Relation-Oriented: Toward Knowledge-Aligned Causal AI

Anonymous authors
Paper under double-blind review

Abstract

In machine learning, we naturally apply an *Observation-Oriented* principle, in which observational variables preexist and set the stage for constructing relationships. While sufficient for traditional models, the integration of AI with big data exposes the *misalignment* between the observational models and our actual comprehension. Contrarily, humans shape cognitive entities defined by relationships, enabling us to formulate knowledge across temporal and hyper-dimensional spaces, rather than being confined to observational constructs. From an innovative *Relation-Oriented* perspective, this study examines the roots of this misalignment within our current modeling paradigm, illuminated by intuitive examples from computer vision and health informatics. We also introduce the *relation-defined representation* learning methodology as a practical implementation of *Relation-Oriented* modeling, supported by extensive experimental validation.

1 Introduction

The prevailing modeling methods premise that relationships are built upon observed entities, designated as model variables and outcomes. This *Observation-Oriented* principle, going back to the 1890s Picard-Lindelof theorem, uses a *logical timeline* t to structure observation timestamps, as denoted by $x_{t+1} = f(x_t)$, to depict variable X's time evolution. Since then, the convention has been to address the temporal dimension as measurements in a constant time lag $\{t, t+1\}$ unit. When modeling a relationship $X \to Y$ as $y_{t+1} = f(x_t)$, the outcome Y is strictly observational, leaving the temporal aspect managed by $f(\cdot)$.

Seeing t as an axis for temporal dimension, conventional models $f(\cdot)$ are limited to linear relationships with respect to t. While maybe sufficient in the past, with the current technological advancements in data collection and Artificial Intelligence (AI), exploring nonlinear relationships has become essential, which requires t to be treated as a distinct computational dimension. However, a single t-axis fails to capture the multifaceted nature of temporal dimensionality, resulting in models misaligned with our cognitive understanding.

This study aims to highlight how the often-overlooked complexity of temporal dimensionality can fundamentally contribute to AI Alignment issues. We propose that shifting toward a *Relation-Oriented* perspective in model construction is vital for accurately integrating "time" as a dimension in our computations.

1.1 Manifestation of Misalignment

Today, AI has displayed capabilities surpassing humans in solely observational learning tasks, such as generating images and Go gaming. Despite these, AI may appear "unintelligent" in comprehending certain relationships 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), particularly evident when leveraging neural networks to address large-scale questions. Such methods are often confined to context-specific applications and encounter difficulties in extending to diverse scenarios. In contrast to other remarkable achievements, AI's capability within the temporal dimension remains notably constrained.

The questions "how to apply deep learning in causality" and "how to simulate hands" may seemly pertain to specific research 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 some relationships 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. Ensuring that AI functions reasonably is becoming increasingly critical to address.

1.2 Division of Cognitive Space

Figure 1 divides our cognitive space into three categories: 1) *observational* space, 2) *temporal* space, and 3) *hyper-dimensional* space. Current models largely operate within 1), defined by observational dimensions. For instance, CNNs (Convolutional Neural Networks) utilize a two-dimensional pixel space to learn associations; a quadrotor's locations can be pinpointed and predicted in three spatial dimensions; LLMs (Large Language Models) function within semantic space along a distinct logical timeline; and medical effects are assessed using observed vital signs in chronological sequence. Some applications (e.g., the last three examples) employ a logical timeline to depict temporal evolution and are often referred to as spatial-temporal analysis in various fields Alkon (1988); Turner (1990); Andrienko (2003). Yet, for the model, a sequence of timestamps serves similarly to an observational attribute, without needing to be temporally significant. Thus, we classify this scenario within *observational* space, including a single timeline to represent the absolute time evolution.

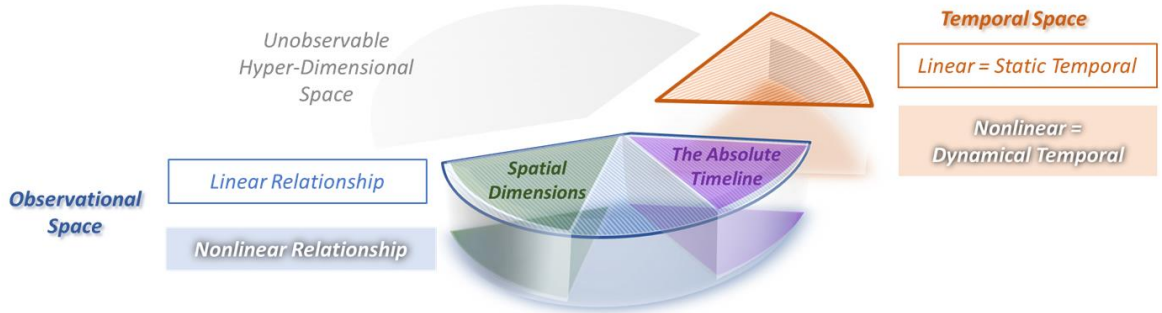


Figure 1: *Observational*, *Temporal*, and *Hyper-Dimensional* spaces, with the former two *Observable*.

In our cognitive framework, the *temporal* space is shaped by ***multiple*** relative logical timelines, in addition to the absolute one. Further insights into it can be found in Section 4. Within the current *Observation-Oriented* modeling paradigm, this multifaceted nature can create *inherent temporal biases* in AI computation. Such biases fundamentally impede AI's ability to handle nonlinearities in the temporal dimension, often with underlying causal significance. In short, they narrow our successful AI applications within *observational* space only, and obstruct causal learning from achieving generalizable success in *observational-temporal* space.

In this paper, "features" refer to potential variables representing the distribution of interest in any dimension. Observational and temporal features are encompassed in observations. Some data, like sequential time series, may integrate both aspects within the *observational-temporal* joint space, as depicted in Figure 1. This joint space may also be referred to as the "*observable* data space", in contrast to the "latent feature space".

Conversely, some knowledge that exists within our comprehension cannot be captured through observations alone. This includes granularity levels we utilize to decode complex phenomena, like how grasping personal traits (a higher individual level) depends on understanding the cultural context (a lower population level). We categorize such *unobservable* ***hierarchies*** within hyper dimensions, forming a *hyper-dimensional* space. This space accommodates relationships present in our knowledge but inaccessible to AI's automatic learning from data. The problem of AI-generated "unrealistic hands" exemplifies this issue (see Section 2.1).

As emphasized by Scholkopf et al. (2021), the recognition of these hierarchies is vital for AI, as it relies on reusable low-level relationships to create generalizable models that function across higher-level learnings. Prevalent VAE (Variational AutoEncoder)-based feature disentanglements, however, fall short in addressing such relation-defined hierarchies. In *observational* space, the challenges may still be manageable; however, in *observational-temporal* space, underlying hierarchical dynamics can lead to significant complications.

1.3 Relation-Defined Representation

Contrary to AI, human understanding is fundamentally *relation-centric*. Relationships serve as indices that point to our mental representations Pitt (2022), shaping our cognitive understanding of observations and temporal events. Through relationships, we can form interconnected knowledge systems in memory. It is not surprising that the cutting-edge methodology to teach AI causal reasoning focuses on "causal representations" Scholkopf et al. (2021), which is, letting the modeled objects represent the causal relations, instead of the observations. This concept indeed provides valuable insight into the pursuit of AI alignment.

Following a similar train of thought, we introduce the ***Relation-Oriented*** principle, coupled with the *relation-defined representation* learning method, devised as an autoencoder-based implementation. By using relations as indices, this approach can extract knowledge-aligned representations from intricate observations to craft suitable model variables with reasoning. More significantly, it enables AI to autonomously identify and represent the desired features across observational, temporal, and hyper dimensions, incorporating causally relevant dynamics without the need to manually recognize logical timelines. Therefore, we regard this method as a vital instrument for progressing toward AI alignment.

Section 2 delves into unobservable knowledge hierarchy issues, followed by Section 3, where existing causality methods are reassessed from a *Relation-Oriented* perspective, highlighting their limitations. Section 4 explores the complexity of temporal dimensionality and our successes in capturing dynamical features. The hierarchical relation-defined representations are formally factorized in Section 5, and Section 6 outlines the proposed implementation methodology, followed by Section 7 to detail the subsequent validation experiments. Finally, Section 8 concludes the significance and target of this study.

2 The Unobservable Hierarchy in Knowledge

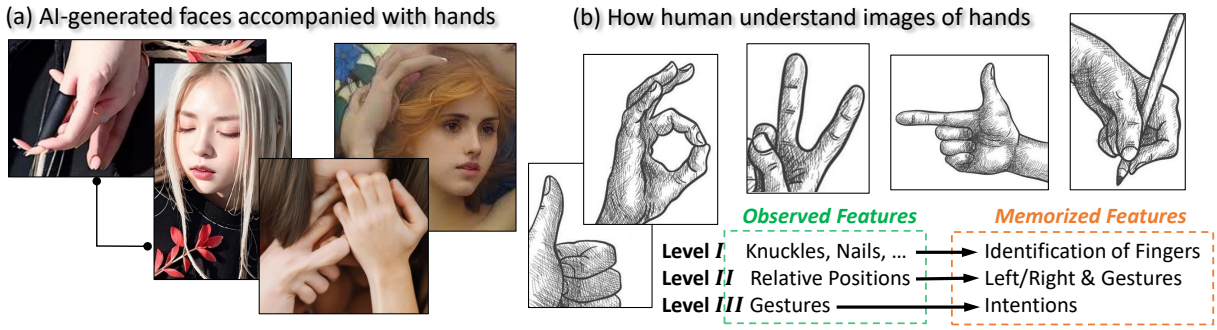


Figure 2: A comparison of AI-generated and human-sketched hand images. AI processes observed features simultaneously, thus treating hands as arbitrary mixtures of finger-like items. For humans, the process is hierarchical, where higher-level recognition relies on lower-level conclusions, even with incomplete views.

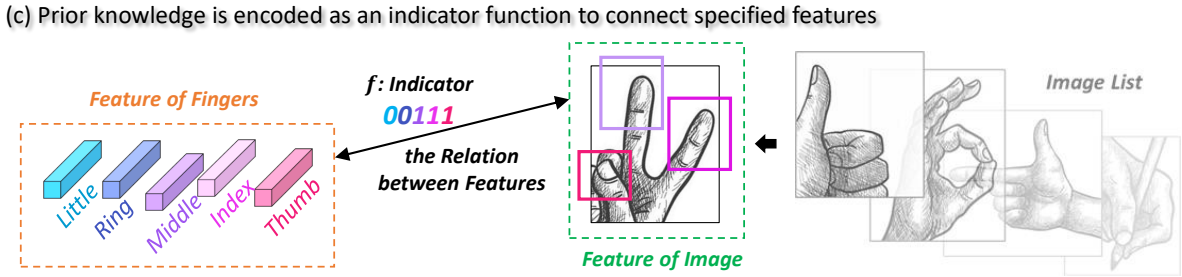


Figure 3: Traditional *Observation-Oriented* decomposation of the relationship in Level **I** knowledge.

Hierarchical structures naturally exist across various fields of knowledge. In this section, we investigate these through two illustrative examples, focused on observational and temporal features, respectively. The first example, grounded in computer vision, emphasizes the direct challenges posed by the unobservable hierarchy in observational learning. The second, stemming from health informatics, highlights the frequently overlooked hierarchical dynamics of causal effects, through the lens of traditional causal learning.

2.1 Hierarchy of Observational Features

Figure 2(a) showcases AI-created hands with vibrant colors but unrealistic shapes, while humans can easily recognize a plausible hand from simple sketches in (b). Indeed, our cognition rapidly decomposes observations into hierarchical features, processed 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 is from prior knowledge but cannot be learned by observations alone. To AI systems, or similarly, to an extraterrestrial without our knowledge, the peculiar hands in Figure 2(a) may seem as reasonable as the actual hands.

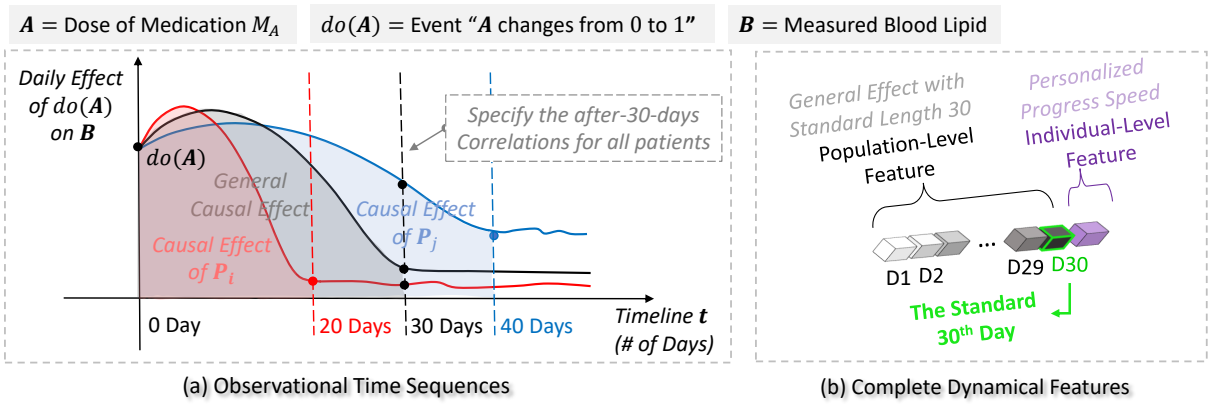
Such unobservable hierarchies may not always create major problems. If features at different levels do not significantly overlap, AI may successfully "distinguish" them. For instance, AI can generate convincing faces because the appearance of eyes is strongly indicative of the facial angle, eliminating the need for AI to recognize "eyes" from "faces". Conversely, similar finger appearances may correspond to various hand gestures, causing AI to misinterpret "hands" as random associations of fingers.

As humans, this hierarchical organization allows us to reuse knowledge across various scenarios efficiently. For example, we can recognize fingers (Level **I** knowledge) from photographs, watercolor paintings, and diverse sketching styles. This repeated identification strengthens both the connections within each level of knowledge and the distinctions between different levels. In essence, our knowledge hierarchy is constructed through relationships rather than mere observations. As the leading technology for feature disentanglement, VAEs can effectively separate observational distributions Burgess (2018), but have limitations in handling knowledge hierarchy without indexing pre-determined relationships.

Refer to Figure 3, where the left side shows well-disentangled finger features, with the right side presenting pencil sketches to recognize. If VAEs are trained using data like watercolor paintings for the disentanglement on the left, they may find it challenging to adapt to the sketches on the right. Although traditional graphical models could establish the inter-relationship between the two sides, such as the indicator function f depicted in Figure 3, they often necessitate fully observed variables on both sides, which may lead to extensive manual labeling, a solution not always feasible. Conversely, the proposed *Relation-Oriented* modeling is to extract relation-defined representations directly from the right-side observations through f as indexing from the left. In this methodology, the function f is not pre-specified; instead, it serves as a filtering mechanism, to exclude color-related aspects from the left side, and keep only relevant information for the right side.

2.2 Hierarchy of Dynamical Temporal Features

Within observational dimensions, AI may learn unobservable knowledge via reinforcement learning Sutton & Barto (2018), guided by human feedback. For example, human approval of five-fingered hands could lead AI to identify fingers autonomously. However, it may not function when involving temporal dimension.



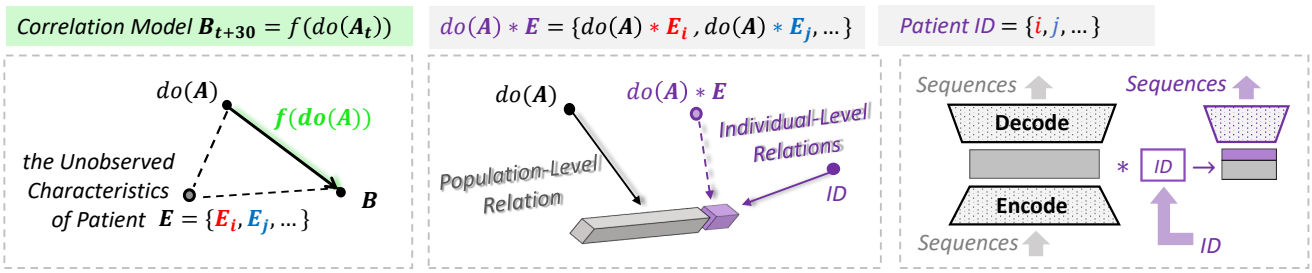
(a) Observational Time Sequences (b) Complete Dynamical Features

Figure 4: 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 in 20 days, while P_j takes 40 days. The conventional medical effect estimation is derived from the patients' after-30-day correlations, instead of their personalized elbows.

Figure 4(a) depicts patients' daily effects on B following $do(A)$, with t indicating the elapsed days. Without loss of generality, 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 effects (e.g., the red and blue curves in (a)) are shaped by two levels of *dynamical* features: 1) the population-level effect sequence with a standard length of 30, and 2) the individual-level progress speed. Evaluating M_A's clinical effectiveness should be based on accurately estimated level 1) dynamic.

Figure 4(b) represents the patients' dynamical effects in a 31-length feature vector, disentangled by two hierarchical levels. Traditional medical effect estimation, often obtained by averaging patients' after-30-days performances, essentially builds a correlation model $B_{t+30} = f(do(A_t))$. It only captures the final step of the standard effect sequence, disregarding the preceding 29 steps, thus only representing a ***static*** fragment of the population-level dynamic. Moreover, it implicitly assumes a normal distribution of patients' personalized time spans around the mean value of 30 days, along with a fundamentally linear effect-decreasing.

2.3 The Elusive Hidden Confounder



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

Figure 5: (a) The traditional causal inference. (b) Hierarchical disentanglement of dynamics using relations as indices. (c) Autoencoder-based generalized and individualized reconstructions of the sequential data.

For individuals like P_i or P_j, the 30th day effect estimated by f has a bias, as P_i has exceeded 100% of the expected full effect, while P_j only achieves about 75%. To account for these individual-level biases, traditional causal DAG (Directed Acyclic Graph) introduces a *hidden confounder* (denoted as E), to represent patients' unobserved personal characteristics, as shown in Figure 5 (a). However, it raises a question - why should we consider a variable not in our model's scope? This inclusion suggests an illogical assertion: "Our model is biased due to ignorance of some aspects we have no need to investigate."

The underlying reason is that while E is unknown, its effect - the individual-level dynamics - is observable. As being excluded by f, it can only be attributed to potentially existent E. Although hidden, E still represents an observational variable and thus could be incorporated by f once revealed. In essence, introducing the hidden confounder is to transform *observed dynamical-temporal* variables into *unobserved observational* ones. This serves to enhance human understanding, though does not necessarily benefit the model.

Figure 5(b) disentangles the dynamical effects hierarchically, using relations as indices. Traditional causal inference views the individual-level effect as caused by the unobserved composite cause $do(A) * E$, which is an interpretable causal relationship, but not directly modelable. Conversely, a *Relation-Oriented* approach treats the relation as an index without requiring modeling, so we can use any observed identifier, such as the patient ID, to pinpoint the individual-level dynamics. Figure 5(c) illustrates the implementation architecture of this *Relation-Oriented* methodology, which realizes two separate reconstruction processes - with, and without the individual-level dynamical temporal features, respectively.

3 Causality on Timeline

The research on causal relationships serves as the probe into the temporal dimension, extending beyond our observational reality. Curiously, one might find it rare to see "incorporation of time" defined as the distinctive factor between causality and mere correlation in causal inference theories. While, as in our modeling context, instead of a logical one, what significance does this distinction hold?

Time series frequently involve a timestamp attribute, logically reflecting the absolute time evolution in reality. For the model, however, this way renders the line between causality and correlation, as it operates within the modeling space we defined, irrespective of the dimensions' temporal significance. While, it does not diminish the importance of the temporal dimension, but rather emphasizes that our current causal learning might not completely align with our intuitive understanding of this dimension, revealing a certain discrepancy. This section is devoted to discussing the present causal modeling from this particular perspective.

3.1 Causality vs. Correlation

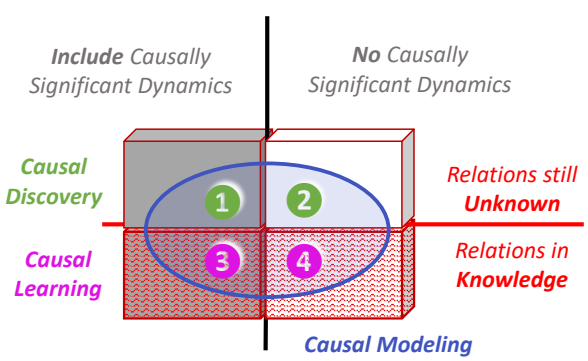
Consider a general model function $Y = f(X; \theta)$, with θ denoting the model parameter. The process of learning θ is irrespective of how we interpret the $X \to Y$ relationship, instead, relying solely on the observations of variable X and outcome Y. A causal relationship comprises two layers: **1)** the basic connection between X and Y that is modeling-significant, and **2)** the roles of cause and effect that are not modeling-significant.

To be specific, for causality $X \to Y$, we can employ the function $Y = f(X; \theta)$ to predict the effect Y, and likewise, use $X = g(Y, \psi)$ to infer the cause X when given Y. Both parameters θ and ψ are equivalently derived from joint probability $\mathbf{P}(X, Y)$, making directionality more a matter of interpretation than a modeling constraint. So, why is "modeling direction" a concern? And why do we even need to specify "causal models"?

The primary reason lies in our current causal models' limitation when addressing causally significant dynamics, which are often unbalanced between cause and effect. For example, in Figure 4, the effect dynamics cannot be fully captured by the correlation model in Figure 5, necessitating a hidden confounder to inform our prior understanding. However, if such dynamics act as the cause, they could be effectively handled with RNNs. For more discussions about addressing dynamics in causality, please refer to Section 4.2.

Lemma 1. Causality vs. Correlation in the modeling context.
- Causality is connection between features, which can be *observational*, *temporal*, and also ***dynamical***.
- Correlation is connection between features, which are ***not dynamical***.

In short, the distinction between causality and correlation lies in the connected critical features in this relationship, instead of the connection itself. The "causal model" concept first emerged within statistical causal inference. Considering the challenges posed by nonlinearity at that time, distinguishing causal relationships and reasonable directions becomes especially crucial for subsequent manual interpretations.



	Modeled Connection	Modeled 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 6: The Current Causal Modeling Context. The rectangle in the left diagram represents all relationships with logical causal significance (these do not necessarily include dynamics, e.g., "raining → wet floor" involves two static temporal features); while the blue circle denotes the potential scope of modeling.

Importantly, a relationship with causal significance in a *logical* context does not automatically translate to "causality" within a *modeling* context. For this reason, the "incorporation of time" alone is not a sufficient condition to validate causal models. A widespread misconception often associates the temporal lag between cause and effect as an indicator of causality. The following example will help to clarify this misconception.

Morgan Spurlock conducts a self-experiment in the documentary film "Super Size Me" (2004). Over a month, Spurlock restricted his diet solely to McDonald's meals, and consecutively recorded the significant deteriorations in his health and physical appearance, which dramatically demonstrated the potential negative health impacts of a fast-food diet. Consider two time sequences: $\mathbf{x} = \{x_1, \ldots, x_{30}\}$ and $\mathbf{y} = \{y_1, \ldots, y_{30}\}$. Here, each x_t signifies the act of eating McDonald's on the tth day, while each y_t corresponds to the fluctuations in his health indices for that day. It is intuitive to recognize that $\mathbf{x} \rightarrow \mathbf{y}$ is a causal relationship, although no time lag exists between $\mathbf{x}$ and $\mathbf{y}$. Yet, a time lag is also conceivable in this case. Suppose another person tried this experiment, but since he had a longer responding digestive system, the same effect denoted by $\mathbf{y}$ emerged with a 5-day delay, creating a 5-day lag between x_1 and y_1. This instance emphasizes that even when both the cause and effect include dynamics, separation by a time-lapse is not a reliable criterion.

3.2 The Current Causal Modeling Context

Figure 6 broadly classifies causal queries into four categories, based on whether incorporate critical dynamical features, and whether already be within existing knowledge. We approach this topic from two aspects: the modeling-significant *basic connection*, and the interpretation-significant *causal direction*.

(1) Aspect of the Basic Connection

The traditional causal inference has made notable advancements in exploring the specific conditions under which dynamical features can be "downgraded" to a level accessible by the employed statistical models. Notably, *do-calculus* Pearl (2012) probes into identifiable events (i.e., treating dynamics as individual entities) to establish variables in the temporal dimension. Essentially, this involves manually converting dynamics into static temporal elements, but it tends to be greater adaptable in handling causes compared to effects (For a more comprehensive discussion, please refer to section 4.2). While for the effect dynamics overlooked by models, if existing knowledge can suggest potential causes, the creation of a *hidden confounder* can enhance comprehension; if not, these dynamics may be dismissed based on the *causal sufficiency assumption*, which could lead to subsequent challenges.

On the other hand, causal discovery methods mainly scan the observational space, incorporating static elements lying on the single absolute timeline (i.e., associated with the timestamp attribute). As a result, if the underlying causal mechanism does not encompass crucial dynamics, 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* that the observed variables fully represent the causal reality.

(2) Aspect of the Causal Direction

Consider observables 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 $X \rightarrow Y$ 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

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}[(\frac{\partial}{\partial\theta}\log p(X,Y;\theta))^2 \mid \theta] = \int_Y\int_X (\frac{\partial}{\partial\theta}\log p(x,y;\theta))^2 p(x,y;\theta)dxdy \\ &= \alpha\int_Y (\frac{\partial}{\partial\theta}\log p(y;x,\theta))^2 p(y;x,\theta)dy + \beta = \alpha\mathcal{I}_{Y|X}(\theta) + \beta, \text{with } \alpha,\beta \text{ constants.}\end{aligned}$$

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

Subsequently, the likelihoods of the estimated parameters $\hat{\theta}$ and $\hat{\psi}$ depend on the amount of information, $\mathcal{I}(\hat{\theta})$ and $\mathcal{I}(\hat{\psi})$. Thus, 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". This interpretation presumes that the cause should be more comprehensively captured in the observations than the effect. While this was a reasonable default setting in previous decades, due to limitations in data collection techniques, it is not necessarily the case in the present era.

In summary, a causal direction purely inferred from observations could be causally meaningful in logic, if satisfying: **1)** the causal relation of interest does not involve significant dynamics, **2)** the observations are known, a prior, to be more informative about the cause than the effect, and **3)** the specified distributions are fundamentally accurate. Technically speaking, in the traditional modeling context, the term "causal model" is not a specific model type. Instead, it designates models that require additional logical interpretations about overlooked dynamics, to bridge comprehension gaps and pave the way for potential improvements.

4 The Overlooked Temporal Dimensionality

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 timestamps, indistinguishable from other observational values. 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.

Lemma 2. *Temporal Dimension* refers to the aggregation of all potential logical timelines, not a single one. A *Temporal Space* in comprehension is the space built by chosen timelines as 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".

In this section, we initially dissect the mechanism of inherent temporal bias within the current AI systems, due to neglecting potential multi-timelines; then, reassess existing methods of learning dynamics; and last, summarize advancements and challenges on our journey towards realizing knowledge-aligned causal AI.

4.1 Inherent Temporal Bias

Overlooking multi-timelines in *structural causal models* (SCMs) can lead to *inherent temporal biases*. These biases substantially limit our ability to fully harness AI's potential in modeling nonlinear dynamics, especially within large-scale causal relationships, which may include more complex multi-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 4 has the redefined DAG depicted in Figure 7(b), with (a) showing the traditional one as a comparison. The edge $do(A) \rightarrow B$ in Figure 7(a) can only (partially) represent population-level effect, thus necessities a hidden confounder to explain the individual-level diversities, while in Figure 7(b), they can be explicitly represented by varying lengths of $\overrightarrow{do(A)\ B}$.

Consider an expanded two-timeline scenario in Figure 8(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 8 (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.

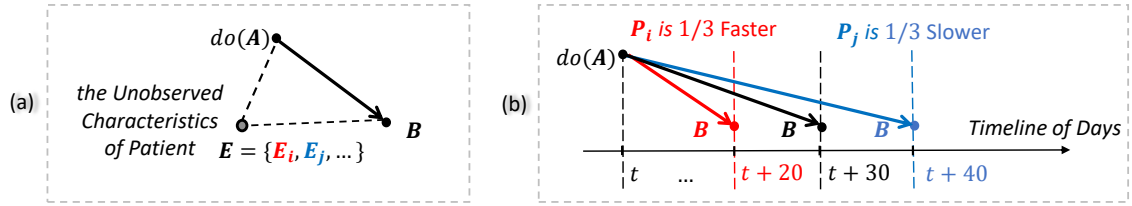


Figure 7: (a) Traditional Causal DAG introducing hidden E. (b) Redefined DAG: the standard black vector signifies the population-level effect, while the individual-level ones are represented by its different scaling.

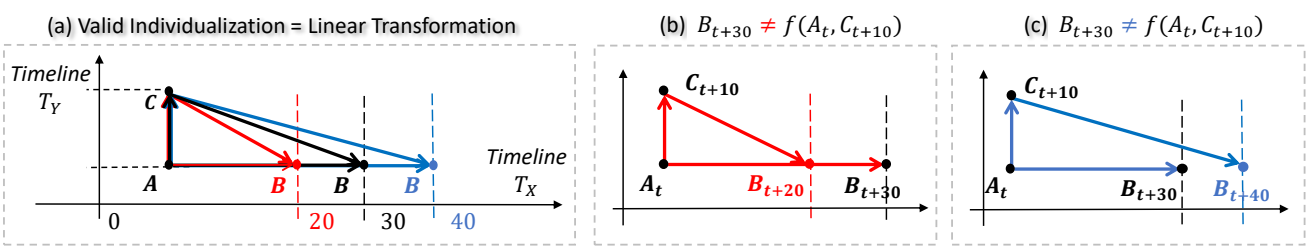


Figure 8: (a) A two-timeline (redefined) DAG space, where a valid individualization presents a linear transformation. (b)(c) Violations of the Markov condition for the prevailing SCM with confounding across timelines.

Notably, in this specific case, the violation may not cause significant issues for RNN models. Given the independence of dynamical features on T_X and T_Y, the SCM can be formulated as $B_{t+30} = f_1(A_t) + f_2(C_{t+10})$. This suggests that the cross-timeline confounding can be broken down into two single-timeline issues, where capturing hierarchical dynamics might challenge statistical models but not RNNs. However, assuming special conditions like independence or non-confounding is impractical. Given that each pair of cause and effect could potentially inhabit a logical timeline, such temporal biases are inherently prone to exponentially accumulate and impact our SCM applications, regardless of the model implementation.

Lemma 3. The traditional SCM may have the inherent temporal bias if containing: **1)** *Confounding* of dynamics across *Multiple* logical timelines, and **2)** Unobservable hierarchy of such confounding.

It is interesting to notice that most of the successful applications instinctively avoid one of the two factors: *confounding* or *multi-timelines*. For statistical causal models, the interpretability allows them to be manually 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 identify, as they often intermingle with biases resulting from the inability to model nonlinearity - They similarly manifest as statistical diversity (or data heterogeneity) among levels, and both can be addressed via de-confounding. Consider Figure 7(a), a correlation model that only captures the population-level effect can mismatch with individuals P_i and P_j, which may not be distinguishable from the model mismatching in Figure 8(b)(c), caused by crossing two timelines.

4.2 Learning Dynamics

For observational variable instance $x \in \mathbb{R}^d$, we frequently encode its temporal evolution as a sequence along the timeline, such as $\{x_t\} = \{x_1, \ldots, x_t, x_{t+1}, \ldots, x_T\}$, without considering t as the $(d+1)$-th dimension. Accordingly, we seldom view x's changing value over time as a "distribution" along the t-axis.

In prediction $x_{t+m} = f(x_t)$ with integer $m > 0$, the potential nonlinearity of function f is only acknowledged within $\mathbb{R}^d$, but leaving the relationship between x_t and x_{t+m} to be ***linear***. It implies that the ***correlation*** model $x_{t+m} = f(x_t)$ can only capture ***static*** temporal elements in sequence $\{x_t\}$. If, however, no suitable (f, m) can be found to ensure this model adequately represents $\{x_t\}$, there may exist unrepresented ***dynamical*** temporal features, which necessitates a ***causality*** model of the form $\{x_\tau\} = f(\{x_t\})$. Here, cause and effect are as two distinct sequences on logical timelines, τ and t, respectively, such that, the relationship between any pair of x_t and x_τ could be ***nonlinear***.

Consider this form of causality model: $y_\tau = f(\{x_t\})$ with $\tau = (t + T) + m$, representing a sequence $\{x_t\}$ causing a static outcome y_τ, where timeline τ is confined as m-timestamp later than timeline t. This form is typically adopted in many dynamical learning methods, such as autoregressive models Hyvärinen et al. (2010) and RNNs Xu et al. (2020). For autoregressive, if f is defined as a linear function, this model is restricted to only capturing *a **single-level** dynamic* of the cause. For instance, in Figure 4(b), let $T = 30$, a linear autoregressive model may capture the population-level complete sequence, but still cannot represent individuals' speeds. Contrarily, RNNs embrace nonlinearity for f, thus potentially can capture both levels of dynamics (without disentanglement).

However, RNNs primarily target dynamics only for the cause, while remaining constrained to a static effect. This limitation makes sense, as it is feasible to designate significant time sequences for the cause, but pinpointing precise start and end timestamps for ensuing effects is challenging. On the other hand, Granger causality Maziarz (2015), a method extensively used in economics, acknowledges the existence of potential multi-timelines in causality. It employs the general form $\{y_\tau\} = f(\{x_t\})$, where t and τ typically represent two distinct timelines, thereby including a sequential effect. However, as primarily a statistical test, it is best suited for handling single-level dynamics and necessitates additional specifications for effect sequences.

A more universally applicable concept of dynamics is the *do-calculus* Pearl (2012); Huang & Valtorta (2012), which refrains from assuming specific time sequences. Instead, it treats *identifiable* temporal events as objects around which it conducts elementary calculus, adopting a more *Relation-Oriented* viewpoint. In the subsequent discussion, we will revisit the three rules of *do-calculus* from a differential-and-integral perspective.

For 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 elementary, 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 is *observational* effect. So, dynamics of cause $X \in \mathbb{R}^d$ can be formulated as:

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

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

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

Fundamentally, the three rules of *do-calculus* address all conceivable conditional independencies between the *observational* and *interventional* effects within the $\{X, Y, Z\}$ graphical system, sidestepping the most general cases (specifiable associated cause $do(x_t) \cdot x_t$ belongs to Rule 2).

Given the substantial flexibility provided by the $do(\cdot)$ format, we can also represent the effect Y dynamically, by introducing timeline τ as an extra dimension, yielding $\mathcal{Y} = \langle Y, \tau \rangle$. Nonetheless, determining the *identifiable* events within $\mathcal{Y}$ still requires manual specification that may not be practical, in contrast to our proposed relation-defined representation learning that can construct $\mathcal{Y}$ automatically.

4.3 Toward Knowledge-Aligned AI

Our quest for causally reasoning AI that can align with our knowledge, involves the process of broadening our modeling techniques from purely observational to include temporal and hyper-dimensional spaces. Referring to Figure 9, our present challenge lies in leveraging AI's potential within structural causal models. The key to this task resides in identifying potential multi-timelines. However, manual identification is not feasible, especially given that both logical timelines and knowledge hierarchies are not directly observable. It is time for us to transition from an *Observation-Oriented* to a *Relation-Oriented* modeling paradigm.

Model	***Principle***	***Cause***	***Connection & Direction***	***Effect***	***Handle Unobservable Hierarchy***	***Capture Dynamics***
Mechanistic or Physical	$Y = f(X; \theta)$	Observational-Temporal $\langle X, t \rangle$	by Knowledge	Observational-Temporal $\langle Y, t \rangle$	Yes	Yes
Relation-Oriented Structural Model	*Given* $\boldsymbol{P}(X, Y)$ & $X \rightarrow Y$	Observational-Temporal $\langle X, t \rangle$	Learn Representation $\hat{Y} = f(X; \theta)$	Observational-Temporal $\langle \hat{Y}, t \rangle$	Yes	Yes
Structural Causal Learning	*Given* $\boldsymbol{P}(X, Y)$ & $X \rightarrow Y$ $Y = f(X; \theta)$	Observational X & Sequential $\{X_t\}$	Connected via θ $X \rightarrow Y$ by Knowledge	Observational & Static Y_t	?	?
Graphical Causal Discovery	*Given* $\boldsymbol{P}(X, Y)$ *Speciy* ϑ *Find* $\mathcal{L}(Y\|X; \vartheta) > \mathcal{L}(X\|Y; \vartheta)$	Observational X	Connected via ϑ $X \rightarrow Y$ by Observed Info	Observational Y	?	No
Common Cause Model	*Given* $\boldsymbol{P}(X, Y\|Z)$	Observational X	Connected 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 et al. (2021)), from more knowledge-driven (top in purple) to more data-driven (bottom in green). Notations: ϑ = joint distribution, θ = conditional distribution, $\langle X, t \rangle$ = augment t as an additional 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 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) - We reinterpret it as capturing ***temporal distributions*** and responding to *conditional* questions. Typically, these models leverage prior knowledge to construct causal Directed Acyclic Graphs (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 efficient solutions for large-scale problems. However, this approach potentially conceals numerous relative timelines, which may result in substantial temporal biases. This is particularly apparent in the limited success of applications that incorporate DAGs into network architectures Luo et al. (2020); Ma (2018), such as Neural Architecture Search (NAS). Schölkopf Scholkopf et al. (2021) succinctly highlights three key

challenges impeding the broader success of causal AI: 1) limited model robustness, 2) insufficient model reusability, and 3) inability to handle heterogeneity. The first two challenges can be associated with inherent temporal biases, while the latter originates from the unobservable knowledge hierarchy.

On the flip side, physical models, which explicitly integrate time as a dimension, able to establish abstract concepts through structural knowledge, may provide insights into our upcoming challenges. The proposed *Relation-Oriented* modeling is designed along these lines, aiming to transcend the observational limitations innate to our prevailing modeling paradigms.

5 Obs-Tmp Representations' Hierarchical Disentanglement

Let $X \in \mathbb{R}^d$ denote an observational variable, with instance sequences in data presenting its dynamics. Suppose a logical timeline t exists, along which, X's sequence can be denoted as $\{x_t\} = \{x_1, \ldots, x_t, x_{t+1}, \ldots, x_T\}$. We aim to devise a latent feature space $\mathbb{R}^L$ for two purposes: **1)** Establish representation of the observational-temporal (obs-tmp) variable $\mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1}$ in $\mathbb{R}^L$, fully capturing observational and dynamical features of the sequence $\{x_t\}$ in data. **2)** Disentangle this representation in $\mathbb{R}^L$ according to desired hierarchy.

Also, we want this hierarchical disentanglement to be indexed via relationships, reflecting different levels of knowledge. The advantage of this form is that the representations we establish, defined and encapsulated by recognized relations, are highly *reusable*, accommodating further structural modeling. Consequently, this method enables the *generalization* of the desired knowledge level across varied contexts.

Consider another observational variable $Y \in \mathbb{R}^b$, and suppose its dynamical sequence is along logical timeline τ, denoted as $\{y_\tau\}$. Given relationship $\mathcal{X} \to \mathcal{Y}$, the proposed ***relation-defined representation*** learning is to establish representation of $\hat{\mathcal{Y}} = \langle \hat{Y}, \tau\rangle \in \mathbb{R}^{b+1}$ in the latent space, which represents a new data sequence $\{\hat{y}_\tau\}$, versioned from $\{y_\tau\}$ by selecting its observational-temporal features that can only be caused by $\mathcal{X}$.

In structural modeling, for example, the causal system $\mathcal{X} \to \mathcal{Y} \leftarrow \mathcal{Z}$ can be disassembled into $\mathcal{X} \to \hat{\mathcal{Y}}$ and $\mathcal{Z} \to \tilde{\mathcal{Y}}$, allowing flexible reuse of both $\hat{\mathcal{Y}}$ and $\tilde{\mathcal{Y}}$ representations. Furthermore, if consider $\mathcal{Y}$ features a two-level hierarchy, with the first determined by $\mathcal{X} \to \mathcal{Y}$ and the second by $(\mathcal{X}, \mathcal{Z}) \to \mathcal{Y}$, we can also establish the representation of $\mathcal{Y}$ as hierarchically disentangled, where the second level builds upon the first, introducing an additional data stream from $\mathcal{Z}$.

Next, we will sequentially factorize three transformation processes between the data and latent space $\mathbb{R}^L$: First, observational features from $\mathbb{R}^d$ to $\mathbb{R}^L$; Second, observation-temporal features from $\mathbb{R}^{d+1}$ to $\mathbb{R}^L$; Last, the relationship $\mathcal{X} \to \mathcal{Y}$, from the joint space $\mathbb{R}^{d+1} \circ \mathbb{R}^{b+1}$ to $\mathbb{R}^L$.

5.1 Factorization of Observational Hierarchy

Assume $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ exhibits an n-level hierarchy. We utilize Θ_i to signify its i-th level observable feature in the data form, while its corresponding latent space feature is denoted as θ_i. Then, we obtain:

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

In this context, θ_i is considered as a vector in $\mathbb{R}^L$, in which only a subset of the L dimensions carries significant value, represented as the subspace $\mathbb{R}^{L_i}$. The hierarchical disentanglement is depicted by the arrangement of these subspaces: $\{\mathbb{R}^{L_1}, \ldots, \mathbb{R}^{L_i}, \ldots, \mathbb{R}^{L_n}\}$. Accordingly, the representation function f_i enables the i-th level transformation between $\mathbb{R}^d$ and $\mathbb{R}^{L_i}$, drawing its potential attributes from all preceding lower-level features.

To illustrate this factorization, consider Figure 2 (b). Let's say θ_1, θ_2, and θ_3 effectively represent the features spanning from Level I to III, exclusively occupying meaningful subspaces $\mathbb{R}^{L_1}$, $\mathbb{R}^{L_2}$, and $\mathbb{R}^{L_3}$. The image can be fully represented as an augmented vector $\langle \theta_1, \theta_2, \theta_3\rangle \in \mathbb{R}^L$. Correspondingly, Θ_1, Θ_2, and Θ_3 are full-sized images containing distinct content. Their cumulative representation can progressively render the complete image. E.g., Θ_1 isolates the fingers' details, while $\Theta_1 + \Theta_2$ broadens to depict the entire hand.

5.2 Factorization of Obs-Tmp Hierarchy

For the time sequence of X, denoted as $\{X_t\} = \{X_t \in \mathbb{R}^d \mid t = 1, \ldots, T\}$, it can be viewed as a d-dimensional vector extending T times, while we aim to represent it in a new form $\mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1}$ in $\mathbb{R}^L$, such that:

$$\mathcal{X} = \sum_{i=1}^{n} \Theta_i, \text{ where } \Theta_i = f_i\big(\theta_i;\ \Theta_1, \ldots, \Theta_{i-1}\big) \text{ with } \Theta_i \in \mathbb{R}^{d+1} \text{ and } \theta_i \in \mathbb{R}^{L_i} \subseteq \mathbb{R}^L \tag{2}$$

It is essential to note that each $\theta_i \in \mathbb{R}^{L_i}$ represents an observational-temporal feature, corresponding to the observable component $\Theta_i \in \mathbb{R}^{d+1}$. The latter can also be represented in its original time sequence form $\{\Theta_t\}_i = \{\Theta_{t_i} \in \mathbb{R}^d \mid t_i = 1, \ldots, T\}$. Hence, we have a collection of logical timelines $\{t_1, \ldots, t_i, \ldots, t_n\}$, *relative* to the *absolute* timeline t in this context. These n relative timelines, while may or may not be distinct, are each individually determined by their corresponding relationships. This allows each t_i to automatically adapt to the specifics of its associated relation, liberating us from recognizing potential timelines. For each level of observation, denoted by the summation $\Theta_1 + \ldots + \Theta_i$, the presented form is consistently an observable time sequence along t, without concerns about inherent temporal biases.

Consider the scenario in Figure 8, with the data matrix comprising four columns: values of A, B, C, and timestamp t. We can represent B as a 2-level hierarchy: the first level is defined by its direct effect from A, $A \to \hat{B}$, while the second level involves an additional stream denoting its indirect effect via C, $A \to C \to \tilde{B}$. The comprehensive effect representation $\hat{B} + \tilde{B}$ presents as two data columns, B and t, aligning with its original form. Thus, unlike traditional SCMs, this approach eliminates the need for specifying B_{30}.

5.3 Factorization of Relation-Defined Hierarchy

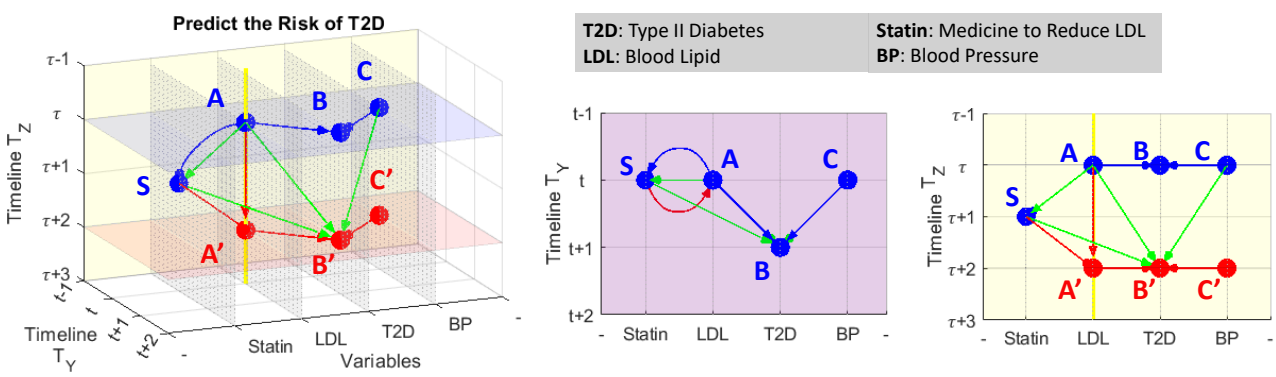


Figure 10: An example DAG in 3D observational-temporal space, where the SCM function $B' = f(A, C, S)$ aims to evaluate Statin's medical effect on reducing the risk of T2D, with two logical timelines $\mathcal{T}_Y$ and $\mathcal{T}_Z$. On $\mathcal{T}_Y$, the step Δt from t to $(t+1)$ allows A and C 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, changing A to A'.

Consider the relationship $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1}$ and $\mathcal{Y} = \langle Y, \tau\rangle \in \mathbb{R}^{b+1}$. Given a collection of n-level hierarchical representation functions for $\mathcal{X}$, denoted as $\mathcal{F}(\vartheta) = \big\{f_i\big(\theta_i\big) \mid i = 1, \ldots, n\big\}$, we aim to define the n relationship functions, collectively referred to as $\mathcal{G}$, to feature $\mathcal{Y} = \mathcal{G}(\mathcal{X})$ as n-level hierarchy. Let the i-th level relationship function be $g_i(\mathcal{X}; \varphi_i)$ with φ_i denoting its parameter. Then, we have:

$$\mathcal{G}(\mathcal{X}) = \sum_{i=1}^{n} g_i(\mathcal{X}; \varphi_i) = \sum_{i=1}^{n} g_i(\Theta_i; \varphi_i) = \sum_{i=1}^{n} g_i\big(\theta_i;\ \Theta_1, \ldots, \Theta_{i-1}, \varphi_i\big) = \mathcal{Y} \tag{3}$$

Therefore, the i-th level relation-defined representation of $\mathcal{Y}$ can be expressed as $g_i(\theta_i; \varphi_i)$ given the preceding $(i-1)$ levels observable features of $\mathcal{X}$ as possible attributes. In other words, the i-th level relationship can be represented as an augmented feature vector in the latent space $\langle \theta_i, \varphi_i\rangle \in \mathbb{R}^L$. Now, let's utilize ϑ_X and ϑ_Y to differentiate the collective hierarchy representations for $\mathcal{X}$ and $\mathcal{Y}$, respectively. The collective relationship $\mathcal{G}$ from $\mathcal{X}$ to $\mathcal{Y}$ can be represented as the latent space relationship $\vartheta_Y = \langle \vartheta_X, \varphi\rangle$, where $\varphi = \{\varphi_1, \ldots, \varphi_n\}$ and $\langle \vartheta_X, \varphi\rangle$ indicates the pairwise augmentations between collection ϑ_X and collection φ.

Let's consider a practical scenario, as depicted in Figure 10. From a differential-and-integral perspective, each differentiable unit of medical effect on LDL, delivered through $\overrightarrow{SA'}$, immediately begins influencing T2D via $\overrightarrow{A'B'}$. Concurrently, the subsequent unit effect is being generated. These two processes occur simultaneously until the medication S is fully administered. Ultimately, the B' we seek to evaluate embodies the total integral influence originating from S. Put another way, the emphasis in this DAG lies not on the absolute values of Δt and $\Delta\tau$, which would represent actual time spans. Rather, it is on the meaning of the nodes, which define what Δt and $\Delta\tau$ represent within this specific context.

In the traditional SCM function $B' = f(A, C, S)$, the variables only signify observational aspects of these nodes. Consequently, the role of B' requires a manual specification for its temporal aspect, to determine a time span for the edge $\overrightarrow{SB'}$. If, by empirical experience, the impact of S lasts around 30 ± 10 days, $\overrightarrow{SB'}$ is probably set to span 30 days, for representing the population-level mean effect at B'.

While, how long $\overrightarrow{SB'}$ spans is not crucial, the critical fact is that assuming a length for $\overrightarrow{SB'}$ essentially fixes the $\Delta t : \Delta\tau$ ratio (consider $\overrightarrow{SB'} = \overrightarrow{SA'} + \overrightarrow{A'B'}$) and consequently sets the shape of triangle ASB' for the entire population. As a result, patients within the ± 10 range risk being attributed an invalid ASB'-triangle and leading to inherent temporal biases, similar to Figure 8(b)(c). More significantly, even if the mean effect at B' is accurately estimated for this particular population, the derived SCM can hardly be generalizable to all scenarios, because the presumed $\Delta t : \Delta\tau$ ratio may not hold true in general.

Contrarily, the proposed relation-defined representation, focusing on the role of B' as dictated by $\overrightarrow{SB'}$, can incorporate its temporal features. This approach obviates the need to define time spans for $\overrightarrow{SA'}$ and $\overrightarrow{A'B'}$, thereby relieving us from concerning the $\Delta t : \Delta\tau$ ratio. Within the 3D observational-temporal space, as depicted in Figure 10, we can consider any desired model individualizations for patients or generalizations for other populations as linear transformations of the green-colored subgraph.

6 Relation-Defined Representation Methodology

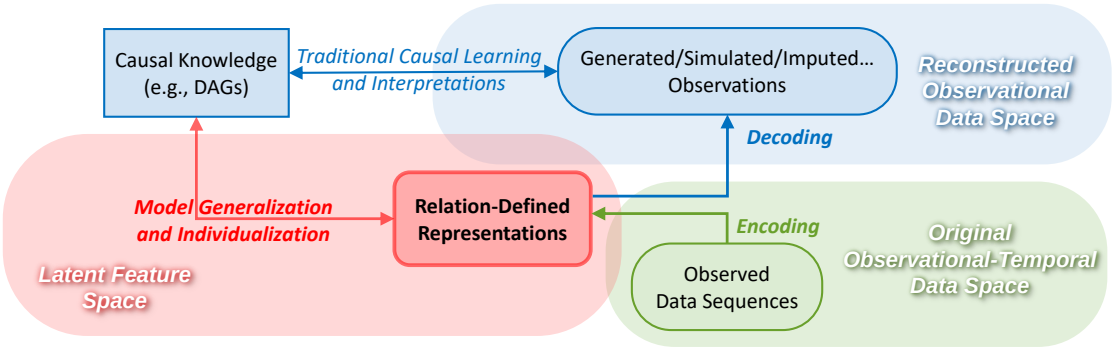


Figure 11: Framework for using *Relation-Defined Representations* to generate apt observations for traditional models, circumventing inherent temporal biases and enriching the observational modeling space.

Essentially, the proposed relation-defined representation is a technique to realize the observational-temporal entities in our cognition as features, allowing AI systems to understand and build models. The inherent temporal biases underscore the observational constraints of traditional methods, while concurrently recognizing their robust ties to our existing knowledge infrastructures. As illustrated in Figure 11, we enable AI to develop generalizable models in latent feature space filled with human-unreadable representations, while harnessing its capabilities to amplify observations for bolstering the effectiveness of traditional models.

In practice, the traditional way of addressing the inherent bias in Figure 10 is randomizing S's usage based on A, to disrupt their conditional dependence, which utilizes selective subpopulation but discards a significant portion. In contrast, the proposed method models $\overrightarrow{SB'}$ and $\overrightarrow{AB'}$ individually without manual adjustments, enabling the counterfactual effects. Conversely, it can simulate the de-confounded observations required by traditional models, to support knowledge-based studies in more comprehensive contexts.

In this section, we propose an autoencoder architecture for implementing relation-defined representation learning; based on it, outline the approach for "stacking" hierarchical levels of representations, to construct graphical models; and, lastly, present a causal discovery algorithm within the latent feature space.

6.1 Designing Higher-Dimensional Feature Representation Autoencoders

Autoencoders, primarily intended for dimensionality reductions Wang et al. (2016), commonly treat all variables (i.e., nodes in the Directed Acyclic Graph, DAG) as aligned observations to reduce attribute matrix dimensionality in structural modeling Luo et al. (2020). However, our focus is on modeling individual relations and "stacking" their representations to construct a DAG within the latent space $\mathbb{R}^L$. This demands a large dimensionality for $\mathbb{R}^L$ to accommodate all potential hierarchical features and to sequentially build L-dimensional node representations in the DAG, hence presenting a substantial technical challenge in facilitating *high-dimensional feature representations*.

Corollary 1. Given a graph G and a data matrix $\mathbf{X}$ column-augmented with all observational attributes of variables in G, along with a column for timestamps (i.e., the absolute timeline), the latent space dimensionality L must be at least as large as $rank(\mathbf{X})$ 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 as Principal Component Analysis (PCA) Baldi & Hornik (1989); Plaut (2018); Wang et al. (2016). Hypothetically, reducing L below $rank(\mathbf{X})$ could 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.

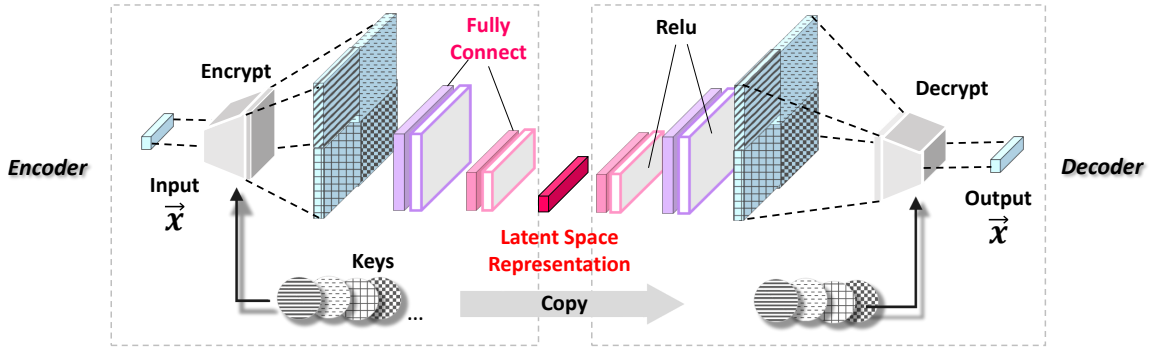


Figure 12: The proposed autoencoder architecture for *Higher-Dimensional* feature representations. The reconstructed observation $\vec{x} \in \mathbb{R}^d$ is denoted and depicted as a vector for clarity.

Figure 12 depicts the proposed autoencoder architecture, which employs symmetrical *Encrypt* and *Decrypt* layers at the input and output, respectively. The Encrypt layer acts as an amplifier, expanding the input vector $\vec{x}$ by extracting its higher-order intrinsic features. Conversely, the Decrypt layer, functioning as a symmetric reducer, restores $\vec{x}$ to its original form. To ensure reconstruction accuracy, the invertibility of these operations is naturally required. Figure 12 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 constants created by the encoder and mirrored by the decoder for reverse decryption. The application of a double-wise expansion *Key* on $\vec{x} \in \mathbb{R}^d$ generates a $(d-1)(d-1)$ length vector. By utilizing multiple *Keys* and augmenting the vectors they produce, $\vec{x}$ can be significantly extended beyond its original length d. The four differently patterned squares in Figure 12 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.

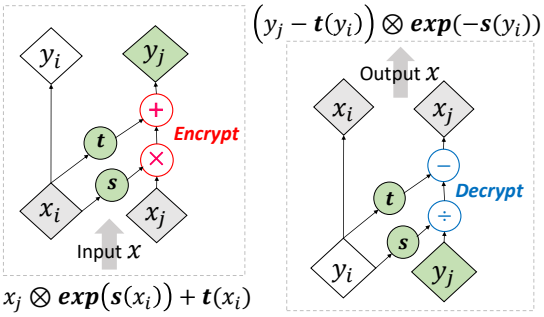


Figure 13: Encrypt (left) and Decrypt (right).

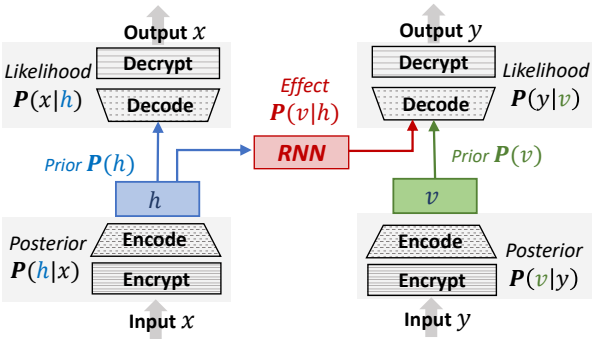


Figure 14: Architecture of Latent Effect.

Figure 13 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. 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 14 portrays the process of connecting $\mathcal{H}$ and $\mathcal{V}$ to represent the relation $\mathcal{X} \to \mathcal{Y}$, while Figure 15 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 14. 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 15 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

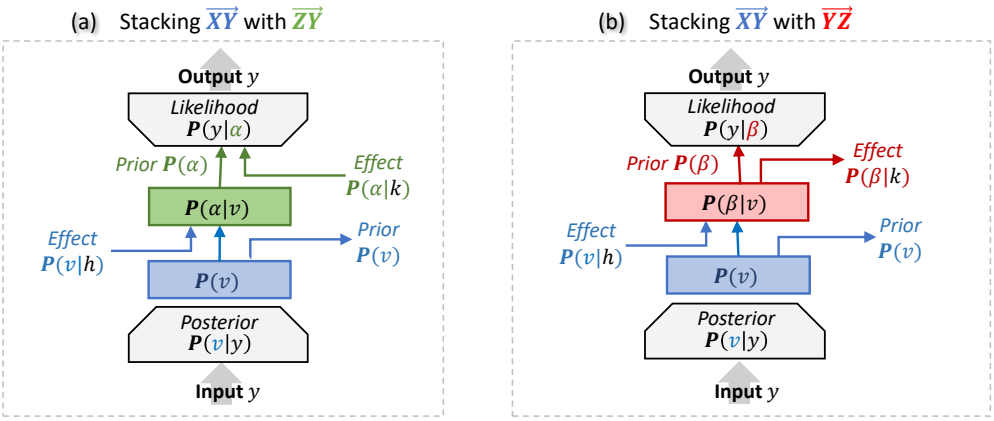


Figure 15: Architecutres of the Relation-Defined Hierarchical Representation.

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 14. 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 $A \rightarrow B$, 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 E = {e_1, ..., e_n}
E = {} ; N_R = {n_0 | n_0 ∈ N, Parent(n_0) = ∅} ;
while N_R ⊂ N do
    Δ = {} ;
    for n ∈ N do
        for p ∈ Parent(n) do
            if n ∉ N_R and p ∈ N_R then
                e = (p, n);  β = {};
                for r ∈ N_R do
                    if r ∈ Parent(n) and r ≠ p then
                        β = β ∪ r
                    end
                end
                δ_e = K(β ∪ p, n) − K(β, n);
                Δ = Δ ∪ δ_e;
            end
        end
    end
    σ = argmin_e(δ_e | δ_e ∈ Δ);
    E = E ∪ σ;  N_R = N_R ∪ n_σ;
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 16 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.

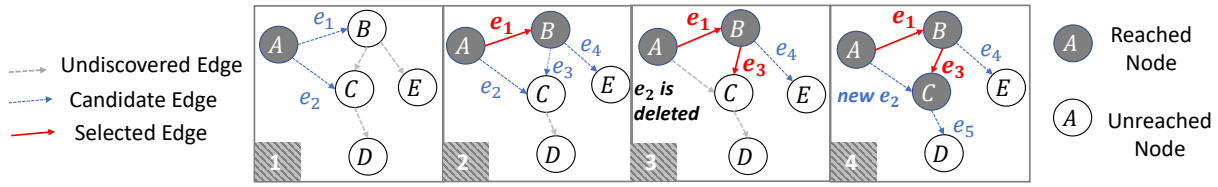


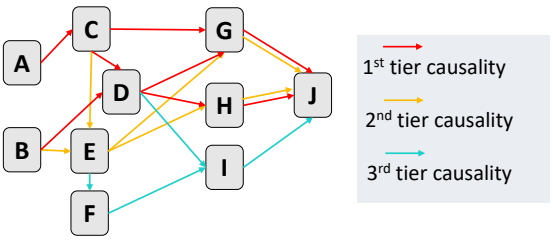
Figure 16: Example of the Latent Space Causal Discovery.

7 Experiments

The experiments aim to validate the efficacy of the proposed *Relation-Oriented* modeling methods 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 complex relationships within 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 17: 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 17 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 17, 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.

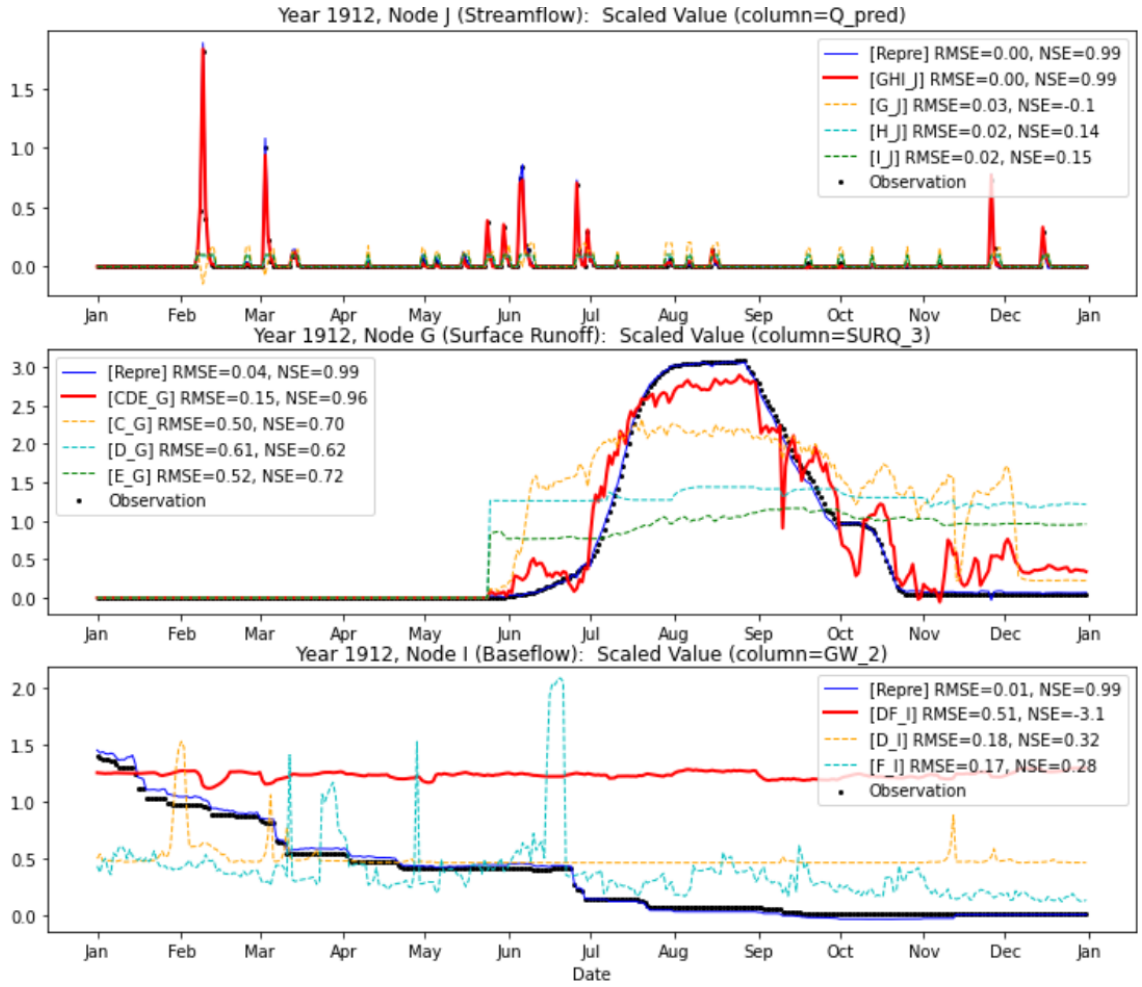


Figure 18: Time series simulation examples for the reconstruction performances' comparison.

Figure 18 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 the ground truth, and red lines hierarchical representations. We employ not only RMSE but also 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 18. 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 17, 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

Our traditional thinking about AI often assumes that incorporating larger, more diverse datasets can lead to more substantial advances. While that indeed powers AI but also creates more significant misalignments that deviate from our knowledge and cognition. As Christian (2020) highlights, AI alignment issues extend beyond technical issues, but point to humans' "blind spots" and "unstated assumptions". We might not differentiate ourselves from AI systems based on our capacity to understand large contexts, but rather on how we perceive the individual relationship.

This study advocates for a "*Relation-Oriented*" modeling principle, which could disrupt our prevalent *Observation-Oriented* modeling convention and align more closely with the relation-centric nature of human comprehension. To be specific, the knowledge nodes we construct in our cognition, are motivated by, and indexed through, relationships that confer them specific meanings under our comprehension context, rather than solely reflecting observations. Through relations, our understanding can surpass observational limitations, spanning: **1)** hyper-dimensions that encompass unobservable knowledge hierarchies, and **2)** temporal spaces where we store dynamic events, governed by multiple logical timelines.

Additionally, this study presents a feasible *Relation-Oriented* modeling technique, extracting relation-defined representations from observations, to instantiate the knowledge nodes in our understanding. AI alignment is never a question with a simple answer, but calls for our interdisciplinary efforts Christian (2020). Through this work, we aim to pave the way toward truly authentic AI and lay the groundwork for future progress.

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.

Irina & Pal Arka & Matthey Loic & Watters Nick & Desjardins Guillaume & Lerchner Alexander Burgess, Christopher P & Higgins. Understanding disentangling in vae. *arXiv preprint arXiv:1804.03599*, 2018.

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-Dickstein, 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				C→ F				D→ E	D→ F				F→ G	G→ H	G→ I	H→ I	
Times of Wrongly Discovered				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