EXPECTATION REASONING FOR EXPLORING FUNCTIONAL FORMS OF DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Network reconstruction becomes particularly challenging when both the interaction structure and the dynamics governing the observations are unknown. We investigate whether an LLM-driven scientific process can explore effective dynamical response laws that improve reconstruction under a fixed graph-reconstruction backend. A closer examination of the research process reveals two limitations of proxy-driven optimization. First, improving the fitting proxy does not necessarily improve downstream network recovery. Second, the proxy-optimized realization provides insufficient evidence to determine whether a proposed dynamical mechanism is actually realized, whether it has a meaningful effect, or whether it remains worth investigating. Thus, successful proxy optimization can support numerical realization without providing an adequate basis for evaluating the scientific status of the proposed mechanism. We introduce Expectation Reasoning (ER), which separates numerical realization, evidence formation, epistemic revision, and solution commitment within the scientific-search process. Across two truth-blind CausalDynamics comparison cohorts totaling 20 cases, the recorded ER trajectories preserve contradictory, partial, and unresolved evidence across research rounds and exhibit a clearer separation between epistemic revision and candidate commitment. Downstream effects remain criterion-dependent: ER reduces fixed-threshold structural error most clearly in the primary cohort, while AUROC and AUPRC do not improve and the resulting reconstructions become substantially sparser. Taken together, our findings support separating scientific evidence and revision from proxy-driven numerical realization, rather than treating optimization outcomes as sufficient scientific decisions.

1 INTRODUCTION

Network reconstruction seeks to recover hidden interactions from observed multivariate dynamics (Zhang et al., 2022; Herdeanu et al., 2026). When the underlying dynamics are unknown, we study a controlled setting in which the network-reconstruction procedure and its configuration are kept fixed, while the effective dynamical response law is treated as a variable to be explored for each reconstruction instance. Our goal is not to recover the true generating dynamics, but to determine whether finding a response law better suited to the current reconstruction instance can improve structural recovery. Large language models (LLMs) provide a flexible way to explore this open function space by proposing candidate response laws, consistent with their emerging use for scientific ideation and automated research (Lu et al., 2024; Si et al., 2025). For each candidate law, numerical optimization searches over its parameter settings, and the fixed network-reconstruction procedure re-estimates the network at each setting. The resulting response law and reconstructed network are then evaluated together by recursively predicting the observed dynamics and computing the mean-squared error (MSE), which serves as the observable fitting signal for the search. Our initial implementation of this pipeline already achieved strong network reconstruction performance, but a closer examination of the research process revealed two limitations of this process that are not apparent from the final reconstruction results alone.

After completing the search with the ground-truth network structure kept hidden throughout, we used the ground truth only for retrospective analysis and identified two recurring limitations. First, prediction MSE remained useful as a fitting signal, but candidates with lower MSE did not consistently correspond to better network reconstruction when evaluated by AUPRC, AUROC, and structural

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.

Hamming distance (SHD). MSE alone was therefore insufficient to determine which response hypothesis should be preferred for the downstream reconstruction task. We refer to this as a proxy–task mismatch. Second, MSE optimization could drive the parameters controlling a proposed functional effect toward values that nearly turn the effect off. Other parameter settings could activate the same effect, and its impact on network reconstruction could change once the network was re-estimated. Suppression in one MSE-preferred setting was therefore insufficient to conclude that the proposed effect was ineffective or not worth further investigation. We refer to this as a proxy–mechanism mismatch. Together, the two mismatches show that the fitting objective alone cannot determine either which response hypothesis is better for downstream network reconstruction or whether a proposed functional effect should be retained, revised, or rejected. More broadly, preference under the fitting objective cannot substitute for experimental evaluation of the hypothesis itself.

These observations motivate Expectation Reasoning (ER). ER complements the existing proxy-based search with an explicit hypothesis–evidence loop for evaluating and revising response hypotheses. At the time a response hypothesis is proposed, ER also specifies the observable effects expected from it and the evidence needed to test them, and fixes both before the corresponding outcomes are examined. The resulting evidence is evaluated against these fixed expectations and carried across research rounds to revise how the hypothesis is assessed, without forcing every new result into an immediate candidate replacement. Only after a pre-specified research horizon does the process select a final response law, still without access to the ground-truth network structure.

We evaluate ER across two truth-blind CausalDynamics cohorts totaling 20 cases (Herdeanu et al., 2026), keeping the ground-truth graph hidden throughout the research process and final selection and revealing it only afterward for evaluation. Across both cohorts, ER produced research trajectories that followed the intended separation of evidence formation, epistemic revision, and solution commitment. Both the initial LLM-driven process and ER remain competitive with the included benchmarks, but ER does not improve all reconstruction metrics: fixed-threshold SHD decreases most clearly in the primary cohort, while AUROC and AUPRC decrease and the resulting reconstructions become substantially sparser. We therefore interpret these results primarily as support for separating evidence-based hypothesis evaluation and revision from proxy-driven numerical realization, rather than as evidence of uniform improvement in network-reconstruction performance.

Our contributions are threefold:

- We empirically identify and characterize two recurring mismatches in LLM-guided, proxy-driven response-law search for network reconstruction: candidates preferred by prediction MSE are not consistently preferred by downstream structural metrics, and a proxy-preferred parameter setting can suppress a proposed functional effect without establishing that the effect itself is ineffective.
- We introduce Expectation Reasoning (ER) for proxy-driven response-law search, which keeps proxy-based numerical realization separate from hypothesis-level evidence evaluation by pairing each proposed response hypothesis with pre-outcome observable expectations and evidence requirements, and carrying the resulting evidence forward for cross-round revision.
- We evaluate ER across two truth-blind CausalDynamics cohorts totaling 20 cases, examining whether its intended separation of evidence formation, epistemic revision, and solution commitment is realized and how it affects downstream reconstruction. ER consistently exhibits this separation, while reconstruction effects remain case- and metric-dependent.

2 METHOD

Expectation Reasoning (ER) builds on an LLM-guided exploration process and reorganizes how proposed Function Families are evaluated and carried forward across research rounds. As illustrated in Figure 1, ER separates three stages: proposing Function Families together with pre-outcome Expectations, evaluating these Expectations against factual evidence, and consolidating the resulting understanding into the Research State. Numerical fitting remains responsible for realizing candidate functions from data, while Expectation–Reality relations are recorded without directly prescribing the next research action or forcing an immediate candidate replacement.

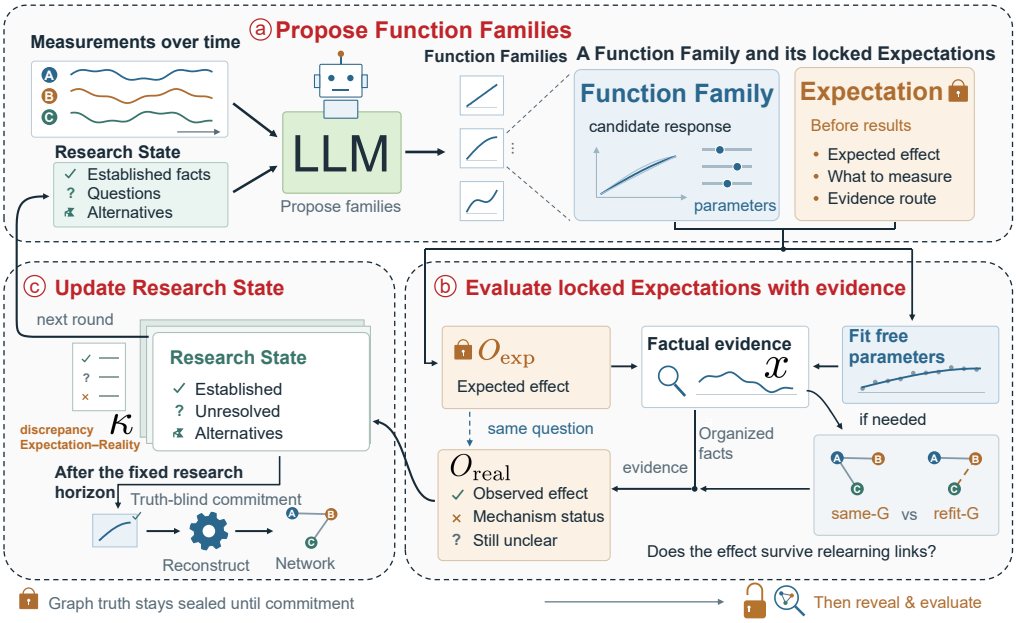


Figure 1: **Operational overview of Expectation Reasoning.** (a) From the observational data and the current Research State, the LLM proposes parameterized Function Families together with their intended effects. Before the corresponding outcomes are observed, these effects are expressed as locked Expectations specifying the expected observation, what should be measured, and the evidence route; a Function Family may carry multiple Expectations. (b) Numerical fitting realizes the free parameters, while experiments produce factual evidence x. When the available evidence is insufficient to answer a locked question, an additional targeted experiment may be introduced; the same-G/refit-G comparison illustrates one such test. Using numerical tools, the LLM organizes the acquired evidence into a Reality O_{real} that answers the same question, and the Expectation–Reality discrepancy is retained as κ. (c) The resulting findings, unresolved questions, and remaining alternatives are consolidated into the Research State and passed to the next round without prescribing its next action. After the fixed research horizon, a truth-blind commitment selects one previously evaluated response law for network reconstruction, after which graph truth is revealed only for evaluation.

2.1 Candidate Responses and Pre-Outcome Expectations

An LLM proposes each candidate as a Function Family with free parameters, together with the effect that the candidate is intended to produce. Numerical tools then fit the free parameters against observable prediction error. As illustrated in Figure 1(a–b), this division allows the LLM to explore different functional structures while leaving their numerical realization to data-driven optimization. We use the Function Family to denote the proposed functional structure and its parameterized realizations, rather than treating any single optimized parameter setting as the hypothesis itself.

The fitted realization, however, need not actually express the intended effect that motivated the candidate. A parameter setting can improve the fitting proxy while weakening or even suppressing that effect. Parameter fitting therefore tells us how a Function Family can be numerically realized under the proxy, but not whether the effect it was introduced to produce is actually expressed in the resulting behavior.

Expectation Reasoning evaluates this intended effect separately from parameter fitting. As illustrated in Figure 1(a), before the corresponding experimental outcome is observed, each intended effect is associated with an Expectation O_{exp}, which states what should be observed if that effect is actually realized. For example, if a candidate is intended to reduce excessive fluctuation, its Expectation specifies that the relevant fluctuation should measurably decrease under an appropriate test. A Function Family may therefore carry multiple Expectations, each corresponding to a different intended effect. These Expectations are not additional optimization objectives or scores for ranking candidates; they specify the scientific questions that subsequent evidence must answer.

2.2 Evidence Formation and Expectation–Reality Comparison

Once an Expectation has been specified, it provides a pre-outcome reference: it states both the scientific question to be examined and what should be observed if the intended effect is realized. The observed content is not required to agree with this reference. As illustrated in Figure 1(b), each Function Family is first examined through a fixed experimental base, which produces a common set of factual observations x. If these observations are sufficient to answer the pre-specified question, they are used directly as evidence. Otherwise, ER introduces a targeted experiment to resolve the remaining evidential gap. The concrete experiment may adapt to facts already obtained, but it remains tied to the original question rather than being reformulated around the observed evidence.

The acquired evidence x may consist of raw measurements, trajectories, or derived numerical results. Such evidence is often more detailed and expressed differently from the corresponding Expectation, and therefore cannot always be compared with O_{exp} directly. Using the available numerical tools, the LLM computes or retrieves the required summaries and organizes the resulting evidence into a Reality O_{real} that answers the same scientific question in a comparable form. For example, an Expectation may state that a candidate effect should produce an approximately linear relation, while the experiment returns a collection of raw response measurements. These measurements can be analyzed and summarized into the Reality that the observed relation is instead distinctly nonlinear. O_{real} therefore does not alter the experimental facts to resemble the Expectation; it expresses, in the same question space, what those facts actually show.

Once O_{exp} and O_{real} answer the same scientific question in comparable form, ER stores their correspondence as κ. Rather than reducing the comparison to a scalar score, κ retains the Expectation, the corresponding Reality, and the discrepancy between them as a structured research record. In the preceding example, it records that an approximately linear relation was expected whereas a distinctly nonlinear relation was observed. At this stage, κ only records the Expectation–Reality relation; it does not determine whether the candidate should be replaced or what the next investigation should be. Those decisions are separated from evidence formation and are addressed in the subsequent research process.

2.3 Research State and Solution Commitment

As illustrated in Figure 1(c), at the end of each research round, the LLM consolidates the accumulated evidence into a Research State that represents the current scientific understanding. This state retains the understanding formed from the available evidence, the recorded Expectation–Reality relations, unresolved uncertainties, and competing explanations that remain plausible. For example, a round may record that an expected effect was not observed while retaining another explanation that remains compatible with the evidence. The Research State describes what is currently understood and what remains open; it does not turn κ into an action rule or prescribe a new Function Family, experiment, or parameter search.

At the beginning of the next round, the Planner reads the Research State together with the observational data and determines what question should be investigated next. The next round therefore inherits a state of knowledge rather than a prescribed next action. The current evidence, uncertainties, and competing explanations provide the context for subsequent reasoning, while the LLM remains responsible for deciding what should be investigated next. In this sense, ER separates epistemic revision from action prescription: one round changes the context from which the next round reasons without directly specifying what the next round must do.

Updating the Research State also does not require the currently retained candidate to change immediately. A research round may change how a Function Family is understood while leaving the currently retained candidate unchanged or leaving the choice unresolved. ER therefore separates epistemic revision from solution commitment rather than forcing every change in scientific understanding into a new candidate choice. Only after the fixed research horizon does the system perform a truth-blind final selection among the previously evaluated Function Families, choosing one for downstream network reconstruction.

3 EXPERIMENTS

3.1 EXPERIMENTAL PROTOCOL AND EVIDENCE ROLES

We use 30 pre-generated Tier-2 coupled CausalDynamics cases (Herdeanu et al., 2026). Each case contains ten observed nodes and ten trajectories of 1,000 time points that share one summary adjacency. The cases have distinct evidential roles. CD001–CD010 are development cases used only to identify and characterize the proxy–task and proxy–mechanism mismatches that motivate ER. CD011–CD020 form the primary truth-blind comparison cohort, with periodic coupling and zero noise, while CD021–CD030 form a second truth-blind cohort with nonlinear coupling and noise level 1.0. We report the two comparison cohorts separately because their data conditions and execution traces differ; 20-case aggregates are descriptive only. Published CausalDynamics reference results are matched to these cases by the full data configuration and source filename.

Each trajectory is partitioned into $\mathcal{D}_{\mathrm{fit}} = 0{:}600$, $\mathcal{D}_{\mathrm{cal}} = 600{:}800$, and $\mathcal{D}_{\mathrm{val}} = 800{:}1000$. For every candidate response law, the graph-fitting algorithm and its configuration remain fixed, while the graph itself is re-estimated for that response. The objective on $\mathcal{D}_{\mathrm{cal}}$ is available during research as the numerical fitting proxy. Structural ground truth and all truth-dependent quantities remain sealed through candidate generation, evidence formation, cross-round state revision, and terminal commitment. Candidate-pool oracles, structural ranks, selection regret, oracle thresholds, and other truth-aware quantities are computed only after unsealing and are used only for post-hoc diagnosis.

The inherited process and ER use the same GPT-5 scientific model, five-round research horizon, data partition, Planner observation cap, SMAC seed (Hutter et al., 2011), and graph-reconstruction backend. The inherited process allocates 256 numerical trial equivalents to each Function Family. ER uses a 176-trial fixed numerical base and may additionally invoke targeted evidence programs. The comparison is therefore protocol-controlled rather than compute-matched. AUROC and AUPRC measure edge-ranking quality, while SHD measures structural error after thresholding at 0.5. Off-diagonal quantities are reported only when localizing changes in cross-node recovery. Paired cases are the unit of statistical comparison; candidate-level quantities are treated as descriptive unless first summarized within case.

3.2 EMPIRICAL RECURRENCE OF TWO PROXY MISMATCHES

We use CD001–CD010 to characterize the two proxy mismatches identified during development and CD011–CD030 to test whether the same candidate-level patterns recur in subsequent research runs (Fig. 2). **Proxy–task mismatch.** Across 134 development candidates, the median case-wise Spearman correlations between the D_{cal} ranking and downstream AUROC, AUPRC, and SHD rankings are -0.065, 0.418, and -0.034, respectively. After balancing the Inherited and ER runs within each of the 20 subsequent cases, the corresponding medians are 0.113, 0.315, and -0.311, with a negative proxy–SHD association in $18/20$ cases. Figure 2(a–c) shows the underlying candidate-level rankings. The predictive proxy therefore retains criterion-dependent downstream information but does not induce a consistent structural ordering. **Proxy–mechanism mismatch.** Figure 2(d–f) shows how local functional gain covaries with the graph refitted for the same candidate. The association between $\mathrm{median}\,|f'(z)|$ and $\|G_{\mathrm{off}}\|_2$ is negative in all $10/10$ development cases and remains negative in all $20/20$ subsequent cases after balancing the two process runs, with median Spearman $\rho = -0.546$ and -0.696, respectively (exact sign test for the subsequent cases, $p = 1.91 \times 10^{-6}$). A complementary shutdown audit using the same family-specific criterion finds suppression in $22/72$ eligible development candidates and $41/154$ eligible candidates across the two CD011–CD020 processes; active realizations are subsequently found for $18/22$ and $36/41$ suppressed candidates, respectively. Thus, proxy-preferred suppression does not establish that the corresponding mechanism is unreachable. Together, the recurrent task-ranking inconsistency, mechanism suppression, and function–graph compensation show that proxy optimization can determine a numerical realization without consistently determining either its downstream structural preference or the evidential status of the mechanism it instantiates.

3.3 DOES ER CHANGE THE RECORDED RESEARCH PROCESS?

We next audit the recorded research process without using graph truth. **Expectation–Reality correspondence.** Across the final-selected ER Function Families, all $60/60$ Expectations in CD011–CD020

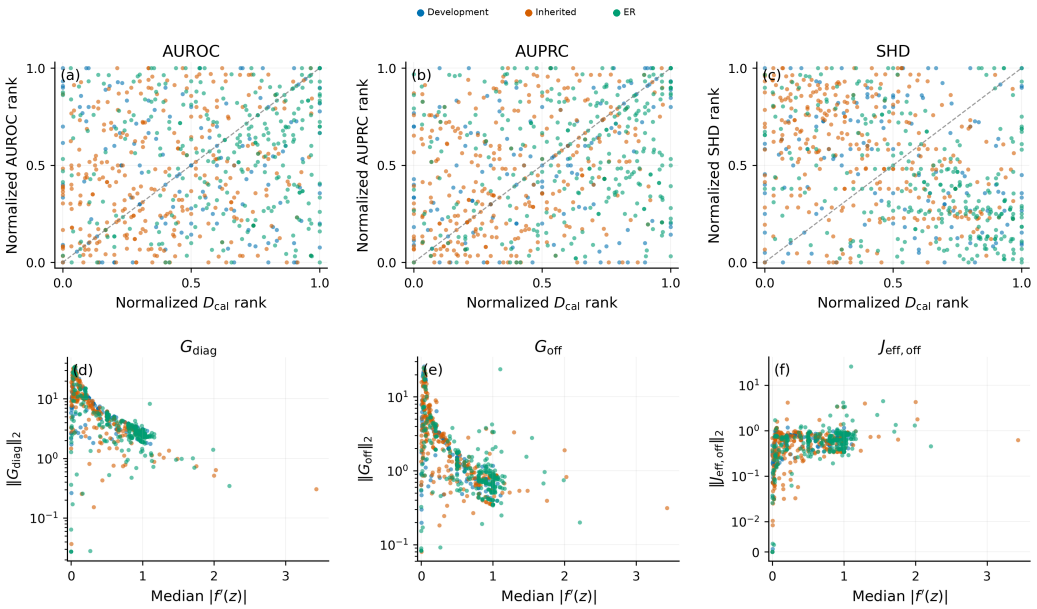


Figure 2: **Candidate-level recurrence of the two proxy mismatches.** Colors denote development candidates from CD001–CD010, subsequent Inherited candidates from CD011–CD030, and subsequent ER candidates from CD011–CD030. Panels (a–c) compare within-case normalized D_{cal} rank with downstream AUROC, AUPRC, and SHD ranks; zero denotes the best candidate and the dashed line identical rankings. Panels (d–f) relate $\mathrm{median}\,|f'(z)|$ to $\|G_{\mathrm{diag}}\|_2$, $\|G_{\mathrm{off}}\|_2$, and $\|J_{\mathrm{eff,off}}\|_2$. Points are candidate-level; statistical summaries are computed at the case level. All truth-dependent quantities are evaluated only after unsealing.

and all $68/68$ in CD021–CD030 preserve the same scientific question through the corresponding Reality and discrepancy record. The resulting evidence is not uniformly confirmatory: $40/60$ (66.7%) and $42/68$ (61.8%) of the Expectations, respectively, retain contradiction, partial agreement, or unresolved evidence (Table 1). The detailed outcome breakdown is reported in the appendix.

Table 1: **Truth-blind ER evidence and continuity audit.** Correspondence is evaluated on the final-selected ER Function Family in each case; continuity covers four cross-round transitions per case. Detailed semantic labels are reported in the appendix.

Audit	CD011–020	CD021–030
Same-question correspondence	60/60	68/68
Non-confirmatory evidence	40/60 (66.7%)	42/68 (61.8%)
Next Research State	32 clear / 8 partial	35 clear / 5 partial
Next Planner questions	27 clear / 13 partial	32 clear / 8 partial
Explicit next-action prescription	0/40	0/40

Epistemic revision versus solution commitment. In CD011–CD020, candidate commitment changes in $18/40$ Inherited transitions and $7/40$ ER transitions, while state revision with unchanged commitment occurs in $22/40$ and $33/40$, respectively (case-level paired Wilcoxon, $p = 0.0625$). The same separation is stronger in CD021–CD030: commitment changes in $20/40$ and $3/40$, while state revision with unchanged commitment occurs in $20/40$ and $37/40$ ($p = 0.0078$). Across both ER cohorts, every audited source discrepancy or uncertainty remains represented at least partially in the next Research State and subsequent Planner questions, while none of the source states explicitly prescribes the next research action (Table 1). The recorded trajectories therefore exhibit both epistemic revision without mandatory solution replacement and cross-round continuity without direct action prescription.

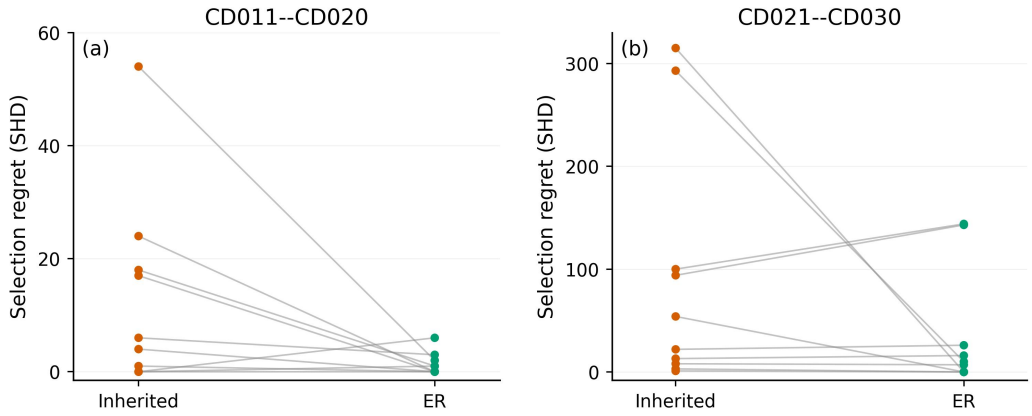


Figure 3: **Post-unsealing within-pool selection regret across the two comparison cohorts.** Each line connects the Inherited and ER regrets for the same case; zero denotes selection of the minimum-SHD candidate in that process's own eligible pool. Panels use separate vertical scales because CD021–CD030 contains substantially larger failures. The processes generate separate pools, so the figure localizes terminal choices within their observed pools rather than comparing two selectors on an identical candidate set.

3.4 Where Do the Terminal Choices Fall in Their Candidate Pools?

After unsealing, we locate each terminal choice within the eligible candidate pool produced by its own process. For case c and process m, let O_c^m be the minimum SHD in that pool and define the post-hoc selection regret as $R_c^m = S_c^m - O_c^m$, where S_c^m is the selected candidate's SHD at the same scoring stage. In CD011–CD020, mean pool-best SHD changes only from 137.5 for Inherited to 135.5 for ER, whereas mean selection regret changes from 12.4 to 1.3. The 13.1-point difference between the same-stage selected means therefore decomposes arithmetically into 2.0 points at the pool minimum and 11.1 points in selection regret, placing 84.7% of the observed difference in the latter term. Figure 3 shows the corresponding case-wise regrets.

The second cohort shows the same separation under substantially larger failures. Mean pool-best SHD remains close (128.5 versus 127.4), while mean selection regret changes from 90.3 to 34.6. Of the 56.8-point difference between selected means, only 1.1 points arise at the pool minimum and 55.7 points in regret. This pattern is less uniform than in the first cohort, as Fig. 3 makes explicit, and therefore does not imply that ER reliably selects a near-oracle candidate. Across both conditions, the post-hoc diagnosis instead places most of the observed SHD difference in where each process commits within its own candidate pool, rather than in a markedly lower observed pool minimum. Because the pools are generated separately, this localization neither establishes equivalent search capacity nor identifies a causal contribution of any individual ER component. Complete pool sizes, rank-based diagnostics, mechanical proxy selections, and case-wise oracle values are reported in the appendix.

3.5 Downstream Consequences and Empirical Boundary

Table 2 reports the fixed terminal reconstructions under the official task metrics and same-instance published CausalDynamics references (Herdeanu et al., 2026). In CD011–CD020, ER lowers SHD in seven cases and ties three relative to Inherited (paired two-sided Wilcoxon, $p = 0.0156$), whereas AUROC and AUPRC favor Inherited and their paired differences are not statistically significant ($p = 0.322$ and $p = 0.193$). In CD021–CD030, the mean SHD direction is preserved but is less uniform, with five ER improvements, one tie, and four regressions ($p = 0.441$); AUROC and AUPRC again favor Inherited ($p = 0.492$ and $p = 0.0645$). ER remains competitive with the published references in both conditions, but the downstream evidence supports a metric-specific structural trade-off rather than uniform reconstruction improvement.

Table 2: **Same-instance network reconstruction across the two comparison cohorts.** Values are case-macro Joint AUROC, AUPRC, and SHD over ten cases. Inherited and ER are the recorded research processes; other rows are official published CausalDynamics references matched by full generation configuration and source filename. Bold marks the best mean in each cohort and metric.

	CD011–CD020			CD021–CD030		
Method	AUROC ↑	AUPRC ↑	SHD ↓	AUROC ↑	AUPRC ↑	SHD ↓
ER	0.7078	0.4001	**136.5**	0.6953	0.3691	162.0
Inherited	**0.7157**	**0.4211**	146.7	**0.7218**	**0.4230**	218.8
PCMCI+(Runge, 2020)	0.7017	0.2832	225.7	0.6550	0.2657	179.5
F-PCMCI(Castri et al., 2023)	0.6907	0.2567	293.4	0.6147	0.2200	200.1
VARLiNGAM(Hyvärinen et al., 2010)	0.5902	0.1874	507.1	0.5624	0.1721	368.0
DYNOTEARS(Pamfil et al., 2020)	0.6026	0.2314	174.0	0.6356	0.2797	148.7
NGC-LSTM(Tank et al., 2021)	0.5006	0.1572	835.9	0.4999	0.1510	840.5
TSCI(Butler et al., 2024)	0.6615	0.2983	156.4	0.5224	0.1814	**145.9**
CUTS+(Cheng et al., 2024)	0.5156	0.1625	286.6	0.5000	0.1510	151.0
RCD(Maeda & Shimizu, 2020)	0.5001	0.1570	157.0	0.5010	0.1519	151.1

The SHD changes must be interpreted together with the fixed threshold and edge-ranking behavior. In CD011–CD020, ER predicts only 13 positive off-diagonal edges across the ten cases, compared with 193 for Inherited, and off-diagonal recall decreases from 6.33% to 0.33%. Its mean off-diagonal SHD of 90.7 is correspondingly close to the 90.0 empty-off-diagonal reference. CD021–CD030 shows the same shift toward fewer predicted edges, from 1,288 to 390, while ER's off-diagonal recall remains only 6.11% and its off-diagonal SHD of 118.0 stays above the empty reference. The lower fixed-threshold SHD should therefore be interpreted primarily as a shift toward stronger sparsification and false-positive suppression, not as uniformly improved recovery of cross-node interactions. Full confusion counts and post-unsealing threshold diagnostics are reported in the appendix.

The second cohort also exposes an evidence-to-commitment boundary. In CD029, ER retains positive matched-control evidence for a local curvature effect together with unresolved amplitude, temporal-fidelity, small-signal, and bias conditions, yet still commits to that family. After unsealing, the selected candidate has SHD 273 while another eligible candidate in the same pool reaches 130. Thus, preserving locally valid evidence and unresolved conditions does not itself guarantee a sound terminal commitment: ER makes the remaining uncertainty explicit, but the present handoff rule does not yet ensure that unresolved conditions constrain commitment strongly enough.

4 Related Work

Function discovery and relational inference. Equation discovery, latent-interaction inference, and recent LLM-guided scientific search provide the immediate setting for our study (Brunton et al., 2016; Kipf et al., 2018). Existing methods already support executable function proposal, numerical parameter fitting, iterative analysis and refinement, and, in closely related settings, joint exploration of symbolic dynamics and latent graphs (Shojaee et al., 2025; Li & Pan, 2026; Liang et al., 2026). ER therefore does not seek novelty in a stronger function-search space or graph estimator. Its question begins after a Function Family has been numerically realized: whether the proxy-preferred response expresses its proposed effect, whether that effect remains consequential after graph refitting, and how the resulting evidence should affect epistemic revision and eventual solution commitment.

Hypothesis validation under proxy-mediated realization. Automated hypothesis-validation systems already translate supplied scientific hypotheses into measurable implications, actively gather evidence, and design prospective falsification tests (Huang et al., 2025). ER does not claim these capabilities as novel. Accurate dynamical prediction also need not identify the underlying interaction topology (Prasse & Van Mieghem, 2022). The difference is that the object evaluated here is first instantiated by proxy optimization and is then coupled to a graph re-estimated for that realization. Proxy fitting may suppress an intended effect, while graph refitting may compensate for changes in the response. ER therefore distinguishes whether an anticipated regime is reachable, whether proxy fitting realizes it, and whether its effect persists after downstream compensation. Expectation–Reality

correspondence prevents the proxy optimum itself from serving as the evidential verdict on the proposed mechanism.

Auditable revision and solution commitment. Recent scientific-agent protocols already maintain persistent hypotheses, attach evidence to them, update hypothesis-level belief or status, and preserve auditable revision histories (Takahara & Mizoguchi, 2026). ER shares this commitment to explicit evidence-linked revision but retains a different state object. Acquired observations are returned to their corresponding pre-outcome Expectations, and the resulting dimension-specific discrepancies remain in the cross-round Research State, allowing supported, contradicted, partial, and unresolved dimensions to coexist. This epistemic state is also separated from task-facing commitment: revising what is scientifically understood need not immediately replace the retained candidate, and a downstream task may eventually require commitment even while parts of the scientific interpretation remain unresolved.

Scientific-agent process evaluation and positioning. Workflow-level evaluation of scientific agents is also established (Chen et al., 2025; Majumder et al., 2025), so process auditing itself is not our methodological contribution. We use it to test whether ER's proposed distinctions are realized in execution and use post-unsealing diagnostics only to characterize their downstream consequences. The distinctive methodological focus of ER is the separation of decision roles when an observable proxy determines the numerical realization of an intermediate hypothesis, the downstream structural objective remains hidden during research, and a compensating downstream model can alter what observations imply. Proxy fitting governs numerical realization, Expectations govern evidence requirements, Reality and discrepancy update Research State, and a separate operation governs solution commitment.

5 Discussion, Limitations, and Outlook

Our results support ER as an operational research-process design for proxy-mediated scientific search, rather than as a domain-independent reasoning framework. In the controlled setting studied here, numerical realization, evidence formation, epistemic revision, and solution commitment can be separated; non-confirmatory evidence remains attached to the questions that generated it while Research State changes without mandatory candidate replacement. The current evidence is limited to unknown-dynamics network reconstruction within one benchmark family, a fixed reconstruction procedure, and one primary LLM research protocol. Because ER is evaluated as a coupled process, the experiments establish system-level behavioral differences rather than causal effects of individual components. Two gaps remain: unresolved evidence may fail to constrain terminal commitment, and improved fixed-threshold structural calibration can coexist with weaker edge-ranking quality. Future work should study evidence-aware commitment, fixed-evidence replays, preservation of ranking evidence under structural calibration, and broader validation across tasks, backends, models, and stochastic executions.

Beyond these extensions, ER raises a broader question about how reasoning may be organized in LLM-based systems. ER points to a complementary axis to answer-oriented reasoning: how evidence changes what a system treats as established, contradicted, uncertain, or viable, and when that epistemic state warrants commitment. The present work provides an initial operational template, not a general theory of reasoning. Future work should test whether separating realization, evidence formation, epistemic revision, and commitment remains useful in open-ended exploration and decision problems where multiple explanations remain viable, evidence arrives incrementally, and no single observable proxy can safely determine the final choice. From this perspective, reasoning may concern not only how intermediate steps produce an answer, but also how evidence regulates epistemic state and when that state justifies commitment.

AI Use Statement

Generative AI tools were used to assist with the development and refinement of the research methodology and experimental protocol, experiment and audit planning, software implementation and debugging, interpretation and organization of experimental results, literature discovery and analysis, scientific figure development, and manuscript drafting and editing. AI-assisted outputs were

treated as intermediate research support rather than as empirical evidence or authoritative scientific judgments. References and citations were manually selected, retrieved, and verified by the authors before inclusion. Experimental claims and numerical results were checked against the underlying run artifacts and analysis outputs. The authors reviewed all AI-assisted material used in the submission and take responsibility for the final text, claims, code, figures, references, and other artifacts.

REFERENCES

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Kurt Butler, Daniel Waxman, and Petar M Djurić. Tangent space causal inference: Leveraging vector fields for causal discovery in dynamical systems. *Advances in Neural Information Processing Systems*, 37:120078–120102, 2024.

Luca Castri, Sariah Mghames, Marc Hanheide, and Nicola Bellotto. Enhancing causal discovery from robot sensor data in dynamic scenarios. In *Conference on Causal Learning and Reasoning*, pp. 243–258. PMLR, 2023.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. Causaldynamics: A large-scale benchmark for structural discovery of dynamical causal models. *Advances in Neural Information Processing Systems*, 38, 2026.

Kexin Huang, Ying Jin, Ryan Li, Michael Y Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. *arXiv preprint arXiv:2502.09858*, 2025.

Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pp. 507–523. Springer, 2011.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pp. 2688–2697. Pmlr, 2018.

Yunlun Li and Sinno Jialin Pan. Robust equation structure learning with adaptive refinement. In *International Conference on Learning Representations*, volume 2026, pp. 70029–70053, 2026.

Xiaoxiao Liang, Juyuan Zhang, Liming Pan, and Linyuan Lü. Interpretable relational inference with llm-guided symbolic dynamics modeling. *arXiv preprint arXiv:2604.12806*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on artificial intelligence and statistics*, pp. 735–745. PMLR, 2020.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4556–4579, 2025.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Bastian Prasse and Piet Van Mieghem. Predicting network dynamics without requiring the knowledge of the interaction graph. *Proceedings of the National Academy of Sciences*, 119(44):e2205517119, 2022.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pp. 1388–1397. Pmlr, 2020.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: scientific equation discovery via programming with large language models. 2024. *URL: http://arxiv. org/abs/2404.18400*, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable ai scientists: A hypothesis evolution protocol for llm agents. *arXiv preprint arXiv:2607.09195*, 2026.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Yan Zhang, Yu Guo, Zhang Zhang, Mengyuan Chen, Shuo Wang, and Jiang Zhang. Universal framework for reconstructing complex networks and node dynamics from discrete or continuous dynamics data. *Physical Review E*, 106(3):034315, 2022.

A Supplementary Results and Experimental Diagnostics

A.1 Cases, sources, and scoring boundaries

The experiments use 30 pre-generated Tier-2 CausalDynamics cases. CD001–CD010 are development cases used for the proxy diagnostics in Section A.2; they are not included in the prospective ER comparison. CD011–CD020 are the primary periodic, zero-noise cohort. CD021–CD030 are a second prospective cohort with nonlinear coupling and noise level 1.0. Each case contains ten trajectories of 1,000 time points over ten observed nodes. The two prospective cohorts are shown separately throughout because their data conditions and execution traces differ. A 20-case summary is descriptive only.

The two local research processes are Inherited and ER. Their Joint AUROC and AUPRC use the official terminal score vectors; Joint SHD uses thresholded terminal predictions at 0.5 and sums errors over the ten graphs in a case. The eight complete external reference methods were taken from the official published CausalDynamics release, matched to the 30 local cases by the full data configuration and source filename. They are reference outputs rather than equal-compute local reruns. The release has a zero-graph fallback on failure, and its scalar result files do not separately flag fallback cases. GRaSP has one matching case; GIN and TCDF have no matching published result, so none enters a complete-cohort mean. Case mappings, hashes, coverage, and external per-case results are supplied in the accompanying provenance files.

A.2 Proxy-Mismatch and Mechanism Diagnostics

The development analysis contains 134 eligible candidates from CD001–CD010. Table 3 reports the main-text rank diagnostics alongside the subsequent run-wise and case-balanced summaries. Calibration-error and structural ranks are oriented so that zero denotes the best candidate. Positive Spearman correlation therefore indicates agreement between the proxy and task orderings. A normalized rank is $(r-1)/(N-1)$ for a candidate with rank r among N eligible candidates. All truth-dependent structural scores and ranks are computed only after unsealing.

Table 3: **Additional proxy-mismatch statistics.** Spearman correlations are summarized within cases or runs as indicated.

Population	AUROC	AUPRC	SHD
Development, case-wise median	-0.065	0.418	-0.034
Subsequent, run-wise median	0.142	0.355	-0.389
Subsequent, case-balanced median	0.113	0.315	-0.311

Table 4 provides the mechanism-suppression audit. A candidate is counted as suppressed when at least one explicit nonlinear mechanism parameter lies inside its family-specific shutdown region after proxy optimization. Active realization records whether the same proposed mechanism can subsequently be instantiated outside that region.

Table 4: **Mechanism suppression and reachability under proxy-preferred realizations.** The CD011–CD020 audit uses the same family-specific shutdown definition as the development analysis.

	Development CD001–010	Inherited CD011–020	ER CD011–020
Eligible mechanism candidates	72	81	73
Suppressed candidates	22 (30.6%)	23 (28.4%)	18 (24.7%)
Active realization found	18/22 (81.8%)	18/23 (78.3%)	18/18 (100%)

CD021–CD030 is excluded from the direct shutdown audit because the corresponding family-specific parameter contracts are not preserved under the same auditable definition. Active-realization rates establish reachability and are not used as a between-process comparison because Inherited and ER obtain this evidence through different routes.

For the function–graph compensation diagnostic, the median case-wise association between $\mathrm{median}\,|f'(z)|$ and $\|G_{\mathrm{off}}\|_2$ is -0.546 in development, -0.746 in the subsequent Inherited runs,

and -0.597 in the subsequent ER runs. After balancing the two processes within each of the 20 subsequent cases, all $20/20$ associations remain negative, with median $\rho = -0.696$.

The gain–graph association is negative in all ten development cases. These diagnostics motivate a targeted realization test: neither a shutdown-region classification nor a negative gain–graph association identifies the generating mechanism.

A.3 Execution Coverage and Process-Audit Details

Table 5 reports file/object coverage, using an executed family as the denominator. CD021–CD030 has 163 executed ER families and three additional PENDING families excluded from that denominator; its additional family-loop executions make its family count different from the primary cohort. In the second cohort, a specialized-experiment file exists for 151 families, including 127 SUCCESS, 21 FAILED, and three FAILED_NON_BLOCKING statuses. Thus, file presence measures execution coverage, not successful evidence acquisition or scientific correctness. All ten terminally selected families in each cohort have the required record objects.

Table 5: **ER protocol execution coverage.** Fractions count executed Function Families; the final-selected column reports coverage of the ten selected families in each cohort. Specialized-experiment files include failed executions.

Recorded object	CD011–020	CD021–030	Final-selected families
Executed Function Families	142	163	10 per cohort
Expectation	141/142	157/163	10/10 in each
Reality	138/142	150/163	10/10 in each
Discrepancy record	138/142	151/163	10/10 in each
Fixed-G and refit-G arms	134/142	141/163	10/10 in each
Specialized-experiment file	139/142	151/163	10/10 in each

Expectation–Reality outcomes. Table 6 records the semantic outcomes of the Expectations attached to the final-selected ER Function Family in each case.

Table 6: **Detailed semantic outcomes of the final-selected ER Expectations.** Categories summarize the recorded evidence state relative to the pre-outcome Expectation.

Outcome	CD011–020	CD021–030
Supported	20	26
Contradicted	16	9
Partial	18	21
Unresolved	6	12
Total	60	68
Non-confirmatory	40/60 (66.7%)	42/68 (61.8%)

The semantic audit asks whether the Reality and discrepancy record continue to address the scientific question fixed by the corresponding pre-outcome Expectation. Supported, contradicted, partial, and unresolved are therefore labels of the recorded evidential relation rather than judgments of ground-truth scientific correctness. Clear cross-round continuation denotes direct semantic preservation of the same discrepancy or uncertainty; partial continuation denotes a broader reformulation that still retains the corresponding scientific issue. Full item-level Expectation–Reality–κ records and transition labels are provided with the supplementary audit files.

Round-end state and incumbent records. At round end, Inherited versus ER records unresolved questions in 1/50 versus 31/50 states in CD011–CD020 and 4/50 versus 35/50 in CD021–CD030. The corresponding incumbent counts are 48/50 versus 19/50 and 46/50 versus 15/50. An incumbent is passed to the next Planner in 38/40 versus 15/40 transitions in the first cohort and 36/40 versus 12/40 in the second. When ER does record an incumbent, it is still the cumulative validation-error leader in 19/19 and 15/15 states. These are properties of process traces; the 50 round-end states are

clustered within ten cases and are not independent statistical units. More unresolved records alone do not imply more accurate judgments.

A.4 TERMINAL OUTCOMES CASE BY CASE

Tables 7 and 8 retain the case-level structure behind the cohort means. In the primary cohort, ER has lower SHD in seven cases and ties three. In the second cohort it improves five, ties one, and worsens four. CD024 and CD028 account for large Inherited errors; CD027 and CD029 show large ER errors. Thus, the second-cohort mean difference of 56.8 SHD points should not be read as a typical per-case gain. These are official terminal Joint metrics, distinct from candidate-pool same-stage scores.

Table 7: Official terminal Joint metrics on CD011–CD020: AUROC↑, AUPRC↑, and SHD↓. Δ is ER minus Inherited SHD.

	Inherited			ER			
Case	AUROC	AUPRC	SHD	AUROC	AUPRC	SHD	Δ
CD011	0.631	0.301	160	0.676	0.340	144	-16
CD012	0.694	0.445	144	0.693	0.445	144	0
CD013	0.679	0.361	142	0.680	0.362	142	0
CD014	0.713	0.458	144	0.703	0.446	132	-12
CD015	0.734	0.380	143	0.729	0.374	142	-1
CD016	0.691	0.407	135	0.692	0.401	135	0
CD017	0.770	0.529	170	0.722	0.414	123	-47
CD018	0.732	0.422	132	0.733	0.423	131	-1
CD019	0.756	0.404	152	0.748	0.382	140	-12
CD020	0.757	0.504	145	0.702	0.416	132	-13

The second cohort's case-wise results follow in Table 8.

Table 8: Official terminal Joint metrics on CD021–CD030: AUROC↑, AUPRC↑, and SHD↓. Δ is ER minus Inherited SHD.

	Inherited			ER			
Case	AUROC	AUPRC	SHD	AUROC	AUPRC	SHD	Δ
CD021	0.736	0.342	133	0.738	0.346	138	+5
CD022	0.675	0.364	142	0.689	0.338	137	-5
CD023	0.752	0.553	110	0.781	0.481	110	0
CD024	0.882	0.628	432	0.750	0.439	120	-312
CD025	0.583	0.325	157	0.589	0.353	146	-11
CD026	0.800	0.455	134	0.800	0.422	140	+6
CD027	0.725	0.441	238	0.720	0.425	278	+40
CD028	0.832	0.546	433	0.767	0.326	150	-283
CD029	0.592	0.195	228	0.485	0.157	273	+45
CD030	0.640	0.381	181	0.636	0.403	128	-53

Case-macro means are reported as (AUROC, AUPRC, SHD). For CD011–CD020, they are (0.7157, 0.4211, 146.7) for Inherited and (0.7078, 0.4001, 136.5) for ER. For CD021–CD030, the corresponding values are (0.7218, 0.4230, 218.8) and (0.6953, 0.3691, 162.0). The descriptive 20-case means are (0.7187, 0.4221, 182.75) and (0.7015, 0.3846, 149.25). The complete eight-method external comparison appears in the main experimental table; its per-case records and source binding are in the accompanying CSVs.

Figure 4 summarizes the paired terminal Joint SHD outcomes across the two comparison cohorts.

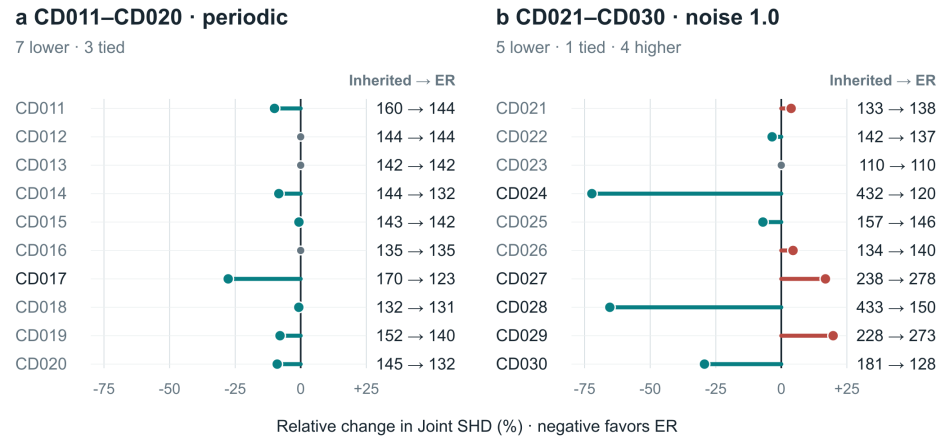


Figure 4: Paired official terminal Joint SHD outcomes. Each process has one fixed terminal prediction per case; lower SHD is better. Tables 7 and 8 give the complete raw metric values.

A.5 Complete Candidate-Pool and Selection Diagnostics

For a process m on case c, let O_c^m be the minimum post-unsealing SHD among its eligible handoff candidates, S_c^m the selected candidate's SHD at the same scoring stage, and $R_c^m = S_c^m - O_c^m$. All quantities in this subsection are computed only after truth unsealing. Pool-best and regret values use the complete eligible handoff pools at the same candidate-scoring stage and must not be mixed with official terminal final-refit scores. Table 9 summarizes the two cohorts.

Table 9: **Complete post-unsealing candidate-pool diagnostics.** N is the number of eligible handoff candidates; regret is selected SHD minus the minimum SHD in the corresponding process's own pool.

Cohort	Process	N	Pool best SHD	Selected SHD	Regret	Rank percentile	Top-3
CD011–020	Inherited	147	137.5	149.9	12.4	33.10%	4/10
	ER	138	135.5	136.8	1.3	7.75%	8/10
CD021–030	Inherited	145	128.5	218.8	90.3	39.10%	4/10
	ER	133	127.4	162.0	34.6	32.74%	5/10

The CD011–CD020 same-stage selected means are 149.9 and 136.8, distinct from the official terminal means of 146.7 and 136.5. ER selects a top-three SHD candidate in 8/10 cases versus 4/10 for Inherited.

Within the primary ER candidate pools, retrospective selection by minimum calibration error, minimum validation error, and the recorded ER commitment gives mean SHD regrets of 49.0, 12.2, and 1.3, respectively. In CD021–CD030, the corresponding D_{cal}-minimum / D_{val}-minimum / recorded regrets are 92.5 / 87.3 / 90.3 for Inherited and 161.3 / 143.5 / 34.6 for ER. These comparisons are post-unsealing diagnostics and do not redefine the online selection objective.

Table 10 gives the primary-cohort case-wise minima, choices, and regrets.

Table 10: **Case-wise eligible-pool SHD diagnostics for CD011–CD020.** O is the minimum SHD in a process's own pool, S is the selected same-stage SHD, and $R = S - O$.

	Inherited			ER		
Case	O	S	R	O	S	R
CD011	144	161	17	144	144	0
CD012	140	144	4	143	143	0
CD013	138	144	6	139	142	3
CD014	126	144	18	127	128	1
CD015	144	144	0	142	143	1
CD016	136	136	0	136	136	0
CD017	129	183	54	123	125	2
CD018	131	132	1	131	131	0
CD019	151	151	0	134	140	6
CD020	136	160	24	136	136	0

Table 11 gives the second cohort's case-wise pool sizes and scores. Its pools contain 145 Inherited and 133 ER candidates (278 in total). For this cohort, candidate final scores agree with official terminal scores. The mean selected difference of 56.8 SHD points partitions arithmetically into a 1.1-point pool-minimum difference and a 55.7-point regret difference. Separately generated pools differ in size and eligibility, so this decomposition neither establishes equivalent search capacity nor identifies a causal contribution of a reasoning component.

Commitment failure. CD029 illustrates a separate evidence-to-commitment limit. Its ER handoff reports a successful matched-control experiment for a curvature effect and a mean 0.117 reduction in refitted off-diagonal graph fraction relative to its linear control. The same handoff retains unresolved amplitude, temporal-fidelity, small-signal, and bias-optional conditions, yet selects that family for network reconstruction. After unsealing, the selected candidate has SHD 273 versus 130 for the best of 18 eligible candidates, placing it last in that pool. This does not show that the local matched-control result was false. It shows that the retained local evidence did not guarantee the terminal structural choice; no frequency claim about this failure type follows from a single case.

Table 11: CD021–CD030 eligible-pool diagnostics. N is the number of eligible handoff candidates; O is the pool minimum; S is selected SHD; $R = S - O$. All values are post-unsealing.

	Inherited				ER			
Case	N	O	S	R	N	O	S	R
CD021	13	120	133	13	17	122	138	16
CD022	16	134	142	8	12	130	137	7
CD023	12	109	110	1	12	110	110	0
CD024	15	117	432	315	14	120	120	0
CD025	16	154	157	3	10	146	146	0
CD026	16	112	134	22	9	114	140	26
CD027	12	138	238	100	14	134	278	144
CD028	15	140	433	293	14	140	150	10
CD029	14	134	228	94	18	130	273	143
CD030	16	127	181	54	13	128	128	0

A.6 Structural Operating Point, Thresholds, and Failure Boundary

Table 12 separates off-diagonal edge recovery from the full Joint task score. Positive-edge and TP/FP/FN counts aggregate the ten cases in each cohort; off-diagonal AUPRC and SHD are case-macro means. Oracle-threshold gaps are post-unsealing diagnostics and were unavailable during research or terminal commitment.

Table 12: **Off-diagonal operating point at the fixed 0.5 threshold.** Oracle-threshold gaps are post-unsealing diagnostics and were unavailable during research or terminal commitment.

	CD011–CD020		CD021–CD030	
Metric	Inherited	ER	Inherited	ER
Off-AUPRC ↑	0.1669	0.1463	0.2014	0.1727
Off-SHD ↓	97.9	90.7	174.6	118.0
Predicted positive edges	193	13	1288	390
True positives (TP)	57	3	221	55
False positives (FP)	136	10	1067	335
False negatives (FN)	843	897	679	845
Off-diagonal recall	6.33%	0.33%	24.56%	6.11%
Fixed-to-oracle SHD gap ↓	12.4	5.3	95.2	34.3
Empty off-diagonal SHD reference	90.0	90.0	90.0	90.0

An empty off-diagonal prediction has mean SHD 90.0 in both cohorts. The first cohort therefore places ER close to this extreme sparsity reference, whereas the second cohort does not: its ER off-diagonal SHD remains 118.0. The near-empty characterization is consequently specific to the primary cohort rather than a general description of ER.

Ranking-specific selection diagnostics. The reduction in SHD selection regret does not extend uniformly to ranking quality. In CD011–CD020, mean Full-AUPRC selection regret changes only from 0.0387 for Inherited to 0.0293 for ER (paired $p = 0.846$), while off-diagonal AUPRC regret changes from 0.0681 to 0.0554. Moreover, over CD012–CD020, the ER-selected candidates have mean off-diagonal AUPRC 0.1443 compared with 0.1740 for the corresponding D_{cal}-best candidates ($p = 0.0391$). These post-hoc diagnostics reinforce that the strong SHD-regret result is specific to the fixed-threshold structural criterion and should not be interpreted as uniform improvement in edge-ranking quality.