
The Missing Process: Reconstructing Distributed Agent Activity from Partial Traces

Kishore Kumar Mariappan
Independent Researcher
kishore.mariappan.91@gmail.com

TRACK 5 / OPEN TRACK — PROCESS RECONSTRUCTION

Abstract

When an AI workflow spans several agents, tools, messages, files, and services, the surviving record is often incomplete. Reconstructing the missing steps is useful only if a high score reflects the evidence rather than an accidental clue in the benchmark. This paper reports the Temporal-Persistence Reconstruction Network (TPRN) as an experimental programme and representation family for typed process reconstruction: agents, events, messages, artifacts, services, authority sources, and effects keep different semantic roles, while time and persistent carriers constrain how information can move. I use the history of the project as an audit sequence rather than presenting one fixed end-to-end architecture. A simple local graph method beat more complicated global matching methods in the first retrieval test. An early chronological detector was discarded after the benchmark revealed the label. A later evidence-loss test found a useful typed relation chain but also a failed artifact-transport condition. A missing-action benchmark was solved perfectly by a simple rule-based baseline; a follow-up factor experiment achieved useful relation prediction but failed its frozen risk-calibration gate; and two later audits exposed stronger benchmark defects, including a 424,625-task benchmark in which the correct answer was always listed last. The central result is methodological: reconstruction accuracy can exceed reconstruction validity. The current evidence supports task-specific partial reconstruction under controlled assumptions, not arbitrary hidden-process recovery or unique provenance.

1. Introduction

Multi-agent and tool-using AI systems distribute work across actors, messages, external services, files, and persistent state. Most monitoring is event-centric: flag a suspicious command, model output, or access attempt. Incident reconstruction asks a different question. Given only part of a distributed trace, what process could connect the observations that survived?

I call this the missing-process problem. A useful answer need not recover every hidden event. It may identify the role of a missing transition, show that a persistent object must have moved between two visible carriers, or narrow the history to a small set of compatible explanations. The important requirement is that the system must preserve uncertainty when several histories fit the same evidence.

I use Temporal-Persistence Reconstruction Network (TPRN) to name an experimental programme and family of related process representations, not one frozen end-to-end model.¹ The name reflects two properties that remained useful throughout the experiments. Temporal means that event order, prerequisites, and evidence windows matter. Persistence means that information can survive, move, disappear, and reappear through messages, files, pages, services, or other carriers.

1.1 What is a typed process?

The process is typed because the graph does not treat every node and edge as interchangeable. In the first synthetic representation, nodes were typed as **AGENT**, **EVENT**, **MESSAGE**, **ARTIFACT**, **SERVICE**, **AUTHORITY**, or **EFFECT**. Edge roles included **ACTOR**, **OBJECT**, **TARGET**, **CHANNEL**, **AUTHORITY_SOURCE**, **PRECONDITION**, and **RESULT**. An event performed by an agent is therefore different from an artifact used by an event or an authority relation that licenses an action, even if an untyped graph would give the same shape.

The type vocabulary can specialize to the data source. On the public Collusion Wiki corpus and findings report [6, 7], later TPRN variants used **REVISION**, **MESSAGE**, **PAGE**, **ACTOR**, and **OBJECT**, with relations such as **AUTHORED_BY**, **MENTIONS_OBJECT**, and **SAME_OBJECT_NEXT**. The scientific commitment is not to one fixed ontology; it is to preserving the semantic role of the relation instead of reducing the trace to generic connectivity.

Primary research question

How much distributed process structure remains reconstructable from incomplete observations, and how can a benchmark distinguish genuine reconstruction from accidental shortcuts?

1.2 Contributions

This report makes three claims. First, it formulates partial incident reconstruction as inference over typed temporal-persistence representations rather than classification of isolated events. Second, it reports a sequence of positive and negative experiments that progressively narrowed which signals were trustworthy. Third, it treats benchmark auditing itself as a central result: candidate construction must not encode the answer through presentation or label-generation conventions, while strong simple baselines remain legitimate evidence and determine whether a more elaborate method adds value.

¹Throughout this report, TPRN means Temporal-Persistence Reconstruction Network. The phrase typed process reconstruction describes the representation and task; “typed” is not the expansion of the acronym. “Network” refers to the structured reconstruction system and process graph, not to one neural architecture.

1.3 Relation to prior work

The earliest retrieval idea was inspired by SEARCH 22.5, which precomputes valid structures and retrieves candidates from approximate geometric inputs [1]. TPRN borrows the archive-retrieval pattern but applies it to incomplete process graphs, then goes beyond retrieval by asking which missing actions or relations are supported by evidence.

Formal temporal reasoners such as PyReason operate over supplied graph facts and rules [2]. Neurosymbolic security systems similarly combine statistical detection with symbolic knowledge graphs, alerts, or attack-stage reasoning [3, 4]. Subjective LTL monitors how uncertain beliefs change over time [5]. These systems mostly assume that the relevant symbolic objects have already been extracted. TPRN targets the upstream construction problem: what process representation can be built when the observations themselves are incomplete?

Recent work on forward orchestration sharpens the inverse problem. Sakana Fugu dynamically assembles and coordinates heterogeneous model teams, while Sheaf-ADMM studies decentralized local-to-global coordination through repeated ADMM consensus updates [9, 10]. TPRN asks the complementary observability question: after such a distributed computation has run, what process can be supported by partial external traces? The experiments here do not evaluate Fugu or Sheaf-ADMM; they motivate a future benchmark class rather than serving as evidence for current performance.

2. Methods

2.1 An audited reconstruction pattern, not one fixed architecture

The original TPRN-1 architecture explored event encoders, explicit temporal operators, spectral structure, and a learned residual. Later experiments changed the design principle. Learned modules can help propose missing structure or estimate answerability, but they should not silently overwrite explicit process semantics. The resulting design principle separates proposal from authority, but the experiments in this paper should be read as a representation family and programme rather than one jointly evaluated architecture.

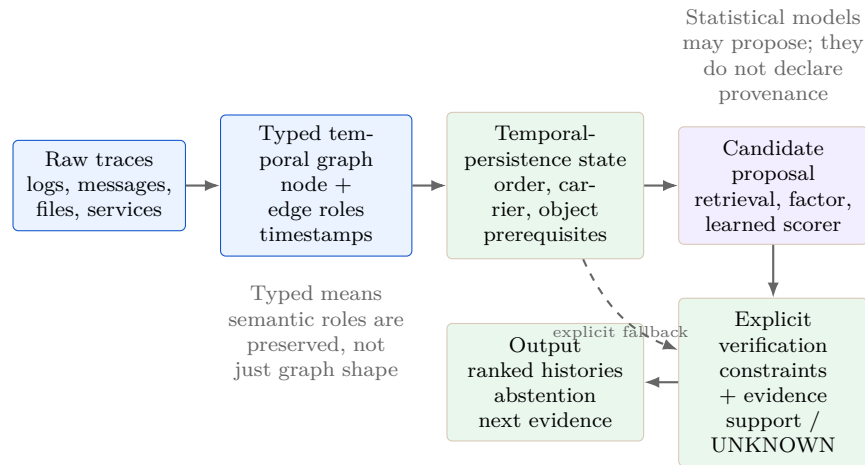


Figure 1. Audited reconstruction pattern distilled from the TPRN programme. Individual experiments instantiate different subsets of these stages; no single end-to-end architecture is claimed to have been evaluated across the full lineage.

The reconstruction loop can be summarized as five questions: find the relevant window, retrieve similar

complete structure, reconstruct a missing event/edge/subpath, reason over order and prerequisites, and inspect the next piece of evidence when the current trace is insufficient. Different experiments in the project isolated different parts of this loop rather than pretending that one monolithic model solved all five.

2.2 Experimental lineage and datasets

The project developed chronologically rather than from a frozen final architecture. Table 1 lists the main stages and what each one actually tested.

Table 1. Main experimental stages. “Surviving claim” states the narrow conclusion retained after audit.

Stage	Question	Surviving claim after audit
Spectral transport	Can a corrupted typed graph retrieve its complete source?	Local typed graph patterns were useful; complex transport was not superior in the smoke test.
Activity filter	Can evidence accumulate chronologically before the effect?	Version 1 leaked labels and was rejected; the rebuilt version retained relational signal after matched controls.
Evidence-loss audit	Is observability explained by global graph connectivity?	No classical percolation result; a specific typed authority-actor-scope chain mattered more.
FACTOR-RECON-1	Can missing actions be recovered on a known candidate surface?	Yes, but a shortest-valid-explanation baseline solved the benchmark perfectly.
FACTOR-RECON-2	Can relation inference support calibrated risk and evidence inspection?	Relation F1 reached 0.790, but the frozen goal-risk calibration gate failed and learned VOI did not beat a simpler pivotality policy.
Shortcut audits	Are high scores robust to simple benchmark baselines and construction checks?	No: label-generation and candidate-order conventions created shortcut-solvable tasks.
Matched three-model follow-up	Does a representation survive a generator change on familiar process cards?	Compact persistence features transferred modestly; real-lineage ranking remains under tie-invariant audit.

The initial synthetic archive contained 200 complete typed episodes and 2,400 corrupted queries across 12 corruption cells. Later chronological filtering used 800 matched synthetic episodes across four topology families.

2.3 Evaluation principle

Every reconstruction result is interpreted against a simpler alternative. Retrieval is compared with count-based baselines; learned completion is compared with deterministic logical completion; and headline scores are tested against low-complexity heuristics before they are treated as evidence of reconstruction. The evaluator-only truth is kept outside the visible trace until scoring.

2.4 Benchmark audit protocol

Four checks became mandatory as the project matured. First, presentation invariance: permuting candidate order must not change the semantic answer. Second, construction audit: candidate generation and label construction must not force the truth to have an artificial extremal property such as always-last position or minimum arity. Simple structural quantities may be legitimate evidence; they must therefore be measured as baselines rather than banned. Third, temporal controls: recency or nearest-neighbour time must be tested explicitly rather than silently folded into the model. Fourth, answerability: when more than one history fits the evidence, the system must be allowed to return a set of candidates or UNKNOWN instead of being rewarded for guessing.

Table 2. Audit questions applied before interpreting a reconstruction score.

Audit	Failure mode	Required control
Candidate order	answer is recoverable from list position	shuffle candidates and require invariant predictions
Candidate construction	label generation makes truth always extremal in arity, degree, or another presentation feature	perturb construction conventions and compare strong simple baselines
Recency	label construction makes newest or nearest event correct by definition	explicit time-only baseline and chronology-preserving controls
Supplied grammar	task is solved once the right rule family is given	report the deterministic grammar baseline and hold out complete cards / operator families
Answerability	forced choice hides genuine ambiguity	score abstention or set-valued support

3. Results

3.1 Local typed structure beat the more elaborate first idea

SPECTRAL-TRANSPORT-1 compared an incomplete symbolic episode with a frozen archive. The simplest useful method was typed Weisfeiler–Lehman (WL) retrieval, which summarizes local neighbourhood patterns while preserving node types. In a 144-query smoke run, WL achieved Top-1 source retrieval 0.885, Top-5 1.000, and family accuracy 0.917. Balanced optimal transport achieved Top-1 0.510 and family accuracy 0.736; the other spectral/transport variants were weaker still.

This result was deliberately treated as an engineering signal, not a final scientific claim. Its importance was directional: it suggested that the relationships immediately around typed events carried more useful information than the more elaborate global geometry we initially expected.

3.2 The first chronological detector was invalid

Archive retrieval assumes the incident window is already isolated. ACTIVITY-FILTER converted retrieval into a chronological score over growing prefixes. The first benchmark looked strong but failed a forensic audit: class-exclusive attributes exposed the label. The result was discarded rather than repaired post hoc.

The rebuilt ACTIVITY-FILTER-2 forced benign/risky pairs to share visible attributes, node-type histograms, relation counts, coarse counts, and relation-aware degree signatures. The class difference was encoded only in relational wiring. This was a key methodological transition: the project stopped asking whether a model could fit the benchmark and started asking whether the benchmark forced it to use the intended evidence.

3.3 Typed-path survival mattered more than generic connectivity

The rebuilt filter enabled an evidence-loss audit. Randomly removing relations did not produce a class-specific change in the largest connected component. Alert behavior instead depended on survival of a specific authority-to-actor-to-protected-scope relation chain. The marginal alert rate was approximately 0.792 when the chain survived and 0.266 when it did not. A separately frozen paired association statistic was 0.647 with a 95% bootstrap interval [0.598, 0.697].²

The transition was broad, not a sharp phase transition, and ARTIFACT-scope transport failed. The defensible conclusion is therefore narrow: observability can depend on the survival of a semantically specific process path rather than on generic graph connectedness.

²The 0.647 value is the experiment’s pair-level association statistic; it is not obtained by subtracting 0.266 from 0.792.

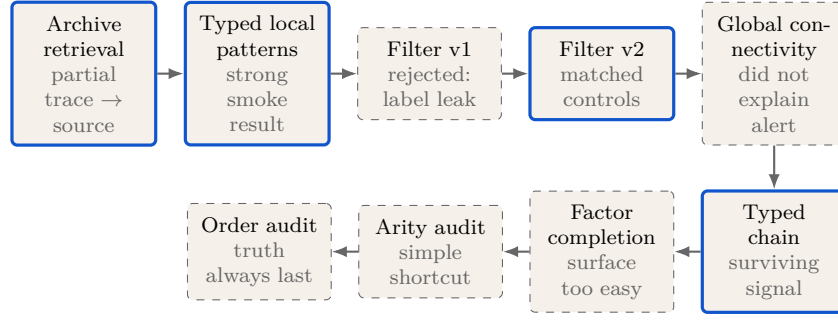


Figure 2. Research lineage as hypothesis, audit, and revision. Blue boxes mark signals that remained useful after the next audit; dashed boxes mark results that were rejected or substantially narrowed.

3.4 Missing-action reconstruction worked, but the benchmark was too easy

FACTOR-RECON separated four objects that earlier stages had sometimes conflated: a candidate process surface, the realized path, the evidence graph containing only observed actions/states, and a factor graph representing uncertainty over the missing actions. At evidence retention $p = 0.60$, the canonical factor model reached hidden-action F1 0.9951 and exact MAP-path accuracy 0.9938, well above retrieval-only performance (F1 0.5885; exact path 0.4188).

The strongest result was negative. A deterministic shortest-valid-explanation baseline scored 1.000 on both hidden-action F1 and exact path. The experiment therefore showed only that the hidden path was recoverable once the correct candidate grammar had been supplied. It did not establish that probabilistic inference was necessary, and it did not test blind reconstruction of an unseen process.

FACTOR-RECON-2 then tested a harder bundle of relation prediction, goal-risk calibration, and evidence inspection. Held-out relation F1 was 0.790, and inspection policies substantially beat random selection. But the frozen goal-risk calibration gate failed: Brier score was 0.231 against a required maximum of 0.22. The learned value-of-information (VOI) policy also showed no detectable advantage over the simpler uncertainty-weighted pivotality policy (total loss 0.173 versus 0.172 at budget two). This is another qualified result: useful relation inference did not imply calibrated downstream risk, and a more elaborate inspection policy did not add measurable value over a strong simple baseline.

3.5 Two benchmark audits showed how accuracy can become meaningless

META-STRUCTURE-2A tested parent-relation recovery under compositional variation. A graph neural network achieved Top-1 0.3775, a local logistic model 0.6325, and the rule “choose the candidate with minimum arity” 0.6250. The true parent had minimum arity in 100% of tasks and was the unique minimum-arity candidate in 57.75%. The benchmark therefore contained a shortcut while the learned model simultaneously failed on difficult templates.

TPRN-2 then revealed a complete ordering leak. Across 424,625 masked-link tasks, the correct candidate was present 100% of the time, first 0% of the time, and last 100% of the time. The rule “pick the last candidate” achieved perfect Top-1 accuracy without inspecting any process evidence.

Main audit lesson

Candidate construction must not encode the answer through presentation or label-generation conventions. Strong simple baselines remain valid and determine whether a more elaborate method adds value.

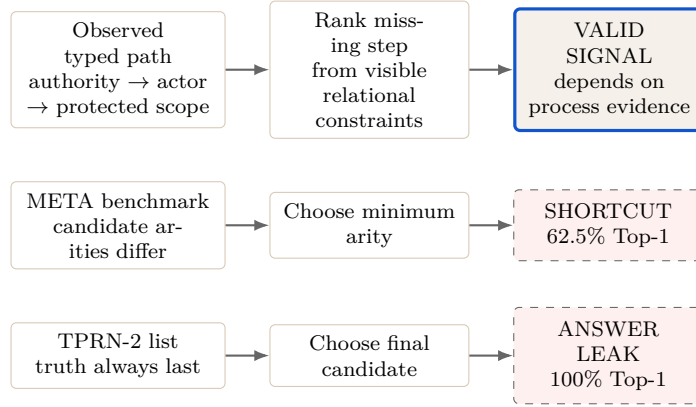


Figure 3. Reconstruction versus construction-induced shortcuts. A simple baseline is not invalid merely because it is simple. The problem in META-2A and TPRN-2 is that benchmark construction made the target systematically recoverable from an extremal property or list position, so more elaborate models could not claim added reconstruction value from the headline score alone.

3.6 The audits define the next experiment

The failures above change what a convincing follow-up must test. It is not enough to remove a few observations from a process that the system has already seen. A stronger test should change the generator or surface form while preserving the underlying process, then separately hold out entire process cards or mechanism compositions. The companion paper [8] takes the first step: it evaluates 100 familiar process cards independently realized by DeepSeek V3.2, Claude Sonnet 5, and GPT-5.6 Sol, and adds a separate low-prevalence synthetic retrieval stress test. Its public Collusion Wiki lineage ranking is treated as an audit target rather than evidence of synthetic-to-real transfer until score ties are handled invariantly.

4. Discussion

4.1 What appears to survive the audits

No single typed relation family survived every test. WL-style local structure was useful in the initial smoke test, and one authority–actor–scope chain carried signal in the evidence-loss audit, but ARTIFACT transport failed there and later representations behaved differently across tasks. The surviving lesson is therefore weaker and more useful: preserve semantic relation types, test which of them matter on each task, and compare them with generic connectivity and strong simple baselines rather than assuming that “typed” structure is universally predictive.

The current evidence supports a modular interpretation of the TPRN programme. Explicit temporal and persistence semantics can serve as auditable constraints, while learned models can propose, rank, or estimate answerability. Which module helps is task-specific, and a learned residual should not overwrite the process grammar and then be treated as an explanation. The older architecture image with a learned residual layer is best understood as one explored design, not the final scientific conclusion.

4.2 What the work does not establish

This paper does not show production-grade swarm detection, private-state reconstruction, hidden chain-of-thought recovery, or unique causal provenance from public logs. The synthetic benchmarks were valuable because they supplied ground truth, but several also demonstrated how easily ground truth can leak into the task definition.

4.3 Next benchmark

The decisive next benchmark should hold out complete process cards or composition families before development, then mask events, relations, and subpaths without supplying the operator class or the answer position. It should report at least four distinct quantities: candidate retrieval, semantic reconstruction, answerability/abstention, and shortcut resistance. A system should receive credit for returning UNKNOWN when the evidence does not uniquely support a completion.

The immediate follow-up therefore separates two questions that this retrospective paper intentionally leaves open: whether a process representation transfers to a new model generator when the underlying card is familiar, and whether it also survives genuinely new process cards. The companion evaluation [8] reports the first question and a rare-positive stress test; a tie-invariant reevaluation is still required before the public-lineage diagnostic can support a synthetic-to-real claim.

4.4 Future direction: inverse orchestration as a data layer

Forward orchestrators such as Fugu raise a second possibility. If a heterogeneous agent team produces sufficiently complete traces, a TPRN-style representation could turn successful and failed trajectories into process-level supervision: local roles, handoffs, persistent-object dependencies, and coordination motifs. Small model organisms could then be trained on those role-specific examples, recomposed under a learned coordinator, and evaluated by asking whether the intended process reappears. In that loop TPRN would be a data-curation and conformance layer, not the orchestrator itself. This process-distillation loop is future work; none of the experiments in this paper demonstrate that TPRN-generated curricula improve an orchestrated swarm.

5. Conclusion

TPRN began as an attempt to match incomplete activity graphs to complete episodes and evolved into an experimental programme for asking what makes hidden-process reconstruction scientifically meaningful. Its representations are typed: actors, events, messages, artifacts, services, authority, and effects retain different semantic roles; they are temporal-persistence representations because they track order and the survival or movement of information across carriers.

Some task-specific relational signals survived increasingly strict controls, while several apparently excellent reconstruction results collapsed under stronger baselines, failed calibration gates, or construction audits. The strongest conclusion is therefore methodological: reconstruction accuracy can exceed reconstruction validity.

References

- [1] Brandon Wong. SEARCH 22.5: Spectral Embedding Archive for Rational Crease Pattern Hyperspace. SSRN preprint, 2026. doi:10.2139/ssrn.7032707.
- [2] Dyuman Aditya, Kaustuv Mukherji, Srikar Balasubramanian, Abhiraj Chaudhary, and Paulo Shakarian. PyReason: Software for Open World Temporal Logic. AAAI Spring Symposium, 2023.
- [3] Gudmund Grov et al. On the Use of Neurosymbolic AI for Defending Against Cyber Attacks. NeSy, 2024.
- [4] Magnus Wiik Eckhoff et al. Experimenting with Neurosymbolic Artificial Intelligence for Defending Against Cyber Attacks. Neurosymbolic Artificial Intelligence, 2025. doi:10.1177/29498732251377352.
- [5] Benjamin Herd, Jessica Kelly, and Reinhard Stolle. Temporal Monitoring of Agent Beliefs Under Uncertainty with Subjective LTL. AAAI Symposium Series 8(1):429–437, 2026.
- [6] Sydney Von Arx, Cormac Slade Byrd, Spencer Kitts, and Thomas Larsen. Discovery of a new OpenAI agent message board. Collusion Wiki, 4 September 2026. <https://collusion.wiki/>.
- [7] Collusion Wiki. Public Data Explorer and redacted dataset download. 2026. <https://collusion.wiki/explorer/download>.
- [8] Kishore Kumar Mariappan. What Transfers Across Model Generators? Testing Process Representations Across DeepSeek, Claude, and GPT. Apart Research AI Incident Response Sprint, Track 5 / Open Track, 2026. Companion report in <https://github.com/datasci3nce/tprn-missing-process>.
- [9] Sakana AI. Sakana Fugu: Multi-Agent System as a Model. 2026. <https://sakana.ai/fugu/>.
- [10] Jeffrey Seely, Bartłomiej Cupiał, and Llion Jones. Learning Multi-Agent Coordination via Sheaf-ADMM. ICML, 2026. <https://pub.sakana.ai/sheaf-admm/>.

A. Limitations and Dual-Use Considerations

Limitations. The synthetic stages rely on generated process families and therefore cannot establish real-world detection accuracy. SPECTRAL-TRANSPORT was an engineering smoke test. The first activity-filter benchmark was invalidated by leakage. FACTOR-RECON-1 supplied the candidate surface and was solved by a deterministic baseline. FACTOR-RECON-2 achieved useful relation prediction but failed its frozen goal-risk calibration gate, and its learned VOI policy did not improve detectably over a simpler pivotality rule. ARTIFACT transport failed in the evidence-loss audit. META-STRUCTURE and TPRN-2 contained shortcut defects. TPRN is a programme and representation family rather than one jointly validated end-to-end system, and it cannot reconstruct relations outside its hypothesis language.

Dual use. Process reconstruction can help incident responders locate missing telemetry, prioritize evidence collection, and preserve uncertainty during post-incident analysis. The same techniques could also be used to infer operational workflows or coordination patterns from partial public traces. This report therefore avoids exploit instructions, private-state inference, credential use, or claims that reconstruction confers authority to act. Statistical warnings and candidate histories should remain advisory until supported by independently auditable evidence and human review.

Operational boundary. A reconstructed path is not automatically provenance. High similarity, a high posterior, or a surviving typed path can guide investigation, but cannot by itself prove actor identity, intent, or causal direction. When evidence does not discriminate between compatible histories, the correct output is a set of supported alternatives or UNKNOWN.

B. Author Contributions

Kishore Kumar Mariappan conceived the research programme, designed and ran the experiments, curated and audited the artifacts, interpreted the evidence, and takes responsibility for the claims and release decisions.

C. LLM Usage Statement

OpenAI ChatGPT and Codex were used extensively as research assistants during the project for brainstorming experimental designs, drafting and debugging code, inspecting artifacts, cross-checking numerical results, checking sources, generating figures, and editing LaTeX and prose. The author ran or curated the reported experiments, reviewed the resulting artifacts, and retained decision authority over benchmark validity, claim boundaries, and interpretation. Reported quantitative claims were checked against frozen protocols, validation reports, raw logs, or machine-readable outputs where those artifacts were available; missing or unrecovered artifacts are stated as such.

D. Notation and Method Glossary

Term	Meaning in this report
TPRN	Temporal-Persistence Reconstruction Network; used here as the name of an experimental programme and related representation family, not one fixed end-to-end architecture.
Typed process	A process graph whose nodes and relations retain semantic roles rather than being treated as generic graph elements.
WL	Weisfeiler–Lehman local graph refinement/retrieval; used as an identity-invariant typed structural baseline.
F+S	Factor + Spectral reconstruction.
IPN	Identifiability-Percolation Network, a typed message-passing model with reconstruction, answerability, and pivotality-related outputs.
Explicit $\rightarrow$ IPN	Explicit process constraints are applied before/around IPN scoring rather than allowing the learned model to replace semantics.
VOI	Value of information: ranking which additional evidence would most reduce uncertainty.
Full hybrid	Explicit $\rightarrow$ IPN $\rightarrow$ factor-based VOI + spectral confirmation.
UNKNOWN	An explicit abstention state when the observed evidence does not justify one completion.