arXiv:2601.02666v1 [cs.AI] 6 Jan 2026

推断因果图时序逻辑公式，加速时序扩展任务中的强化学习

Hadi Partovi Aria[1],**Zhe Xu**[1*]

[1]物质、运输与能源工程学院，亚利桑那州立大学，美国{hpartovi, xzhe1}@asu.edu

摘要

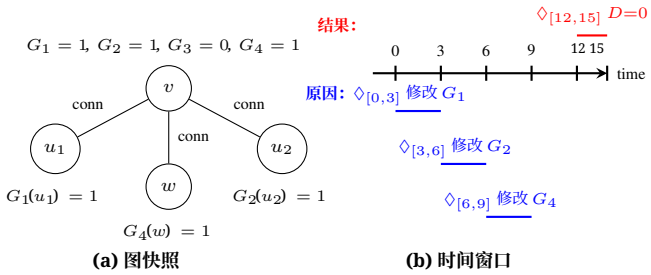
决策任务通常在具有时空动态特性的图上展开。黑箱强化学习往往忽视局部变化如何沿网络结构传播，从而限制了样本效率和可解释性。我们提出 GTL-CIRL，这是一种闭环框架，可同时学习策略并挖掘因果图时序逻辑（因果GTL）规范。该方法利用鲁棒性塑造奖励，在结果未能实现时收集反例，并采用高斯过程（GP）驱动的贝叶斯优化来优化参数化原因模板。GP模型能够捕捉系统动态中的时空相关性，从而实现对复杂参数空间的高效探索。基因网络和电力网络中的案例研究表明，与标准强化学习基线方法相比，该方法学习速度更快，且行为更清晰、更易验证。

引言

网络中的强化学习（RL）面临传播延迟的问题。标准马尔可夫决策过程（MDP）掩盖了这些依赖关系。我们提出因果图时序逻辑，以编码这种结构，从而实现高效学习。

相关工作：以往研究将因果推理或时序逻辑与强化学习相结合（Partovi Aria等人，2025；Bareinboim等人，2020；Aksaray等人，2016）。近期研究利用基于时序逻辑的因果图（Corazza等人，2024）或图时序逻辑（GTL）进行分类和综合（Xu等人，2019；Cubuktepe、Xu和Topcu，2020），但很少有研究将图结构提升为一等表示，或将因果公式与策略协同优化。我们在动态图中打通了探索与因果综合之间的闭环。

贡献： (1) 一种耦合学习框架（GTL-CIRL），能够同时发现因果GTL公式并训练策略，利用反例轨迹在图拓扑和时间两个层面逐步完善因果结构。(2) 为带有GTL鲁棒性奖励的Q学习收敛性，以及具有有界遗憾的因果模板贝叶斯优化提供理论保证。(3) 一种反例生成方法，通过有策略地探索因果关系的边界条件来提高公式精度。**示例动机：**考虑设计一种基因疗法来治疗一种复杂疾病。该疾病可能由生物单元（BU，即细胞）内特定组合的突变基因导致。一个简单的做法可能是只修正其中一个基因。然而，这种干预能否成功，往往取决于空间和时间背景。在空间上，BU的邻居可能需要处于有利状态，例如表达某些能够促成基因编辑成功的蛋白质。在时间上，可能需要按特定顺序、在精确的时间窗口内进行一系列编辑，疗法才能生效。图1展示了这种复杂性。中央BU v 具有导致疾病的基因模式。要使干预奏效，至少需要一个邻居提供空间支持，且该邻居的特定基因处于活跃状态。此外，还需要按时序进行一系列干预：先修改 G_1，再修改 G_2，最后修改G_4，每一步都需在不同的时间间隔内完成。只有当这一精确的时空事件序列发生时，预期结果——疾病缓解——才会随后显现。因果GTL提供了一种形式化语言，用于表达这类复杂规则；配备图时序逻辑的强化学习智能体则能够学习这些错综复杂的因果关系。



(a) 图快照 **(b) 时间窗口**

图 1：用于动机基因调控示例的图时序逻辑。(a) 一个生物单元 v 及其邻居u_1， u_2， w。(b) 时间窗口编码有序编辑（原因）和结果窗口（结果）。

预备知识

图时序逻辑：图时序逻辑（GTL）是一种用于表达随时间演化的网络（图）属性的形式化语言（Xu等人，2019）。它通过添加能够表达图中节点之间空间关系的构造，扩展了传统时序逻辑

*通讯作者 版权所有 © 2026，美国人工智能促进协会（www.aaai.org）。保留所有权利。

。原子命题与鲁棒性：原子节点命题 π 是形如 $f(x) \geq c$ 的谓词，其中 f 将节点标签映射为实数，$c \in \mathbb{R}$ 是阈值。对于离散状态（例如基因状态），我们将其映射为实数（例如，活跃为 1.0，不活跃为 0.0），或使用有符号距离度量。鲁棒度 $\rho(\tau, \varphi, t)$ 量化了轨迹 τ 在时间 t 满足 φ 的程度。对于 $\pi \equiv f(x) \geq c$，$\rho(\tau, \pi, t) = f(x_t) - c$。鲁棒性递归传播：$\rho(\tau, \neg\varphi, t) = -\rho(\tau, \varphi, t)$，$\rho(\tau, \varphi_1 \wedge \varphi_2, t) = \min(\rho(\tau, \varphi_1, t)\ \rho(\tau, \varphi_2, t))$。对于时序算子，$\rho(\tau, \Diamond_{[a,b]}\varphi, t) = \max_{t' \in [t+a, t+b]} \rho(\tau, \varphi, t')$，而 $\rho(\tau, \Box_{[a,b]}\varphi, t) = \min_{t' \in [t+a, t+b]} \rho(\tau, \varphi, t')$。我们使用 GTL 的 (F, G)-片段，仅采用有界最终（$\Diamond_{[a,b]}$）和始终（$\Box_{[a,b]}$）算子。这确保了任意公式的时域都是有限的。

设 $G = (V, E)$ 为一个无向图，其中 V 是有限节点集，E 是有限边集。我们用 $\mathcal{X}$ 表示节点标签集，用 $\mathcal{Y}$ 表示边标签集。图 G 上的图时序轨迹是一个元组 $g = (x, y)$，其中 $x: V \times \mathbb{T} \to \mathcal{X}$ 为每个时间索引 $k \in \mathbb{T}$ 下的每个节点 $v \in V$ 指定一个节点标签；$y: E \times \mathbb{T} \to \mathcal{Y}$ 为每个时间索引 $k \in \mathbb{T}$ 下的每条边 $e \in E$ 指定一个边标签。GTL 公式 φ 的语法可定义为：$\varphi := \pi \mid \exists^N(\bigcirc_{\rho_n} \cdots \bigcirc_{\rho_1})\varphi \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Diamond_{[a,b]}\varphi \mid \Box_{[a,b]}\varphi$，其中 π 是原子节点命题，ρ_i 是边命题，$\exists^N(\bigcirc_{\rho_n} \cdots \bigcirc_{\rho_1})\varphi$ 表示在邻居操作下至少存在 N 个满足 φ 的节点；$\neg$ 和 $\wedge$ 分别表示否定和合取；$\Diamond_{[a,b]}$ 和 $\Box_{[a,b]}$ 分别是有界最终和始终算子。

因果GTL： 因果GTL扩展了图时序逻辑，用于形式化网络中的因果关系，通常表示为：

$$\Phi := \mathrm{do}(\phi_c) \rightsquigarrow \phi_e \tag{1}$$

其中，ϕ_c 表示原因公式，ϕ_e 表示结果公式。算子 $\mathrm{do}(\phi_c)$ 表示一种干预，其中 ϕ_c 中的变量被强制设为满足 ϕ_c 的值，从而覆盖默认动态。这些公式使用图时序逻辑算子：$\Diamond_{[a,b]}\phi$（最终）表示 ϕ 在 $[a, b]$ 内某个时刻成立；$\Box_{[a,b]}\phi$（始终）表示 ϕ 在整个 $[a, b]$ 期间成立。为量化因果关系的强度，因果GTL基于经验数据引入了充分性、必要性和存在性的度量。充分性程度定义如下（Partovi Aria 和 Xu，2025）：

$$S(\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}_+|} \sum_{\tau \in \mathcal{D}_+} \rho(\tau, \phi_e, t \mid \mathrm{do}(\phi_c)) \tag{2}$$

其中，$\mathcal{D}_+$ 是数据集 $\mathcal{D}$ 中满足 $\rho(\tau, \phi_c, 0) > 0$ 的轨迹子集。必要性程度为：

$$N(\theta; \mathcal{D}) = -\frac{1}{|\mathcal{D}_-|} \sum_{\tau \in \mathcal{D}_-} \rho(\tau, \phi_e, t \mid \mathrm{do}(\neg\phi_c)) \tag{3}$$

其中，$\mathcal{D}_-$ 是子集，在该子集中，$\rho(\tau, \phi_c, 0) < 0$。存在度用于量化原因公式在数据集中出现的频率：

$$E(\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} e^{-(\rho(\tau, \phi_c(\theta), t) - \rho(\tau^*, \phi_c(\theta), t))} \tag{4}$$

其中，τ^* 是标称参考轨迹。此处，$\rho(\tau, \phi, t)$是鲁棒度，用于量化 τ 在时间 t满足ϕ的程度。

基于GTL目标的Q学习

Q学习是一种无模型强化学习算法，为了适用于GTL目标，需要通过奖励函数进行调整，使其与GTL满足准则相一致。为处理GTL规范固有的历史依赖性，系统被建模为一个 τ-MDP，其中每个状态 $s^\tau \in \Sigma^\tau$ 表示原始MDP中 τ个连续状态组成的序列，τ由内层公式 ϕ的时域决定。对于因果GTL规范 Φ，鲁棒度用于量化轨迹满足公式的程度。基于鲁棒性的奖励函数定义如下：

$$R(s^\tau_{t+1}) = \begin{cases} e^{\beta\rho(s^\tau_{t+1}, \phi)}, & \text{if } \Phi \text{ is } \Diamond_{[0,T]}\phi \\ -e^{-\beta\rho(s^\tau_{t+1}, \phi)}, & \text{if } \Phi \text{ is } \Box_{[0,T]}\phi \end{cases} \tag{5}$$

其中$\beta > 0$ 是一个缩放参数。对于形式为 $\Diamond_{[0,T]}\phi$的Φ，指数奖励会鼓励在轨迹中的某个时刻最大化鲁棒性。对于形式为 $\Box_{[0,T]}\phi$的Φ，负指数会惩罚较低的鲁棒性。随后针对这一 τ-MDP设置调整Q学习算法，以优化动作价值函数$Q(s^\tau, a)$，其更新规则如下：

$$\begin{aligned} Q(s^\tau_t, a) \leftarrow Q(s^\tau_t, a) + \alpha[R(s^\tau_{t+1}) \\ + \gamma \max_{a'} Q(s^\tau_{t+1}, a') - Q(s^\tau_t, a)] \end{aligned} \tag{6}$$

其中，$R(s^\tau_{t+1})$ 是下一τ状态的基于鲁棒度的奖励。

GTL-CIRL框架

本节介绍 GTL-CIRL，这一框架可联合优化强化学习策略和因果GTL公式。算法 2 概述了主要流程：智能体探索环境，通过 Q学习更新策略，同时收集反例轨迹。算法1分析这些轨迹，计算充分性（S）、必要性（N）和存在性（E）分数，以量化候选原因公式对观测结果的解释效果。在算法1中，π'_c 表示一种干预策略，它通过修改状态或动作序列来强制满足（或否定）原因条件 ϕ_c，从而生成反事实轨迹 τ'。这些指标为贝叶斯优化提供指引，以改进因果公式，形成策略学习与因果发现相互促进的反馈回路。

因果公式 $\phi_c(\theta)$ 的参数 θ（例如电网中的 0.90 等阈值或时间界限 $[a, b]$）是通过高斯过程优化的连续变量。高斯过程探索参数空间，以最大化目标 $J(\phi_c)$，使系统能够发现最能解释数据的精确取值。

系统会持续评估原因与结果公式的轨迹鲁棒性$\rho(\tau, \phi_c, t)$ 以及 $\rho(\tau, \phi_e, t)$。当一条轨迹违反结果公式（$\rho(\tau, \phi_e, t) \leq 0$）时，该轨迹会作为反例存储在缓冲区 $\mathcal{CE}$（算法 2，第 12 行）中。这些反例对于

算法1：评估充分性、必要性和存在性

初始化 空 列表 充分性分数， 必要性分数、存在性分数
对 $i = 1$ 到 I **执行**
  获取 $\tau \in \mathcal{CE}$
  在 $do(\pi_c')$ 下生成反事实 τ'
  计算 $\rho(\tau',\ \phi_c,\ t)$ 和 $\rho(\tau',\ \phi_e,\ t)$
  如果 $\rho(\tau',\ \phi_c) > \epsilon_{d_1}$ **则**
    将 $\rho(\tau',\ \phi_e)$ 添加到充分性分数中
  结束条件判断
  如果 $\rho(\tau', \phi_c) < -\epsilon_{d_2}$，**则**
    将 $\rho(\tau', \phi_e)$ 添加到必要性分数
  结束条件判断
  将 $\rho(\tau', \phi_c)$ 添加到存在性分数
结束循环
$S \leftarrow$充分性分数的均值
$N \leftarrow e^{-(\text{Mean}(\texttt{necessity scores}))}$
$E \leftarrow e^{-(\text{Mean}(\texttt{existence scores}))}$
返回$(S,\ N,\ E)$

计算指导公式细化的充分性、必要性和存在性度量至关重要。公式细化过程会优化目标函数：

$$\sup J(\phi_c) = -E + \lambda_S S + \lambda_N N \tag{7}$$

其中， S、 N和 E 分别表示充分性、必要性和存在性度量。贝叶斯优化通过维护目标函数的概率模型，引导搜索以改进原因公式。高斯过程模型使用径向基函数核：

$$k(x, x') = \exp\left(-\frac{1}{2l^2}||x - x'||^2\right) \tag{8}$$

其中， $l > 0$ 是长度尺度参数，用于确定函数的平滑度。

算法 2：GTL-CIRL

初始化 $Q(s^\tau,\ a) \leftarrow 0$、策略 π、高斯过程模型、 $\mathcal{C} \leftarrow \emptyset$
设置 $\phi_c \leftarrow \phi_c^0$
对于 $k = 1$至 K **执行**
  重置 $\mathcal{E}$，获取 s_0，初始化 $\tau_{\text{cur}} \leftarrow \emptyset$
  对于 $t = 0$至 $T - 1$ **执行**
    选择 $a_t \sim \pi(s_t^\tau)$（ϵ-贪心）
    执行 a_t，观察 s_{t+1}，更新 τ_{cur}
    计算 $\rho(\tau_{\text{cur}}, \phi_c, 0)$， $\rho(\tau_{\text{cur}}, \phi_e, 0)$
    计算奖励
    更新 $Q(s_t^\tau, a_t)$ 和 $\pi(s_t^\tau)$
    如果 $\rho(\phi_e, \tau) \leq 0$ **则**
      添加 τ_{cur} 到 $\mathcal{CE}$
    结束 if
  结束 for
  使用算法1计算 $S,\ N,\ E$
  $\phi_c^{k+1} \leftarrow \arg\max_{\phi_c}(-E + \lambda_S S + \lambda_N N)$
  使用 $S,\ N,\ E$ 更新高斯过程模型
结束循环
返回 $(\phi_c,\ \pi^*)$

反例生成

我们通过扰动结果违反附近的状态并展开轨迹，合成具有信息量的反例从这些扰动条件中得出轨迹。保留满足原因但破坏结果的候选项，以丰富缓冲区，为充分性/必要性/存在性估计及后续公式细化提供依据。这里，反例是指任何违反结果公式的轨迹 τ。这类轨迹位于因果关系的决策边界附近，因此最能为更新 ϕ_c的参数提供信息。将基于 GP 的优化集中于这些失败案例，可以更清晰地区分原因对结果真正具有充分性/必要性的情形与不具备这些性质的情形。具体而言，我们在当前状态向量 s_t 中，对单个状态变量 v_i（例如电压、负载或基因表达水平）施加小幅扰动范围 $\epsilon > 0$内的扰动。在算法3中，此过程具体实现为对每个 v_i 施加有限对称集合的偏移量 $\delta \in \{-\epsilon, \epsilon\}$，因此扰动集合是有限的，且其元素均被完整枚举。算法3详细说明了这一过程。

算法3：反例生成

要求： 当前状态 s_t、动作 a_t、公式 ϕ_c， ϕ_e、扰动范围 ϵ
确保： 反例集合 $\mathcal{CE}$
初始化 $\mathcal{CE} \leftarrow \emptyset$
$\tau_{\text{base}} \leftarrow$GetCurrentTrajectory()
如果 $\rho(\phi_e,\ \tau_{\text{base}},\ t) \leq 0$ **则**
  对于所有状态变量 v_i 在 s_t **执行**
    对所有 $\delta \in \{-\epsilon, \epsilon\}$ **执行**
      $s_t' \leftarrow$扰动状态(s_t, v_i, δ)
      $\tau' \leftarrow$ 模拟轨迹(s_t', a_t)
      如果 IsValidCounterexample(τ', ϕ_c, ϕ_e) **则**
        $\mathcal{CE} \leftarrow \mathcal{CE} \cup \{\tau'\}$
      结束如果
    结束循环
  结束循环
结束条件判断
返回 $\mathcal{CE}$

理论分析

本节为我们的GTL-CIRL框架提供理论保证。

定理1（GTL鲁棒性奖励下Q学习的收敛性）。 对于具有*GTL*奖励（式*5*）的有限 τ-*MDP*，在满足*Robbins-Monro*条件时， Q_k 以概率*1*收敛到 Q^* 。

概述。 M^τ在 $[-1, 1]$中是有限的，且奖励有界。贝尔曼算子 T 是 γ-压缩映射。根据随机逼近理论（Tsitsiklis 1994），$Q_k \to Q^*$。 □

定理 2（遗憾界）。 采用高斯过程置信上界进行选择时，简单遗憾 $r_K \leq \sqrt{\frac{C\beta_K\gamma_K}{K}}$，其中 $\gamma_K = O((\log K)^{d+1})$ 。

示意图。遵循 Srinivas 等人，2012，累积遗憾 $R_K \leq \sqrt{C\beta_K K\gamma_K}$。简单遗憾为 $r_K \leq R_K/K$。 □

实现与实验

基线方法

我们将 GTL-CIRL 方法与两种基线方法进行比较：**(1)标准强化学习**：不含任何因果或时序逻辑组件的原始 Q学习，仅依赖环境奖励；**(2) 反事实强化学习**：通过对假设情景建模来进行反事实分析，从而评估尚未探索的动作所获得的奖励。

基因调控网络

我们将 GTL-CIRL 应用于一个模拟的基因调控网络，其中的生物单元（BUs）具有四种基因状态（G_1、G_2、G_3、G_4）。目标是学习能够将疾病进展降低至零的干预措施。智能体必须满足一条隐藏的因果规则：识别具有特定突变模式的生物单元，验证邻居支持，并在规定时间执行基因修饰。其因果GTL公式为：

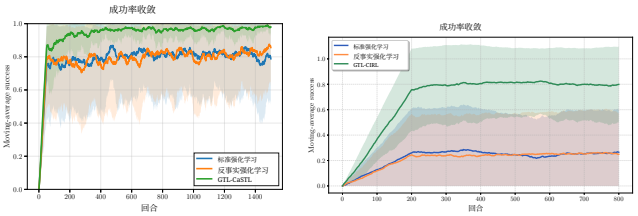
$$\begin{aligned}\Phi :=& \text{do}\Big(\exists v \in V : \Big[\Box_{[0,10]}(G_1(v)=1 \wedge G_2(v)=1 \\ & \wedge G_4(v)=1 \wedge G_3(v)=0) \wedge \exists^1(\bigcirc_{\text{conn}}(G_1(u)=1 \\ & \vee G_2(u)=1 \vee G_4(u)=1))\Big] \\ & \wedge \Diamond_{[0,3]}(\text{ModifyG}_1(v)=0) \wedge \Diamond_{[3,6]}(\text{ModifyG}_2(v)=0) \\ & \wedge \Diamond_{[6,9]}(\text{ModifyG}_4(v)=0)\Big) \\ & \rightsquigarrow \Diamond_{[12,15]}(\text{DiseaseProgression}=0)\end{aligned} \tag{9}$$

电网管理

为了在另一个领域评估我们的方法，我们基于 IEEE 14节点测试系统实现了一个电网管理环境。智能体必须通过控制发电机输出和负荷削减来维持整个网络的电压稳定性，同时避免级联故障和停电。该环境的因果结构通过以下因果GTL公式形式化：

$$\begin{aligned}\Phi :=& \text{do}\Big(\Box_{[0,0]}(V<0.90) \wedge \neg\exists^1(\bigcirc_{P>0})(V \geq 0.90)\Big) \\ & \rightsquigarrow \Diamond_{[1,5]}(G>0)\end{aligned} \tag{10}$$

该公式描述了在什么情况下，没有高电压邻居的低电压母线需要在一定时间窗口内增加发电量。图2显示，在这两个领域中，GTL-CIRL 都通过利用因果结构加速收敛，优于基线方法。



（a）基因调控网络　　（b）电网管理

图2：GTL-CIRL 在两个领域中的成功率与基线方法的对比。

结论

我们提出了 GTL-CIRL，将强化学习与因果GTL挖掘相结合。它利用鲁棒性奖励、反例和遗传编程驱动的优化，揭示时空结构，从而加快学习。在基因调控网络和电网管理领域开展的实验表明，与标准强化学习相比，该方法提高了效率和可解释性。

致谢

本研究部分由美国国家科学基金会资助，项目编号为 CNS 2304863、CNS 2339774、IIS 2332476，以及美国海军研究办公室项目 N00014-23-1-2505。

参考文献

Aksaray, D.；Jones, A.；Kong, Z.；Schwager, M.；以及 Belta, C.，2016。用于稳健满足信号时序逻辑规范的Q学习。载于*2016 IEEE* 第*55*届决策与控制会议（*CDC*），6565–6570。

Bareinboim, E.; Correa, J. D.; Ibeling, D.; 和 Icard, T. 2020. 关于 Pearl 的层级结构与因果推断基础。载于概率与因果推断：*Judea Pearl* 的著作，507–556。

Corazza, J.; Aria, H. P.; Neider, D.; 和 Xu, Z. 2024. 通过在环境中融入时序因果关系知识，加速强化学习。载于 Locatello, F.; 和 Didelez, V. 编，第三届因果学习与推理会议论文集，第 236 卷机器学习研究论文集，643–664。PMLR。

Cubuktepe, M.; Xu, Z.; 和 Topcu, U. 2020. 面向因子化马尔可夫决策过程的策略综合及图时序逻辑规范。arXiv:2001.09066。

Partovi Aria, H.；Kim, H.；Meshkat Alsadat, S.；以及 Xu, Z. 2025。《学习因果动力学与奖励机：利用扩展时序任务实现更快速强化学习的框架》。载于*2025* 年第五届国际计算机、控制与机器人会议（*ICCCR*），519–525。

Partovi Aria, H.；以及 Xu, Z. 2025。《挖掘因果信号时序逻辑公式，以实现利用时序任务的高效强化学习》。载于 Pappas, G.；Ravikumar, P.；以及 Seshia, S.A. 编，《style id='1'>神经符号系统国际会议论文集》，机器学习研究论文集第 288 卷，524–542。PMLR。

Srinivas, N.；Krause, A.；Kakade, S.M.；以及 Seeger, M. W. 2012。《老虎机设定下高斯过程优化的信息论遗憾界》。*IEEE* 信息论汇刊，58(5)：3250–3265。

Tsitsiklis, J. N. 1994. 异步随机逼近与Q学习。机器学习，16(3): 185–202.

Xu, Z.; Nettekoven, A. J.; Julius, A. A.; and Topcu, U. 2019. 用于分类与识别的图时序逻辑推断。arXiv:1903.09714.