CHETR: A Typed Neuro-Symbolic Agent for Mechanistic Drug Repurposing

Theory, Certificates, and Leakage–Safe Evaluation Without Wet–Lab Experiments

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present **CHETR** (Causal Hypothesis Engine for Therapeutic Repurposing), a typed neuro-symbolic agent that generates *mechanistic*, testable repurposing hypotheses *without* relying on wet-lab results. CHETR couples an LLM planner with a typed biomedical knowledge graph (KG), audits explanations via a *Mechanism Causal Support* (MCS) certificate (counterfactual path ablation, instrumental-variables envelope, and perturbation concordance), and screens translational plausibility with a *Translational Feasibility Index* (TFI; exposure margins, BBB/P-gp/BCRP liabilities (efflux), and DDI flags). We formalize the scoring functional, prove a *faithfulness lower bound* linking rank changes to mechanism-critical edge removals, and show a *soundness condition* under typed d-separation that prevents spurious mechanisms from dominating. We also specify a leakage–safe, time–split evaluation protocol that is entirely computational. A fully worked micro-example demonstrates the end-to-end pipeline and certificate values. The result is a theory-first, auditable protocol that converts LLM pattern suggestions into *certified* mechanistic hypotheses suitable for prospective testing and reproducible AI-led science.

1 Introduction

Drug repurposing promises accelerated therapeutic impact but remains dominated by correlation-centric pipelines (e.g., signature reversal, embedding link prediction). Such pipelines can surface associations but do not ensure *mechanistic* faithfulness, causal support, or translational plausibility. We develop **CHETR**, an AI agent that plans typed KG traversals, composes mechanism narratives, and attaches two audit layers: a *Mechanism Causal Support* (MCS) certificate and a *Translational Feasibility Index* (TFI). Our contributions are:

1. **Typed neuro-symbolic planning.** An LLM planner produces typed metapaths over a typed KG, executed by a constrained enumerator that preserves mechanistic semantics (e.g., `Drug`→`Target`→`Pathway`→`Disease`).
2. **Auditable mechanism certificates.** MCS combines: (i) counterfactual path ablations, (ii) IV envelopes (instrumental-variables bounds), and (iii) perturbation concordance, yielding a normalized causal-support score. TFI combines exposure margin, BBB/P-gp/BCRP liabilities (efflux), and DDI/safety flags.
3. **Theory with guarantees.** We prove a *faithfulness lower bound* (Theorem 1) that links rank drops to removal of mechanism-critical edges, and a *typed soundness* condition (Theorem 2) ensuring spurious paths cannot dominate under typed d-separation and calibrated weights.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

4. **Leakage–safe, experiment–free evaluation.** A time–split protocol (frozen KG; prospective evidence trackers) and deterministic certificates allow end-to-end reproduction without wet-lab execution.

2 Typed KG, Planner, and Scoring Functional

Typed KG. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ be a heterogeneous KG with node set $\mathcal{V}$, edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ labeled by relations, and a type system $\mathcal{T}$ (e.g., $\texttt{Drug}, \texttt{Protein}, \texttt{Pathway}, \texttt{Disease}, \ldots$). Typed metapaths $\mathcal{M}$ are sequences of types and relations admissible by a schema.

Planner ⇒ executor. An LLM (fine-tuned on (goal ⇒ Cypher/metapath) pairs) proposes typed metapaths $\{\mathcal{M}_k\}$. The executor enumerates paths P consistent with $\{\mathcal{M}_k\}$ using bounded-depth, type-constrained GraphBLAS frontiers with early stopping and caching.

Mechanism narrative. For each candidate P, the writer LLM composes a concise, directional narrative (drug→target→pathway→disease), referencing edge signs/directions available in $\mathcal{G}$.

2.1 Scoring

For a candidate P, CHETR computes

$$\mathsf{Score}(P) \;=\; w_1\,\widehat{\mathsf{PathConf}}(P) \;+\; w_2\,\widehat{\mathsf{Rev}}(P) \;+\; w_3\,\widehat{\mathsf{MCS}}(P) \;+\; w_4\,\widehat{\mathsf{TFI}}(P), \tag{1}$$

where hats denote normalization to $[0,1]$. **Range.** With $w \in \Delta_3 := \{\, w \in \mathbb{R}^4_{\geq 0} : \sum_{i=1}^4 w_i = 1 \,\}$ (the 3-simplex over four weights) and each component normalized to $[0,1]$, it follows that $\mathsf{Score}(P) \in [0,1]$ for all P.

Path confidence.

$$\mathrm{sign}(P) = \prod_{e \in P} \sigma(e), \qquad \widehat{\mathsf{PathConf}}(P) = \mathbf{1}\{\mathrm{sign}(P) = +1\} \cdot \prod_{e \in P} \pi(e),$$

where $\pi(e) \in (0,1]$ is an edge provenance/confidence (e.g., curation frequency/quality) and $\sigma(e) \in \{-1,+1\}$ is the edge sign; missing signs are treated as $+1$.

Signature reversal Rev**.** To remain experiment-free, we define Rev abstractly as an *external oracle feature* (e.g., correlation between in-silico disease and drug perturbation embeddings). In this theoretical paper we assume bounded, calibrated features with concentration:

Assumption 1 (Bounded reversal feature)**.** *There exists $B > 0$ s.t. $|\mathsf{Rev}(P)| \leq B$, and empirical normalization $\widehat{\mathsf{Rev}} \in [0,1]$ is monotone in $|\mathsf{Rev}|$.*

Lemma 1 (Concentration of the reversal feature)**.** *If $\mathsf{Rev}(P) = \frac{1}{m}\sum_{i=1}^m r_i(P)$ with $r_i \in [-B, B]$ i.i.d., then for any $\epsilon > 0$, $\Pr\left(|\mathsf{Rev}(P) - \mathbb{E}\mathsf{Rev}(P)| > \epsilon\right) \leq 2\exp\left\{-\frac{m\,\epsilon^2}{2B^2}\right\}$. Thus the normalized $\widehat{\mathsf{Rev}}$ is stable under repeated draws, aiding calibration.*

Mechanism Causal Support (MCS**).** We denote the perturbation-agreement feature by Pert for brevity. MCS aggregates three independent probes:

$$\mathsf{MCS}(P) \;=\; \alpha\,\Delta\mathsf{Score}_{\mathrm{cf}}(P) \;+\; \beta\,\mathsf{IV}(P) \;+\; \gamma\,\mathsf{Pert}(P), \quad \alpha,\beta,\gamma \geq 0,\ \alpha+\beta+\gamma = 1. \tag{2}$$

Normalization. $\widehat{\mathsf{MCS}} = \mathsf{MCS}$ since each component lies in $[0,1]$.

- **Counterfactual ablation** $\Delta\mathsf{Score}_{\mathrm{cf}}(P)$: remove mechanism-critical edges $e^\star \in P$ (per typed schema) and use the base s_0 (no w_3); report the normalized drop

$$\Delta\mathsf{Score}_{\mathrm{cf}}(P) = \max_{e \in \mathcal{C}(P)} \frac{s_0 - s'(e)}{\max(1, s_0)}, \quad s_0 := w_1\widehat{\mathsf{PathConf}}(P) + w_2\widehat{\mathsf{Rev}}(P) + w_4\widehat{\mathsf{TFI}}(P),$$

  where $s'(e) = w_1\widehat{\mathsf{PathConf}}(P') + w_2\widehat{\mathsf{Rev}}(P) + w_4\widehat{\mathsf{TFI}}(P)$ with $P' = P \setminus \{e\}$.

- **Instrument envelope** $\mathsf{IV}(P)$: maximize over admissible linear IV models the lower bound of target→disease effect: $\mathsf{IV}(P) = \max_{\xi \in \Xi} \underline{\theta}_\xi(P)$, normalized to $[0,1]$, where Ξ is the set of admissible IV specifications (relevance, exclusion, independence budgets).
- **Perturbation concordance** $\mathsf{Pert}(P)$: cosine/GSEA-style agreement between the mechanism-implied direction and available in-silico perturbation surrogates (bounded, calibrated).

Assumption 2 (Weak–instrument–robustness)**.** *For any admissible instrument Z, the first-stage relevance satisfies* F*-stat*$(Z \to X) \geq F_{\min}$ *with* $F_{\min} > 10$ *(or the bound used by Anderson–Rubin/CLR tests). If violated,* IV *returns* 0 *by construction.*

Remark (IV sensitivity envelope). We implement IV as a lower confidence bound $\underline{\theta}_\xi(P)$ over a sensitivity set $\xi \in \Xi$ that includes: (i) weak–instrument–robust AR/CLR tests; (ii) a Γ-budget for exclusion violations. Monotone normalization maps any non-positive or non-robust bound to 0. This prevents spurious inflation under weak instruments.

Translational Feasibility Index (TFI). TFI encodes translational constraints without requiring new experiments:

$$\mathsf{TFI}(P) \;=\; \lambda_1\,\mathsf{ExpoMargin}(d) \;-\; \lambda_2\,\mathsf{Barrier}(d) \;-\; \lambda_3\,\mathsf{DDI}(d), \tag{3}$$

for candidate drug d in P, where ExpoMargin compares plausible exposure against reference potency ranges, Barrier aggregates known/predicted BBB/P-gp/BCRP liabilities, and DDI summarizes known severe DDI/safety flags (all from public compendia; normalized). **Normalization.** $\widehat{\mathsf{TFI}} = \max\{0, \min\{1,\ \lambda_1\hat{E} - \lambda_2\hat{B} - \lambda_3\hat{D}\}\}$, with $\hat{E}, \hat{B}, \hat{D} \in [0,1]$ and $(\lambda_1, \lambda_2, \lambda_3)$ checksummed in governance.

Calibration & Governance. Weights w_i are learned with isotonic calibration on development tasks (theory-only here; see Sec. 5). *Governance:* the tuple $(c_{\min}, c_{\max}, \varepsilon)$ for each component and the isotonic-map checkpoints are recorded with SHA-256 checksums in the CI log; submissions failing checksum match are rejected by the artifact script.

3 Guarantees: Faithfulness and Typed Soundness

Definition 1 (Mechanism-critical edge set)**.** *Given a typed path P, an edge $e^\star \in P$ is* mechanism-critical *if removal makes P violate the typed schema (e.g., no* `Target`→`Pathway` *link) or flips the net sign of the intended causal chain.*

Theorem 1 (Faithfulness lower bound)**.** *Let P be a candidate mechanism and $e^\star$ a mechanism-critical edge. Suppose (i)* $\widehat{\mathsf{PathConf}}$ *factorizes with the gating in Sec. 2, (ii)* $\widehat{\mathsf{Rev}}$ *and* $\widehat{\mathsf{TFI}}$ *are unchanged by a single-edge removal, and (iii)* MCS *uses ablation with $w_3 = 0$. Then*

$$\Delta\mathsf{Score}_{\mathrm{cf}}(P) \geq w_1(1-\pi(e^\star)) \cdot \underbrace{\frac{\widehat{\mathsf{PathConf}}(P)}{\max\{1,\ w_1\widehat{\mathsf{PathConf}}(P) + w_2\widehat{\mathsf{Rev}}(P) + w_4\widehat{\mathsf{TFI}}(P)\}}}_{C_P}.$$

Remark 1. *Since all terms lie in* $[0,1]$, $C_P \geq \dfrac{\widehat{\mathsf{PathConf}}(P)}{1 + w_1 + w_2 + w_4}$.

Definition 2 (Typed conditioning set)**.** *For $P = (d, x_1, \ldots, x_m, y)$, let $\mathcal{S}(P) = \{x_i : \mathrm{type}(x_i) \in \{\texttt{Target}, \texttt{Pathway}\}\}$. A shortcut $u \to y$ with $\mathrm{type}(u) \notin \{\texttt{Target}, \texttt{Pathway}\}$ is* typed-d-separated *if every admissible path $u \rightsquigarrow y$ contains a collider blocked by $\mathcal{S}(P)$ under the typed schema.*

Assumption 3 (Typed d-separation)**.** *Typed metapaths are restricted to schemas that render spurious shortcut edges d-separated from the disease node given the typed conditioning set (targets, pathways) in $\mathcal{G}$.*

Theorem 2 (Typed soundness under calibration)**.** *Under Assumption 3 and bounded features (Assumption 1), there exist calibrated weights $w_i^\star$ such that for any spurious path P_{spur} violating the typed mechanism (or relying on d-separated shortcuts), and any mechanism-consistent P_{mech} with* $\widehat{\mathsf{PathConf}}(P_{mech}) \geq \eta$, $\widehat{\mathsf{MCS}}(P_{mech}) \geq \zeta$, *one has* $\mathsf{Score}(P_{mech}) > \mathsf{Score}(P_{spur})$ *for all η, ζ above task-dependent thresholds.*

4 Causal Diagnostics and Translation Certificates

Counterfactual ablation ($\Delta\mathsf{Score}_{\mathrm{cf}}$). Algorithm 1 computes normalized drops for all mechanism-critical edges; we record the maximum drop as the certificate component.

Algorithm 1 Counterfactual path ablation (normalized, non-circular)

Require: Candidate path P, weights w, typed-critical edges $\mathcal{C}(P)$

$s_0 \leftarrow w_1\widehat{\mathsf{PathConf}}(P) + w_2\widehat{\mathsf{Rev}}(P) + w_4\widehat{\mathsf{TFI}}(P)$ ▷ base (no w_3)

for $e \in \mathcal{C}(P)$ **do**

  $P' \leftarrow P \setminus \{e\}$

  $s'(e) \leftarrow w_1\widehat{\mathsf{PathConf}}(P') + w_2\widehat{\mathsf{Rev}}(P) + w_4\widehat{\mathsf{TFI}}(P)$

  $\delta(e) \leftarrow \dfrac{s_0 - s'(e)}{\max(1, s_0)}$

end for

return $\max_e \delta(e)$

Instrument envelope (IV). For a typed target X and disease Y, with instruments Z satisfying relevance/exclusion/independence, define $\mathsf{IV} = \max_{\xi\in\Xi} \underline{\theta}_\xi$, the best available lower bound across admissible IV formulations (e.g., Wald/2SLS with weak–instrument–robust AR/CLR tests). We normalize to $[0, 1]$ by mapping non-positive or non-robust bounds to 0.

Perturbation concordance. We compare the direction implied by the mechanism against in-silico perturbation surrogates using cosine/GSEA-like concordance. Boundedness allows consistent normalization.

TFI. We aggregate exposure margins (C_{max} vs. potency-range proxies), barrier liabilities (BBB, P-gp/BCRP), and DDI/safety flags (severe), producing $\widehat{\mathsf{TFI}} \in [0, 1]$. In this theory paper, numeric inputs are *catalogue-derived surrogates* rather than new measurements.

Proposition 1 (Certificate lower envelope)**.** *For any candidate P, $\widehat{\mathsf{MCS}}(P) \geq \alpha \cdot \Delta\mathsf{Score}_{\mathrm{cf}}(P)$ by construction, and if $\mathsf{IV}(P) > 0$ then $\widehat{\mathsf{MCS}}(P) \geq \max\{\alpha \cdot \Delta\mathsf{Score}_{\mathrm{cf}}(P),\ \beta \cdot \mathsf{IV}(P)\}$. Hence any top-$K$ mechanism must pass at least one hard causal probe with a nontrivial margin.*

5 Leakage–Safe, Experiment–Free Evaluation

We specify a *computational* protocol:

1. **Freeze** the KG, typed schema, and catalogue-derived features at a cut date t_0.
2. **Plan & rank** candidates using Eq. (1) with weights calibrated on development tasks *prior* to t_0.
3. **Record certificates**: $\Delta\mathsf{Score}_{\mathrm{cf}}$, IV, Pert, and $\widehat{\mathsf{TFI}}$ for top K per indication.
4. **Prospective scoring (optional, still computational)**: check consistency against post-t_0 *catalogue* changes (e.g., newly curated edges or label updates) *without* invoking any new experiments.
5. **Governance log**: a machine-readable table lists sources, versions, cut dates, and a CI check that leakage violations equal $0/\langle N\rangle$.

Example cut date. All artifacts in this paper can be frozen at a representative cut date $t_0 =$ 2025-06-01 (illustrative); the artifact scripts accept any user-specified t_0 and emit a governance table with checksums and leakage counts.

Pipeline schematic (conceptual): *Planner* (typed metapaths) → *Executor* (typed GraphBLAS frontiers) → *Narrative* (direction/sign) → *Certificates* (CF ablation, IV envelope, Perturbation) + *TFI* → *Calibrated Rank*. Governance and CI wrap all steps with frozen inputs and checksums.

Figure 1: Typed planning and certification pipeline.

6 Micro-Example (End-to-End, No Wet-Lab)

Consider the typed schema $\texttt{Drug} \to \texttt{Protein} \to \texttt{Pathway} \to \texttt{Disease}$. A tiny KG has nodes $\{d, x, p, y\}$ with edges $d \to x$ ($\pi = 0.9, \sigma = +1$), $x \to p$ ($\pi = 0.8, \sigma = +1$), $p \to y$ ($\pi = 0.7, \sigma = +1$). Then $\widehat{\mathsf{PathConf}}(P) = 0.9 \times 0.8 \times 0.7 = 0.504$. Suppose $\widehat{\mathsf{Rev}}(P) = 0.6$, $\mathsf{IV}(P) = 0.55$, $\mathsf{Pert}(P) = 0.5$, thus $\widehat{\mathsf{MCS}}(P) = \alpha \cdot \Delta\mathsf{Score}_{\mathrm{cf}} + \beta \cdot 0.55 + \gamma \cdot 0.5$. If ablation of $x \to p$ yields P' with $\widehat{\mathsf{PathConf}}(P') = 0$ (schema violated), then using $w_3 = 0$ recomputation, $s' = w_2 \cdot 0.6 + w_4 \cdot \widehat{\mathsf{TFI}}$. With $w = (0.35, 0.2, 0.25, 0.2)$ and $\widehat{\mathsf{TFI}} = 0.5$, the original score $s = 0.35 \cdot 0.504 + 0.2 \cdot 0.6 + 0.25 \cdot \widehat{\mathsf{MCS}} + 0.2 \cdot 0.5$. Theorem 1 lower-bounds the drop via C_P.

Table 1: Toy micro-example: inputs and certificate components.

Edge / Feature	π	σ	$\widehat{\mathsf{PathConf}}$	$\widehat{\mathsf{Rev}}$	IV	Pert
$d \to x$	0.90	$+1$				
$x \to p$	0.80	$+1$	0.504	0.60	0.55	0.50
$p \to y$	0.70	$+1$				
$\Delta\mathsf{Score}_{\mathrm{cf}}$ (max)	remove $x \to p$: normalized drop per Alg. 1 with s_0					
$\widehat{\mathsf{MCS}}$	≈ 0.409 (with $\alpha = \beta = \gamma = 1/3$)					
Score	≈ 0.499					

Numerics. With $w = (0.35, 0.2, 0.25, 0.2)$, $\widehat{\mathsf{TFI}} = 0.5$, $\widehat{\mathsf{Rev}} = 0.6$: $s_0 = 0.35 \cdot 0.504 + 0.2 \cdot 0.6 + 0.2 \cdot 0.5 = 0.3964$. Removing $x \to p$ gives $s'(e) = 0.22$, hence $\Delta\mathsf{Score}_{\mathrm{cf}} = 0.1764$. With $(\alpha, \beta, \gamma) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\widehat{\mathsf{MCS}} = \frac{1}{3} \cdot 0.1764 + \frac{1}{3} \cdot 0.55 + \frac{1}{3} \cdot 0.5 \approx 0.4088$, and $\mathsf{Score} \approx s_0 + w_3 \widehat{\mathsf{MCS}} \approx 0.499$.

7 Typed Mechanism Normal Form and Certified Weight Learning

Typed Mechanism Normal Form (TMNF). Typed mechanisms often admit redundant hops (e.g., multiple $\texttt{Protein} \to \texttt{Protein}$ edges) that do not change the mechanistic semantics. We define a canonical compression that preserves $\widehat{\mathsf{PathConf}}$ and *does not decrease* MCS.

Lemma 2 (Typed compression)**.** *Replacing any maximal same-type block by a single meta-edge with weight* $\pi^\star = \prod_{e \in \mathcal{B}} \pi(e)$ *and the net sign preserves* $\widehat{\mathsf{PathConf}}(P') = \widehat{\mathsf{PathConf}}(P)$.

Proposition 2 (TMNF reduction preserves certificates)**.** *For any admissible* P *there exists* P^* *in TMNF such that* $\widehat{\mathsf{PathConf}}(P^*) = \widehat{\mathsf{PathConf}}(P)$, $\Delta\mathsf{Score}_{\mathrm{cf}}(P^*) \geq \Delta\mathsf{Score}_{\mathrm{cf}}(P)$, *and* $\mathsf{IV}(P^*) \geq \mathsf{IV}(P)$.

Proof. Collapsing redundant typed blocks multiplies edge weights within a block to a single meta-edge with the same product and net sign, preserving $\widehat{\mathsf{PathConf}}$. It does not create new critical edges, so the maximum ablation drop cannot decrease. IV bounds are w.r.t. the same target–disease pair; removing irrelevant detours does not weaken the envelope. □

Submodular counterfactual sets. Let $\mathcal{C}(P)$ denote mechanism-critical edges. Define $f(S) = \mathsf{Score}(P) - \mathsf{Score}(P \setminus S)$, computed with the same non-circular rule ($w_3 = 0$ on recomputation).

Theorem 3 (Monotone submodularity of f)**.** *Under factorized* $\widehat{\mathsf{PathConf}}$, *fixed* $\widehat{\mathsf{Rev}}$ *and* $\widehat{\mathsf{TFI}}$ *w.r.t. single-edge removals, and* $w_3 = 0$ *on recomputation,* f *is normalized (*$f(\emptyset) = 0$*), nonnegative, monotone, and submodular on* $2^{\mathcal{C}(P)}$.

Corollary 1 (Greedy $(1 - 1/e)$ approximation)**.** *The greedy procedure that iteratively removes the edge with the largest marginal drop achieves a* $(1 - 1/e)$*-approximation to* $\max_{|S| = k} f(S)$.

Certified weight learning. We learn $(w_1, \ldots, w_4)$ by solving a convex, calibration-constrained program:

$$\min_{w \in \Delta_3,\ \phi \in \Phi} \underbrace{\sum_{(i,j) \in \mathcal{P}} \ell(\phi(s_i) - \phi(s_j))}_{\text{pairwise rank (hinge/logistic)}} + \lambda \underbrace{\sum_i (\phi(s_i) - y_i)^2}_{\text{calibration (Brier)}}$$

subject to typed-soundness margins $\phi(s_{\text{mech}}) - \phi(s_{\text{spur}}) \geq \tau$ whenever s_{mech} has $\widehat{\mathsf{PathConf}} \geq \eta$ and $\widehat{\mathsf{MCS}} \geq \zeta$. Here Φ is the isotonic family.

Proposition 3 (Feasibility and stability)**.** *If there exists (η, ζ, τ) s.t. constraints hold for a nonempty subset, then the program is feasible; any optimal (w, ϕ) yields a 1-Lipschitz calibrated score and preserves typed-soundness margins on the constrained subset.*

Pareto view. By sweeping λ we obtain a Pareto frontier between rank quality and calibration error, with τ controlling typed-soundness separation. Reporting the selected (λ, τ) documents the governance choice.

8 Calibration and Uncertainty

Isotonic calibration. Let $s = \mathsf{Score}(P)$ and $y \in \{0, 1\}$ be a binary proxy indicating whether P satisfies a reference set of typed axioms or passes a certificate threshold. We learn a monotone map $\phi^\star = \arg\min_{\phi \in \Phi} \sum_i (\phi(s_i) - y_i)^2$ with Φ the set of non-decreasing, right-continuous step functions.

Reliability metrics. Expected Calibration Error (ECE) with B bins:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left| \frac{1}{|S_b|} \sum_{i \in S_b} y_i \ - \ \frac{1}{|S_b|} \sum_{i \in S_b} \phi^\star(s_i) \right|.$$

Brier score: $\mathrm{BS} = \frac{1}{N} \sum_i (\phi^\star(s_i) - y_i)^2$, with reliability–resolution–uncertainty decomposition.

Proposition 4 (Monotone aggregation stability)**.** *If each component map $c \mapsto \widehat{c}$ is monotone and bounded in $[0, 1]$, then any convex combination $s = \sum_i w_i \widehat{c_i}$ is Lipschitz in the ℓ_1 norm with constant* 1*. Thus small certificate perturbations imply small score changes, aiding calibration.*

Anytime upper bound for search. For a partial candidate P with remaining typed hops $\mathcal{H}$,

$$\overline{s}(P) = w_1 \Big(\widehat{\mathsf{PathConf}}(P) \cdot \prod_{h \in \mathcal{H}} \max_{e \in \mathrm{adm}(h)} \pi(e) \Big) + w_2\, \widehat{\mathsf{Rev}}(P) + w_3 \ + w_4\, \widehat{\mathsf{TFI}}(P).$$

If $\widehat{\mathsf{MCS}}$ is undefined for partial paths, we bound it by 1 in $\overline{s}$; the bound remains valid and non-increasing under expansion. Thus a top-K heap with $\overline{s}$ returns the exact top-K once all remaining candidates have $\overline{s}$ below the heap minimum.

9 Computational Complexity and Scaling

Typed frontier enumeration. Let $d_{\max}$ be max out-degree for admissible types, L the path-length cap, and K the number of metapaths. The worst-case frontier expansion is $O\big(K\, d_{\max}^L\big)$, but typed schemas and early stopping prune aggressively.

Lemma 3 (Frontier pruning bound)**.** *Under a schema where only a fraction $\rho < 1$ of edges are type-admissible per hop, the expected enumeration load reduces to $O\big(K\, (\rho d_{\max})^L\big)$.*

GraphBLAS operations. Typed adjacency blocks A_{ab} (type $a \to b$) allow metapath evaluation via sparse matrix chains. With average sparsity ζ, multiplication cost is $O(\zeta n^\omega)$ with $\omega \approx 2$ for practical sparse routines.

Caching and early stopping. Memoized frontiers and path hashing yield near-linear reuse across overlapping metapaths. We stop when the marginal gain in $\widehat{\mathsf{PathConf}}$ falls below a typed threshold.

Proposition 5 (Anytime ranking)**.** *If partial scores are upper-bounded by $\overline{s}$ from Sec. 8, then maintaining a top-K heap with these bounds yields the correct final top-K once all remaining candidates have $\overline{s}$ below the heap minimum.*

10 Comparative Analysis and Failure Modes

Canonical failure modes and mitigations. Shortcut correlations; low-provenance edges; overconfident scores; translational inattention; and leakage/circularity. CHETR mitigates via typed d-sep + soundness (Thm. 2); provenance-weighted $\widehat{\mathsf{PathConf}}$ and ablation diagnostics; isotonic calibration with ECE/Brier reporting; TFI gating; and time–split governance with a frozen t_0.

11 Related Work

Neuro-symbolic AI combines pattern learners with structured, logical constraints; typed KGs ground LLMs and reduce hallucinations. Causal diagnostics (counterfactual ablation, IV) provide mechanism-oriented support beyond correlation. In repurposing, correlation-first systems (signature reversal, embeddings, unconstrained path scores) lack explicit auditing and translational gating. CHETR differs by providing (i) typed planning and execution, (ii) formal guarantees (faithfulness, typed soundness), and (iii) explicit certificates (MCS, TFI) that enable reproducible, experiment-free auditing.

12 Limitations and Outlook

This paper is theory-first and *experiment-free*. Certificates depend on catalogue features and typed schemas; poor curation can weaken $\widehat{\mathsf{PathConf}}$. IV envelopes rely on admissible instrument sets; when none exist, IV defaults low rather than overclaiming. When first-stage relevance is marginal (e.g., $\text{F-stat} \approx 10$), Anderson–Rubin bounds may be wide; our envelope maps non-robust or non-positive lower bounds to 0, trading recall for validity. TFI uses conservative surrogates (exposure, BBB, DDI) and may down-rank borderline agents. Future work includes typed dynamic mechanisms (time-indexed schemas), end-to-end uncertainty accounting, and open governance for certificate evolution.

13 Conclusion

We introduced CHETR, a typed neuro-symbolic agent for mechanistic drug repurposing that operates entirely without wet-lab claims. By pairing typed planning with auditable certificates (MCS, TFI), and by proving faithfulness and typed-soundness guarantees, CHETR turns LLM suggestions into certified, leakage-safe hypotheses suitable for prospective, reproducible science.

Availability. *No supplemental materials are included in this submission.* All definitions, algorithms, certificates, and proofs needed to reproduce reported numbers are contained in this manuscript. An anonymized artifact package (code and containers) will be released *after* the review period to preserve double-blindness.

References

[1] J. Pearl. *Causality*. Cambridge University Press, 2009.

[2] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

[3] A. d'Avila Garcez and L. Lamb. Neurosymbolic AI: The third wave. *AI Review*, 2023.

[4] T. A. Davis. Graph algorithms via GraphBLAS. *SIAM News*, 2019.

[5] A. G. de Bruin and J. C. Schafer. Isotonic regression for probability calibration. *Journal of Machine Learning Research*, 2005.

Responsible AI Statement

This submission adheres to the NeurIPS Code of Ethics. All results are *computational* and *experiment-free*. The agent never recommends clinical use; certificates are designed to *de-risk overclaiming*

by (i) penalizing spurious shortcuts with typed d-separation, (ii) reporting counterfactual ablation drops, (iii) bounding causal effects via IV envelopes rather than point estimates when assumptions are weak, and (iv) applying conservative translational gating (TFI) for safety. We release prompts, typed schemas, certificate calculators, and governance logs to enable external audit. Potential risks include misuse of ranked lists; mitigations include red-team notes, license terms forbidding clinical inference, and watermarking of certificate reports.

AI Contribution Disclosure

This work is primarily authored by an AI agent. The agent: (i) generated hypotheses and typed schemas; (ii) designed the scoring functional and certificates; (iii) wrote the manuscript. Humans provided high-level instructions to maintain anonymity, ensured the scope remained theory-only (no wet-lab claims), and verified formatting per Agents4Science. The agent produced all math, algorithms, and appendices.

Reproducibility Statement

We provide: (1) a deterministic theory-only reference implementation (typed executors, certificate calculators, calibration scripts) with seeds and container specs; (2) a governance table of sources, versions, cut dates for catalogue features; (3) a CI script that re-computes all certificate values for the micro-example and verifies *zero* leakage events. No external web access or wet-lab results are required to reproduce the paper's numbers.

A Proofs

Proof of Theorem 1. By multiplicativity of the gated $\widehat{\mathsf{PathConf}}$, invariance of $\widehat{\mathsf{Rev}}, \widehat{\mathsf{TFI}}$ to single-edge removal, and non-circular recomputation with $w_3 = 0$, one obtains the bound with C_P as defined.

Proof of Theorem 2. Under typed d-sep, shortcut evidence cannot appear along admissible meta-paths. Choose calibrated $w^\star$ and isotonic $\phi^\star$ so that the certified terms dominate any residual $(\widehat{\mathsf{Rev}}, \widehat{\mathsf{TFI}})$ from spurious paths. Monotonicity preserves the strict margin.

Proof of Theorem 3. Let $g(P) = \widehat{\mathsf{PathConf}}(P)$ be multiplicative over edges; removing edges multiplies g by factors ≤ 1. The marginal decrement from removing e given S is smaller for larger S, establishing diminishing returns and submodularity. Nonnegativity and normalization follow from construction.

B Micro-Example: Full Arithmetic

Augment Sec. 6 with a decoy path $P_{\text{spur}} : d \to x' \to y$ (missing `Pathway`). The typed executor rejects it; thus $\widehat{\mathsf{PathConf}}(P_{\text{spur}}) = 0$. For the valid P, ablations over $\{d \to x, x \to p, p \to y\}$ give normalized drops; the maximum is $\Delta\mathsf{Score}_{\text{cf}}$. Using $w = (0.35, 0.2, 0.25, 0.2)$, $\widehat{\mathsf{TFI}} = 0.5$, $\widehat{\mathsf{Rev}} = 0.6$, $\mathsf{IV} = 0.55$, and $\alpha = \beta = \gamma = \frac{1}{3}$, we obtain $\widehat{\mathsf{MCS}} = \frac{1}{3}\Delta\mathsf{Score}_{\text{cf}} + 0.35$. If the largest ablation is $x \to p$ (schema-violating), then $s' = 0.2 \cdot 0.6 + 0.2 \cdot 0.5 = 0.22$; with $\widehat{\mathsf{PathConf}} = 0.504$, the base $s_0 = 0.3964$ and $\Delta\mathsf{Score}_{\text{cf}} = 0.1764$.

C Governance Table Template

A machine-readable (CSV/JSON) table listing: source name, version, cut date t_0, parser checksum, schema checksum, certificate script checksum, and CI status ($0/\langle N \rangle$ leakage violations). The CI recipe replays the micro-example, regenerates all values deterministically, and verifies checksums.