarios (i.e., generalization or individualization). Figure 10 illustrates how this approach conceals AI's black-box nature within the latent space, managing human-indecipherable feature representations while simultaneously enhancing traditional models by refining observations, such as simulating counterfactual effects.

This section introduces a specialized autoencoder architecture crucial for implementing this approach, outlines the method for hierarchical representation disentanglement in constructing graphical models, and presents a causal discovery algorithm for the latent feature space.

6.1 Invertible Autoencoder for Higher-Dimensional Representation

Autoencoders are generally used for dimensionality reduction, often aligning all observational variables as data attributes for this purpose in structural modeling Wang et al. (2016). However, our objective diverges. We aim to model individual relationships to disentangle variables' representations and simultaneously "stack" them to form a DAG within the latent space, $\mathbb{R}^L$. This space must be large enough to accommodate potential

relationships in the form of $\vartheta_Y = \langle \vartheta_X, \varphi \rangle$. This poses a substantial technical challenge, as we need to achieve higher-dimensional representation extraction for individual variables.

Remark 2. Given a causal graph G with data matrix $\mathbf{X}$ column-augmented by all nodes' attributes, the latent space dimensionality L must satisfy $L \geq rank(\mathbf{X})$ to adequately represent G.

Remark 2 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 adequate but causally more significant latent space through better alignment Jain et al. (2021) (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.

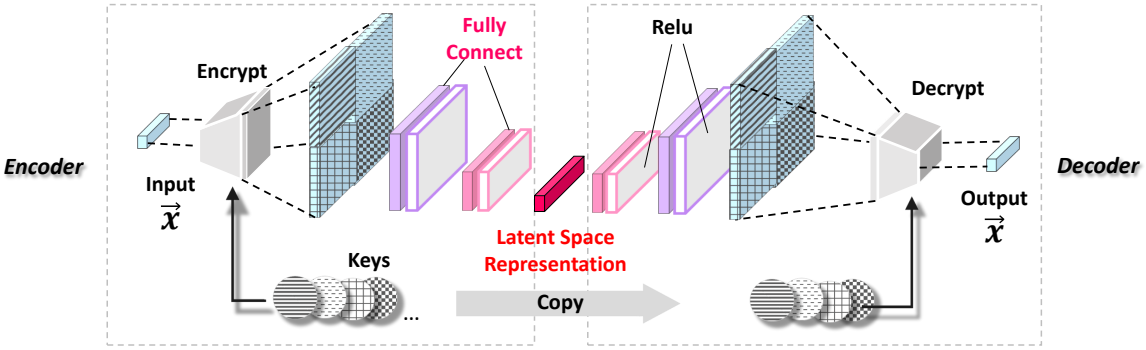


Figure 11: *Invertible* autoencoder architecture for extracting *higher-dimensional* representations.

Figure 11 depicts the proposed autoencoder architecture, featured by the symmetrical *Encrypt* and *Decrypt* layers. Encrypt amplifies the input vector $\vec{x}$ by extracting its higher-order associative features; conversely, Decrypt symmetrically reduces dimensionality 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* associative feature expansion, where each pair of *two* digits from $\vec{x}$ are encoded to form a new digit, by associating with a randomized constant *Key*, which is created by the encoder and mirrored by the decoder. A double-wise expansion on $\vec{x} \in \mathbb{R}^d$ generates a $(d-1)(d-1)$ length vector. By using multiple *Keys* and augmenting the derived vectors, $\vec{x}$ can have a significantly extended dimensionality.

The four differently patterned blue squares represent the vectors expanded by four distinct *Keys*, with the grid patterns indicating their "signatures". Each square visualizes a $(d-1)(d-1)$ length vector (not signifying a 2-dimensional vector). In a similar way, higher-order extensions, such as *triple-wise* ones across every three digits, can also be employed by appropriately adapting *Keys*.

Figure 12 depicts the encryption and decryption processes used to expand a digit pair (x_i, x_j), where $i \neq j \in 1, \ldots, d$. The encryption function $f_\theta(x_i, x_j) = x_j \otimes exp(s(x_i)) + t(x_i)$ is defined by two specific elementary functions, $s(\cdot)$ and $t(\cdot)$. The parameter θ, serving as a *Key*, consists of their respective weights, $\theta = (w_s, w_t)$.

Specifically, the encryption of (x_i, x_j) transforms x_j into a new digit y_j using x_i as a selected attribute. The decryption process symmetrically employs the inverse function f_θ^{-1}, defined as $(y_j - t(y_i)) \otimes exp(-s(y_i))$. Notably, this approach sidesteps the need to calculate s^{-1} or t^{-1}, allowing $s(\cdot)$ and $t(\cdot)$ to be flexibly specified as needed for nonlinear transformations. This design is inspired by the pioneering work of Dinh et al. (2016) on invertible neural network layers that utilize bijective functions.

By collectively representing all f_θ functions as $\mathcal{F}(X; \vartheta)$, where ϑ encompasses all parameters, the Encrypt and Decrypt layers can be denoted as $Y = \mathcal{F}(X; \vartheta)$ and $X = \mathcal{F}^{-1}(Y; \vartheta)$, respectively. The source code for both Encrypt and Decrypt is provided [1], along with a comprehensive experimental demo.

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

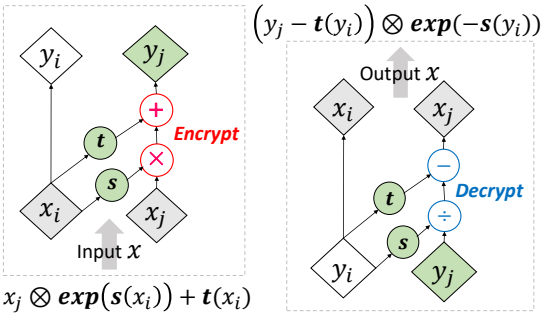


Figure 12: Encrypt (left) and Decrypt (right).

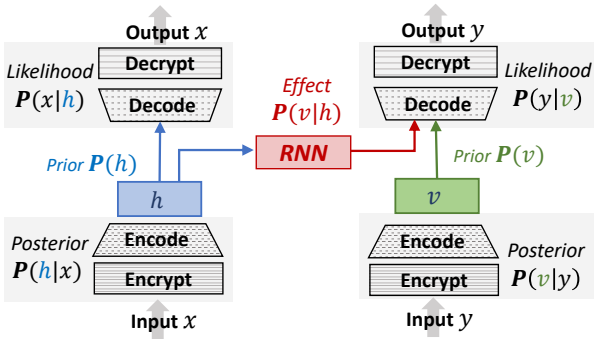


Figure 13: Relationship model architecture.

6.2 Stacking Hierarchical Representations to form SCM

Consider a causal system comprising three variables $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, each with corresponding representations $\{\mathcal{H}, \mathcal{V}, \mathcal{K}\} \in \mathbb{R}^L$ initially extracted by three separate autoencoders. Figure 13 illustrates the process of linking $\mathcal{H}$ and $\mathcal{V}$ to model the relationship $\mathcal{X} \to \mathcal{Y}$. Additionally, Figure 14 depicts how two modeled relationships related to $\mathcal{Y}$ are stacked to form a hierarchically disentangled representation.

Consider instances x and y for the relationship $\mathcal{X} \to \mathcal{Y}$, which are represented as h and v in $\mathbb{R}^L$. To estimate the latent dependency $\mathbf{P}(v|h)$, we use an RNN, as shown in Figure 13, to explicitly include the temporal features of h. For now, we suppose $\mathcal{V}$ can capture potential dynamics autonomously, expecting future refinements. Each iteration of the learning process involves three optimizations:

1. Optimizing encoder $\mathbf{P}(h|x)$, RNN model $\mathbf{P}(v|h)$, and decoder $\mathbf{P}(y|v)$ to reconstruct $x \to y$ relation.
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 distance in $\mathbb{R}^L$, and RNN acts as a bridge to traverse this distance, thereby informatively modeling the relation $x \to y$.

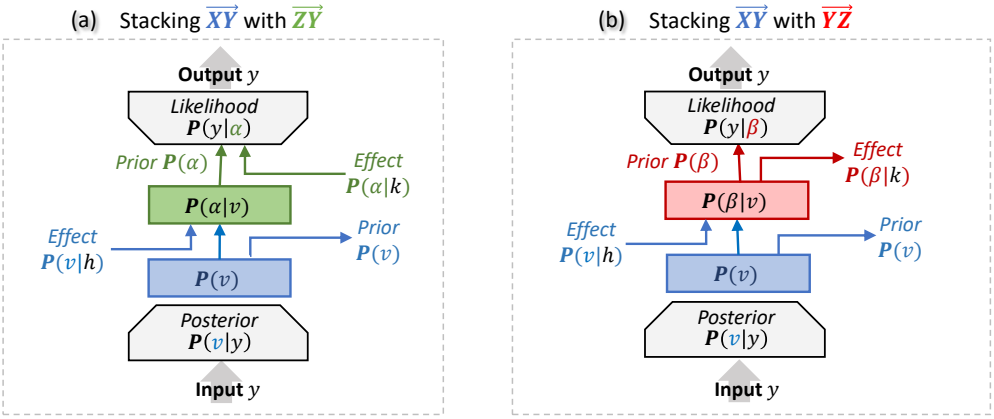


Figure 14: Architecutres of the relation-defined hierarchical disentanglement.

Figure 14 presents two stacking scenarios for $\mathcal{Y}$ within the $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$ system, according to different causal directions. Given the established $\mathcal{X} \to \mathcal{Y}$ relationship in $\mathbb{R}^L$, the left-side architecture completes $\mathcal{X} \to \mathcal{Y} \leftarrow \mathcal{Z}$ structure, while the right-side caters to $\mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$. By stacking an additional representation layer, hierarchical disentanglement is formed, allowing for various input-output combinations (denoted as $\mapsto$) based on practical needs. For instance, in the left-side setup, $\mathbf{P}(v|h) \mapsto \mathbf{P}(\alpha)$ signifies the $\mathcal{X} \to \mathcal{Y}$ relationship, while $\mathbf{P}(\alpha|k)$ suggests $\mathcal{Z} \to \mathcal{Y}$. On the right side, $\mathbf{P}(v) \mapsto P(\beta|k)$ indicates the $\mathcal{Y} \to \mathcal{Z}$ relationship with $\mathcal{Y}$ as input; conversely, $\mathbf{P}(v|h) \mapsto P(\beta|k)$ signifies the causal chain $\mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$.

Causal relationships of known edges can be sequentially stacked using existing causal DAGs in domain knowledge. Additionally, this approach aids in discovering causal structures within the latent space by identifying potential relationships among the initial variable representations.

6.3 Causal Discovery in Latent Space

Algorithm 1 outlines the heuristic procedure for identifying edges among the initial variable representations. We use Kullback-Leibler Divergence (KLD) as a metric to evaluate the strength of causal relationships. Specifically, as depicted in Figure 13, KLD evaluates the similarity between the RNN output $\mathbf{P}(v|h)$ and the prior $\mathbf{P}(v)$. Lower KLD values indicate stronger causal relationships due to closer alignment with the ground truth. Although Mean Squared Error (MSE) is a common evaluation metric, its susceptibility to data variances Reisach et al. (2021); Kaiser & Sipos (2021) led us to prefer KLD, using MSE as a secondary measure. In the graphical representation context, we refer to variables A and B in the edge $A \rightarrow B$ 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 $\beta \rightarrow n$
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 illustrates the causal structure discovery process in latent space over four steps. Two edges, (e_1 and e_3), are sequentially selected, with e_1 setting node B as the starting point for e_3. In step 3, edge e_2 from A to C is deselected and reassessed due to the new edge e_3 altering C's existing causal conditions. The final DAG represents the resulting causal structure.

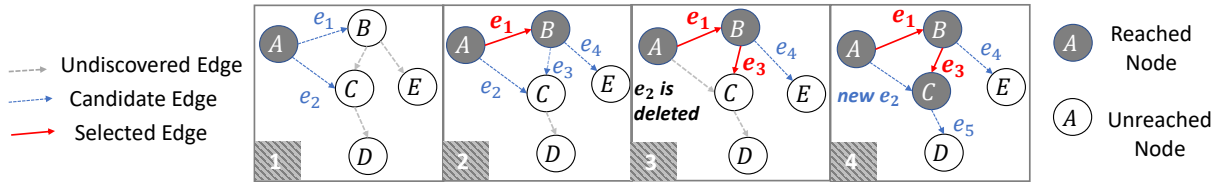


Figure 15: An example of causal discovery in the latent space.

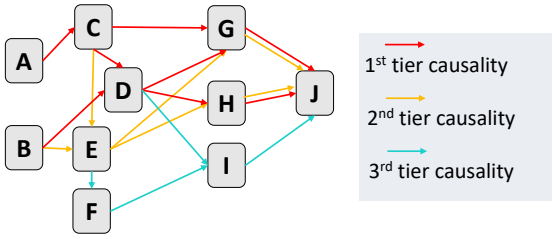
7 Efficacy Validation Experiments

The experiments aim to validate the efficacy of the *relation-defined representation* learning method in three areas: 1) extracting higher-dimensional representations with the proposed autoencoder architecture, 2) hierarchically establishing relation-defined representations, and 3) discovering DAG structure in latent space. A full demonstration of the experiments conducted in this study is available online [2].

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

We use a synthetic hydrology dataset for our experiments, a common resource in the field of hydrology. The task focuses on predicting streamflow based on observed environmental conditions like temperature and precipitation. The application of relation-defined representation learning aims to create a streamflow prediction model that is generalizable across various watersheds. While these watersheds share a fundamental hydrological scheme governed by physical rules, they may exhibit unique features due to unobserved conditions such as economic development and land use. Current models based on physical knowledge, however, often lack the flexibility to fully capture multiple levels of dynamical temporal features across these watersheds.

To evaluate robustness and generalizability, health informatics data would be optimal due to their complex confounding dynamics across multiple timelines. However, empirical constraints prevented us from accessing such data for this study. For validating the inherent bias, please refer to previous work Li et al. (2020).



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: Hydrological causal DAG: routine tiers organized by descending causal strength.

7.1 Hydrology Dataset

In hydrology, deep learning, particularly RNN models, has gained favor for extracting observational representations and predicting streamflow Goodwell et al. (2020); Kratzert (2018). For our experiments, we employ the Soil and Water Assessment Tool (SWAT), a comprehensive system grounded in physical modules, to generate dynamically significant hydrological time series. We focus on a simulation of the Root River Headwater watershed in Southeast Minnesota, covering 60 consecutive virtual years with daily updates.

Figure 16 displays the causal DAG employed by SWAT, complete with node descriptions. The hierarchy of hydrological routines is color-coded based on their contribution to output streamflow. Surface runoff (1st tier) significantly impacts rapid streamflow peaks, followed by lateral flow (2nd tier). Baseflow dynamics (3rd tier) have a subtler influence. Our causal discovery experiments aim to reveal these underlying relationships from the observed data.

7.2 Higher-Dimensional Variable Representation Test

In this test, we have a total of ten variables (or nodes), each requiring a separate autoencoder for initializing a higher-dimensional representation. Table 1 lists the statistics of their post-scaled (i.e., normalized) attributes, as well as their autoencoders' reconstruction accuracies. Accuracy is assessed in the root mean square error (RMSE), where a lower RMSE indicates higher accuracy for both scaled and unscaled data.

The task is challenging due to the limited dimensionality of the ten variables - maxing out at just 5 dimensions and the target node, J, having just one attribute. To mitigate this, we duplicate their columns to a consistent 12 dimensions and add 12 dummy variables for months, resulting in a 24-dimensional input. A double-wise extension amplifies this to 576 dimensions, from which a 16-dimensional representation is extracted via the autoencoder. Another issue is the presence of meaningful zero-values, such as node D (Snowpack in winter), which contributes numerous zeros in other seasons and is closely linked to node E (Soil Water). We tackle this by adding non-zero indicator variables, called *masks*, evaluated via binary cross-entropy (BCE).

Despite challenges, RMSE values ranging from 0.01 to 0.09 indicate success, except for node F (the Aquifer). Given that aquifer research is still emerging (i.e., the 3rd tier baseflow routine), it is likely that node F in this synthetic dataset may better represent noise than meaningful data.

Table 1: Statistics of variable attributes and performances of the variable representation test.

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 space causal discovery test.

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

7.3 Hierarchical Relation-Defined Representations Test

Table 3 presents the results of the relation-defined representation learning. We use the term "single-effect" to describe the accuracy of a specific result node when reconstructed from a single cause node (e.g., $B \rightarrow D$ and $C \rightarrow D$), and "full-effect" for the accuracy when all its cause nodes are stacked (e.g., $BC \rightarrow D$). To provide context, we also include baseline performance scores based on the initial variable representations. During the relation learning process, the result node serves two purposes: it maintains its own accurate representation (as per optimization no.2 in 6.2) and helps reconstruct the relationship (as per optimization no.1). Both aspects are evaluated in Table 3.

The KLD metrics in Table 3 indicate the strength of learned causality, with a lower value signifying stronger. For instance, node J's minimal KLD values suggest a significant effect caused by 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 "full-effect" are moderate compared to their "single-effect" scores, suggesting a lack of informative associations among the cause nodes. In contrast, for nodes G and H, lower "full-effect" KLD values imply capturing meaningful associative effects through hierarchical stacking. The KLD metric also reveals the most contributive cause node to the result node. For example, the proximity of the $C \rightarrow G$ strength to $CDE \rightarrow G$ suggests that C is the primary contributor to this causal relationship.

Figure 17 showcases reconstructed time series, for the result nodes J, G, and I, in the same synthetic year to provide a straightforward overview of the hierarchical representation performances. Here, black dots represent the ground truth; the blue line indicates reconstruction via the initial variable representation, and the "full-effect" representation generates the red line. In addition to RMSE, we also employ the Nash–Sutcliffe model efficiency coefficient (NSE) as an accuracy metric, commonly used in hydrological predictions. The NSE ranges from -∞ to 1, with values closer to 1 indicating higher accuracy.

The initial variable representation closely aligns with the ground truth, as shown in Figure 17, attesting to the efficacy of our proposed autoencoder architecture. As expected, the "full-effect" performs better than the "single-effect" for each result node. Node J exhibits the best prediction, whereas node I presents a challenge. For node G, causality from C proves to be significantly stronger than the other two, D and E.

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

7.4 Latent Space Causal Discovery Test

The discovery test initiates with source nodes A and B and proceeds to identify potential edges, culminating in the target node J. Candidate edges are selected based on their contributions to the overall KLD sum (less gain is better). Table 6 shows the order in which existing edges are discovered, along with the corresponding

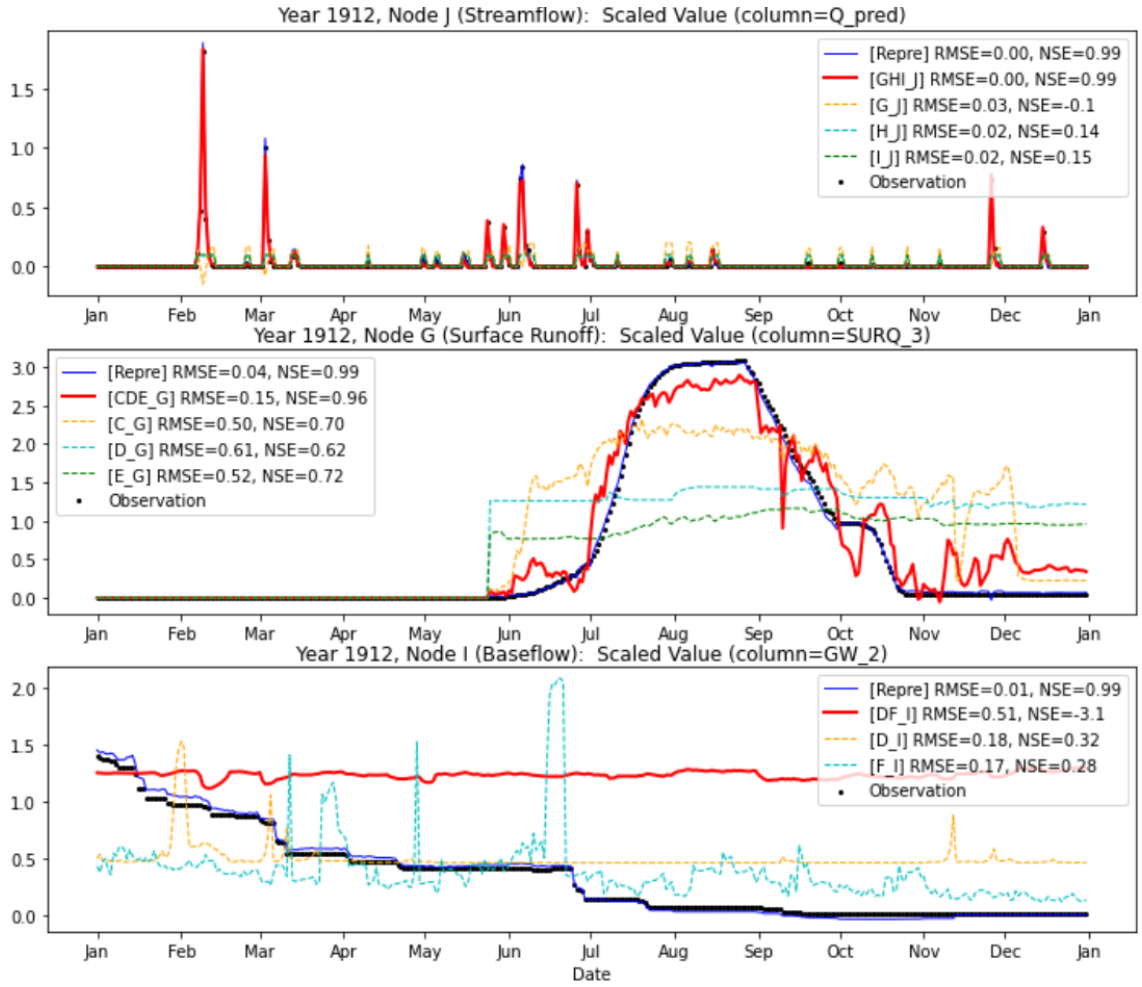


Figure 17: Reconstructed time series, via hierarchically stacked relation-defined representations.

KLD sums and gains after each edge is included. Color-coding in the cells corresponds to Figure 16, indicating tiers of causal routines. The arrangement underscores the efficacy of this latent space discovery approach.

A comprehensive list of candidate edges evaluated in each discovery round is provided in Table 4 in Appendix A. For comparative purposes, we also performed a 10-fold cross-validation using the conventional FGES discovery method; those results are available in Table 5 in Appendix A.

8 Conclusions

Driven by the misalignment issues between causal knowledge and established causal models in widespread AI applications, this study examines fundamental limitations of the dominant *Observation-Oriented* learning paradigm. In response, we advocate for a novel *Relation-Oriented* paradigm, inspired by the relation-centric nature of human knowledge, and complemented by a practical approach of *relation-defined representation* learning, with demonstrated efficacy.

The concept of a "hyper-dimension" is initially proposed, as an accommodation for unobservable knowledge. We subsequently build a comprehensive framework of dimensionality, to offer more intuitive insights into relationship learning. The discrepancy, between our comprehension of "time" and the single timeline used in our causal models, inherently causes misalignment, and results in model generalizability issues.

Relation-Oriented reflects the process of human understanding, aims to mitigate AI misalignment, paving the way toward causally interpretable AGI. Constructing AGI is a long-term, intricate process requiring extensive work within interdisciplinary efforts, and we seek to lay a foundation for its future advancements.

Table 3: Performances of the relation-defined representations, sorted by the result node.

Result Node	Variable Representation (Initial)			Cause Node	Variable Representation (in Relation Learning)			Relationship 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

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.

Stuart J & Angelini Gianni D & Blackstone Eugene H Benedetto, Umberto & Head. Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6): 1112–1117, 2018.

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

Cristobal Pagán Coulson, Seana & Cánovas. Understanding timelines: Conceptual metaphor and conceptual integration. *Cognitive Semiotics*, 5(1-2):198–219, 2009.

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.

Ursula Fuller, Colin G Johnson, Tuukka Ahoniemi, Diana Cukierman, Isidoro Hernán-Losada, Jana Jackova, Essi Lahtinen, Tracy L Lewis, Donna McGee Thompson, Charles Riedesel, et al. Developing a computer science-specific learning taxonomy. *ACm SIGCSE Bulletin*, 39(4):152–170, 2007.

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.

Marco Huang, Yimin & 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.

Phu & Sorooshian Soroosh & Hsu Kuo-lin Ombadi, Mohammed & Nguyen. 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.

Francesco & Bauer Stefan & Ke Nan Rosemary & Kalchbrenner Nal & Goyal Anirudh & Bengio Yoshua Scholkopf, Bernhard & Locatello. 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.

& Max Tegmark Wes, Gurnee. Language models represent space and time, 2023.

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.

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

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