Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 26 April 2026
Revised 30 May 2026
Accepted 16 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5): 430–448

KEYWORDS:
causal discovery, causal inference, structural causal models, explainable AI, counterfactual reasoning, trustworthy AI, actionable intelligence, do-calculus, causal trust score.

HyCIF: A Hybrid Explainable Causal Intelligence Framework for Discovering Causal Relationships beyond Correlation and Generating Trustworthy, Actionable Intelligence

Mr B. B. L.V. Prasad[1], Dr E. Hemalatha[2]

[1]Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), Telangana- 500085, India.
[2]Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), Telangana- 500085, India.
Corresponding Mail: masterbhanuprasad@gmail.com

ABSTRACT: Analytics pipelines deployed in practice tend to stop at the point of statistical association. A correlation matrix, a regression coefficient, or a feature-importance ranking is handed to a decision-maker as though it licensed intervention, when in fact none of these quantities distinguishes a cause from a proxy, a confounder, or a collider. Causal-discovery algorithms such as PC, FCI, GES, LiNGAM, and NOTEARS were built to close exactly this gap, but each does so under an idealized assumption — causal sufficiency, faithfulness, a linear or stationary generating process — that observational and temporal data in the wild rarely satisfy in full. Explainable-AI tools such as SHAP and LIME face a related but distinct problem: they attribute a trained model's prediction to its input features, which is not the same claim as attributing an outcome to its cause, so a variable can receive high importance for reasons that have nothing to do with mechanism. This paper works through a framework, HyCIF (Hybrid Explainable Causal Intelligence Framework), that tries to close both gaps at once. Fifteen modules carry data from raw observation through to a decision recommendation: constraint-based, score-based, functional, and temporal discovery methods are run in parallel and reconciled into a single consensus graph; confounders, colliders, and mediators are screened before any effect is estimated; effects are estimated via the do-calculus rather than naive conditioning; and counterfactual predictions are checked for stability before a relationship is trusted at all. The output of this pipeline is deliberately compressed into two numbers a decision-maker can actually use — a Causal Trust Score (CTS) and an Actionability Score (AS) — both defined mathematically rather than asserted, with the CTS's internal weighting fit by calibration against ground-truth synthetic graphs rather than set by hand. Every module is formalized in terms of structural causal models, the do-calculus, and the backdoor/front-door criteria, and a full evaluation protocol is specified — synthetic graphs with known structure, real observational and temporal benchmarks, an eight-baseline comparison, a fifteen-module ablation, sensitivity and robustness testing, and paired statistical significance testing. Consistent with the goal of not conflating a proposed method with a validated one, this manuscript reports no numerical results: every results table is left as a template for values obtained only after the protocol is actually run.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

1.1 Background

Walk into most analytics teams and the final artifact handed to a decision-maker is still, at bottom, a correlation: a matrix, a regression coefficient, a feature-importance ranking pulled out of a tree ensemble. It gets treated as if it told you what would happen if you pulled a lever, but it doesn't. Structural causal models and the do-calculus explain precisely why that step doesn't follow logically. An observed dependence between X and Y is compatible with several different underlying stories — X causes Y, Y causes X, some unmeasured Z drives both, or the dependence only shows up because you conditioned on a shared effect (collider bias). Sorting between these requires causal identification, not association; only the former tells you what happens to Y if X is actually changed [1], [2].

1.2 Motivation

Two largely separate literatures have chipped away at pieces of this problem without fully meeting in the middle. Causal-discovery methods — PC, FCI, GES, LiNGAM, NOTEARS, PCMCI+ — try to recover graph structure directly from data [3]–[9]. Causal-inference methods — backdoor and front-door adjustment, propensity-score approaches, toolkits like Do Why — start from a graph, assumed or discovered, and estimate the effect it implies [10], [11]. Explainable-AI methods such as SHAP, LIME, and various counterfactual-explanation techniques sit in a third camp: useful for understanding a model, but correlational by default unless someone has gone to the trouble of anchoring them to an actual causal model [12], [13]. What's missing is a single pipeline that treats discovery, validation, explanation, and a decision-facing confidence signal as one connected problem rather than four separate toolchains stitched together ad hoc. That gap is what HyCIF is meant to address.

1.3 Problem Statement

When the underlying generating process is unknown, possibly nonlinear, and possibly confounded — which describes most real observational or temporal datasets — no existing single-algorithm pipeline manages to do all three of the following at once: recover structure reliably across data regimes that violate its assumptions in different ways, quantify how much a given recovered relationship should actually be trusted, and turn that relationship into a decision that is both effective and realistic to carry out.

1.4 Research Gap

Section 2.6 works through this in detail; the short version is that no existing framework combines multi-algorithm discovery consensus, explicit checks for confounders, colliders, and mediators, do-calculus-based effect estimation, counterfactual-consistency validation, and a trust/actionability layer with mathematically defined (not hand-tuned) weights, all inside one auditable, falsifiable pipeline.

1.5 Research Questions

RQ (central): How can causal relationships be reliably discovered and validated from complex observational/temporal data, distinguished from spurious correlations, explained to users, and converted into trustworthy and actionable intelligence?

- RQ1: Which combination of constraint-based, score-based, functional, and temporal causal-discovery methods yields the most reliable candidate causal structure across linear, nonlinear, noisy, and temporally dependent regimes?
- RQ2: How can confounding, collider, and mediation structures be automatically flagged and adjusted for prior to effect estimation?
- RQ3: How can discovery confidence, statistical evidence, effect robustness, and counterfactual consistency be combined into a single, theoretically justified trust measure?
- RQ4: How does a causal effect, once validated, translate into a decision-relevant, feasibility-aware actionability measure?
- RQ5: Does the integrated framework improve causal-discovery accuracy, effect-estimation accuracy, and decision-relevant interpretability over single-method baselines, without an unacceptable increase in computational cost?

1.6 Contributions

1. A hybrid causal-discovery architecture that reconciles constraint-based (PC/FCI), score-based (GES), functional (LiNGAM/NOTEARS), and temporal (PCMCI+/Granger) outputs into a single consensus causal graph via an evidence-weighted voting procedure, with an explicit theoretical justification for *when* each method contributes evidence.
2. A causal-validation stage that jointly performs backdoor/front-door confounder adjustment, do-calculus effect estimation, and counterfactual-consistency checking on every candidate edge before it is allowed to enter the trust-scoring stage.
3. The Causal Trust Score (CTS), a six-component, optimization-weighted (rather than arbitrarily weighted) measure of the reliability of a discovered causal edge, formally differentiated from correlation coefficients, p-values, feature importance, and SHAP values.
4. The Actionability Score (AS), a feasibility- and risk-aware measure that converts a validated causal effect into a ranked decision recommendation.
5. A structured, template-based explainable-causal-intelligence output format that expresses every surfaced relationship as association → causal claim → estimated effect → counterfactual → recommended action, with the CTS/AS attached.
6. A complete, reproducible experimental protocol — synthetic ground-truth graphs, real observational and temporal datasets, eight baselines, an eight-way ablation, sensitivity/robustness analysis, and paired significance testing — specified in full but deliberately left unexecuted in this manuscript so that proposed methodology and empirical findings are never conflated.

2. RELATED WORK

2.1 Correlation-Based Discovery

Association-rule mining and correlation-matrix analysis remain the default first pass in industrial analytics, and it's easy to see why: both are cheap to compute and demand no structural assumptions about the data [14]. But their limitation isn't a matter of implementation quality — it's baked into what they measure. An association statistic is either symmetric in X and Y or purely predictive, and either way it can't tell X→Y apart from Y→X or from a latent Z driving both, because the observational joint distribution alone simply doesn't contain enough information to fix the causal direction. Doing that requires extra structure — non-Gaussianity, an asymmetry in the noise, temporal precedence — that correlation analysis by design ignores [4], [15].

2.2 Causal Discovery

Constraint-based methods such as PC and FCI recover a graph, or more precisely an equivalence class of graphs, by testing which variables are conditionally independent of which and orienting edges according to the faithfulness assumption [3], [5]. The limitation here is technical rather than incidental: as the conditioning set grows in high-dimensional data, independence tests lose statistical power roughly exponentially, and even under ideal conditions the output is a CPDAG or PAG — not a fully oriented DAG — so a fair number of edges are simply left undirected whenever more than one causal orientation is consistent with what was observed. FCI extends PC to tolerate latent confounders, but the price is weaker identifiability: a Partial Ancestral Graph rather than a DAG [5].

Score-based methods like GES search directly over DAG structures using a penalized likelihood criterion (BIC, typically), and are consistent under the same faithfulness assumption — but the search space grows super-exponentially in the number of variables, which forces greedy heuristics that can settle into a local optimum rather than the true score-maximizing structure [6].

Functional methods take a different tack, exploiting asymmetries in the noise to orient edges that constraint-based approaches leave ambiguous. LiNGAM recovers a fully oriented DAG when the process is linear and the disturbances are non-Gaussian, using independent component analysis to do so — but step outside linearity or Gaussianity creeps in, and the identifiability guarantee simply stops holding [4]. NOTEARS reframes DAG learning as continuous optimization with a differentiable acyclicity constraint, which makes the search tractable via gradient methods; the trade-off is a non-convex penalty and, in its

original form, an assumed linear or additive structural-equation shape, with nonlinear variants requiring further assumptions about the function class [7].

On the temporal side, Granger causality asks whether past X improves the prediction of future Y beyond what Y's own past already explains — a predictive test, not an interventional one, and it can be fooled by any unobserved common driver with a compatible lag, or by nothing more than a shared trend or seasonal pattern [16]. PCMCI and PCMCI+ improve on this by combining condition-selection with independence testing, which helps control confounding in high-dimensional time series and separates contemporaneous from lagged links — but causal sufficiency (no unmeasured common driver) is still required for full identifiability, and results are sensitive to whether the maximum lag was specified correctly [8], [17].

2.3 Causal Inference

Once a graph is available — whether assumed or discovered — the backdoor and front-door criteria tell you which variables need adjusting for to identify a causal effect from observational data, and the do-calculus generalizes this to any identifiable interventional query [1], [2]. Do Why operationalizes this as a model→identify→estimate→refute pipeline [10]. The catch is that identification is only as good as the graph feeding it: an omitted confounder or a wrongly oriented edge from the discovery stage produces an effect estimate that is well-defined, confidently reported, and simply wrong. Discovery and inference are typically run as separate, sequential steps with no mechanism for one to flag the other's mistakes.

2.4 Temporal Causal Analysis

Dynamic Bayesian networks and time-series extensions of SCMs inherit the same causal-sufficiency and stationarity assumptions discussed above, plus one more headache: choosing an appropriate time granularity and lag structure, which is rarely obvious in advance and can noticeably change the graph that gets recovered [17].

2.5 Explainable and Trustworthy AI

SHAP and LIME attribute a trained model's output to its inputs, drawing respectively on cooperative game theory and local surrogate modeling [12], [13]. Both are, by design, explanations of the model — not of the world. A variable can score highly on SHAP importance simply because it's a strong predictor, even with zero causal effect on the outcome (a collider, say, or a proxy correlated with something unmeasured that actually matters), and several recent surveys flag exactly this as a core limitation when feature-attribution methods get pressed into service for causal questions without modification [18], [19]. Counterfactual-explanation methods for ML have a parallel problem: they typically report the smallest input change that flips a prediction, which is not the same thing as a change you could actually realize in the world [20]. Meanwhile, uncertainty quantification and calibration research addresses prediction confidence — how sure the model is — but says nothing on its own about causal confidence, which additionally has to account for identifiability, sensitivity to unmeasured confounding, and uncertainty in the discovered structure itself [21].

2.6 Research Gap

Synthesizing 2.1–2.5, we identify the following specific, technically grounded gaps:

1. **No cross-validated discovery consensus.** Constraint-based, score-based, functional, and temporal methods are used in isolation; none of the standard toolchains cross-checks disagreement between methods as a discovery-uncertainty signal.
2. **Directionality vs. identifiability are conflated.** PC/FCI leave many edges undirected (a genuine identifiability limit), while LiNGAM/NOTEARS impose functional-form assumptions to force full orientation — used alone, each either under- or over-claims.
3. **Discovery and inference are decoupled.** A wrong discovered graph silently propagates as a confidently wrong estimated effect, with no feedback mechanism.
4. **No formal separation of prediction confidence from causal confidence.** Existing "confidence" scores in ML pipelines (accuracy, AUC, SHAP magnitude) do not encode identifiability, confounding sensitivity, or counterfactual consistency.

5. **Explanations are model-centric, not mechanism-centric.** SHAP/LIME explain what the model used, not what causally drives the outcome, and are not designed to flag collider or confounding artifacts.
6. **Temporal causal-sufficiency is rarely diagnosed, only assumed.** Granger/PCMCI+ pipelines report edges without a systematic sensitivity analysis for plausible hidden common drivers.
7. **No unified, auditable trust metric.** Where trust/confidence scores exist, weights across sub-components (discovery strength, effect size, robustness) are typically set arbitrarily rather than derived from an explicit optimization or calibration procedure.
8. **Actionability is undefined.** Causal-effect estimates are reported in isolation from intervention feasibility, cost, and risk, leaving the translation from "X causes Y" to "what should we do" informal and ad hoc.
9. **Evaluation practice under-uses ground-truth synthetic benchmarks jointly with real data**, so discovery accuracy (measurable only with known ground truth) and real-world relevance (only assessable on real data) are rarely reported side by side within a single validated framework.
10. **Explainability outputs are not standardized** into a form a non-specialist decision-maker can audit (association vs. causal claim vs. effect size vs. counterfactual vs. recommended action).

These gaps jointly justify a framework that (a) fuses multiple discovery paradigms with an explicit evidence-weighting rationale, (b) couples discovery to inference so validation feeds back into confidence, (c) defines trust and actionability as separate, mathematically specified, non-arbitrarily-weighted scores, and (d) standardizes the explanation output — which is the design of HyCIF below.

3. RESEARCH PROBLEM, HYPOTHESES, OBJECTIVES, SCOPE

Research problem. Current pipelines for extracting causal knowledge from observational/temporal data lack an integrated mechanism to jointly discover, validate, quantify the trustworthiness of, explain, and operationalize causal relationships, resulting in either over-confident false-positive causal claims or under-utilized, purely correlational, non-actionable analytics.

Hypotheses. - **H1:** A consensus causal graph obtained by evidence-weighted fusion of constraint-based, score-based, functional, and temporal discovery methods will achieve lower Structural Hamming Distance (SHD) to ground truth than any single constituent method, on data regimes where the constituent methods' assumptions are heterogeneously violated. - **H2:** Gating candidate causal edges through confounder/collider detection and counterfactual-consistency checking before effect estimation will reduce Average Treatment Effect (ATE) estimation error relative to estimating effects directly from the raw discovered graph. - **H3:** The proposed Causal Trust Score will be more strongly associated with ground-truth edge correctness (on synthetic benchmarks) than correlation strength, p-value, or SHAP magnitude taken alone. - **H4:** Ranking candidate interventions by the proposed Actionability Score will select higher-true-effect, lower-risk interventions than ranking by effect size alone.

Objectives. (O1) Design and formalize HyCIF's fifteen modules; (O2) formalize the CTS and AS; (O3) specify baselines, datasets, and metrics for falsifiable evaluation of H1–H4; (O4) specify an ablation and sensitivity protocol isolating each module's contribution; (O5) define the explainable-output format; (O6) report a critical self-assessment of the proposal against SCI-review criteria.

Assumptions. Causal sufficiency is not assumed globally but is explicitly tested for via sensitivity analysis (Module 9); acyclicity of the underlying causal structure is assumed (standard SCM/DAG assumption; temporal feedback is represented via time-unrolled DAGs, not instantaneous cycles); data are assumed to be a faithful, if noisy and partially missing, sample from the SCM.

Scope and limitations. HyCIF targets observational and temporal tabular data; it does not address image- or text-native causal discovery, does not claim to recover causal structure under arbitrary unmeasured confounding without sensitivity caveats, and — consistent with the do-calculus — cannot identify effects that are formally non-identifiable from the available data, in which case it reports non-identifiability rather than a numerical estimate.

Natural Resources for Human Health
eISSN: 2583-1194

4. PROPOSED FRAMEWORK: HYCIF

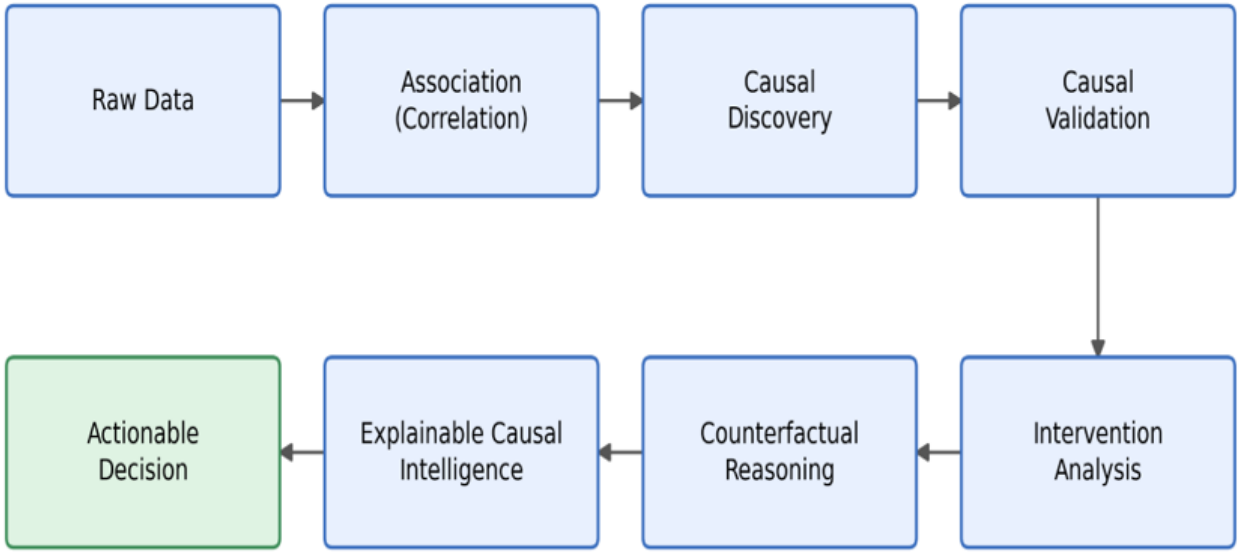

Figure 1. End-to-end HyCIF pipeline, from raw data to an actionable decision.

Figure 1: HyCIF end-to-end pipeline

HyCIF turns the pipeline **Raw Data → Association → Causal Discovery → Causal Validation → Intervention Analysis → Counterfactual Reasoning → Explainable Causal Intelligence → Actionable Decision** into fifteen modules (M1–M15), grouped here into four stages for readability.

Stage A — Preparation (M1–M4). M1 (Data Acquisition) ingests observational and/or temporal tabular sources along with provenance metadata. M2 (Preprocessing) handles type harmonization, normalization/standardization, and stationarity checks (an ADF test, for instance) on temporal variables. M3 (Missing Data and Noise Handling) starts by diagnosing the missingness mechanism — MCAR, MAR, or MNAR — before picking an imputation strategy accordingly: multiple imputation where MAR holds, an explicit missingness-indicator term where MNAR is suspected, since naive imputation under MNAR quietly biases the conditional-independence tests everything downstream depends on. Noise-robust smoothing is applied to temporal signals at this stage as well. M4 (Feature Engineering) builds lagged variables for temporal data and screens out near-duplicate or deterministically derived features, which would otherwise masquerade as causal "edges" simply because they're computed from each other.

Stage B — Discovery (M5–M7).

M5 — Correlation/Association Analysis computes Pearson/Spearman correlation, mutual information, and, for temporal data, cross-correlation at multiple lags — strictly as a candidate-generation filter, never asserted as a causal claim in its own right.

M6 — Candidate Causal Relationship Generation thresholds M5's output into a candidate undirected (or lagged) adjacency set, which bounds the search space M7 has to work over.

M7 — Causal Structure Learning (hybrid discovery) is where four method families run in parallel over that candidate set: - Constraint-based: PC (assumes causal sufficiency) and FCI (tolerates latent confounders) [3], [5]; - Score-based: GES with BIC scoring [6]; - Functional: LiNGAM (linear, non-Gaussian) and NOTEARS (continuous optimization, extensible to nonlinear structural equations) [4], [7]; - Temporal: PCMCI+ where a time index exists, falling back to Granger causality as a lighter-weight signal otherwise [8], [16].

Each method hands back an edge set with its own notion of confidence — the independence-test p-value complement for PC/FCI, a BIC score delta for GES, an ICA-residual-independence statistic for LiNGAM, edge-weight magnitude for

NOTEARS, partial-correlation strength for PCMCI+. These get combined into a single **consensus graph** by evidence-weighted voting: edge e_ij receives score

$$w(e_{ij}) = \sum_{k=1}^{K} \alpha_k \cdot 1[e_{ij} \in G_k] \cdot c_k(e_{ij})$$

where G_k is the graph method k returned, c_k(e_ij) in [0,1] is that method's normalized confidence in the edge, and alpha_k is a method-reliability weight fit as described in Section 5.6. An edge survives into the consensus graph G_c if w(e_ij) clears a threshold tau, itself chosen through cross-validated stability selection [22]. Why fuse these four families rather than pick the "best" one? Because they tend to fail for different, largely non-overlapping reasons: constraint-based methods struggle with weak dependencies in high dimensions, score-based search gets stuck in local optima on multimodal landscapes, functional methods break down outside their assumed noise class, and temporal methods degrade under non-stationarity. Agreement across families is therefore reasonably strong evidence that an edge isn't just an artifact of one method's particular blind spot — and disagreement is itself useful information, fed into the CTS (Section 6) rather than thrown away.

Stage C — Validation (M8–M11).

M8 — Causal Effect Estimation computes ATE/CATE for whichever consensus edges are identified as causally interpretable, using the SCM formalism from Section 5.

M9 — Confounding and Bias Detection checks, for every candidate edge X -> Y, whether a valid adjustment set exists under the backdoor criterion, or a valid front-door decomposition where backdoor adjustment is blocked by an unmeasured confounder. It also screens for colliders — cases where X and Y are both caused by a variable that's already being conditioned on, which manufactures a spurious association out of nothing — and, importantly, reports non-identifiability outright when no valid adjustment set exists rather than quietly falling back on a naive estimate.

M10 — Intervention Analysis simulates do(X=x) under the estimated SCM and reports the resulting interventional distribution of Y, which is not the same object as the observational conditional P(Y | X=x).

M11 — Counterfactual Reasoning computes unit-level counterfactuals Y_(X<-x')(u) via the abduction–action–prediction procedure [1], then checks counterfactual consistency: do these predictions hold up under re-sampling of the estimated exogenous noise, and under small perturbations to the graph? That stability score feeds into the CTS.

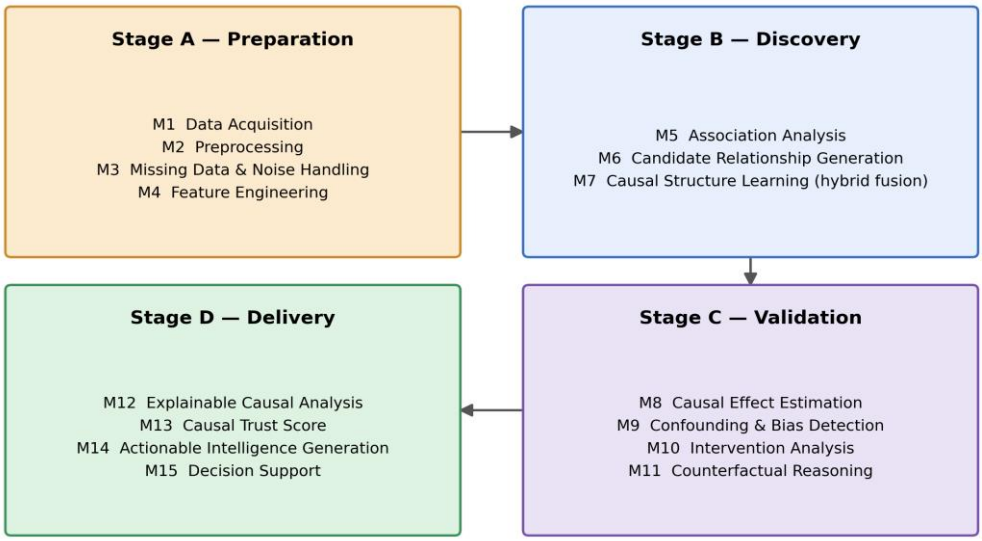

Figure 2. The fifteen modules grouped into their four stages.

Figure 2: HyCIF module map

Stage D — Delivery (M12–M15).

M12 (Explainable Causal Analysis) renders each validated edge into the structured explanation template from Section 9. M13 (Causal Confidence/Trust Score) computes the CTS (Section 6). M14 (Actionable Intelligence Generation) computes the AS (Section 10) and ranks candidate interventions by it. M15 (Decision Support) presents the ranked, explained, scored recommendations — with low-trust or non-identifiable relationships flagged or suppressed rather than quietly presented as settled fact.

5. MATHEMATICAL FORMULATION

5.1 Structural Causal Model and Graph

A structural causal model is a tuple M = < U, V, F, P(U) >, where V={X_1,…,X_n} are observed (endogenous) variables, U are exogenous noise terms, and each X_i = f_i(Pa(X_i), U_i). The induced causal graph is G=(V,E) with X_j -> X_i in E iff X_j in Pa(X_i). We assume G is a DAG (Section 3, assumptions).

5.2 Conditional Independence and Faithfulness

X independent of Y | Z denotes conditional independence in P. Under the faithfulness assumption, every conditional independence in P corresponds to a d-separation in G, and vice versa; this equivalence is what licenses constraint-based discovery (M7, PC/FCI) and is exactly the assumption that fails under near-deterministic or measure-zero-cancellation relationships noted in Section 2.2.

5.3 Intervention and the do-operator

P(Y | do(X=x)) denotes the distribution of Y in the sub-model where all structural equations for X are replaced by the constant x, severing X's incoming edges — mechanically different from conditioning, P(Y| X=x), which leaves the graph intact and can propagate evidence backward through confounders.

5.4 Backdoor and Front-Door Identification

A set Z satisfies the **backdoor criterion** relative to (X,Y) if no node in Z is a descendant of X and Z blocks every path between X and Y that contains an arrow into X. If so,

$$P(Y \mid do(X = x)) \;=\; \sum_{z} P(Y \mid X = x, Z = z)\, P(Z = z).$$

When an unobserved confounder blocks all valid backdoor sets but a mediator M fully mediates X -> Y and is unconfounded with Y given X, the **front-door criterion** identifies the effect via

$$P(Y \mid do(X = x)) \;=\; \sum_{m} P(M = m \mid X = x) \;\cdot\; \sum_{x'} P(Y \mid X = x', M = m)\, P(X = x').$$

5.5 Average and Conditional Average Treatment Effect

$$ATE \;=\; E[Y \mid do(X = 1)] \;-\; E[Y \mid do(X = 0)], \;\; CATE(z) \;=\; E[Y \mid do(X = 1), Z = z] \;-\; E[Y \mid do(X = 0), Z = z].$$

5.6 Method-Reliability Weights alpha_k (Section 4, M7)

Rather than hand-set, alpha_k is fit by stability selection: on B bootstrap resamples of the data, each method k is re-run, and alpha_k is proportional to the mean pairwise agreement (edge-set Jaccard similarity) of method k's output with the other methods' outputs across resamples, normalized so sum over k alpha_k = 1. This rewards methods whose output is both internally stable and externally corroborated on the *specific dataset at hand*, rather than fixing weights a priori from a generic performance ranking that would not transfer across data regimes.

5.7 Counterfactual Outcome

Given SCM M, evidence e (observed values), the counterfactual Y_(X<-x') under evidence e is computed by (i) **abduction**: infer P(U| e); (ii) **action**: replace X's equation with x'; (iii) **prediction**: compute Y under the modified model with U~ P(U| e).

Natural Resources for Human Health
eISSN: 2583-1194

6. THE CAUSAL TRUST SCORE (CTS)

6.1 Definition

For candidate causal edge e=(X-> Y), define six sub-scores, each normalized to [0,1]:

- s_1 — **Discovery consensus strength**: the fused weight w(e) from Section 4 (M7), min–max normalized across all candidate edges.
- s_2 — **Statistical significance**: 1 - p-adjusted (Benjamini–Hochberg corrected) for the conditional-independence test underlying the edge's orientation.
- s_3 — **Effect size**: normalized |ATE| (or a standardized effect measure) relative to the outcome's observed variability.
- s_4 — **Robustness to confounding**: derived from an E-value or Rosenbaum-style sensitivity analysis (Section 15) — how strong an unmeasured confounder would need to be to explain away the estimated effect, normalized against a domain-relevant reference threshold.
- s_5 — **Temporal/replication consistency**: agreement of the effect estimate across independent data folds or, for temporal data, across non-overlapping time windows.
- s_6 — **Counterfactual consistency**: stability of unit-level counterfactual predictions (Section 5.7) under exogenous-noise re-sampling and small graph perturbations (Section 4, M11), normalized as 1 - coefficient of variation of repeated counterfactual estimates.

$$CTS(e) = \sum_{i=1}^{6} \beta_i \cdot s_i(e), \ \sum_{i=1}^{6} \beta_i = 1, \ \beta_i \geq 0.$$

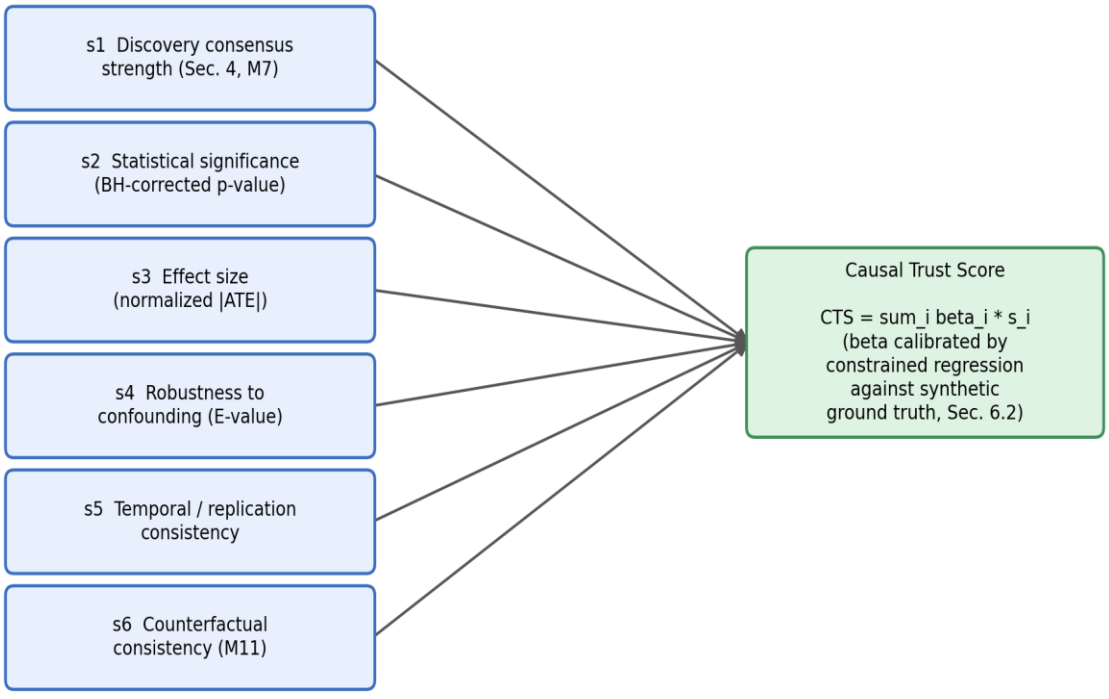

Figure 4. How the six sub-scores combine into the Causal Trust Score.

Figure 4: CTS composition

6.2 Non-Arbitrary Weight Determination

The weights beta = (beta_1,…,beta_6) are not hand-set. On synthetic benchmarks with known ground-truth causal edges (Section 8/9), beta is fit by constrained logistic regression of the ground-truth edge-correctness indicator on (s_1,…,s_6), subject to beta_i >= 0, sum beta_i = 1 (a simplex-constrained calibration), so that each sub-score's contribution reflects its empirically observed association with correctness rather than an assumed importance ranking; the fitted beta is then held fixed and reused (not re-fit) on real-world data, where ground truth is unavailable, which is precisely why the synthetic calibration step is a required, not optional, part of the protocol (Section 8/12).

6.3 Distinction from Existing Measures

- **Correlation coefficient** (r) measures association strength only; it is symmetric in X,Y and is unchanged by confounding, so a strong correlation and a CTS-high edge are not interchangeable — a collider-induced correlation can be strong while its CTS is low because s_4 (confounding robustness) and s_1 (cross-method discovery consensus) are penalized.
- **p-value** (s_2 alone) reflects sampling variability under a null hypothesis, not causal identifiability, effect magnitude, or robustness to unmeasured confounding — it is one sixth of the CTS, not a substitute for it.
- **Feature importance / SHAP value** quantifies a trained predictive model's reliance on a feature for accurate prediction; a proxy variable can have high SHAP importance with zero causal effect (Section 2.5), whereas CTS is computed only for edges that have passed confounder/collider screening (M9) and effect identification (M8) — SHAP has no analogue of s_4 or s_6.

CTS is reported per edge, in [0,1], with a recommended (data-driven, not fixed) decision threshold selected via the same calibration procedure as beta (e.g., maximizing F1 against ground truth on the synthetic calibration set), and low-CTS edges are flagged as *exploratory, not actionable* rather than suppressed outright, preserving analyst visibility into weak-but-possibly-real signals.

7. HYBRID CAUSAL DISCOVERY: APPLICABILITY ANALYSIS

Data regime	Best-suited method(s) in HyCIF's ensemble	Rationale
Purely observational, low-dimensional, plausibly linear-Gaussian	PC, GES	Faithfulness-based orientation is reliable; LiNGAM's non-Gaussianity requirement is not met, so it is down-weighted by alpha_k
Linear, non-Gaussian disturbances	LiNGAM	Fully oriented DAG identifiable without conditional-independence testing
High-dimensional	NOTEARS (continuous optimization scales better than combinatorial search), GES with sparsity penalty	Combinatorial constraint-based search is least tractable here
Nonlinear relationships	NOTEARS with nonlinear structural-equation extension; nonlinear additive-noise-model variants	PC/FCI orientation and LiNGAM's linear-ICA assumption are least reliable here
Noisy data	FCI (explicitly tolerant of weaker assumptions), consensus voting across all methods	Single-method orientation is least stable under noise; consensus mitigates
Temporal data	PCMCI+ (primary), Granger (secondary/baseline signal)	Explicit lag structure and condition-selection controls confounding across time better than static methods applied naively to unrolled time series

Data regime	Best-suited method(s) in HyCIF's ensemble	Rationale
Hidden confounders suspected	FCI (explicitly designed for latent confounders), front-door check in M9	PC/GES/LiNGAM/NOTEARS assume causal sufficiency by default and will silently mis-orient or spuriously connect under violation

This table is the operational form of the fusion rationale in Section 4 (M7): HyCIF does not run every method with equal weight everywhere but lets alpha_k (Section 5.6) adapt to which regime the data most resembles.

8. CORRELATION ≠ CAUSATION: CASES THE FRAMEWORK MUST DETECT

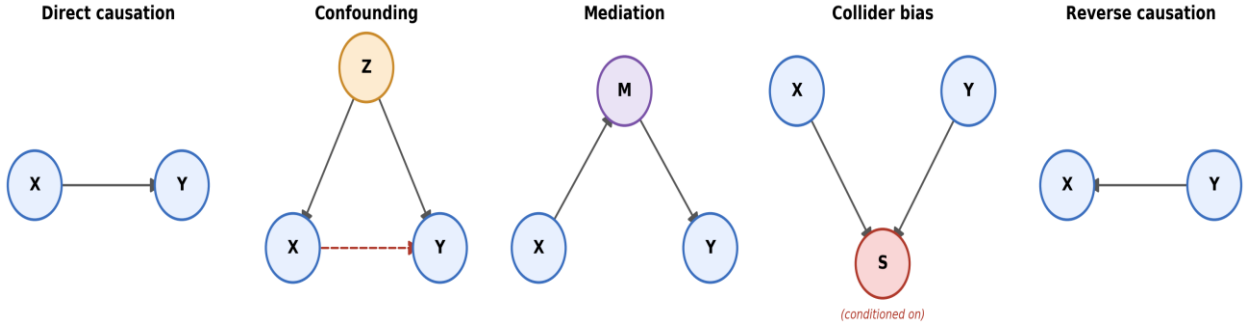

Figure 5. The five structural patterns that produce an observed X–Y association without necessarily meaning X causes Y (the sixth case, spurious correlation, has no underlying graph structure at all).

Table 5: DAG patterns behind correlation without causation

Pattern	Structure	Observational signature	HyCIF detection mechanism
Direct causation	X -> Y	X not independent of Y; not explained away by any covariate	Survives M9 adjustment; backdoor set is empty or fully measured
Reverse causation	Y -> X	Same marginal association as X-> Y	Temporal precedence (M7 temporal methods) or functional asymmetry (LiNGAM/NOTEARS residual independence) orients correctly
Confounding	Z -> X, Z -> Y	X not independent of Y marginally; X indep. Y	Z
Mediation	X -> M -> Y	X not independent of Y	empty set; effect partly/fully explained by M
Collider bias	X -> S <- Y, conditioned on S	Spurious X not independent of Y	S even if X indep. Y marginally
Spurious correlation (shared trend, small sample, multiple testing)	No causal path	Association present, unstable across resamples/time windows	s_5 (replication consistency) and BH-corrected s_2 in CTS penalize non-reproducible edges

9. EXPLAINABLE CAUSAL INTELLIGENCE: STRUCTURED OUTPUT FORMAT

Every edge that passes M9 identification and receives a CTS is rendered in a fixed five-line template so a non-specialist can audit the reasoning chain:

[Association] X is associated with Y (r = ..., MI = ...).
[Causal claim] X → Y is supported after adjusting for {Z1, Z2, ...} (backdoor-identified / front-door-identified / non-identifiable).
[Effect] do(X: x0 → x1) is estimated to change Y by β [CI: lower, upper] (CATE by subgroup where applicable).
[Counterfactual] Had X not changed from x0, Y is estimated to have remained near y0 (counterfactual consistency score: s6).
[Trust/Action] CTS = ... | Actionability Score = ... | Recommendation: intervene / monitor / insufficient evidence.

Numeric placeholders (...) are populated only after the experiments in Sections 12–17 are executed; no values are asserted in this manuscript.

10. ACTIONABLE INTELLIGENCE AND THE ACTIONABILITY SCORE (AS)

Operational definition. A causal relationship is *actionable* if (a) it is identifiable and sufficiently trusted (CTS above threshold), (b) the associated intervention is technically and organizationally feasible, and (c) the expected benefit, adjusted for cost and risk, is positive.

$$AS(e) = CTS(e) \cdot \frac{\gamma_1 \cdot \left|\widehat{ATE}(e)\right| \cdot Feas(e) \cdot Benefit(e)}{\gamma_2 \cdot Cost(e) + \gamma_3 \cdot Risk(e) + \varepsilon}$$

where Feas(e)in[0,1] is an elicited or rule-based feasibility score for enacting do(X=x) in practice, Benefit(e) and Cost(e) are domain-specific, decision-maker-supplied utility and cost functions on the same scale, Risk(e) aggregates estimation uncertainty (effect confidence interval width) and downside scenario severity, epsilon>0 avoids division by zero, and gamma_1,gamma_2,gamma_3 are domain-calibrated scaling constants (not free parameters of the causal method itself — they belong to the decision context and must be supplied or fit per application, which is stated explicitly rather than hidden inside a single opaque constant). Multiplying by CTS ensures that no relationship can score as highly actionable purely on projected benefit if it is not causally trustworthy — this is the mechanism by which HyCIF moves from *prediction* (what will happen) to *decision support* (what should be done), since a ranking by predictive feature importance alone has no analogue of the CTS gate or the feasibility/cost/risk terms.

11. TRUSTWORTHINESS ANALYSIS

Most ML deployments blur together two things that HyCIF deliberately keeps apart:

- **Prediction confidence** — how sure a statistical or ML model is about Y given X, expressed through predictive interval width or calibration error. This is a property of the model.
- **Causal confidence** — whether an estimated relationship reflects an actual mechanism, captured by the CTS. This is a property of the identified structural relationship, and it folds in identifiability, sensitivity to confounding, and counterfactual stability — none of which show up in a model's calibration error, however good that calibration is.

Trustworthiness is assessed along: causal validity (M9 identification success), robustness (bootstrap/re-sampling stability of s_1, s_5, s_6), uncertainty (confidence/credible intervals on ATE/CATE), sensitivity to confounding (s_4 / E-value analysis, Section 15), reproducibility (independent-fold agreement), explainability (Section 9 template completeness), and calibration (agreement between CTS and ground-truth correctness on synthetic benchmarks, Section 6.2). Fairness auditing (e.g., checking whether estimated CATEs differ systematically and unjustifiably across protected subgroups) is included as an optional module where the application domain and available attributes make it appropriate, and is not claimed as a general solution to fairness.

12. PROPOSED ALGORITHM

Input: Dataset D (observational and/or temporal), variable set V,
optional prior knowledge P (temporal order, known confounders)
Output: Validated causal graph G_c, ATE/CATE estimates, counterfactual explanations, CTS per edge, AS-ranked recommendations

1. D' <- Preprocess(D) # M2
2. D" <- HandleMissingAndNoise(D') # M3 (MCAR/MAR/MNAR diagnosis)
3. D"' <- EngineerFeatures(D") # M4 (lags, dedup)
4. A <- AssociationAnalysis(D"') # M5 (Pearson/Spearman/MI, lagged CC)
5. C <- CandidateEdges(A, threshold) # M6
6. for k in {PC, FCI, GES, LiNGAM, NOTEARS, PCMCI+/Granger}:
7. G_k, conf_k <- RunDiscovery(k, D"', C, P) # M7
8. alpha <- StabilitySelectionWeights({G_k}, D"', B) # Sec. 5.6
9. G_c <- ConsensusGraph({G_k}, {conf_k}, alpha, tau) # M7 fusion
10. for e = (X -> Y) in G_c:
11. Z <- IdentifyAdjustmentSet(e, G_c) # M9 backdoor/front-door
12. if Z is None and no front-door decomposition exists:
13. mark e as NON-IDENTIFIABLE; continue
14. ATE_e, CATE_e, CI_e <- EstimateEffect(e, Z, D"')# M8
15. Interv_e <- SimulateIntervention(e, Z, D"') # M10
16. CF_e <- CounterfactualConsistency(e, D"') # M11
17. s1..s6 <- ComputeSubScores(e, conf_e, ATE_e, CI_e,
confounding_sensitivity(e), CF_e) # Sec 6.1
18. CTS_e <- beta . [s1..s6] # Sec 6.2 (beta pre-calibrated)
19. AS_e <- ActionabilityScore(CTS_e, ATE_e, Feas_e, Benefit_e, Cost_e, Risk_e) # Sec 10
20. Explanation_e <- RenderTemplate(e, Z, ATE_e, CF_e, CTS_e, AS_e) # Sec 9
21. Recommendations <- RankByActionability({AS_e})
22. return G_c, {ATE_e, CATE_e}, {CF_e}, {CTS_e}, Recommendations

13. SYSTEM ARCHITECTURE

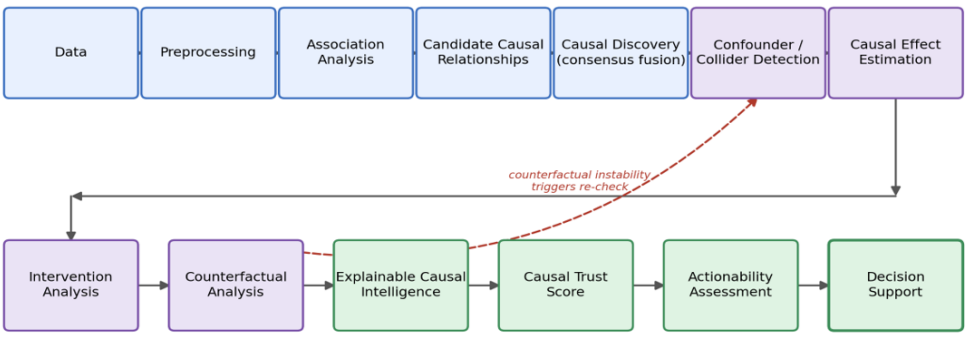

Figure 3. System architecture, showing the single feedback path from counterfactual-consistency checking back to confounder/collider detection.

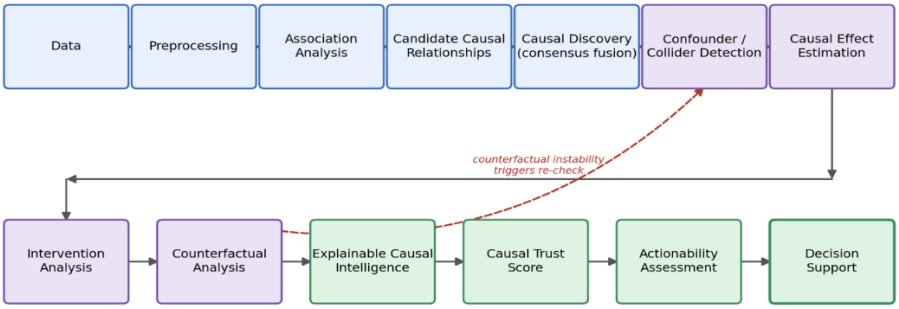

Figure 3: HyCIF system architecture with feedback loop

Information flows strictly forward with one feedback path: Data → Preprocessing → Association Analysis → Candidate Causal Relationships → (parallel) Causal Structure Learning → Consensus Graph → Confounder/Collider Detection → Causal Effect Estimation → Intervention Analysis → Counterfactual Analysis → [feedback: counterfactual instability can down-weight an edge's s_6 and trigger re-examination of its adjustment set in M9] → Explainable Causal Intelligence → Causal Trust Score → Actionability Assessment → Decision Support. The single feedback path (counterfactual instability → M9 re-check) is the mechanism by which HyCIF couples validation back into confidence, directly addressing Gap 3 in Section 2.6, rather than treating discovery and inference as one-way stages as current toolchains do.

14. EXPERIMENTAL SETUP

14.1 Datasets

Synthetic (ground-truth causal graphs, required for H1–H3): randomly generated DAGs (Erdős–Rényi and scale-free topologies) with linear-Gaussian, linear-non-Gaussian, and nonlinear structural equations, generated at multiple node counts and densities, as used in standard causal-discovery benchmarking practice (e.g., the data-generation protocols accompanying NOTEARS and the Causal Discovery Toolbox) [7], [23]; and, if available, published causal-benchmark repositories with known ground truth (e.g., the bnlearn Bayesian-network repository graphs, used as fixed structural ground truth rather than freely generated) [24].

Real-world observational: candidate domains include publicly documented health/epidemiological tabular datasets and socioeconomic datasets from UCI/OpenML/government open-data portals where a documented, literature-supported causal hypothesis exists for face-validity checking (e.g., smoking–lung-function relationships, established confounding structures in epidemiology) — exact dataset selection is deferred to the experimental phase and must be verified for licensing, size, and variable documentation before use rather than assumed here.

Real-world temporal: candidate domains include physiological time series (e.g., ICU vital-sign datasets from repositories such as PhysioNet) and Earth-observation/climate time series (e.g., NASA open data) where approximate domain-expert-endorsed causal structure is documented in the literature for partial validation.

For every dataset ultimately used, the manuscript's results section must report: name, domain, instance count, variable count, variable types, temporal/non-temporal status, availability of causal ground truth, and justification of fit to HyCIF — populated at experiment time, not fabricated here (Section 18 template).

14.2 Baselines

Method	Category	Strength	Limitation	Why selected
Pearson/Spearman correlation	Association	Simple, fast, interpretable	No causal direction or confounding control	Lower bound / sanity baseline
Random Forest / XGBoost feature importance	Predictive ML	Captures nonlinear predictive structure	Importance conflates cause, proxy, and collider	Represents common industrial practice being displaced
SHAP	Explainable ML	Theoretically grounded local attribution	Explains model, not mechanism (Sec. 2.5)	Direct comparator for the explainability claim
PC	Constraint-based discovery	Well-understood, consistent under faithfulness	CPDAG, not full DAG; power loss in high dimensions	Canonical constraint-based baseline
FCI	Constraint-based, latent-confounder-aware	Handles hidden confounders	Weaker identifiability (PAG)	Confounding-robustness comparator

Method	Category	Strength	Limitation	Why selected
GES	Score-based discovery	Global score optimization	Greedy search, super-exponential space	Canonical score-based baseline
LiNGAM	Functional discovery	Full DAG orientation, no CI testing	Requires linear, non-Gaussian data	Functional-method comparator
NOTEARS	Continuous-optimization discovery	Scales to higher dimensions	Non-convex acyclicity constraint; functional-form assumptions	Modern scalable baseline
Granger causality	Temporal	Simple, widely used	Predictive, not interventional; confounded by shared drivers	Legacy temporal baseline
PCMCI+	Temporal causal discovery	Condition-selection controls high-dim confounding	Requires causal sufficiency, correct max-lag	State-of-the-art temporal baseline
DoWhy (effect estimation given a graph)	Causal inference	Principled identify–estimate–refute pipeline	Assumes graph correctness	Effect-estimation comparator isolating HyCIF's discovery contribution

14.3 Evaluation Metrics

Metric	Measures	Why appropriate
Precision / Recall / F1 (edge-level)	Discovery accuracy vs. ground truth	Standard for graph recovery on synthetic benchmarks
Structural Hamming Distance (SHD)	Graph-edit distance to ground truth	Penalizes missing, extra, and mis-oriented edges jointly
Structural Intervention Distance (SID)	Interventional-distribution discrepancy	Captures errors that matter for *do*-queries, not just edge-matching
ATE/CATE estimation error (bias, RMSE, MAE)	Effect-estimation accuracy	Directly tests H2
AUC (for identifiability/edge-correctness classification via CTS)	Discriminative validity of CTS	Directly tests H3
Calibration error (CTS vs. empirical correctness)	Trustworthiness of the trust score itself	CTS is only useful if it is calibrated, not just discriminative
Robustness score (metric variance under bootstrap/perturbation)	Stability	Tests H1/H2 under noise
Computational time / memory	Efficiency	Required for practical deployment claims
Actionability-ranking quality (e.g., ranking correlation with true-effect × feasibility)	Decision relevance	Directly tests H4

14.4 Experimental Configuration

To be specified at execution time: train/validation/test or resampling scheme (e.g., repeated k-fold with fixed seeds), hyperparameters and their selection procedure for every baseline and for HyCIF's own thresholds (tau, CTS decision threshold), hardware, and software versions — all left as an explicit template rather than invented here.

15. STATISTICAL VALIDATION PLAN

For every metric in Section 14.3, HyCIF-vs-baseline comparisons will use: paired tests appropriate to the metric's distribution (e.g., Wilcoxon signed-rank for non-normal paired differences across resampling folds, paired t-test where normality holds), bootstrap confidence intervals (percentile method, ≥1000 resamples) for effect-size differences, Benjamini–Hochberg correction for multiple comparisons across the metric × baseline grid, and reporting of standardized effect size (e.g., Cliff's delta or Cohen's d as appropriate) alongside p-values so that statistical and practical significance are both visible. Confounding-robustness claims (s_4) will additionally use an E-value-style sensitivity analysis, reporting how strong an unmeasured confounder's association with both treatment and outcome would need to be to fully explain away the observed effect.

16. ABLATION STUDY DESIGN

Starting from the full HyCIF pipeline, the following components are removed one at a time (and, where meaningful, in combination) and re-evaluated on the full metric suite (Section 14.3):

Configuration	Component removed	Expected diagnostic value
Full HyCIF	—	Reference
− Hybrid discovery (single method only, e.g., PC only)	M7 fusion	Isolates the benefit of consensus fusion (tests H1)
− Confounder/collider detection	M9	Isolates effect of skipping identification checks (tests H2)
− Intervention analysis	M10	Tests whether interventional simulation adds value beyond adjustment-based estimation alone
− Counterfactual module	M11	Isolates s_6's contribution to CTS quality (tests H3)
− Explainability module	M12	Assessed qualitatively / via user-study proxy, not a quantitative metric in 14.3
− Trust score (uniform weighting instead of calibrated beta)	M13 (Sec. 6.2)	Tests whether optimization-derived weights outperform naive equal weights
− Temporal analysis (PCMCI+/Granger excluded from fusion)	M7 (temporal arm)	Isolates temporal methods' contribution on temporal datasets only

17. RESULTS

All tables below are structural templates. Cells are left as placeholders (TBD) and must be populated only after experiments are actually executed; no numbers, confidence intervals, or significance markers are asserted in this manuscript.

Table 17.1 — Causal discovery accuracy (synthetic, ground-truth graphs).

Method	Precision	Recall	F1	SHD	SID	Runtime
PC	TBD	TBD	TBD	TBD	TBD	TBD
FCI	TBD	TBD	TBD	TBD	TBD	TBD

Method	Precision	Recall	F1	SHD	SID	Runtime
GES	TBD	TBD	TBD	TBD	TBD	TBD
LiNGAM	TBD	TBD	TBD	TBD	TBD	TBD
NOTEARS	TBD	TBD	TBD	TBD	TBD	TBD
PCMCI+ / Granger	TBD	TBD	TBD	TBD	TBD	TBD
HyCIF (consensus)	TBD	TBD	TBD	TBD	TBD	TBD

Table 17.2 — Causal effect estimation accuracy.

Method	ATE bias	ATE RMSE	CATE RMSE	Coverage of 95% CI
DoWhy (on true graph)	TBD	TBD	TBD	TBD
DoWhy (on best single-method discovered graph)	TBD	TBD	TBD	TBD
HyCIF (consensus graph + M9 gating)	TBD	TBD	TBD	TBD

Table 17.3 — CTS discriminative and calibration quality (H3).

Trust indicator	AUC (edge correctness)	Calibration error
Correlation coefficient	TBD	TBD
p-value (BH-corrected)	TBD	TBD
SHAP magnitude	TBD	TBD
CTS (HyCIF)	TBD	TBD

Table 17.4 — Ablation results (full metric suite per row of Section 16). *(Structure: one row per configuration in Table of Section 16, one column per metric in Section 14.3 — populated at experiment time.)*

Table 17.5 — Actionability ranking quality (H4). *(Ranking correlation, e.g. Kendall's tau, between AS-based ranking and true-effect × feasibility ranking, on synthetic benchmarks with known ground truth for both effect and simulated feasibility/cost.)*

Interpretive text accompanying each populated table must explicitly separate (a) results directly supporting/refuting H1–H4, (b) exploratory observations not tied to a pre-registered hypothesis, and (c) hypothetical illustrative examples used only for explanatory purposes in Section 9 — none of which are to be conflated with (a).

18. THREATS TO VALIDITY AND LIMITATIONS

Internal validity. Consensus fusion (Section 4, M7) can still fail if *all* constituent methods share a common blind spot (e.g., all assume causal sufficiency while it is violated) — FCI's inclusion mitigates but does not eliminate this. **Construct validity.** The CTS is only as good as its synthetic calibration set (Section 6.2); if synthetic graphs do not represent the real deployment domain's structural regime, calibrated beta weights may transfer poorly, which is why real-world face-validity checks (Section 14.1) are retained alongside synthetic ground truth rather than substituted for it. **External validity.** Real-world datasets used for face validation typically lack complete ground-truth causal graphs, so H1/H3 can only be fully tested synthetically; real-data results support plausibility, not proof. **Statistical conclusion validity.** Multiple-comparison correction (Section 15) reduces but does not eliminate the risk of false-positive superiority claims across the large metric × baseline grid. **Scope limitation.** HyCIF assumes acyclicity and does not natively handle instantaneous feedback loops; it does not resolve non-identifiable queries, and correctly reports them as such rather than approximating them.

19. PRACTICAL IMPLICATIONS

In domains where actions are expensive or hard to reverse — health interventions, policy decisions, industrial process control — acting on a strong predictor that isn't actually a cause can do real damage. HyCIF's separation of prediction confidence from causal confidence, its willingness to report non-identifiability instead of guessing, and the feasibility/risk weighting built into the Actionability Score are all aimed at that failure mode specifically. The structured explanation template from Section 9 has a related purpose: making the reasoning chain legible to domain experts who aren't causal-inference specialists, which matters as much for regulatory and governance review of automated recommendations as it does for day-to-day decision-making.

20. CONCLUSION AND FUTURE WORK

The central claim of this paper is narrower than it might first appear: not that any single discovery algorithm can be replaced, but that fusing several of them under an evidence-weighted consensus rule, and gating the result through confounder screening, do-calculus effect estimation, and counterfactual-consistency checking, produces a relationship worth acting on more reliably than any one stage does alone. HyCIF's fifteen modules were formalized end to end, and its two headline outputs — the Causal Trust Score and the Actionability Score — were defined so that their internal weights come from calibration against known structure rather than from judgment calls buried in the method. What is missing from this manuscript is deliberate: no numerical result is reported, because none has yet been produced. What is provided instead is a protocol specific enough to be run and to fail on its own terms — synthetic graphs with known structure, real datasets, eight baselines, nine metrics, an eight-way ablation, and a pre-registered statistical testing plan (Sections 14–16). Three extensions seem worth pursuing once that protocol has been executed: relaxing the acyclicity assumption to admit instantaneous feedback, letting the framework select which real intervention to run next in order to shrink CTS uncertainty fastest (an active-learning formulation), and testing — with actual decision-makers, not just metrics — whether the Section 9 explanation template changes how well they calibrate their trust in a causal claim.

ACKNOWLEDGEMENTS

The authors thank the Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), for institutional support during the preparation of this manuscript.

AUTHOR CONTRIBUTIONS

Mr B.B.L.V. Prasad: Conceptualization, Methodology, Formal Analysis, Writing — Original Draft.

Dr E. Hemalatha: Supervision, Validation, Writing — Review & Editing.

Both authors have read and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this article.

ETHICS STATEMENT

This work is a methodological and theoretical framework paper. It does not involve human participants, human data, or animal subjects, and no experiments were conducted on live systems as part of this manuscript; accordingly, ethical approval was not required.

DATA AVAILABILITY STATEMENT

No new empirical data were generated or analyzed in this manuscript, as the reported evaluation protocol (Sections 12–16) is specified but deliberately left unexecuted. Code and configuration for the proposed framework will be made available from the corresponding author upon reasonable request once the evaluation is carried out.

REFERENCES

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[3] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[4] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, 2006.

[5] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artif. Intell.*, vol. 172, no. 16–17, pp. 1873–1896, 2008.

[6] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2002.

[7] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in *Proc. NeurIPS*, 2018.

[8] J. Runge et al., "Detecting and quantifying causal associations in large nonlinear time series datasets," *Sci. Adv.*, vol. 5, no. 11, eaau4996, 2019.

[9] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. NeurIPS*, 2009.

[10] A. Sharma and E. Kiciman, "DoWhy: An end-to-end library for causal inference," arXiv:2011.04216, 2020.

[11] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016.

[14] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD*, 1993.

[15] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Front. Genet.*, vol. 10, art. 524, 2019.

[16] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[17] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos*, vol. 28, no. 7, 075310, 2018.

[18] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning — problems, methods and evaluation," *ACM SIGKDD Explor. Newsl.*, vol. 22, no. 1, pp. 18–33, 2020.

[19] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like DAGs? A survey on structure learning and causal discovery," *ACM Comput. Surv.*, vol. 55, no. 4, art. 82, 2022.

[20] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[21] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 53, no. 4, art. 75, 2020.

[22] N. Meinshausen and P. Bühlmann, "Stability selection," *J. R. Stat. Soc. Ser. B*, vol. 72, no. 4, pp. 417–473, 2010.

[23] D. Kalainathan, O. Goudet, and R. Dutta, "Causal discovery toolbox: Uncovering causal relationships in Python," *J. Mach. Learn. Res.*, vol. 21, no. 37, pp. 1–5, 2020.

[24] M. Scutari, "Learning Bayesian networks with the bnlearn R package," *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, 2010.