TIME-AWARE NEURO-SYMBOLIC ABDUCTION WITH STRUCTURED CONFORMAL PREDICTION FOR BIOMEDICAL CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Unraveling deep mechanistic pathways underlying disease progression and drug action is a core pursuit of modern pharmacology and targeted therapeutics. However, transforming sparse and asynchronous clinical observations into statistically verifiable causal structures remains a long-standing challenge.Existing paradigms relying heavily on static knowledge graphs or large language models (LLMs) suffer from critical limitations: structural hallucinations due to insufficient causal constraints, search explosion in open-domain literature mining, rapid knowledge obsolescence for dynamic clinical scenarios, and the absence of distribution-free statistical safety bounds for causal predictions.In this paper, we introduce TRACE, a unified time-aware neuro-symbolic abductive framework integrated with structured conformal prediction, which transforms the passive candidate path validation to active open-domain causal discovery. Concretely, a time-aware hierarchical topological representation module that aligns sparse and asynchronous clinical observations into explicit causal subgraphs is first designed to resolve the mismatch between static knowledge representations and dynamic clinical data. To mitigate search explosion and semantic inconsistency in open-domain evidence retrieval, we propose an active state-graph backward search algorithm to guide multi-hop causal chain assembly across heterogeneous literature streams.Then, we incorporate macro-skill discovery and temporal update mechanism with continual learning objectives, which reduces online computational overhead and mitigates knowledge aging by continuously distilling verified reasoning trajectories.Furthermore, we establish a structured conformal prediction module tailored for multi-hop directed acyclic graphs, which defines path-level joint non-conformity scores via atomic triple decoupling and constructs distribution-free confidence bounds with finite-sample coverage guarantees, addressing the lack of statistical reliability in existing causal path predictions.We validate TRACE across four benchmark datasets: MIMIC, ADE-Time, CoMAGC, and FB15k-Subset. Experimental results demonstrate that TRACE achieves an absolute MRR gain of up to 10.4% over state-of-the-art baselines, while strictly maintaining the target 95% statistical coverage rate guaranteed by the conformal prediction module, reducing average prediction set sizes by over 50.0% via tighter confidence bounds, and accelerating inference speed by 54% through macro-skill reuse.

1 INTRODUCTION

Deciphering biomedical causal pathways is critical for precision medicine, target identification and pharmacovigilance (Himmelstein et al., 2017; Toonsi et al., 2024). Clinicians routinely encounter sparse, asynchronous patient observations such as lab anomalies and post-treatment adverse events. Translating these fragmented signals into verifiable biological mechanisms is vital for drug-safety monitoring, yet open-domain causal discovery from raw clinical data remains difficult.

As illustrated in Fig. 1, current systems face four core bottlenecks addressed by TRACE: (i) *Structural hallucinations*: LLMs and static knowledge graphs produce plausible yet biologically invalid causal chains due to confounder bias (Kícíman et al., 2023; Pearl, 2009; Ji et al., 2024). (ii) *Search-space explosion*: Multi-hop abductive backward reasoning from phenotypes expands combinatorially; static

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

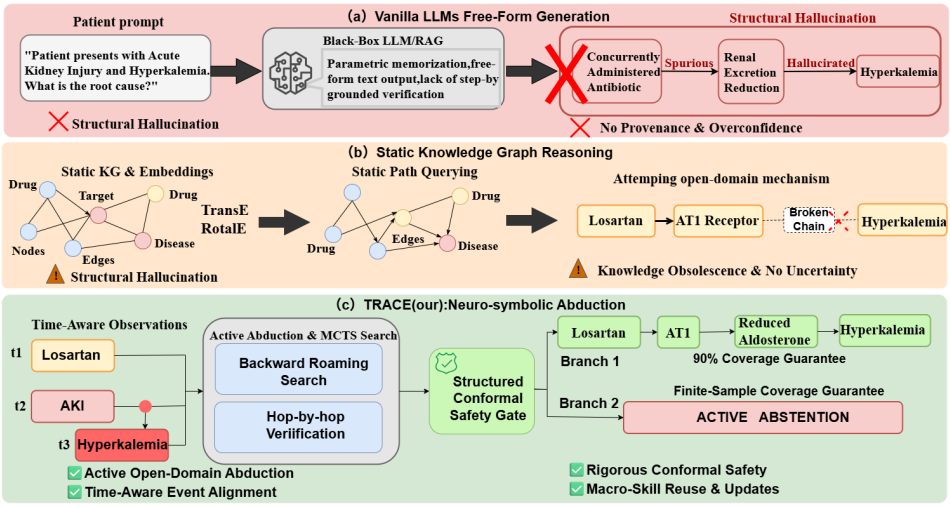


Figure 1: Comparison of biomedical causal discovery paradigms: (a) LLM generation prone to structural hallucinations; (b) static KG reasoning limited by incompleteness and obsolescence; (c) TRACE: neuro-symbolic active abduction with conformal safety and abstention.

knowledge graphs suffer relational incompleteness for unrecorded mechanisms (Bhagavatula et al., 2020; Du et al., 2021). (iii) *Knowledge obsolescence*: Biomedical evidence evolves continuously, while offline models lack dynamic updating and falsification of superseded relations (Radwan et al., 2026; Zhang et al., 2025b). (iv) *Absence of formal safety bounds*: Conventional uncertainty estimates lack distribution-free finite-sample guarantees for causal graph outputs (Abdar et al., 2021; Angelopoulos & Bates, 2021).

Recent work, CaPeR (Anonymous, 2025), performs hop-by-hop literature verification for pre-existing KG candidate paths but is restricted to passive auditing: it cannot explore unknown upstream causes, nor model temporal dynamics or provide conformal statistical guarantees. TRACE overcomes these limits by enabling active abductive discovery with formal safety constraints.

The contributions of this work are summarized below:

1. **Time-aware topological alignment**: Builds local causal subgraphs from asynchronous clinical inputs using temporal-decay clustering and KG-infused Bayesian structure learning, grounding subsequent abductive reasoning.
2. **MCTS-guided backward exploration**: Combines Monte-Carlo Tree Search, UMLS concept normalization and cross-literature transfer priors to assemble multi-hop causal trajectories over open-domain literature.
3. **Macro-skill continual dynamization**: Encapsulates verified recurring reasoning sub-paths as reusable macro-skills; time-decay reweighting distills evidence to lower inference latency and counter knowledge obsolescence.
4. **Structured conformal safety gate**: Defines path-level non-conformity scores via atomic triple decoupling for multi-hop causal DAGs, delivering distribution-free coverage guarantees and active abstention for unsafe outputs.

2 Related Work

2.1 Biomedical Knowledge Graph and Causal Reasoning

Biomedical knowledge graphs (BKGs) integrate drugs, genes, diseases and phenotypes to support causal analysis for pharmacovigilance and drug repurposing (Li et al., 2024a). Multiple public resources such as Hetionet, UMLS, DrugBank, DisGeNET and CTD aggregate multi-source biomedical relational data (Himmelstein et al., 2017; Wishart et al., 2018; Piñero et al., 2020; Davis et al., 2021).

Knowledge-graph embedding methods such as RotatE learn low-dimensional relational embeddings but encode statistical correlation rather than true causality, yielding high false-positive rates for biomedical causal tasks (Sun et al., 2019; Yao et al., 2021; Zheng et al., 2023). Graph neural networks improve relational representation but are built on correlation-driven training objectives and homophily assumptions, and cannot guarantee directed causal inference.

Retrieval-augmented LLM systems including G-Retriever, GraphRAG and MedRAG combine external knowledge with generative reasoning (He et al., 2024; Edge et al., 2025; Zhao et al., 2025). However, retrieval and reasoning are decoupled, preventing hop-by-hop cross-evidence validation for multi-hop causal chains.

Neuro-symbolic methods integrate symbolic constraints with neural models to boost interpretability (Garcez et al., 2019; D'Amato, 2020). Chain-of-Knowledge (Li et al., 2024b), Paths to Causality (Susanti & Färber, 2025) and domain-specific biomedical neuro-symbolic pipelines improve reasoning, yet remain confined to reasoning over pre-stored KG content and cannot discover unmapped mechanisms. CaPeR (Anonymous, 2025) verifies predefined candidate paths via literature evidence but lacks active exploration, temporal modeling and conformal safety. TRACE moves beyond passive verification toward open-domain abduction.

2.2 Abductive and Causal Discovery

Abductive reasoning infers the most plausible root cause for given observations and is central for biomedical mechanistic discovery (Bhagavatula et al., 2020). Existing LLM abduction work focuses on hypothesis generation (Zheng et al., 2025; Gao et al., 2025; Prenosil, 2024), yet seldom performs closed-loop literature validation and often treats timestamped clinical events as static snapshots.

Traditional causal discovery algorithms (PC, LiNGAM) reconstruct DAGs from observational data but require strong distributional assumptions and large sample sizes that are hard to satisfy for clinical datasets (Glymour et al., 2019; Shimizu et al., 2006). MCTS- and RL-based graph search methods improve structure learning (Zhu et al., 2020; Pei et al., 2022) but suffer efficiency bottlenecks for sparse high-dimensional clinical observations. Query-driven KG reasoning systems solve fixed logical queries but are not designed for open-ended abduction with unknown root causes (Ren et al., 2020; Yasunaga et al., 2021).

2.3 Continual Learning and Skill Discovery

Continual learning mitigates catastrophic forgetting when acquiring new knowledge (French, 1999; Kirkpatrick et al., 2017). Biomedical continual-learning frameworks such as CMKL and MedKGent adapt to evolving knowledge graphs (Radwan et al., 2026; Zhang et al., 2025b).

Skill discovery in reinforcement learning compresses repeated trajectories into reusable macro-actions to reduce long-horizon decision cost (Sutton et al., 1999; Rho & Smith, 2025). For LLM agents, skill discovery accelerates tool use and task decomposition (Shinn et al., 2023). Nevertheless, skill discovery for evidence-grounded biomedical graph reasoning remains underexplored. TRACE realizes macro-skill encapsulation for verified causal reasoning paths paired with time-decay knowledge cleansing to discard superseded biomedical assertions.

3 Methodology

TRACE consists of four tightly coupled components: (1) base verification engine for literature-grounded path scoring; (2) Stage I: time-aware hierarchical topological representation to construct causal subgraphs from asynchronous clinical observations; (3) Stage II: MCTS-driven backward abductive search for open-domain multi-hop causal trajectory assembly; (4) Stage III: macro-skill discovery and temporal dynamic update to reduce inference overhead and handle knowledge obsolescence; (5) Stage IV: structured conformal prediction with active abstention providing distribution-free statistical safety guarantees. The overall architecture is illustrated in Fig. 2.

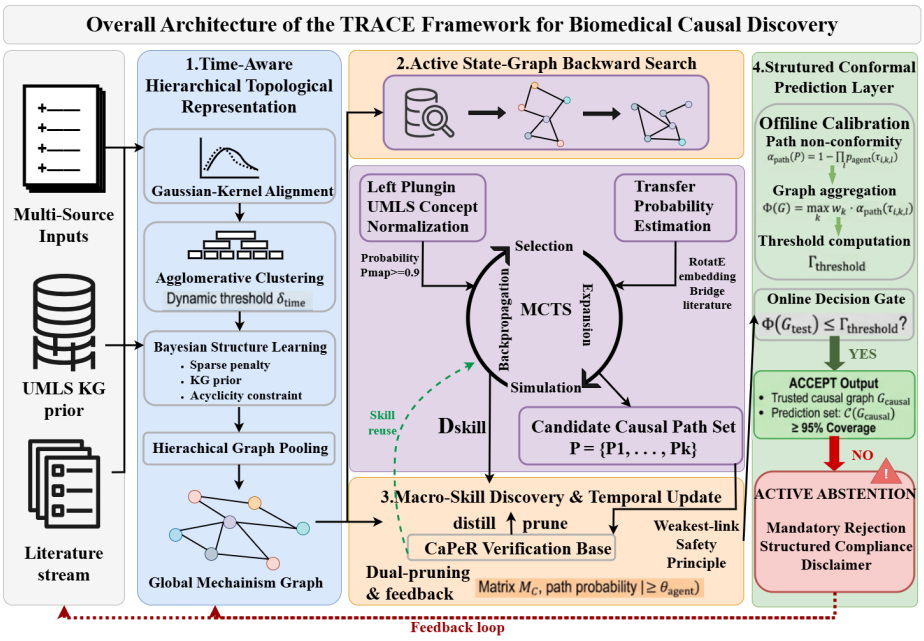


Figure 2: Overall architecture of TRACE, integrating time-aware representation, MCTS backward search, macro-skill dynamization and structured conformal safety gate.

3.1 Base Verification Engine: Dual-Pruning and Hard Negative Feedback

The base verifier supplies evidence-aware path scores used for MCTS rollout reward calculation and conformal non-conformity evaluation. Given N candidate causal paths and M retrieved literature snippets, we build a cross-modal consistency matrix $M_C \in \mathbb{R}^{N \times M}$ via bilinear attention:

$$M_C(i,j) = \text{softmax}\left(\frac{h_{topo}(i)^T W_A h_{sema}(j)}{\sqrt{d}}\right) \tag{1}$$

where $h_{topo}(i)$ denotes topological embedding of path i, $h_{sema}(j)$ is semantic embedding of literature snippet j, and W_A is learnable projection matrix.

Bidirectional dual pruning operates over this matrix: paths with high topological plausibility but insufficient literature support are filtered as structural false positives; paths with moderate topological ranking but strong empirical evidence are retained. Hard-negative samples (high-ranked paths rejected by literature auditing) are fed back to fine-tune the verifier with amplified contrastive loss:

$$\mathcal{L}_{feedback} = -\log \frac{\exp(S_{pos}/\tau)}{\exp(S_{pos}/\tau) + \beta \cdot \exp(S_{hardneg}/\tau)} \tag{2}$$

$\beta \geq 1$ is hard-negative penalty weight, τ temperature parameter. The trained verifier outputs atomic triple verification probability $p_{agent}(\tau)$ for each causal triple τ, which serves as the fundamental scoring primitive for downstream modules.

3.2 Stage I: Time-Aware Hierarchical Topological Representation

Raw clinical observations are sparse, asynchronous and irregularly sampled. Direct static graph modeling ignores temporal precedence and yields spurious causal edges from mere co-occurrence. Our Stage I transforms timestamped observation set $O = \{(o_i, t_i)\}_{i=1}^{N}$ into a global observational causal graph G_{obs} in three steps: temporal-semantic similarity calculation, event clustering, local Bayesian DAG reconstruction and hierarchical graph merging.

Table 1: Algorithm 1: Stage I: Time-Aware Topological Representation Construction

Require: Timestamped observations $O = \{(o_i, t_i)\}_{i=1}^{N}$, prior W_{prior}, bandwidth σ_t, threshold δ_{time}, penalty λ
Ensure: Global observational causal graph $G_{obs} = (V_{obs}, E_{obs})$

Step 1: Temporal-Semantic pairwise similarity
  for each event pair $1 \le i < j \le N$
    $S_{time}(i,j) \leftarrow \exp\left(-(t_i - t_j)^2/(2\sigma_t^2)\right)$
    $S_{sem}(i,j) \leftarrow \cos(\text{PubMedBERT}(o_i), \text{PubMedBERT}(o_j))$
    $S_{total}(i,j) \leftarrow \alpha \cdot S_{time}(i,j) + (1-\alpha) \cdot S_{sem}(i,j)$
  end for

Step 2: Agglomerative event clustering
  Initialize clusters $C_i = \{o_i\}$
  Merge pairs with $\max S_{total} > \delta_{time}$ iteratively until convergence
  Output event clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$

Step3: Local Bayesian DAG reconstruction and graph integration
  for each cluster $C_k \in \mathcal{C}$
    Reconstruct local DAG $g_k = (V_k, E_k)$ maximizing $\mathcal{L}(g_k)$ under sparsity and acyclicity constraints
  end for
  $V_{obs} \leftarrow \bigcup_{k=1}^{K} V_k$
  $E_{obs} \leftarrow \bigcup_{k=1}^{K} E_k \cup \{\text{temporal transitive cross-cluster edges from } KG\}$

return G_{obs}

We combine Gaussian temporal decay and PubMedBERT semantic embedding to compute pairwise event similarity:

$$S_{total}(i,j) = \alpha \exp\left(-\frac{(t_i - t_j)^2}{2\sigma_t^2}\right) + (1-\alpha)\frac{z_{o_i} \cdot z_{o_j}}{\|z_{o_i}\|\|z_{o_j}\|}. \tag{3}$$

$\alpha \in [0,1]$ balances temporal and semantic contribution. Agglomerative hierarchical clustering groups events into coherent clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$ using the combined similarity metric.

For each cluster C_k, we reconstruct local DAG $g_k = (V_k, E_k)$ with background knowledge graph prior W_{prior}. Score-based structure search optimizes the objective with sparsity penalty and DAG acyclicity trace constraint:

$$\begin{aligned} \mathcal{L}(g_k) = \log P(C_k \mid g_k, W_{prior}) - \lambda \|E_k\|_0 \\ \text{s.t.} \quad \text{Tr}(e^{A_k}) - d_k = 0 \end{aligned} \tag{4}$$

Local subgraphs are aggregated into global G_{obs}: vertices are the union of cluster nodes; inter-cluster directed edges are added when background KG supports a path and source event precedes target temporally:

$$\begin{aligned} E_{obs} = \bigcup_{k=1}^{K} E_k \cup \{(u,v) \mid u \in C_a, v \in C_b, \\ \text{Path}_{KG}(u,v) \neq \emptyset \wedge t_u < t_v\} \end{aligned} \tag{5}$$

Algorithm 1 summarizes the complete procedure.

3.3 Stage II: Active State Graph Backward Roaming Search

Starting from observed clinical phenotypes in G_{obs}, TRACE executes a backward abductive search toward upstream root causes. To navigate the exponentially expanding search space without combinatorial explosion, we integrate UMLS concept normalization, transfer probability estimation, and MCTS heuristic search.

3.3.1 UMLS CONCEPT NORMALIZATION FOR TERMINOLOGY ALIGNMENT

To resolve medical terminology variation (e.g., "MI" vs. "myocardial infarction"), text spans s_{text} are normalized to UMLS Concept Unique Identifiers (CUIs) via a bi-encoder architecture:

$$P_{map}(c_u \mid s_{text}) = \frac{\exp(z_{text} \cdot z_{cui}/\tau_{norm})}{\sum_{c \in C_{UMLS}} \exp(z_{text} \cdot z_c/\tau_{norm})} \tag{6}$$

where z_{text} and z_{cui} are PubMedBERT representations of the candidate mention and canonical UMLS strings. Only candidate mappings exceeding a strict threshold ($P_{map} \geq 0.9$) are added to the search state graph.

3.3.2 CROSS-LITERATURE CAUSAL TRANSFER PROBABILITY

For frontier node v_t, reverse queries ("what upstream factors cause v_t?") retrieve upstream candidate nodes $v_{cand} \in V_{upstream}$. The transfer probability $P_{trans}(v_{cand} \to v_t)$ combines RotatE topological embeddings $z(\cdot)$ and bridging literature evidence D_{bridge}:

$$P_{trans}(v_{cand} \to v_t) = \sigma\Bigg(W_{trans} \cdot [z(v_{cand}) \parallel z(v_t)] \cdot \frac{1}{|D_{bridge}|} \sum_{d \in D_{bridge}} h_{doc}(d) \Bigg) \tag{7}$$

Candidates with $P_{trans} < \theta_{trans}$ are pruned early to constrain the search branching factor.

3.3.3 MCTS-DRIVEN HEURISTIC BACKWARD EXPLORATION

TRACE optimizes multi-hop trajectory assembly using a customized MCTS engine. MCTS systematically executes four operations across iterations: (1) **Selection**: Traverse tree nodes via UCT policy $UCT(s,a) = Q(s,a) + c_{puct} P_{trans}(a) \frac{\sqrt{N(s)}}{1+N(s,a)}$; (2) **Expansion**: Generate candidate predecessors via LLM reverse querying and UMLS filtering; (3) **Simulation**: Perform rollout trajectories guided by transfer probabilities P_{trans}; (4) **Backpropagation**: Evaluate completed paths using the base verifier $R_{eval} = \text{Base_Verify}(P_{rollout})$ and update node values.

Full detailed pseudocode for the MCTS-driven backward search engine is provided in Appendix B.2.

3.4 STAGE III: MACRO-SKILL DISCOVERY AND TEMPORAL DYNAMIC UPDATE

Repeated sub-paths frequently appear across different patient reasoning tasks. We encapsulate verified high-frequency sub-trajectories as reusable macro-skills to avoid redundant verification overhead. Each supporting document d_k associated with a skill is assigned time-decay evidence weight:

$$\omega_{time}(d_k) = \exp\left(-\lambda \cdot (t_{now} - t_d)\right) + \beta_{impact} \cdot I(d_k) \tag{8}$$

where $\lambda > 0$ is decay constant, t_d is publication time, $\beta_{impact} \cdot I(d_k)$ accounts for journal impact factor. Skill confidence score is re-weighted with aggregated time-decayed evidence:

$$C_{skill}^{(temporal)} = C_{skill} \cdot \frac{\sum_{d \in D_{skill}} \omega_{time}(d)}{|D_{skill}|} \tag{9}$$

Skills falling below threshold θ_{skill} trigger re-auditing against newly arrived literature. If supporting evidence no longer holds, the macro-skill is purged to eliminate outdated causal assertions. During MCTS expansion, validated macro-skills are directly invoked to jump over known sub-chains, drastically cutting per-task inference cost.

3.5 STAGE IV: STRUCTURED CONFORMAL PREDICTION WITH ACTIVE ABSTENTION

We design a DAG-aware structured conformal prediction module for multi-hop causal graphs, delivering finite-sample distribution-free coverage guarantees under exchangeability assumption. The pipeline contains offline calibration phase and online test phase.

Offline Calibration Phase: Given calibration dataset $D_{cal} = \{(G_i, y_i)\}_{i=1}^{n}$. Each causal graph G_i is decomposed into multiple causal paths $P_{i,k}$; every path is further decomposed into atomic triples $\tau_{i,k,l}$. Path-level non-conformity score is computed from triple-wise verification probability $p_{agent}(\tau)$:

$$\alpha_{path}(P_{i,k}) = 1 - \prod_{l} p_{agent}(\tau_{i,k,l}). \tag{10}$$

Graph-level non-conformity score aggregates all path scores with weight $w_{i,k}$:

$$\Phi(G_i) = \max_{k} w_{i,k} \cdot \alpha_{path}(P_{i,k}). \tag{11}$$

We compute conformal quantile threshold at target significance level ϵ:

$$\Gamma_{threshold} = \text{Quantile}\left(\{\Phi(G_i)\}_{i=1}^{n}, \frac{\lceil (n+1)(1-\epsilon) \rceil}{n}\right) \tag{12}$$

Online Testing & Decision Gate: For test candidate causal graph G_{test}, compute its graph-level non-conformity $\Phi(G_{test})$. If $\Phi(G_{test}) \leq \Gamma_{threshold}$, output prediction set $\mathcal{C}(G_{test})$ with guaranteed $1-\epsilon$ marginal coverage. Otherwise, trigger *Active Abstention*: reject this low-confidence causal graph and return compliance disclaimer instead of risky causal conclusions.

Complete pseudocode for the structured conformal prediction with active abstention module can be found in Appendix B.2.

4 Experiments

We investigate five research questions: (RQ1) abductive discovery performance; (RQ2) conformal coverage and prediction-set efficiency; (RQ3) component ablation; (RQ4) inference efficiency; (RQ5) qualitative interpretability. All experiments run on Xeon Gold 6330, RTX3090 24GB, 64GB RAM.

4.1 Experimental Setup

Four biomedical benchmarks are adopted in our experiments: **MIMIC**, derived from MIMIC-IV EHR (Johnson et al., 2023) for timestamped clinical abductive root-cause inference; **ADE-Time**, a temporally annotated adverse-drug-event dataset constructed from FAERS, DrugBank and UMLS (Zhang et al., 2022); **CoMAGC**, a cancer-gene multi-hop causal mechanism benchmark (Anonymous, 2025); and **FB15k-Subset**, a filtered FB15k subset for conformal coverage evaluation (Bordes et al., 2013). Our baselines span graph embedding, graph-RAG, prompt-based reasoning, causal abduction and conformal structured prediction, including RotatE (Sun et al., 2019), G-Retriever (He et al., 2024), GraphRAG (Edge et al., 2025), CoK (Li et al., 2024b), PTC (Susanti & Färber, 2025), LogiDynamics (Zheng et al., 2025), CtrlHGen (Gao et al., 2025), CSP (Zhang et al., 2025a), and CaPeR (Anonymous, 2025).

4.1.1 Metrics

Abductive performance: MRR, Hit@1, Hit@10, with

$$\text{Hit@}K = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{1}\big(\text{rank}_i \leq K\big). \tag{13}$$

Conformal metrics: empirical coverage rate and average prediction-set size. Full hyperparameter list is in Appendix B.3.

4.2 Main Abductive Reasoning Results

Tab. 2 summarizes abductive performance across MIMIC, ADE-Time and CoMAGC. TRACE consistently outperforms all competing baselines across all datasets. Static embedding RotatE yields weak performance because it encodes correlation rather than causal mechanism. RAG and prompt-based methods improve over embedding baselines but are limited by their inability to synthesize unrecorded

Table 2: Abductive reasoning performance comparison across three biomedical dynamic datasets.

Method	MIMIC			ADE-Time			CoMAGC		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
RotatE	0.187	11.2 %	32.4 %	0.224	13.5 %	37.8 %	0.203	12.1 %	34.1 %
G-Retriever	0.364	24.3 %	48.7 %	0.418	28.6 %	55.3 %	0.412	27.8 %	53.5 %
GraphRAG	0.341	22.1 %	46.2 %	0.395	26.4 %	52.1 %	0.386	25.9 %	50.8 %
CoK	0.403	27.8 %	53.1 %	0.441	31.2 %	58.4 %	0.389	25.4 %	50.1 %
PTC	0.476	35.1 %	61.3 %	0.508	38.4 %	64.9 %	0.493	36.7 %	63.4 %
LogiDynamics	0.492	37.5 %	62.8 %	0.521	39.5 %	66.8 %	0.552	42.6 %	69.1 %
CtrlHGen	0.504	38.9 %	64.1 %	0.567	43.5 %	70.9 %	0.531	40.2 %	67.4 %
CaPeR	0.518	40.2 %	65.5 %	0.551	41.8 %	68.7 %	0.524	39.5 %	65.8 %
TRACE	**0.608**	**48.7 %**	**76.3 %**	**0.647**	**52.3 %**	**80.5 %**	**0.621**	**49.8 %**	**78.1 %**

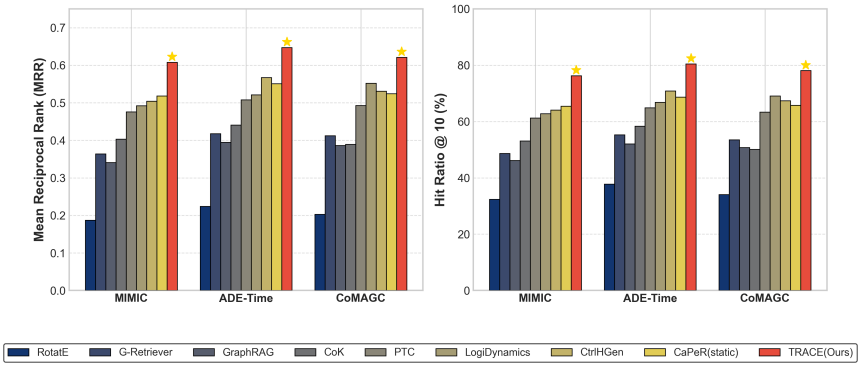


Figure 3: Main abductive reasoning performance (MRR and Hit@10) on MIMIC, ADE-Time, CoMAGC.

multi-hop causal pathways. Task-specific abduction baselines achieve stronger results but are hindered by greedy local search and static graph assumptions. TRACE benefits from temporal topological grounding, MCTS open-domain exploration and macro-skill reuse, yielding substantial MRR and Hit@K gains.

4.3 Conformal Prediction Reliability Analysis

We evaluate conformal coverage and prediction-set size at target $95\,\%$ coverage ($\epsilon = 0.05$) on FB15k-Subset and MIMIC, comparing against fixed Top-5 heuristics and vanilla CSP. Results are shown in Tab. 3. Heuristic top-5 outputs fail to satisfy coverage requirements. Vanilla CSP meets coverage but produces bloated prediction sets. TRACE strictly respects the $95\,\%$ coverage target while halving average prediction-set size, achieving better compactness without sacrificing statistical safety.

Table 3: Coverage and prediction set size comparison at 95 % target confidence.

Method	FB15K-Subset		MIMIC	
	Coverage	Set Size	Coverage	Set Size
CtrlHGen (Top-5)	86.4 %	5.00	81.2 %	5.00
CaPeR (Top-5)	85.1 %	5.00	79.8 %	5.00
CSP	95.7 %	6.84	96.1 %	8.37
TRACE	**95.4 %**	**3.42**	**95.2 %**	**4.15**

4.4 Ablation Studies

Ablation experiments on MIMIC evaluate each core module, with results summarized in Tab. 4. Removing temporal alignment yields the largest MRR drop and substantially inflates prediction-set size, confirming the critical importance of temporal ordering for clinical causal reasoning. Hierarchical graph aggregation, UMLS normalization and MCTS exploration each contribute meaningful performance gains. Disabling macro-skills moderately degrades performance by limiting effective search depth. Removing the conformal safety gate drastically degrades empirical coverage, demonstrating that heuristic ranking cannot replace formal statistical safety guarantees.

Table 4: Ablation study results on MIMIC dataset.

Model Variant	MRR	Hit@10	Coverage	Set Size
Full Model	**0.608**	**76.3 %**	95.2 %	**4.15**
w/o Temporal Module	0.534	66.8 %	94.8 %	7.62
w/o Hierarchical Graph	0.561	69.4 %	95.0 %	5.88
w/o UMLS Normalization	0.552	67.9 %	94.9 %	6.31
w/o MCTS (greedy search)	0.547	67.2 %	95.1 %	5.74
w/o Macro-skills	0.583	73.1 %	95.1 %	4.62
w/o Conformal Prediction	0.595	74.1 %	82.7 %	5.00

Supplementary analyses Parameter-sensitivity scans over α, inference-efficiency benchmarks and qualitative ADR case study are presented in Appendix A.1, Appendix A.2, and Appendix A.3. These additional experiments further validate the trade-offs of conformal threshold selection, macro-skill latency reduction and model interpretability.

5 Discussion

TRACE delivers layered formal guarantees: its conformal prediction provides finite-sample distribution-free coverage under clinical observation exchangeability with ϵ-tunable error bounds; transfer-prior-accelerated MCTS backward search probabilistically converges to optimal causal trajectories given sufficient rollouts; and its neuro-symbolic design enables full literature traceability for every inferred causal edge to support clinical auditing and regulatory compliance. For biomedical AI, we show conformal safety can be integrated into neuro-symbolic pipelines without performance degradation; macro-skill discovery offers a general multi-hop agent reasoning optimization to cut compute cost without full retraining; and time-decay continual distillation mitigates knowledge drift while preserving core biological knowledge. Limitations include potential exchangeability violations under severe temporal distribution shifts requiring adaptive conformal calibration; absent multimodal support for medical imaging and whole-genome sequencing; entity-alignment bottlenecks for UMLS-absent rare diseases; higher MCTS computational cost precluding ultra-low-latency real-time deployment; and benchmark curation bias, which demands prospective physician-guided clinical validation prior to hospital deployment.

6 Conclusion

This paper introduced TRACE, a time-aware neuro-symbolic abduction framework augmented with structured conformal prediction, shifting biomedical causal discovery from passive path auditing toward active open-domain mechanistic exploration. TRACE integrates time-aware topological observation alignment, MCTS-guided backward open-domain search, macro-skill continual dynamization, and structured conformal prediction with active abstention. Comprehensive evaluations across four biomedical benchmarks demonstrate substantial improvements over state-of-the-art baselines, while preserving formal conformal statistical coverage guarantees. Macro-skill reuse reduces inference overhead and suppresses structural hallucinations.

Future work will pursue two directions: extending TRACE toward multimodal biomedical inputs and designing dynamic adaptive conformal calibration under distribution shifts; conducting prospective real-world clinical validation for pharmacovigilance and disease-mechanism discovery workflows.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language polishing and readability improvement. We have not used generative AI tools for algorithm design, experimental implementation, dataset construction, and result analysis, and other tasks are not applicable to this work. Additionally, we did not use generative AI tools for other recommended disclosure tasks. We have reviewed all AI-assisted work. We checked LLM-generated text for potential plagiarism through manual literature survey. All AI-polished manuscript contents were revised and verified by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study utilizes publicly available, de-identified biomedical benchmark datasets. All datasets comply with their respective original data use agreements. Our proposed TRACE framework is designed for clinical decision support rather than autonomous clinical diagnosis. Model outputs must always be reviewed by qualified clinical professionals before making clinical decisions. We do not introduce new human subject experiments in this work. We have mitigated potential risks of biased causal reasoning by incorporating literature-grounded verification and conformal active abstention mechanisms to reject high-uncertainty outputs. Privacy, security, and research integrity concerns have been considered throughout the design of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility of this work, all four benchmark datasets utilized in our experiments are publicly available open-access resources, as described in the data availability description within the main text. The core algorithmic pipeline of TRACE is described in full detail via algorithm listings in Section 3 and Appendix B.2. Hyperparameter configurations are provided in Appendix B.3. For theoretical claims regarding conformal prediction coverage guarantees, please refer to Section 3.5. For ablation analysis verifying component contributions, see Section 4.4. This paragraph does not contain full reproduction details but points readers to main text, appendix and supplementary materials for complete reproduction resources.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

Anonymous. Caper: Cascaded path-evidence reasoning for biomedical causal verification. *PRCV 2025*, 2025.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, et al. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, et al. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 1–9, 2013.

Claudia D'Amato. Machine learning for the semantic web: A neuro-symbolic approach. *Semantic Web*, 11(1):1–3, 2020.

Allan Peter Davis, Caren J Grondin, Robin J Johnson, Debra Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database (ctd): update 2021. *Nucleic Acids Research*, 48(D1):D899–D906, 2021.

Li Du, Ming Ding, et al. Abductive reasoning over knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 702–712, 2021.

Dylan Edge, Huy Trinh, Nanyun Cheng, et al. From local to global: A graph rag approach to query-focused summarization, 2025.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Yang Gao, Jun Bai, Tianyu Zheng, et al. Controllable logical hypothesis generation for abductive reasoning in knowledge graphs, 2025.

Artur d'Avila Garcez, Marco Gori, Luis C Lamb, et al. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6(4):611–632, 2019.

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

Xiaoxin He, Yijun Tian, Yifei Sun, et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Daniel S Himmelstein, Antoine Lizee, Christine Hessler, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

Ziwei Ji, Nayeon Lee, Rita Frieske, et al. Survey of hallucination in large language models. *ACM Computing Surveys*, 56(11):1–38, 2024.

Alistair EW Johnson, Lucas Bulgarelli, Tom J Pollard, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9, 2023.

Emre Kícíman, Robert Ness, Amit Sharma, et al. Causal reasoning and large language models: Opening a new frontier, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Ming Li, Hao Wang, Liang Zhao, et al. A review of knowledge graph-based clinical decision support systems. *Journal of Biomedical Informatics*, 152:104628, 2024a.

Xinyu Li, Ruochen Zhao, Yew Ken Chia, et al. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *International Conference on Learning Representations (ICLR)*, 2024b.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Yixin Pei et al. Structure learning of bayesian networks using monte carlo tree search. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23343–23355, 2022.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020.

George A Prenosil. Neuro-symbolic ai for auditable cognitive information extraction from medical reports. *Journal of Medical Informatics*, 5:100123, 2024.

Youssef A Radwan, Yan Li, Quan Qing, et al. Cmkl: Modality-aware continual learning for evolving biomedical knowledge graphs, 2026.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *International Conference on Learning Representations (ICLR)*, 2020.

Seung Rho and Lisa Smith. Language guided skill discovery. *Journal of Machine Learning Research*, 26(123):1–32, 2025.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, et al. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10):2003–2030, 2006.

Noah Shinn, Federico Cassano, Ashwin Gopinath, et al. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, et al. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuliana Susanti and Michael Färber. Paths to causality: Finding informative subgraphs within knowledge graphs for knowledge-based causal discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2897–2908, 2025.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

Sondos Toonsi, Paul N Schofield, and Robert Hoehndorf. Causal knowledge graph analysis identifies adverse drug effects. *Briefings in Bioinformatics*, 25(6):bbae512, 2024.

David S Wishart, Yannick D Feunang, An C Guo, Eric J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

Liuyi Yao, Zhixuan Chu, Sheng Li, et al. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46, 2021.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, et al. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 535–546, 2021.

Boyu Zhang et al. Conformal structured prediction. In *International Conference on Learning Representations (ICLR)*, 2025a.

Dong Zhang, Zhen Wang, Zhi-Zhong Li, et al. Medkgent: A large language model agent framework for constructing temporally evolving medical knowledge graph, 2025b.

Ying Zhang, Xiaowei Chen, et al. Extracting adverse drug events from clinical notes. *Journal of the American Medical Informatics Association*, 29(10):1812–1820, 2022.

Xin Zhao, Sijia Liu, Samuel Y Yang, et al. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot, 2025.

Chao Zheng et al. Identifying causal mechanisms through structural causal models. *Journal of Machine Learning Research*, 24(112):1–42, 2023.

Tianyu Zheng, Jiawei Cheng, Chenghao Li, et al. Logidynamics: Unraveling the dynamics of inductive, abductive and deductive logical inferences in llm reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

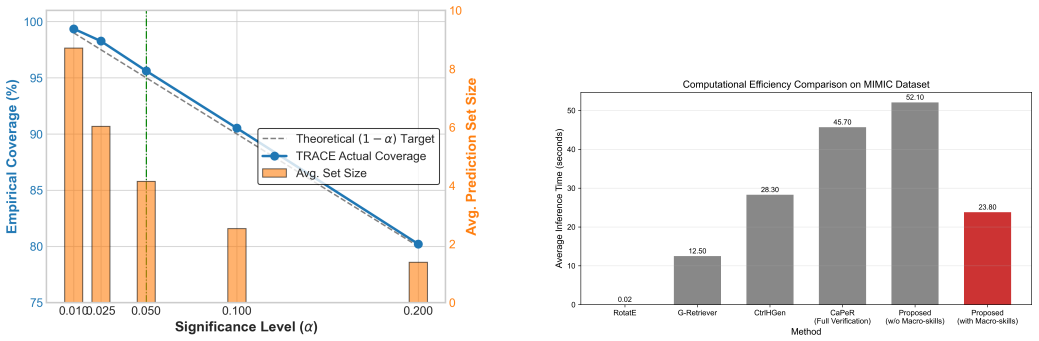
Shengyu Zhu et al. Causal discovery with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

A Supplementary Experimental Analysis

A.1 Parameter Sensitivity Analysis

To evaluate the empirical robustness of our structured conformal prediction layer, we perform parameter sensitivity experiments on MIMIC by sweeping significance level $\alpha \in [0.01, 0.20]$.

Across all α thresholds, empirical coverage consistently meets or exceeds the theoretical target $1 - \alpha$. As α increases from 0.01 to 0.20, average prediction-set size decreases monotonically from 8.71 to 1.38. At $\alpha = 0.05$, we obtain $95.20\,\%$ coverage with average set size 4.15, the recommended operating point for general clinical deployment. For high-stakes scenarios, stricter $\alpha = 0.025$ can be used for higher safety margins at the cost of larger output sets.



(a) Parameter sensitivity of α

(b) Computational efficiency comparison

Figure 4: Experimental analysis on MIMIC dataset: (a) Impact of significance level α on empirical coverage and prediction set size, and (b) Computational efficiency comparison across different methods highlighting latency reduction by TRACE.

A.2 Computational Efficiency Analysis

We report per-task inference latency on MIMIC across different paradigms. RotatE achieves extremely low latency 0.02s/task but poor causal reasoning accuracy. Vanilla TRACE without macro-skill suffers 52.10s/task latency. Macro-skill encapsulation reduces mean per-task latency down to 23.80s ($54.3\,\%$ reduction). TRACE also outperforms CtrlHGen (28.30s) and CaPeR (45.70s) in wall-clock runtime. The 24s/task footprint is suitable for offline batch pharmacovigilance workflows; real-time emergency scenarios would require further engineering optimizations such as pre-computed skill repositories and hierarchical caching.

A.3 Case Study: ADR Discovery

We examine a representative ADE-Time clinical scenario: inpatient develops acute kidney injury (AKI) and hyperkalemia ten days after starting losartan plus a co-administered antibiotic. TRACE ranks the pathway:

Losartan $\rightarrow$ *AT1 receptor antagonism* $\rightarrow$ *Reduced aldosterone secretion* $\rightarrow$ *Impaired renal potassium excretion* $\rightarrow$ *Hyperkalemia* $\rightarrow$ *Acute kidney injury*.

Pathway confidence is 0.92, well inside the $95\,\%$ conformal region. Three alternative competing hypotheses are retained in the prediction set. When temporal module is ablated, TRACE mis-attributes adverse events to the co-administered antibiotic. Without conformal safety gate, only a single top-1 hypothesis is output with over-inflated confidence, discarding competing clinical mechanisms. This case illustrates TRACE's ability to produce temporally consistent, evidence-grounded causal chains together with calibrated uncertainty.

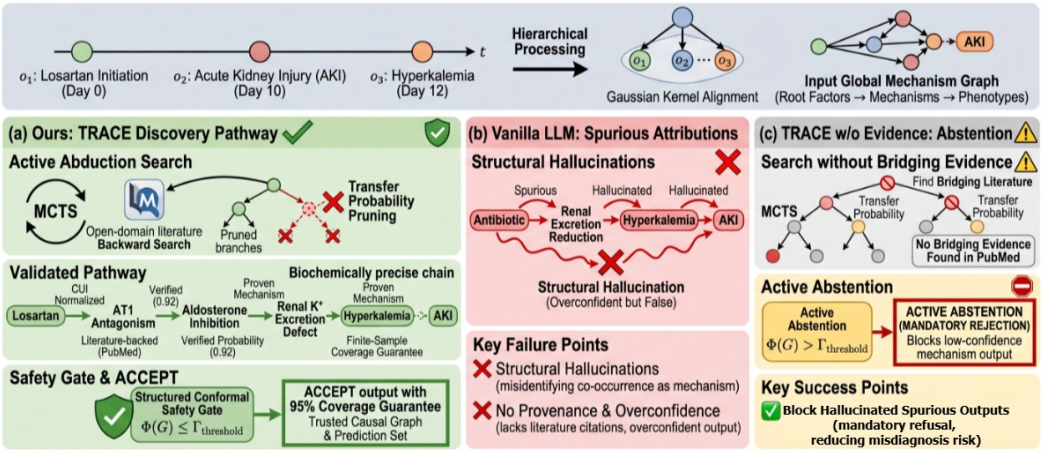


Figure 5: Real-world case study for Adverse Drug Reaction (ADR) abduction. TRACE successfully reconstructs the mechanistically plausible pathway (Losartan → AKI & Hyperkalemia), whereas vanilla LLM generates structural hallucinations and Conformal Abstention blocks low-confidence output.

B Additional Experimental Details

B.1 Full Algorithm Pseudocode

B.2 Full Algorithm Pseudocode

Table 5: Algorithm 2: Stage II: MCTS-driven Active Backward Search Engine

Require: Discovery graph G_{obs}, root targets V_{obs}, visit threshold N_{exp}, transfer threshold θ_{trans}, search budget T_{mcts}
Ensure: Top-K candidate causal paths $\mathcal{P} = \{P_1, \ldots, P_K\}$

Initialize root state s_0 with observed phenotype nodes V_{obs}
for iteration $= 1$ **to** T_{mcts}
  Selection: Traverse search tree from s_0 to leaf s_{leaf} using UCT policy
  if $N(s_{leaf}) \geq N_{exp}$ and s_{leaf} is not fully expanded
    Expansion: Query LLM for upstream causes $V_{upstream}$
    Filter candidates via $P_{map} \geq 0.9$ and $P_{trans} \geq \theta_{trans}$
    for each candidate $v_{cand} \in V_{upstream}$
      Compute $P_{trans}(v_{cand} \to v_{leaf})$
      Add child node if threshold satisfied
    end for
  end if
  Simulation: Rollout complete path $P_{rollout}$ guided by P_{trans}
  Evaluation: Compute reward $R_{eval} \leftarrow$ Base_Verify$(P_{rollout})$
  Backpropagation: Update $N(s, a)$ and $Q(s, a)$ along search path
end for
return Top-K causal trajectories $\mathcal{P}$ sorted by value descent

Table 6: Algorithm 3: Stage IV: Structured Conformal Prediction with Active Abstention

Require: Calibration dataset $D_{cal} = \{(G_i, y_i)\}_{i=1}^{n}$, test candidate G_{test}, significance level $\epsilon \in (0, 1)$
Ensure: Acceptance output decision or active abstention triggering

Phase 1: Offline Calibration Phase
for each calibration sample $i = 1$ **to** n
  for each path $P_{i,k} \in G_i$
    Compute path non-conformity: $\alpha_{path}(P_{i,k}) \leftarrow 1 - \prod_l p_{agent}(\tau_{i,k,l})$
  end for
  Aggregate graph non-conformity: $\Phi(G_i) \leftarrow \max_k w_{i,k} \cdot \alpha_{path}(P_{i,k})$
end for
Compute quantile threshold: $\Gamma_{threshold} \leftarrow \text{Quantile}\left(\{\Phi(G_i)\}_{i=1}^{n}, \frac{\lceil (n+1)(1-\epsilon) \rceil}{n}\right)$
Phase 2: Online Testing Phase
Compute graph non-conformity score $\Phi(G_{test})$ for test candidate G_{test}
if $\Phi(G_{test}) \leq \Gamma_{threshold}$
  return ACCEPT: prediction set $\mathcal{C}(G_{test})$ with $1 - \epsilon$ coverage guarantee
else
  return ABSTAIN: Trigger mandatory rejection and structured compliance disclaimer
end if

B.3 Full Hyperparameter Settings

MCTS exploration constant $c_{explore} = 1.414$, MCTS iterations per task $= 2000$, temporal decay window $\sigma_t = 14$ days, inverse rank fusion $k = 60$, self-consistency samples $M = 5$, agent verification threshold $\theta_{agent} = 0.7$. Default conformal target coverage $1 - \alpha = 0.95$ with $15\,\%$ calibration split. Macro-skill confidence threshold $\theta_{skill} = 0.85$. All hyperparameters are tuned on held-out validation set.