Validated cross-source synthesis of structural links in a compiled medical knowledge base: an independent literature check with a discrimination control

Youssef Ghaly

yg@gatmedi.org

Gacrux Advanced Technologies in Medicine https://orcid.org/0009-0002-9554-3691

Research Article

Keywords: cross-source synthesis, medical knowledge base, knowledge representation, retrieval versus synthesis, clinical decision support, AI hallucination, provenance, source independence, computational discovery

Posted Date: July 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10427287/v1

Additional Declarations: The authors declare potential competing interests as follows: The author is the founder of GATMEDI, the company developing the engine evaluated in this study.

Abstract

A knowledge base earns its value less from the facts it stores than from the connections it can compose across them — connections that no single source states. Composing such links, however, risks fabrication, so the property that matters is not that links are surfaced but that they are correct. We report an independently validated test of cross-source synthesis by a proprietary engine operating over a compiled medical knowledge base. The behavioural claim, fixed in advance, is that given an entity the engine surfaces structural connections that are not stated in any single ingested source and are independently verifiable against external literature the engine did not use. Because the production setting — patient data — cannot be exposed, the capability was validated on a checkable surrogate, textbook medicine, with the operation asserted identical and only the input differing. Validation was performed by an architecturally independent check against published literature; the engine did not grade its own output. A representative sample of nine composed links, spanning mechanistic, clinical, and knowledge-to-compute domains and each drawn from multiple independent sources, was examined: nine of nine were confirmed correct. A discrimination control — a real cross-organ pattern the engine recognised but deliberately withheld — showed the synthesis is selective, not indiscriminate. The links are correct, not novel: established literature is the validation instrument, which is what shows the composition is right rather than hallucinated. The capability was demonstrated at an early stage of the knowledge base's compilation, well below its intended scale. The claim is reliable cross-source synthesis, not discovery; validated synthesis of this kind is the precondition for computational discovery, which we frame as forward outlook, not a result.

1. Introduction

A medical knowledge base is valuable less for the facts it stores than for the connections it can compose across them — the structural relationships that hold between entities described in different sources, which no single source states. Retrieval surfaces text relevant to a query; synthesis composes a relationship that is present in none of the retrieved documents individually. The difficulty is that a system able to compose such connections is equally able to fabricate them, and a plausible-looking but wrong connection is worse than none. The property that must be demonstrated is therefore not that connections are surfaced but that the connections surfaced are correct, and that false ones are withheld.

This distinction is where artificial intelligence matters most in medicine. The medical literature and a single patient's record already exceed what any clinician can integrate, and the high-value task is not retrieving a fact but composing the connections latent across fragmented sources. A general-purpose large language model can generate fluent medical text, but it composes and asserts connections without a means of separating the correct from the plausible-but-wrong, and without a reliable signal for when it should decline; in a setting where an incorrect connection can harm a patient, fluency is not sufficient. What a regulated clinical environment requires is different: an output that is auditable and reliable — traceable to sources, checkable, and accompanied by the discipline to withhold when the

evidence is not there. The engine evaluated here is built to that requirement, and this study tests the property on which such reliability depends: whether the connections it composes are correct, and whether it withholds those it should not assert.

By synthesis we mean composing a structural relationship that no single source states. As a concrete example, the self-templating misfolding of a protein — in which a misshapen form propagates its conformation to normal copies — is described separately for prion disease, for the amyloid-β of Alzheimer's disease, for the α-synuclein of Parkinson's disease, and for the islet amyloid of type-2 diabetes, in sources that do not cross-reference one another. The engine composed the single shared mechanism across them, and the independent literature confirms exactly this 'prion-like' class. That composition — correct, and absent from any one source — is what this study means by synthesis, and its significance is that it is the operation that underlies discovery.

The claim is behavioural and is treated as a black box: given an entity, the engine surfaces structural connections not stated in any single ingested source that an independent check against external literature confirms as correct. The literature is the measuring instrument, not a competitor, and the claim concerns the engine's synthetic behaviour, not the novelty of any medical fact; this is explicitly not a discovery claim, because the connections are established principles, and their being established is what makes them usable as ground truth. The production setting — placing a patient's presentation against the knowledge base and surfacing the connections latent in it — cannot be exposed without breaching confidentiality and revealing the proprietary engine, so the capability is validated on a checkable surrogate domain, with the operation the same and only the input differing. How the engine composes a connection, and how the knowledge base is built, are proprietary and are not described here. The work was performed at an early stage of compilation, well below the intended scale of the knowledge base, and the claims are bounded accordingly.

2. Methods

2.1 Scope of disclosure

This report is method-blind by design. It describes inputs, outputs, and external validation, and phrases the engine's action in terms of observable behaviour — the engine surfaces a connection, composes it across sources, or withholds it — not in terms of how a connection is produced. How the engine composes a connection, and how the knowledge base is built, are proprietary and out of scope; the deliberate withholding of the mechanism is a condition of the report.

2.2 Evidence base and validator independence

The links examined were produced by the engine in the course of its normal operation, rather than generated on demand for this evaluation. This strengthens the claim: the connections arose as the knowledge base grew, with nothing steering toward a demonstration, which is a stronger statement than that the engine can produce such links when asked. Validation was performed after the fact by an

architecturally independent check against published literature, with the engine playing no part in grading its own output; ground truth was not controlled by the producer, and misses were to be reported. Each validated connection was drawn from multiple independent sources that do not cross-reference one another — the basis for the claim that the connection is stated in no single source.

2.3 Sample, metrics, and control

A representative sample of nine composed links was drawn from approximately twenty-four such links the engine had produced, chosen to span three domains: mechanistic principles, clinical disease architecture, and a knowledge-to-compute bridge. For each link the metrics were the validation rate against external literature (a link counted as confirmed only if published literature articulates the same structural relationship) and the number of independent sources across which it was composed. A discrimination control was specified: because an engine that only ever asserts a connection proves nothing, the evaluation required evidence that the engine withholds a connection where over-asserting one would reduce confidence. Coverage was noted qualitatively — the capability was demonstrated at an early stage of compilation, well below the intended scale — with its dating.

3. Results

3.1 Positive arm

All nine links were confirmed correct against external literature (nine of nine), across the three domains and each composed across multiple independent sources (two to six; Table 1, Fig. 1). The confirmed links included mechanistic principles such as a pathway-distributed enzyme-deficiency architecture (in which loss of function produces both toxic accumulation and downstream deficiency), bounded positive-feedback amplifier cascades, and conserved molecular machinery under divergent regulation; clinical architectures such as an asymptomatic precursor state recurring across malignancies, templated conformational conversion across protein-aggregation diseases, a promiscuous cardiac off-target liability, and iatrogenic phenocopy of natural disease; and one knowledge-to-compute bridge. In each case the linked facts had entered the knowledge base from sources that did not individually state the connection, yet the composed link matched independent published knowledge.

Table 1
The nine composed links and their independent external confirmation. Each was composed across multiple independent sources; all nine were confirmed correct. The links are established principles — their being established is what makes the external literature a valid ground-truth check.

#	Structural link	Domain	Ind. sources	External literature confirms
1	Model-informed dosing shared across narrow-therapeutic-index drugs	compute bridge	3	Established MIPD field; shared popPK/TDM anatomy
2	Pathway-distributed enzyme-deficiency cluster	mechanistic	5	Inborn errors of metabolism (Garrod)
3	Bounded positive-feedback amplifier cascades	mechanistic	4	Coagulation / complement / caspase cascades
4	Gradient as universal energy currency	mechanistic	2	Chemiosmotic coupling; active transport
5	Conserved machinery, divergent regulation	mechanistic	5	Conserved core with diverged modulatory structures
6	Precursor state as universal disease architecture	clinical	5	MGUS→myeloma precursor paradigm
7	Templated conformational conversion	clinical	2	Prion-like self-propagation across aggregation diseases
8	hERG as universal off-target safety liability	clinical	4	Promiscuous hERG block → acquired long-QT
9	Iatrogenic phenocopy of natural disease	clinical	6	Drug-induced parkinsonism / lupus via native machinery

3.2 Discrimination control

The discrimination control was met. A genuine cross-organ pattern — fibrosis as a final common pathway across lung, liver, kidney, and heart, which the external literature confirms is real — was recognised by the engine and deliberately deferred, logged as withheld to keep the sweep high-confidence. The engine did not miss a real pattern; it recognised one and chose not to over-assert it. An engine that filed every pattern it encountered would prove nothing, so the presence of a recognised-but-withheld pattern establishes that the synthesis is selective rather than indiscriminate.

3.3 Interpretation and coverage

The links are correct rather than novel, and this is the validation mechanism rather than a weakness. Each is an established principle in published literature; that the composed link checks out against independent published knowledge in every case, when the linked facts came from sources that did not individually state the connection, is precisely what shows the composition to be correct rather than fabricated. The capability was demonstrated at an early stage of the knowledge base's compilation, well

below the intended scale, with the continuous and patient-data trajectory lying beyond the current corpus rather than within it.

4. Discussion

The result establishes reliable cross-source synthesis: correct structural relationships composed from sources that do not individually state them, confirmed against independent literature in every case examined, together with evidence that the engine withholds a connection it could have asserted. The behaviour that distinguishes the capability is observable from the outside and does not depend on any account of how it works. A retrieval system returns text relevant to a query; the engine returns a structural link absent from any single retrieved source, and the link is independently correct. That is synthesis rather than lookup, and it is the property a retrieval system cannot supply — and, as set out in the Introduction, the property a regulated clinical use of artificial intelligence requires.

The same operation is intended to apply to patient data — placing a presentation against the knowledge base and surfacing the connections latent within it — which is why the capability was validated on checkable medical knowledge rather than on patients: the generalisation is stated, not demonstrated here, and rests on the operation being the same with only the input differing. The claim is bounded to the current scale of the knowledge base and to the sample examined, and it concerns synthesis reliability, not the discovery of anything previously unknown.

5. Limitations

1. The result is a representative sample, not a census: nine of approximately twenty-four composed links were checked, across three domains, rather than every link.
2. The links are correct but not novel; the result is about synthesis reliability, not originality, and no discovery is claimed.
3. No temporal precedence is claimed: the knowledge base was committed in bulk, so any prediction-before-confirmation ordering is not independently verifiable and is excluded from the claim.
4. Validation targeted multi-source links, where the property of not being stated in any single source is strongest; single-source cross-chapter links were excluded from this sample.
5. The capability was validated on a surrogate domain (textbook medicine); its application to patient data is asserted on the basis of an identical operation, not demonstrated here.
6. Validation was a single independent literature-check pass; it was not repeated by multiple independent assessors.
7. The full-multimodal coverage denominator is an order-of-magnitude estimate dominated by genomic and imaging modalities, and is reported as such rather than as a precise figure.

6. Significance and outlook

This section places the result in the context of the wider programme. It is separated deliberately: the validated claim above is bounded to what was demonstrated — reliable, selective synthesis of established relationships — and the paragraphs here describe direction, and should be read as context, not results.

Validated synthesis is the precondition for computational discovery. The operation demonstrated here — composing a correct structural relationship that no single source states — is, run forward against what the literature has not yet connected, the same operation by which a novel but correct connection would be surfaced. What this study establishes is that the composition step is reliable and selective; a discovery workflow built on it would stand on that reliability. What is shown here is retrospective: connections composed by the engine, then confirmed against literature that already exists. That is validation of the capability, not discovery, and the distinction is kept deliberately.

Applied prospectively, a capability of this kind bears on problems where connecting fragmented evidence is the bottleneck. These are directions the validated capability motivates, not results reported here; each would require its own prospective, pre-registered validation.

The honest form of the claim is therefore narrow: the engine composes correct structure from sources that do not contain it, and withholds where it should, demonstrated in hindsight against known medicine. That reliability is what a prospective discovery workflow would stand on; the workflow itself is not attempted here.

7. Conclusion

An independent literature check confirmed nine of nine composed structural links, each drawn from sources that did not individually state the connection, together with a discrimination control in which a real pattern was recognised and deliberately withheld. Reliable, selective cross-source synthesis is thereby demonstrated on checkable medical knowledge at an early stage of compilation, well below the intended scale; the generalisation to patient data is the same operation on a sensitive input, and the extension to prospective discovery is a defined next step — both stated as outlook, not demonstrated here.

Declarations

Competing interests

The author is the founder of GATMEDI, the company developing the engine evaluated in this study.

Funding

This work was funded internally by GATMEDI. It received no external funding.

Data and code availability

The engine and the underlying compiled knowledge base are proprietary and are not available. The composed structural links were validated against external, published literature cited in the References (all with digital object identifiers), which is publicly available; the engine and the compiled knowledge base are not.

Ethics

This study used only computational synthesis validated against published literature. It involved no human subjects and no patient data, and did not require ethics approval.

References

1. Wicha SG, Märtson A-G, Nielsen EI, Koch BCP, Friberg LE, Alffenaar J-W, Minichmayr IK. From therapeutic drug monitoring to model-informed precision dosing for antibiotics. Clin Pharmacol Ther. 2021;109(4):928-941. doi:10.1002/cpt.2202.
2. Garrod AE. The incidence of alkaptonuria: a study in chemical individuality. Lancet. 1902;160(4137):1616-1620. doi:10.1016/S0140-6736(01)41972-6. (See also: Garrod AE. Inborn Errors of Metabolism. London: Hodder & Stoughton; 1909.)
3. Macfarlane RG. An enzyme cascade in the blood clotting mechanism, and its function as a biochemical amplifier. Nature. 1964;202:498-499. doi:10.1038/202498a0. Davie EW, Ratnoff OD. Waterfall sequence for intrinsic blood clotting. Science. 1964;145(3638):1310-1312. doi:10.1126/science.145.3638.1310.
4. Mitchell P. Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. Nature. 1961;191(4784):144-148. doi:10.1038/191144a0.
5. Vandekerckhove J, Weber K. Actin amino-acid sequences: comparison of actins from calf thymus, bovine brain, and SV40-transformed mouse 3T3 cells with rabbit skeletal muscle actin. Eur J Biochem. 1978;90(3):451-462. doi:10.1111/j.1432-1033.1978.tb12624.x.
6. Kyle RA, Therneau TM, Rajkumar SV, et al. Prevalence of monoclonal gammopathy of undetermined significance. N Engl J Med. 2006;354(13):1362-1369. doi:10.1056/NEJMoa054494.
7. Jucker M, Walker LC. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. Nature. 2013;501(7465):45-51. doi:10.1038/nature12481.
8. Sanguinetti MC, Tristani-Firouzi M. hERG potassium channels and cardiac arrhythmia. Nature. 2006;440(7083):463-469. doi:10.1038/nature04710.
9. Conn H, Jankovic J. Drug-induced parkinsonism: diagnosis and treatment. Expert Opin Drug Saf. 2024;23(12):1503-1513. doi:10.1080/14740338.2024.2418950.
10. Vaglio A, Grayson PC, Fenaroli P, et al. Drug-induced lupus: traditional and new concepts. Autoimmun Rev. 2018;17(9):912-918. doi:10.1016/j.autrev.2018.03.016.
Each reference is the external, independent literature used as ground truth for the corresponding

composed link (links 9 and 10 both support link 9 of Table 1); all are given with digital object identifiers. Bibliographic details are stated only where read directly from a source.

Figures

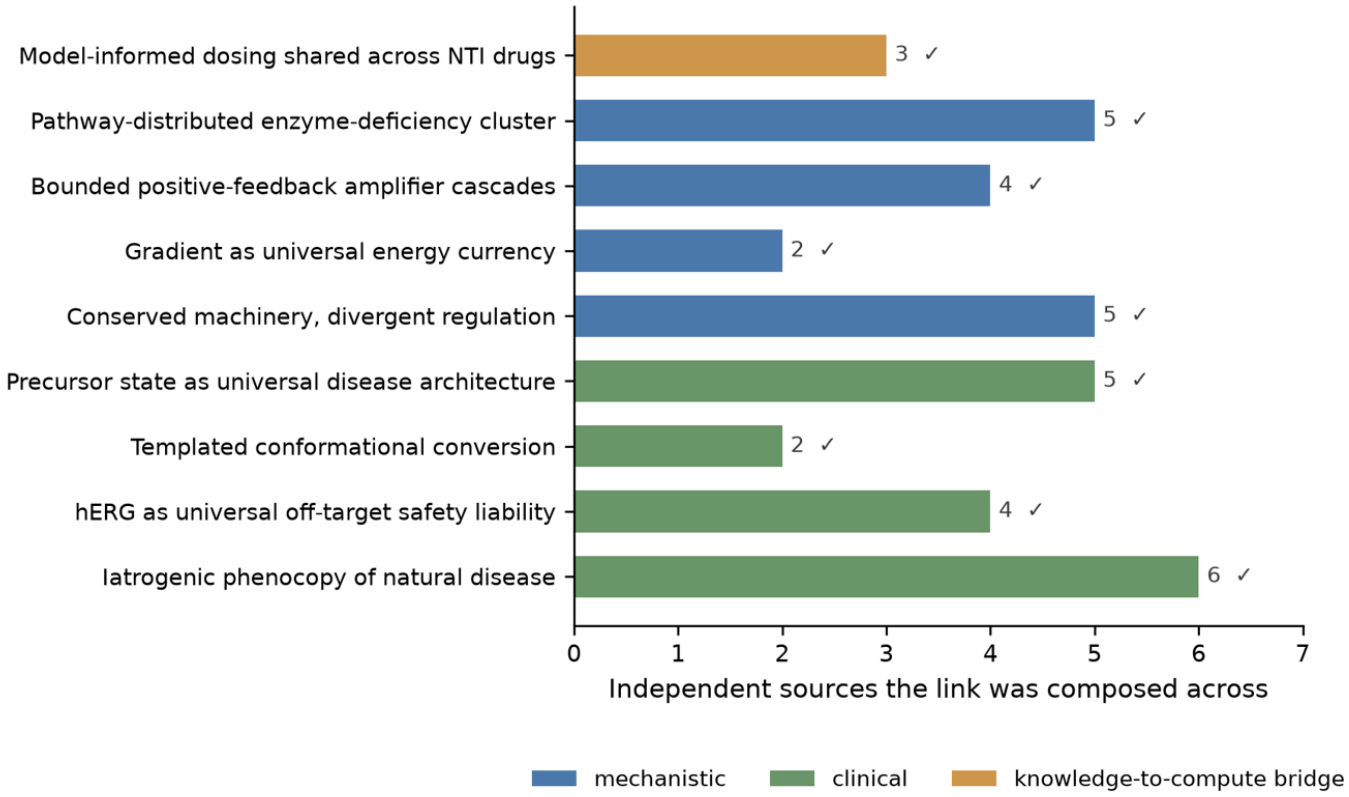

Figure 1

Independent sources composed per validated link, by domain. Each of the nine links (all externally confirmed, marked ✓) was composed across multiple independent sources — the evidence that the link is not stated in any single source.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx