arXiv:2603.04473v1 [stat.ML] 4 Mar 2026

基于词典的模式熵因果方向发现

Harikrishnan N B[1] **Shubham Bhilare**[1] **Nithin Nagaraj**[2] **Aditi Kathpalia**[3]

[1]计算机科学与信息系统系，BITS PilaniK K比尔拉果阿校区，NH 17B，旁路公路，Zuarinagar，403726，果阿，印度，harikrishnannb@goa.bits-pilani.ac.in，2025proj022@goa.bits-pilani.ac.in[2]复杂系统项目，国家高级研究所，印度科学研究所校区，班加罗尔，560012，卡纳塔克邦，印度，nithin@nias.res.in

[3]印度理工学院马德拉斯分校应用力学与生物医学工程系，印度金奈 IIT P.O.，邮编 600036，aditi@iitm.ac.in

摘要

从时间观测数据中发现因果方向，对于符号序列而言尤其具有挑战性，因为这类序列通常缺少函数模型和噪声假设。我们提出一种新颖的基于字典的模式熵（DPE）框架，用于推断因果方向以及驱动效应变量变化的具体子模式。该框架整合了算法信息论（AIT）与香农信息论。因果关系被解释为候选原因中紧凑的、基于规则的模式的涌现，这些模式能够系统地约束结果。 DPE 构建特定于方向的字典，并利用基于熵的度量量化其影响，从而在确定性模式结构与随机变异性之间建立有依据的联系。通过最小不确定性准则推断因果方向，即选择表现出更强且更一致的模式驱动组织结构的方向。如表7所总结， DPE 在多种合成系统中始终表现可靠，包括延迟比特翻转扰动、AR(1)耦合、一维偏斜帐篷映射和稀疏过程，其表现优于或不逊于其他基于AIT的方法（ ETC_E、 ETC_P、 LZ_P）。在生物和生态数据集中，该方法表现具有竞争力；而在特定基因组场景中，其他方法则展现出优势。总体而言，结果表明，最小化模式层面不确定性能够为因果发现提供一个稳健、可解释且适用范围广泛的框架。

1 引言

现代人工智能系统，尤其是最先进的机器学习模型，主要用于根据经验数据估计从输入特征到目标结果的函数映射 $f\colon \mathcal{X} \to \mathcal{Y}$。这些模型经过优化，能够捕捉统计规律和关联，从而最大限度地提升预测性能。然而，这类关联通常并不意味着存在因果关系。学习到的变量间函数依赖，并不能证明某一特征会对结果产生因果影响，也可能只是由于混杂或潜在因素而与结果相关。

识别因果关系通常需要进行干预。在标准因果推断框架中，干预查询通过 $do(\cdot)$ 算子形式化。具体而言，通过比较不同 x取值下的干预分布 $P(Y \mid do(X = x))$，可以刻画 X 对 Y的因果效应。随机对照试验（RCT）被广泛视为确立因果效应的黄金标准，因为随机化能够确保治疗分配与潜在结果之间的统计独立性，从而减轻混杂。然而，在许多领域，开展随机对照试验在技术上不可行、经济上不切实际，或在伦理上不允许。例如，为了检验吸烟对癌症发病率的影响而让个人吸烟，会违背基本的伦理原则。在这种情况下，研究人员必须依靠观测数据，例如比较吸烟者和非吸烟者的历史记录，来推断因果关系。

从观测数据中进行因果发现，旨在无需受控干预即可恢复底层因果结构的某些方面。用于从观测数据中进行因果发现的方法可大致分为以下几类：

- 基于图（基于约束与评分） [17； 2]
- 函数因果模型（FCM） [15； 7]
- 算法信息论（AIT） [1]

- 信息论（基于香农理论） [14]
- 时间与预测（时间序列） [5]
- 动力系统（状态空间） [18]

基于图的方法 [13], 以结构因果模型（SCM）为基础，旨在通过利用条件独立关系（基于约束的方法）或在可能的图的空间中优化全局模型评分（基于评分的方法）来恢复因果结构。与之密切相关但概念上有所不同的是函数因果模型（FCMs） [15； 7], 这类方法通过利用数据生成过程中的不对称性来识别因果方向，例如加性噪声模型中原因与噪声之间的独立性，或线性模型中的非高斯性假设。这两类方法通常都试图在明确的建模假设下重构因果图或有向关系，不过两种方法的假设和可识别性条件有所不同。

相比之下，若干框架无需明确指定结构方程即可推断因果方向。算法信息论方法以柯尔莫哥洛夫复杂度 [4; 23； 22] 和算法马尔可夫条件 [8], 为基础，后者断言，真实因果方向对联合分布的描述比反向因果方向更简短 [1]。在实际应用中，这一原则通过最小描述长度（MDL）表述 [6] 以及基于压缩的复杂度度量来实现。

代表性方法包括ORIGO，它通过最小化基于回归的描述长度来推断因果方向 [1]；ERGO，一种基于最小描述长度（MDL）原则的熵回归框架 [21]；以及压缩复杂度因果性（CCC），它利用压缩努力（Effort-To-Compress，ETC）来量化时间序列中的有向影响 [10]。相关方法采用Lempel–Ziv复杂度以及基于语法的压缩度量，通过差异可压缩性而非显式函数建模来检测不对称性 [19]。总体而言，这些方法将因果关系归因于能够对数据给出更简约描述的方向。

信息论方法量化有向统计依赖或有向信息流 [14], ，而时间与预测框架（如格兰杰因果关系） [5], 则根据按时间排序的观测中的非对称可预测性推断因果关系。最后，动力系统方法，包括状态空间重构和收敛交叉映射 [18],，利用耦合非线性系统的性质，从重构动力学中检测因果影响。综合来看，这些方法表明，基于观测数据的因果发现并非由单一统一框架所支配，而是反映了因果推断的多种互补视角。

尽管方法各不相同，所有基于观测数据进行因果发现的方法都依赖于一些约束可识别性的假设。图模型和函数模型对条件独立性、函数形式或噪声结构作出假设，而基于压缩的方法和信息论方法则依赖于对算法复杂度或信息流的近似。尤其是，信息论度量需要准确估计熵及相关量的概率 [11],，这通常需要较长的序列或较大的样本量。在数据稀缺或结构高度复杂的观测场景中，这类全局估计可能变得不可靠，从而限制其实际适用性。

在本研究中，我们提出了一种结合算法信息理论（AIT）与香农信息论（IT）的方法，用于从含噪观测数据中进行因果发现。

- **算法视角：** 受 AIT 启发，我们将因果关系视为“支配程序”，而非统计相关性。该方法提取重复子模式的字典，这些子模式代表驱动结果的潜在机制单元。本质上，我们旨在找到一种最精简的、基于模式的描述，说明变量如何相互影响或触发彼此。

- **信息论验证：** 为考虑真实数据中的噪声，我们使用香农熵来衡量所提取字典的可靠性。我们的度量指标 响应确定性（R-flip）量化了 X 中的子模式对 Y 变化的预测能力，取值范围为 0（无变化）到 1（确定变化）。0 和 1 均表示确定性。

总体而言，该框架揭示了数据中的确定性因果结构：影响值为 0 或 1 表示完全确定的结果，即诱导响应变量的熵为零（没有意外或不确定性）。中间值则表示仍存在不确定性，反映了这种关系中的随机成分。较低的不确定性（或较少的意外）表明该方向相对于反向方向具有更强的确定性影响。

本文其余部分安排如下。第 2 节介绍所提出的方法，并通过一个完整示例加以说明。第 3 节描述实验设置，并比较所提出的方法与其他基于 AIT 的因果发现算法的性能。第 4 节讨论结果及其意义。第 5 节概述当前框架的局限性以及未来工作方向。最后，第 6 节总结全文。

2 方法

本节将详细介绍所提出的基于字典的模式熵方法，用于从两个变量中进行因果发现。

2.1 基于字典的模式熵（DPE）

设 $X=(x_1, x_2, \ldots\ldots, x_N)$ 和 $Y=(y_1, y_2, \ldots\ldots, y_N)$ 表示两个长度相等的符号序列 N，其中每个符号 x_i 和 y_i 都取自有限字母表 $\mathcal{A}$（例如，二进制序列中的 $\mathcal{A}=\{0,1\}$）。本研究旨在直接根据观测到的符号模式推断 X 和 Y 之间的因果关系，而不假设明确的概率模型、函数形式或长序列渐近性。

我们提出了一种称为基于字典的模式熵（DPE）的因果推断框架，该框架利用字典构建和模式敏感的信息度量来刻画符号序列之间的因果不对称性。所提出的 DPE 框架有两个目标：

1. 通过识别由不对称模式驱动的信息变化，确定 X 和 Y之间的因果方向，即判断 X 中的变化是否会引起 Y中的结构性变化，反之亦然。

2. 在推断出的因果序列中识别并归因于特定模式，这些模式导致了结果序列中可观察到的符号转变，从而能够在模式层面归因因果强度。

本节重点介绍所提出的方法，并通过一个完整示例来理解为解决第2.1节所述问题而开发的方法。

2.1.1 因果方向识别

如果 X 中反复出现的有限模式能够系统性地决定 Y的转移行为，我们就称 X 对 Y 具有因果性。在此定义下，因果方向性可归结为：识别一个序列中的有限模式，这些模式一旦出现，就能确定性地固定另一个序列的转移行为。以下步骤将这一原则具体化，用于确定两个变量 X 和 Y之间的因果方向。文中还提供了一个完整示例进行说明。

设输入序列为二进制字符串 $X=(x_1, x_2, \ldots, x_n)$，$x_i \in \{0,1\}$。相应的输出序列记为 $Y=(y_1, y_2, \ldots, y_n)$，$y_i \in \{0,1\}$。序列 Y的构造旨在反映模式1101在 X中出现的情况，并允许重叠实例。也就是说，每当子序列1101出现在 X中时，便会在 Y的相应位置引发一次符号转变。形式上，输出定义为

$$y_i = \begin{cases} 1, & \text{if } (x_{i-3}, x_{i-2}, x_{i-1}, x_i) = (1,1,0,1), \\ 0, & \text{otherwise.} \end{cases}$$

因此，只要输入序列 X 中包含子序列1101，输出 Y 就取值为1，即使这些出现彼此重叠；否则，输出仍为0。考虑以下符号序列：

$$X = 01110111101001100111010110100 1$$

模式集定义如下：

$$\text{Patterns} = \{1101\}$$

相应的输出序列为：

$$Y = 00000100001000000000100001000$$

1. **步骤 1：初始化**。我们定义一个字典，将 X 中对应于 Y发生变化的位置的片段（或模式）存储为 $G_{X\to Y}$。类似地，我们将 $G_{Y\to X}$ 定义为一个字典，用于存储 X发生变化时 Y 的片段。初始时，两个字典均为空：

$$G_{X\to Y} = G_{Y\to X} = \{\}.$$

2. **步骤 2：构造** $G_{X\to Y}$

如果在位置 $k>1$处满足以下条件，则称 Y发生了一次比特翻转

$$y_k \neq y_{k-1}.$$

字典 $G_{X\to Y}$通过从左向右扫描 Y来构建。令 s 表示当前起始索引，初始化为 $s=1$。

每当位置 k发生比特翻转时，我们就提取子串

$$(x_s, x_{s+1}, \ldots, x_k)$$

X中的子串，其中明确包含与 Y中的比特翻转对齐的符号 x_k。该子串会被插入$G_{X\to Y}$。随后，起始索引更新为 $s=k+1$，并继续扫描，直到 Y 中出现下一次比特翻转。重复的子串在字典中只存储一次。以下面的说明性示例为例：

$$Y = 00000100001000000000100001000,$$

$$X = 01110111101001100111010110100 1.$$

我们用红色突出显示 Y 中发生比特翻转的子模式，以及记录在 $G_{X\to Y}$ 中的 X 对应子模式。

Y中的第一次比特翻转：

$Y =$ **000001** 00001000000000100001000

$X =$ **011101** 11101001100111010110100 1

Hence, $G_{X\to Y} = \{011101\}$.

Y中的第二次比特翻转：

Y = 000001 **00001** 000000000100001000
X = 011101 **11101** 001100111010110100 1
现在， $G_{X\to Y} = \{011101, 11101\}$。

Y中的第三次比特翻转：

Y = 00000100001 **00000000001** 00001000
X = 01110111101 **00110011101** 01101001因此，
$G_{X\to Y} = \{011101, 11101, 00110011101\}$。

Y中的第四次比特翻转：

Y = 000001000010000000000 1 **00001** 000
X = 011101111010011001110 1 **01101** 001因此，最终的字典为：$G_{X\to Y}$ = $\{011101, 11101, 00110011101, 01101\}$。总之，$G_{X\to Y}$对 X 中与 Y中的比特转换在时间上对齐的子模式进行编码。这个字典可以解释为从 X 到 Y的有向影响映射，描述了 Y的变化如何取决于 X的历史模式。

3. **步骤 3：构建** $G_{Y\to X}$
如步骤 2 中所述，每当 X 中发生比特翻转时，Y 的相应子序列就会存储在 $G_{Y\to X}$ 中。对于所考虑的示例， $G_{Y\to X}$ $=\{00, 10, 000, 10000\}$

4. **步骤 4：因果模式提取。**
对于 $G_{X\to Y}$ 中的每对子模式和 $G_{Y\to X}$ 中的子模式，我们执行基于同或的滑动比较，以识别高度相似的区域。形式上，给定两个长度分别为 n_1 和 n_2 的二进制序列 p_1 和 p_2 （$n_1 \le n_2$），我们将 p_1 从左到右滑过 p_2，并计算重叠比特之间的按位同或。如果在任意位置同或结果产生两个或更多连续的1，则提取 p_2 中对应的子序列，并将其作为潜在的因果子模式存储在 模式字典 中。数学上，两个比特 a 和 b 之间的同或运算定义为：

$$a \,\text{XNOR}\, b = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{otherwise.} \end{cases}$$

示例： 考虑两个子模式 $p_1 = 01101 \in G_{X\to Y}$和 $p_2 = 00110011101 \in G_{X\to Y}$。将 p_1在 p_2 上滑动并执行按位同或运算，可得到多个对齐。当两个或多个连续的同或运算输出为 1 时，我们从 p_2 中提取对应的子序列，并将其存入模式字典。下表展示了 p_1的滑动过程。

根据表 1a 和 1b，我们找出了两个子模式 $p_1 = 01101 \in$之间的公共子序列 $G_{X\to Y}$和 $p_2 = 00110011101 \in G_{X\to Y}$，依据的是异或非比较结果中连续的 1。将从这一对（p_1， p_2）中提取出的所有公共子序列的集合记为 $\mathcal{P}(p_1, p_2)$。形式化地，

$$\mathcal{P}(p_1, p_2) = \{\, s \subseteq p_1 \mid s \text{ occurs in } p_2 \text{ whenever XNOR}(p_1, p_2) \text{ yields two or more consecutive 1s} \,\}.$$

对于给定示例，$\mathcal{P}(p_1, p_2) =$ $\{0110, 01, 011, 1101\}$。对所有不同的子模式对（p_i， p_j）重复此过程，得到 $\in G_{X\to Y} \times G_{X\to Y}$与 $i \neq j$，我们将所有成对提取结果的并集定义为整体模式字典：

$$\mathcal{P}_{X\to Y} = \bigcup_{\substack{i,j \\ i\neq j}} \mathcal{P}(p_i, p_j).$$

因此，对于所考虑的示例， $\mathcal{P}_{X\to Y}$ = $\{01, 011, 0110, 011101, 11, 110, 1101, 11101\}$和 $\mathcal{P}_{Y\to X} = \{00, 000, 10\}$。

5. **步骤 5：响应确定性（R_{flip}）。**
对于每个提取出的常见模式，我们首先计算其在候选原因序列中的总频率（例如在分析 $X \to Y$ 时记为 X）。所有出现均予以考虑，包括重叠实例。

对于给定模式，我们将 X 中的每一次出现与 Y 中的对应窗口对齐，并检查在该对齐区域内 Y 是否发生符号转变（比特翻转）。设 N_{occ} = 为该模式在 X 中的总出现次数， N_{flip} = 为与 Y 中比特翻转相关联的对齐出现次数。随后将该模式的响应确定性定义为

$$R_{\text{flip}} = \frac{N_{\text{flip}}}{N_{\text{occ}}}.$$

该比率量化了 X 中特定模式的出现在多大程度上诱发 Y 中的符号转变。接近 1 的值表明该模式会持续触发 Y 的变化，而接近 0 的值表明它会持续保持 Y 的状态。中间值则反映部分影响或随机影响。

示例：考虑模式‘01’。
$X = 01110111101001100111010110100 1.$
‘01’出现了九次。Y 中对应的窗口如下所示 Y = 000001000010000000000010000101000。在 X 中‘01’的九次出现中， Y中只有四个窗口发生了变化，五个窗口没有变化。01 的 R_{flip} 为 4 9。下表（表 2）计算了$\mathcal{P}_{X\to Y}$中所有共有模式的全部 R_{flip} 。

表 1：通过滑动和异或非比较提取公共子序列的示意图，连续匹配以红色突出显示标为红色。

(a) $p_1 = 01101$ 在 $p_2 = 00110011101$上的滑动

	位位置										
p_2	0	0	1	1	0	0	1	1	1	0	1
幻灯片 0	0	1	1	0	1						
幻灯片 1		0	1	1	0	1					
幻灯片 2			0	1	1	0	1				
幻灯片 3				0	1	1	0	1			
幻灯片 4					0	1	1	0	1		
幻灯片 5						0	1	1	0	1	
幻灯片 6							0	1	1	0	1

(b) 每张幻灯片的异或非结果（1 ⇒ 位匹配）

	位位置										
p_2	0	0	1	1	0	0	1	1	1	0	1
幻灯片 0	1	0	1	0	0						
幻灯片 1		1	1	1	1	0					
幻灯片 2			0	1	0	1	1				
幻灯片 3				0	0	0	0	1			
幻灯片 4					1	0	1	0	1		
幻灯片 5						1	1	1	0	0	
幻灯片 6							0	1	1	1	1

表 2：模式统计，展示变化次数、无变化次数及比率（来自 $X \to Y$）。

模式	变化计数	无变化计数	比率
01	4	5	0.444
011	1	4	0.200
0110	0	2	0.000
011101	2	0	1.000
11	1	8	0.111
110	0	5	0.000
1101	4	0	1.000
11101	3	0	1.000

表 3：模式统计，显示变化数、无变化数及其比率（来源：$Y \to X$）。

模式	变化计数	无变化计数	比率
00	10	11	0.476
000	13	3	0.8125
10	3	1	0.75

同样，我们计算了 $\mathcal{P}_{Y \to X}$ 中所有共有模式的 R_{flip}（表 3）。

6. **第 6 步：计算加权熵。**

X 和 Y 这两个二元序列之间的因果关系，基于<style id='3'>最小不确定性原则进行推断。其核心思想是，真实的因果方向应表现出更具确定性的模式级影响，因此具有更低的不确定性。

对于每个提取的模式 $p \in \mathcal{P}$，我们计算<style id='2'>加权二元熵，用于量化与 p影响相关的不确定性。

模式 p的加权熵定义为

$$H_w(p) = W_p \, H_b(r_p) \quad \text{bits}, \tag{1}$$

其中，r_p（模式 p 的 R_{flip} 值）表示模式 p 的翻转贡献比率，而 $H_b(\cdot)$ 是二元熵函数

$$H_b(r_p) = -\big[r_p \log_2 r_p + (1-r_p)\log_2(1-r_p)\big]. \tag{2}$$

权重 W_p 表示模式 p 出现频率的归一化频率，其定义为

$$W_p = \frac{C_p}{N - L + 1}, \tag{3}$$

其中，C_p 是假定因果序列中模式p 出现的总次数，N 是序列长度，L 是模式的长度。量 $N - L + 1$ 表示假定因果序列中长度为 L 的所有可能子串的总数。为评估总体方向，我们计算模式集 $\mathcal{P}$上的<style id='7'>平均加权熵：

$$\bar{H} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} H_w(p) \quad \text{bits}. \tag{4}$$

通过比较汇总的不确定性来确定因果方向：

$$\bar{H}_{X \to Y} \quad \text{and} \quad \bar{H}_{Y \to X}.$$

¯得出较小 H 值的方向被推断为因果方向，因为它对应着较低的模式级不确定性和更强的确定性结构。

7. **步骤 7：方向比较与因果判定**¯使 H 取值更低的方向对应于最小不确定性，因而反映出对因果变量对效应变量之可预测影响具有更高的置信度。因此，该方向被推断为因果方向。

$$\text{Verdict} = \begin{cases} X \to Y & \text{if } \bar{H}_{X\to Y} < \bar{H}_{Y\to X} \\ Y \to X & \text{if } \bar{H}_{X\to Y} > \bar{H}_{Y\to X} \\ \text{Independent} & \text{if } \bar{H}_{X\to Y} = \bar{H}_{Y\to X} \end{cases}$$

我们用下面的演算示例来说明步骤 7 和步骤 8 所涉及的计算。让我们考虑前面示例中提到的 X 和 Y 的符号序列。

$$Y = 00000100001000000000100001000,$$

$$X = 01110111101001100111010110100 1.$$

我们使用以下公式:

$$H_w(p) = W_p \cdot (-[r_p \log_2(r_p) + (1 - r_p) \log_2(1 - r_p)])$$

为了说明该方法，我们计算 $H_b(r_p)$、$H_w(p)$、$\bar{}$和 $H_{X\to Y}$，并假设 X 为因果变量， Y 为效应变量。计算 $H_{Y\to X}$的$\bar{}$时，也可以遵循相同步骤。考虑模式 $p = 01 \in \mathcal{P}_{X\to Y}$。

1. **二元熵计算（$H_b(p)$）：** 对于模式 p =01，比率为 $r_p = 0.444$ （表 2 ）。 H_b(0.444) ≈ 0.991 比特

2. **频率权重计算（W_p）：** 计数为$C_p = 9$，总长度为 $N = 30$，模式长度为$L = 2$:

$$W_p = \frac{9}{30 - 2 + 1} = \frac{9}{29} \approx 0.310$$

3. **加权熵（H_w）：**

$$H_w(01) = W_p \cdot H_b(r) = 0.310 \cdot 0.991 = 0.307 \text{ bits}$$

对 $\mathcal{P}_{X\to Y}$ 中所有其余模式执行相同的计算。

$\bar{}$4. **平均加权熵（$\bar{H}$）：** 集合 $P_{X\to Y}$ 的总平均加权熵等于所有单个 H_w 值之和除以唯一模式的总数（$|P_{X\to Y}| = 8$）：

$$\bar{H}_{X\to Y} = \frac{1}{8} \sum_{p\in P} H_w(p)$$

$$\bar{H}_{X\to Y} = \frac{0.307 + 0.129 + 0 + 0 + 0.156 + 0 + 0 + 0}{8}$$

$$\bar{H}_{X\to Y} = 0.074 \text{ bits}$$

可以按照相同的步骤计算$\bar{H}_{Y\to X}$。

5.**因果判定** 对于所考虑的示例，$\bar{H}_{X\to Y} = 0.074$ 和 $\bar{H}_{Y\to X} = 0.411$，我们观察到:

$$\bar{H}_{X\to Y} < \bar{H}_{Y\to X}$$

因此，因果方向被识别为 $X \to Y$。

图 1 将模式转移结构可视化为一个有向网络。每个节点对应于从源序列中提取的一个唯一子串，其关联的加权熵量化了目标序列中诱导转移行为的不确定性。较低的熵值表示确定性控制，而较高的值则反映诱导转移中的变异性。双向构造（$X \to Y$和 $Y \to X$）可用于评估确定性结构中的不对称性，这构成了因果方向识别的基础。

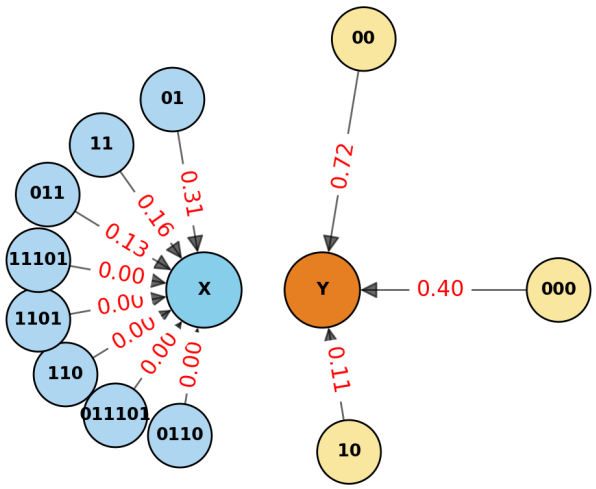

图1：符号模式的有向网络，展示 X 与 Y之间的确定性结构。节点表示从驱动序列中提取的不同子串，边则编码它们相对于目标序列的加权熵。加权熵值接近0表明该模式在目标中诱导出高度确定性的转移行为，而较大的值则表明诱导转移中的不确定性增加。该图同时展示两个方向 $X \to Y$和 $Y \to X$，从而能够比较方向确定性。

3 实验与结果

本节将讨论 DPE 因果框架的实验设置及其与文献中基于近期算法信息论的现有方法的比较性能分析，例如 Lempel-Ziv 惩罚（LZ_P） [19], 压缩努力效能（ETC_E） [19], 压缩努力惩罚（ETC_P） [19]。

表 4：模式的熵和权重（来自 $X \to Y$）

模式（p）	比率（r）	权重（W_p）	二元熵 $H_b(r_p)$（比特）	加权熵 $H_w(p)$（比特）
01	0.444	0.310	0.991	0.307
011	0.200	0.179	0.722	0.129
0110	0.000	0.074	0.000	0.000
011101	1.000	0.080	0.000	0.000
11	0.111	0.310	0.503	0.156
110	0.000	0.179	0.000	0.000
1101	1.000	0.148	0.000	0.000
11101	1.000	0.115	0.000	0.000
平均加权熵			$\bar{H}_{X\to Y}$	**0.074**

表 5：模式的熵和权重（来自 $Y \to X$）

模式（p）	比率（r）	权重（W_p）	二元熵 $H_b(r_p)$（比特）	加权熵 $H_w(p)$ **(比特)**
00	0.476	0.724	0.999	0.723
000	0.8125	0.571	0.696	0.397
10	0.750	0.138	0.811	0.112
平均加权熵			$\bar{H}_{Y\to X}$	**0.411**

3.1 延迟位翻转对模式因果方向的影响：

设序列 X 中重复次数最多的模式为模式1101，Y为序列 X中模式1101出现后发生延迟比特翻转的序列。两个序列的长度均为100。对于从0到6的每种延迟，我们生成了1000对序列。理想情况下，因果方向应从 $X \to Y$ 出发。

- **因果关系：** $X \to Y$
- **模式：** 1101
- **延迟：** $k = 2$

$$X = \ldots x, \mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}, x_{i+1}, x_{i+2}, \ldots$$
$$Y = \ldots y, 0, 0, 0, 0, 0, \mathbf{1}, y_{i+3} \ldots$$

比特串示例 x，$y \in \{0,1\}^*$:

$$X = \ldots \mathbf{1101} \ldots$$
$$Y = \ldots 0000\mathbf{00}1 \ldots$$

3.1.1 结果：延迟位翻转对模式因果方向的影响：

如实验 3.1 所述，本实验的真值为 $X \to Y$（X 导致 Y）。DPE 能够正确预测因果方向，平均准确率为在 k 从 0 到 6 的范围内达到 99%，LZ_p 表现出相似的结果，准确预测了 97.9% 的配对，这一点可以在图 2 中清楚地看到。虽然 ETC_P 仅能在 57% 的配对中正确预测因果方向，但 ETC_E在此类情况下完全失效。

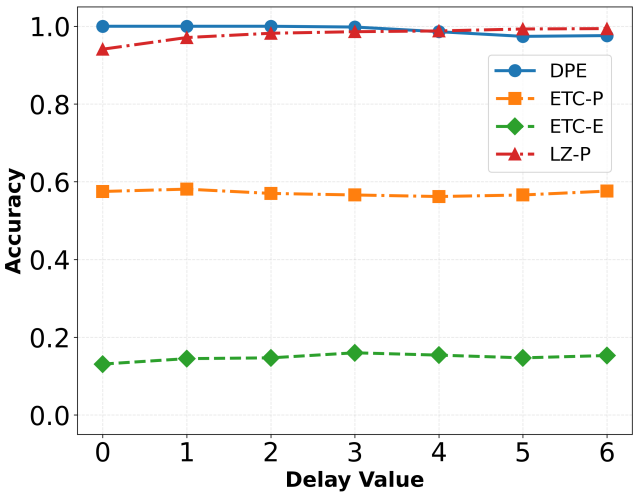

图 2：延迟位翻转对因果方向检测的影响：DPE、ETC_P、ETC_E 和 LZ_P 的准确率与比特翻转延迟的关系。

3.2 合成单向耦合

自回归（AR）模型广泛用于描述时变过程，这类过程线性依赖于自身过去以及其他过程的过去 [16]。我们使用 AR 模型进行单向因果推断，并按如下方式模拟一阶自回归过程（AR(1)）：

其中， X_t 和 Y_t 分别为因变量过程和自变量过程。

$$X_t = aX_{t-1} + \eta Y_{t-1} + \epsilon_{x,t}$$
$$Y_t = bY_{t-1} + \epsilon_{y,t}$$

X_t 和 Y_t 的系数固定为 $a = b = 0.8$。我们为 X_t 和 Y_t 生成了长度为的时间序列。耦合强度 ϕ 以 0.05 为步长，从 0 变化到 0.95，并加入了强度为 $\nu = 0.01$ 的加性噪声，该噪声取自标准正态分布。对于每个 ϕ值，我们生成了 2,000 次独立试验，每次试验包含长度为 1500 的 X_t, Y_t 对序列，并移除了前 500 个瞬态值。每个离散时间点都使用等宽分箱策略编码为二进制序列。随后，针对每次试验，使用全部四种模型估计因果方向性和强度，对总计 40,000 次试验进行了全面评估。

3.2.1 结果：AR 过程合成单向耦合：

对于 DPE 因果发现框架，当 $\phi = 0.05$ 为 0.20 时，模型开始表现更好，准确率从 55.15% 快速升至 91.95%。从 $\phi = 0.2$ 到 0.35，模型可靠性提升，准确率超过 98.5%。最后，当耦合强度达到 0.40 或更高时，模型的准确率几乎达到完美，在 99% 到 100% 之间。对于所有大于 0 的 ϕ 值， DPE 的表现优于 ETC_E和 ETC_P。 DPE 的表现与 LZ_P相近。

图4所示方向性熵度量的变化进一步解释了这一表现。图中绘出了¯ ¯平均熵 $H_{X\to Y}$和 $H_{Y\to X}$ 与耦合强度的关系。这两个熵值之间的差异证实，随着耦合强度增大，该模型识别真实因果驱动因素的能力也不断增强。

3.3 稀疏过程

令 $T = \{1, 2, \ldots, n\}$ 表示离散时间索引集合，其中 $n = 2000$。令 $Z_1(t)$ 和 $Z_2(t)$ 为定义在 $t \in T$ 上的两个随机过程。令 $T_1 \subseteq T$ 为一个子集，满足 $|T_1| = k$，其中 T_1 的元素从 T 中均匀随机无放回地选取。定义 $T_2 = \{t+1 \mid t \in T_1,\ t+1 \in T\}$。因此， T_2 由 T_1 中元素的紧随其后的时间索引组成，并限制在 T 范围内。令 $T_1^c = T \setminus T_1$ 和 $T_2^c = T \setminus T_2$ 分别表示 T_1 和 T_2 的补集。潜在过程定义如下：

$$Z_1(t) = \alpha Z_1(t-1) + \epsilon_1(t),$$
$$Z_2(t) = \beta Z_2(t-1) + \gamma z_1(t-1) + \epsilon_2(t),$$

其中， $\epsilon_1(t)$ 和 $\epsilon_2(t)$ 是相互独立的高斯噪声项，取自 $\mathcal{N}(0,\ 0.1)$。参数固定为 $\alpha = 0.8,\quad \beta = 0.08,$ $\gamma = 0.75$。观测到的稀疏序列 $z_1(t)$ 和 $z_2(t)$ 定义如下：

$$z_1(t) = \begin{cases} Z_1(t), & t \in T_1, \\ 0, & t \in T_1^c, \end{cases}$$

$$z_2(t) = \begin{cases} Z_2(t), & t \in T_2, \\ 0, & t \in T_2^c. \end{cases}$$

分析时，稀疏序列按以下方式进行二值化：

$$z_1^{(b)}(t) = \begin{cases} 1, & z_1(t) \neq 0, \\ 0, & z_1(t) = 0, \end{cases}$$

$$z_2^{(b)}(t) = \begin{cases} 1, & z_2(t) \neq 0, \\ 0, & z_2(t) = 0. \end{cases}$$

稀疏度参数 k 以5为步长，从5变化到50。对于 k的每个取值，均进行100次独立试验。

3.3.1 结果：稀疏过程

对于稀疏数据的二值离散化，如图5所示， DPE 在所有稀疏度取值下的表现均达到100%的准确率；而 LZ_P、 ETC_E和 ETC_P在预测大多数序列比例相互独立时，表现有所下降。

3.4 耦合一维偏斜帐篷映射

为评估该框架在非线性时间序列数据上的有效性，我们采用一维偏斜帐篷映射的耦合驱动变量—响应变量配置。映射函数 $T(x,\ b)$ 是一个分段线性混沌映射，定义如下：

$$T(x,b) = \begin{cases} x/b & \text{if } 0 \le x < b \\ (1-x)/(1-b) & \text{if } b \le x < 1 \end{cases}$$

用于生成驱动变量（D_t）和响应变量（R_t）时间序列的控制方程遵循单向耦合方案：

$$D_t = T_1(D_{t-1}, b_1)$$
$$R_t = (1-\eta)T_2(R_{t-1}, b_2) + \eta D_t$$

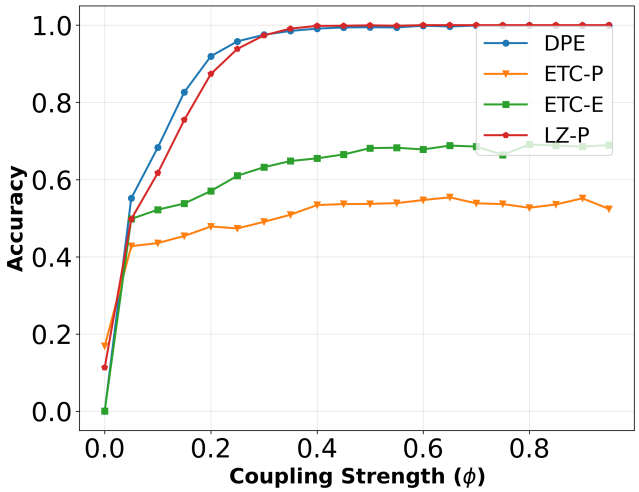

图3：合成单向耦合：DPE、ETC_E、ETC_P和LZ_P的准确率随耦合强度变化。

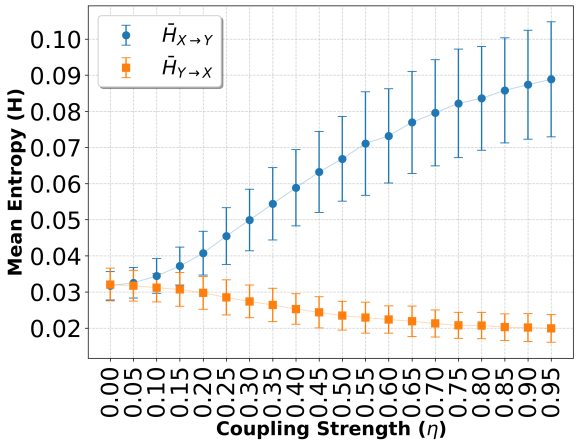

图4：合成单向耦合：对于 DPE 因果发现框架，平均熵（$H_{X\to Y}$和 $H_{Y\to X}$）随耦合强度（ϕ）变化的情况。两条曲线之间的间距不断增大，表明该框架区分驱动变量和响应变量的能力有所提升。

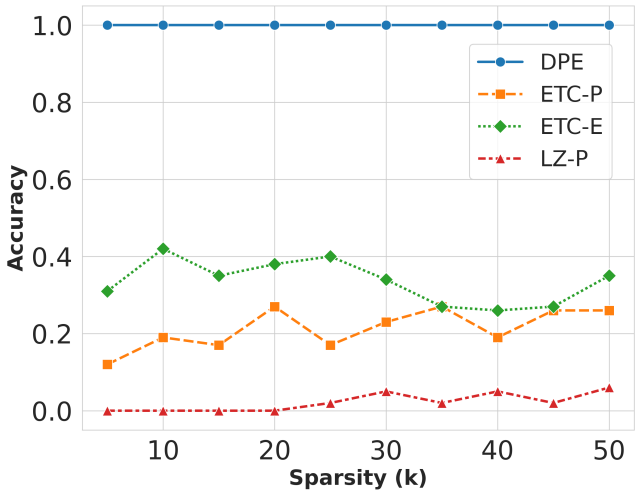

图5：稀疏过程：比较DPE、ETC_E、ETC_P和LZ_P在不同稀疏度（k）下的准确率。

其中，D_t代表驱动变量（原因），R_t代表响应变量（结果）。请注意，响应系统在时间t的状态会受到主系统同期状态D_t的影响。驱动系统的映射偏斜参数固定为$b_1 = 0.35$，响应系统的映射偏斜参数固定为$b_2 = 0.76$。耦合系数η从0.0变化到0.9，步长为0.1。对于η的每个取值，进行2000次独立试验，每次试验生成长度为1500的D_t和S_t时间序列数据。我们移除前500个瞬态值。每次试验的初始条件均从均匀分布$U(0,1)$中抽取。

3.4.1 结果：耦合一维偏斜帐篷映射：

DPE 在耦合强度（η）大于0时，能够检测出正确的因果方向，表现优于所有模型，总体准确率达到90%。而 LZ_P 的总体准确率为62.84%，优于 ETC_E（22.48%）和 ETC_P（49.88%）。从时间序列的角度来看，耦合系数为0.9时，时间序列会趋于同步，导致许多方法无法检测因果关系；但 DPE 仍能以100%的准确率检测出因果方向。

3.5 SARS-CoV-2基因组因果分析

本实验旨在评估，国家特定的初始序列（CW）还是全球参考序列（RS）SARS-CoV-2（GenBank登录号：$NC_045512.2$）[3]是导致后续国内病毒演化的主要原因。我们使用来自GenBank数据库中19个国家的17,567条高质量核苷酸序列构成的数据集[12]，检验零假设（H_0）：全球共识序列（RS）导致所有后续国内序列；并将其与备择假设（H_1）进行比较：某个国家境内报告的首条序列（CW）具有更强的影响力[19]。我们采用数值映射（A=1、C=2、G=3、T=4），在随机抽取的子集中评估来自中国、法国和印度等不同国家的基因组序列。

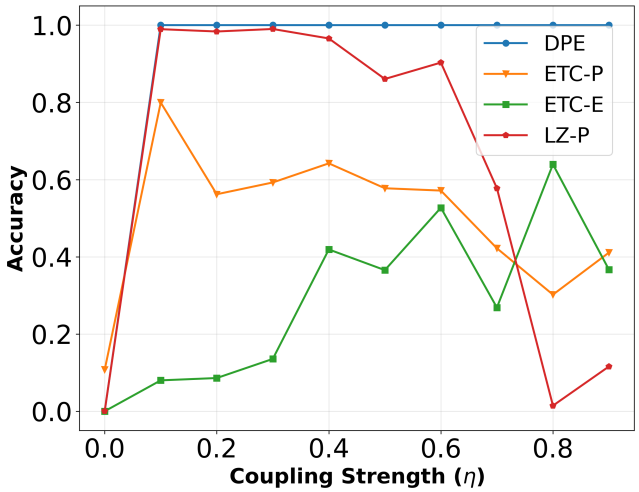

图 6：耦合一维偏斜帐篷映射： DPE、ETC_E、 ETC_P和 LZ_P在不同耦合强度下的准确率。

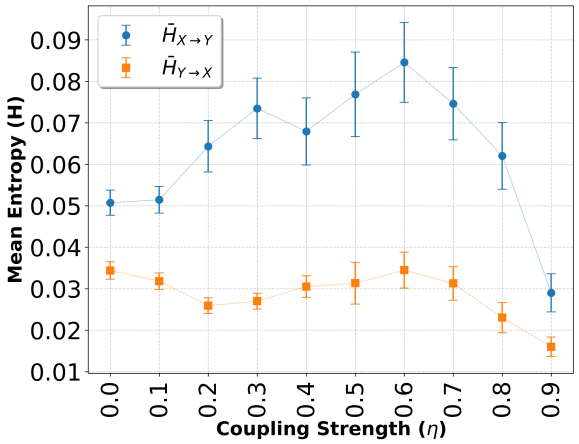

图 7：耦合一维偏斜帐篷映射：在耦合一维偏斜帐篷映射实验中，使用 DPE因果发现框架得到的平均熵¯ ¯值（$H_{X\to Y}$和 $H_{Y\to X}$）。

3.5.1 结果：SARS-CoV-2基因组因果分析：

我们分析了SARS-CoV-2共识序列（RS）是否“导致”所有其他序列。在19个国家中，16个国家对于DPE和 ETC_P，11个国家对于 ETC_E，以及19个国家对于 LZ_P，至少有5%的序列支持 H_0 假设所指示的方向。图8显示，在 ETC_P和 LZ_P模型下，各国观察到的比例通常较高。

按国家统计，由 CW“导致”的序列比例高于由 RS“导致”的比例。 DPE 显示，在 19 个国家中， ETC_E为 10 个国家、14 个国家，而 ETC_P和LZ_P均为 3 个国家。各国的比例见图 8。

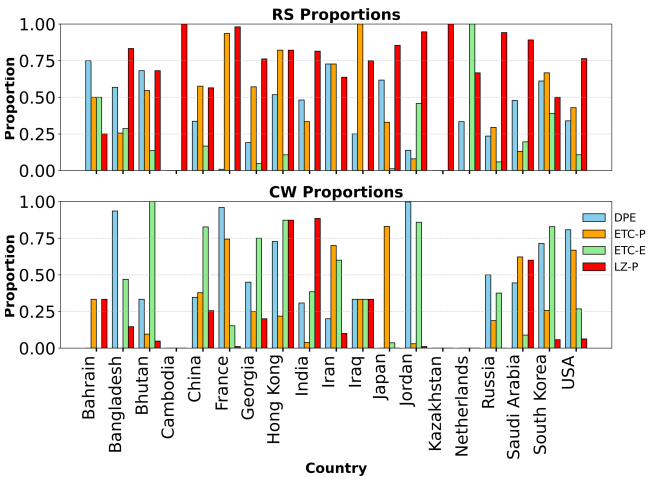

图 8：SARS-CoV-2 的基因组因果分析： DPE、 ETC_E、 ETC_P 和 LZ_P 对全局（RS）与局部（CW）进化驱动因素的敏感性比较总结。

3.6 捕食者-猎物系统

DPE 在一个真实世界的捕食者-猎物系统数据集 [20][9] 上进行了评估。该数据集包含 71 个数据点，记录了捕食者（Didinium nasutum）和猎物（草履虫，Paramecium aurelia）的种群（图 9）。捕食者种群直接影响猎物种群，随后自身又受到猎物种群变化的影响，这表明该系统存在双向因果关系。预计捕食者对猎物的因果影响方向强于相反方向。去除最初的 9 个暂态后，剩余的 62 个数据点用于本次分析。

3.6.1 结果：捕食者-猎物系统

DPE、 ETC_E和 LZ_P正确识别了从捕食者到猎物的主导因果方向。移除最初九组瞬态配对，并分析剩余的 62 组样本配对后， DPE 为 $Predator \to Prey$ 方向计算出的加权熵低于表 6 中所示的反向 $Prey \to Predator$，这表明目标方向上的因果确定性更高（$Predator \to Prey$）。所有模型都检测出，从捕食者到猎物的因果关系强度更高。

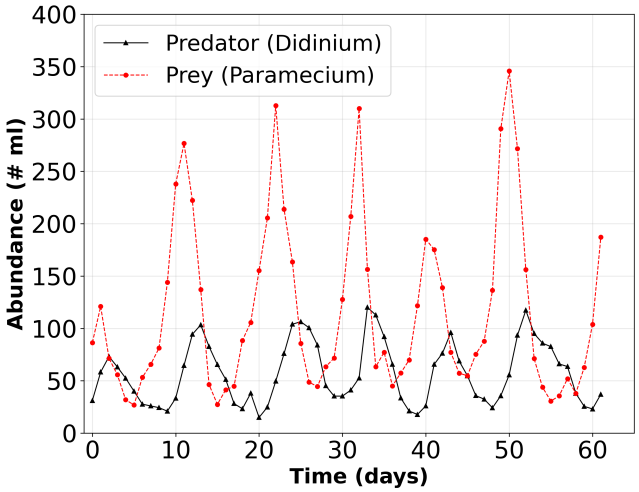

图 9：捕食者（Didinium nasutum）和猎物（Parameciumaurelia）的种群动态。

表 6：捕食者-猎物系统：不同模型下捕食者-猎物相互作用的因果指标比较。

模型	方向	强度	捕食者 → 猎物	猎物 →预测值
DPE	预测值 →猎物	0.1125	0.1700	0.2825
ETC_P	预测值 → 猎物	1.0000	0.0000	1.0000
ETC_E	预测值 → 猎物	0.0778	0.8000	0.7222
LZ_P	预测值 → 猎物	5.0000	2.0000	7.0000

4 讨论

表7总结了不同实验设置下模型的可靠性，而非报告确切的性能值。如果某个框架在多次试验中的平均准确率至少达到 80%，则认为该框架可靠。因此，该表将模型在每种实验条件下的表现是否稳定作为二元指标。这种呈现方式突出了各方法有望稳健发挥作用的领域，同时避免过度解读细微的数值差异。

表7：实验可靠性总结。勾号（✓）表示该框架在相应实验的各次试验中平均准确率达到 $\geq 80\%$，而（×）则表示未达到。该表反映的是可靠性，而非具体的性能值。

实验	DPE	ETC_E	ETC_P	LZ_P
延迟位翻转	✓	×	×	✓
AR(1)耦合	✓	×	×	✓
一维偏斜帐篷映射	✓	×	×	×
稀疏过程	✓	×	×	×
SARS-CoV-2（RS 与 CW）	×	×	✓	✓
捕食者-猎物	✓	✓	✓	✓

5 局限性与未来工作

本研究未明确考虑混杂变量的存在。具体而言，所提出的框架假设，观测到的非对称模式级影响直接源于候选原因。未来可进一步研究如何识别潜在混杂因素。一种可能的方向是检测在两个序列中均起共同触发作用的模式，这表明可能存在同时影响 X 和 Y 的未观测变量。

另一个有前景的方向是将该框架扩展为因果发现的反事实表述。具体而言，我们旨在探究从假定的原因序列中移除某个特定触发模式，是否会改变结果序列中观察到的转移。这样的表述有助于我们通过模式级干预评估因果影响，从而增强该方法的可解释性和稳健性。

对于所提出的 DPE 方法，在单向耦合实验和一维斜帐篷映射实验中，对应于 $\eta = 0$ 或 $\phi = 0$ 的情形均表示系统之间确实相互独立。然而，所提出的方法以及所有用于比较的基线方法都未能正确识别这一独立状态，反而检测出虚假的方向性影响。这一局限表明，用于推断独立性的当前判定规则仍需进一步研究。尤其是在有限样本情形或确定性混沌系统中，用于区分微弱因果影响与真正独立性的判据可能不够严格。为解决这一问题，未来工作应纳入替代数据分析和正式的统计显著性检验，以评估所推断因果方向的强度和可靠性。建立适当的零模型和假设检验程序，将有助于区分真实因果结构与由数据长度有限、共享动力学或底层映射的内在复杂性所产生的伪象。

6 结论

在本研究中，我们提出了一种新颖的基于字典的模式熵（DPE）框架，用于从时序观测数据中进行因果发现。除了推断因果方向外，所提出的方法还能够识别影响结果变量变化的可解释子模式（算法单元）。这种双重能力使 DPE 区别于现有的算法信息论方法。从概念上讲，该框架处于算法信息论（AIT）与香农信息论的交汇处。从算法角度来看，因果关系被解释为候选原因中简洁、基于规则的模式的涌现，这些模式系统性地约束结果。

通过构建重复出现的模式字典，捕捉这些算法结构。从信息论角度来看，基于熵的度量量化了各模式所蕴含的确定性，从而将结构化规则提取与观测数据中的随机变异性联系起来。表7汇总的比较结果阐明了 DPE 相较于其他基于AIT的方法（ETC_E、 ETC_P 和 LZ_P）的相对表现。以80%的平均准确率作为判定可靠适用性的阈值时， DPE 在合成动力系统（延迟位翻转、一阶自回归耦合、一维偏斜帐篷映射）和结构化稀疏过程中均表现稳定，在大多数受控设置下优于或持平于其他方法。值得注意的是， DPE 是所考察的合成实验中唯一在所有实验里都保持可靠性的方法。在生态数据中，各方法的表现更为接近；而在SARS-CoV-2突变分析中，其他基于模式的方法则展现出优于 DPE的竞争优势。总体而言，结果表明， DPE 能够在具有结构化模式传递和确定性影响的系统中实现稳健、稳定的因果检测，同时在真实场景中也具有竞争力。这些发现表明， DPE 是一个通用且可解释的因果发现框架，尤其适用于因果关系通过可识别的算法子模式而非纯粹的全局复杂性度量呈现出来的动力系统。

代码可用性

所有实验的源代码，包括所提出的 DPE 框架的实现，均可在 https://github.com/i-to-the-power-i/dpe-causal-discovery 上公开获取。

作者贡献

概念构思：Harikrishnan N B；方法：Harikrishnan N B、Aditi Kathpalia、Shubham、Nithin Nagaraj；软件：Shubham Bhilare（最终实现）、Harikrishnan N B（初始原型）；验证：Shubham Bhilare、Harikrishnan N B、Aditi Kathpalia、Nithin Nagaraj；形式分析：Harikrishnan N B、Shubham Bhilare、Aditi Kathpalia、Nithin Nagaraj；调查：Shubham Bhilare、Harikrishnan N B、Aditi Kathpalia、Nithin Nagaraj；撰写——原稿准备：Harikrishnan N B 和 Shubham Bhilare；撰写——审阅与编辑：Harikrishnan N B、Shubham Bhilare、Aditi Kathpalia 和 Nithin Nagaraj；经费获取：Harikrishnan N B。

致谢

Harikrishnan N. B. 谨此感谢总理早期职业研究资助项目提供的经费支持（项目编号：ANRF/ECRG/2024/004227/ENS）。

参考文献

[1] Kailash Budhathoki 和 Jilles Vreeken。通过压缩进行因果推断。载于 *2016* 年 *IEEE* 第 *16* 届国际数据挖掘会议（*ICDM*），第 41–50 页。IEEE，2016 年。

[2] David Maxwell Chickering。利用贪心搜索进行最优结构识别。机器学习研究期刊，3（11 月）：507–554，2002 年。

[3] 国际病毒分类委员会冠状病毒科研究小组。严重急性呼吸综合征相关冠状病毒物种：对2019新型冠状病毒进行分类并将其命名为SARS-CoV-2。《自然·微生物学》，5(4):536–544，2020年。doi: 10.1038/ s41564-020-0695-z。[4] Dhruthi、Nithin Nagaraj和Harikrishnan Nellippallil Balakrishnan。使用Lempel–Ziv复杂度进行因果发现与分类。《数学》，13 (20):3244，2025年。[5] Clive WJ Granger。通过计量经济学模型和交叉谱方法研究因果关系。《计量经济学：计量经济学会杂志》，37 (3):424–438，1969年。[6] Peter D Grünwald。最小描述长度原理。麻省理工学院出版社，2007年。[7] Patrik Hoyer、Dominik Janzing、Joris M Mooij、Jonas Peters和Bernhard Schölkopf。基于加性噪声模型的非线性因果发现。神经信息处理系统进展，21，2008年。[8] Dominik Janzing和Bernhard Schölkopf。利用算法马尔可夫条件进行因果推断。《*IEEE*信息论汇刊》，56(10):5168–5194，2010年。[9] Christian Jost和Stephen P. Ellner。检验捕食者-猎物动力学中的捕食者依赖性：一种非参数方法。《伦敦皇家学会会刊*B*辑：生物科学》，267(1453):1611–1620，2000年。[10] Aditi Kathpalia和Nithin Nagaraj。基于数据的干预方法：复杂度-因果性度量。《*PeerJ*计算机科学》，5:e196，2019年。[11] Nithin Nagaraj。信息论方法用于因果学习的问题。*arXiv*预印本 *arXiv:2110.12497*，2021年。[12] 美国国家生物技术信息中心。GenBank概览。https://www.ncbi.nlm.nih.gov/genbank/，2026年。访问日期：2025年11月24日。

[13] Leland Gerson Neuberg。因果关系：模型、推理与推断，Judea Pearl著，剑桥大学出版社，2000年。《计量经济学理论》，19(4):675–685，2003年。[14] Thomas Schreiber。测量信息传递。《物理评论快报》，85(2):461，2000年。[15] Shohei Shimizu、Patrik O Hoyer、Aapo Hyvärinen、Antti Kerminen和Michael Jordan。用于因果发现的线性非高斯无环模型。《机器学习研究杂志》，7(10)，2006年。[16] Robert H. Shumway和David S. Stoffer。时间序列分析及其应用：*R*示例。施普林格统计学丛书。施普林格国际出版公司，Cham，第4版，2017年。doi: 10.1007/978-3-319-52452-8。[17] Peter Spirtes、Clark Glymour和Richard Scheines。《因果、预测与搜索》。麻省理工学院出版社，纽约州纽约市，第2版，2000年。[18] George Sugihara、Robert May、Hao Ye、Chih-hao Hsieh、Ethan Deyle、Michael Fogarty和Stephan Munch。在复杂生态系统中检测因果关系。《科学》，338(6106):496–500，2012年。[19] Pranay Sy和Nithin Nagaraj。使用压缩复杂度度量进行因果发现。《生物医学信息学杂志》，117:103724，2021年。doi: 10.1016/j.jbi.2021.103724。[20] Brendan G. Veilleux。Didinium与草履虫之间捕食相互作用的分析。硕士论文，阿尔伯塔大学，埃德蒙顿，加拿大，1976年。[21] Jilles Vreeken。基于信息方向的因果推断。载于*2015*年美国工业与应用数学学会国际数据挖掘会议论文集，第909–917页。美国工业与应用数学学会，2015年。[22]Liang Wendong、Simon Buchholz和Bernhard Schölkopf。通过压缩涌现的算法因果结构。*arXiv*预印本 *arXiv:2502.04210*，2025年。[23]Sascha Xu、Sarah Mameche和Jilles Vreeken。按拓扑序进行信息论因果发现。载于第*28*届人工智能与统计国际会议，2025年。