DeepRoot: A KG-Coordinated Multi-Agent System for Therapeutic Reasoning over Historical Medical Texts

Zijian (Carl) Ma [*1,2] **Sean J. Wang** [*1,2] **Sijbren Kramer** [*1,2] **Li Erran Li** [3]

Abstract

Historical medical archives and traditional medicines hold immense potential for drug discovery and remain a primary source for current drug development. However, pre-ontological prose and idiosyncratic taxonomies prevent the standardization and medical modernization of the data for use in current biomedical pipelines. Furthermore, no existing LLM agent system, whether tool-calling, retrieval-augmented, or agentic deep-research, can convert such text into verifiable drug-discovery leads at scale. We close this gap with DeepRoot, a multi-agent LLM system that jointly builds and utilizes a verified knowledge graph, showing that grounding and reasoning—often conflated—are separable axes the system can compose for therapeutic reasoning. Applied to the *Shen Nong Ben Cao Jing*, DeepRoot recovers 10 of 21 held-out compound–disease treatment pairs at R@20 (47.6% vs. 4.8% for a raw corpus LLM and $\sim 2.4\%$ random) and dominates an LLM-as-judge audit for reasoning quality over baseline LLMs and LLMs with direct tool-call access to the same APIs DeepRoot itself queries. Tool-using LLMs hallucinate evidence on 87% of claims, versus 7–10% for DeepRoot. Graph-only inference hallucinates 0% but ranks lowest on reasoning coherence; DeepRoot KG + LLM is the only condition to win on both axes, pointing toward a route for systematic mining and repurposing of historical medical knowledge.

1. Introduction

Natural products—chemical compounds synthesized by living organisms—remain the leading source of approved drugs and provide scaffolds for developing more potent derivatives (Newman & Cragg, 2020; Koehn, 2012). Many natural products have been uncovered through mining traditional medicines, including morphine from opium poppies and the antimalarial artemisinin, with the latter isolated by Tu Youyou after consulting a 4th-century Chinese medical text (Tu, 2011; Brook et al., 2017).

ML, DL, and LLM approaches for mining historical medical texts at scale have been reported before but treat the text as pure input-classification problems without a reasoning trace grounded in verified biological evidence or mechanism ontologies (Li et al., 2024; Hui et al., 2020; Liu et al., 2025; Dai et al., 2024). In parallel, multi-agent LLM systems leverage a shared knowledge graph (KG) for coordinated reasoning (Ghafarollahi & Buehler, 2025; Rasmussen et al., 2025), but only qualitatively: they neither ablate the graph against agent decomposition, nor evaluate on the regimes we target—historical clinical cases where traditional text lacks clean ontological anchors, and discovery problems with sparse ground truth.

Building on these advances, we introduce **DeepRoot** (Figure 1), a multi-agent LLM pipeline where agents collectively construct and reason over a shared KG (Neo4j). Closest to our work is OpenTCM (He et al., 2025), which uses a Graph-RAG architecture for LLM reasoning. However, its construction relied on expert oversight and pure LLM-generated outputs. **DeepRoot Assembly** agentically populates the knowledge graph via seven specialized agents that combine LLM canonicalization with strict verification against curated biomedical databases. **DeepRoot Discovery** then employs critic and discovery agents, leveraging Neo4j Cypher walks for subgraph traversal to evaluate therapeutic claims and identify potential therapeutics that are mechanistically grounded in the KG.

2. Methods and KG construction

2.1. Dataset and grounding sources

Corpus. We evaluate on the *Shen Nong Ben Cao Jing* materia medica, segmented into 71 chunks. The corpus catalogues plants, animals, and minerals (*sources*), maladies, preparation methods, and claimed therapeutic uses.

[*]Equal contribution [1]Department of Bioengineering [2]Stanford University [3]Amazon AWS AI. Correspondence to: Zijian (Carl) Ma <mazijian@stanford.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026), Seoul, South Korea. PMLR 306, 2026.
Copyright 2026 by the author(s).

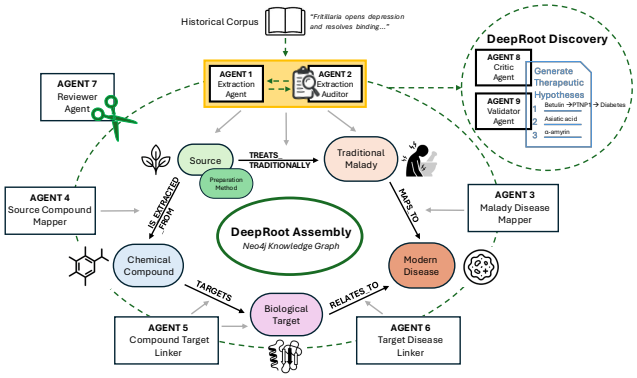

Figure 1. Schematic of DeepRoot. Graph nodes and edges are represented by rounded rectangles and black arrows. Gray arrows indicate creation of specific nodes and edges by particular agents.

External grounding. Every entity is verified against curated biomedical databases. Sources are linked to compounds via COCONUT2.0 (Chandrasekhar et al., 2025) (natural products) and PubChem (Kim et al., 2023) (chemicals); compounds are linked to molecular targets and clinical indications via ChEMBL (Mendez et al., 2019); protein targets are linked to diseases via Open Targets (Ochoa et al., 2021), and pathogenic-organism targets via NCBI Taxonomy (Schoch et al., 2020) with OLS4 (McLaughlin et al., 2025). Modern disease nodes are anchored to ICD-10, MeSH, SNOMED, MONDO, and DOID identifiers via NLM and EBI lookup services.

2.2. Knowledge graph schema

The graph has six node types and seven edge types (Figure 1). A therapeutic claim is *verifiable* when its mechanistic loop closes: a `Source` treats a `Traditional_Malady` that maps to a `Modern_Disease`; the source contains a `Chemical_Compound` that targets a `Biological_Target` which itself relates to that same `Modern_Disease`. Identity for compounds is the RDKit-computed InChIKey and identity for targets is the curated ChEMBL ID, so equivalent entities arriving from different routes collapse onto the same node. Full schema is tabulated in Table S1.

Assembly. Seven specialized agents populate the graph in dependency order: an *extractor* emits Source, Malady, and Preparation nodes from raw text; an *auditor* canonicalizes sources and archives evidence spans that fail substring verification against their source chunk; three *linkers* ground audited entities to compounds, molecular targets, and target-to-disease associations using the databases above; a *malady-to-disease mapper* follows a generate-then-verify protocol in which LLMs propose canonical names and ontology codes are recovered only by tolerant exact match, eliminating hallucinated identifiers; and a *reviewer* archives orphans and off-domain entities.

Resulting graph. On *Shen Nong Ben Cao Jing*, Assembly yields **21,111 active nodes** (415 sources, 294 maladies, 129 modern diseases, 18,012 compounds, 2,211 targets, 50 preparations) and **52,467 active edges** (32,909 `IS_EXTRACTED_FROM`, 16,696 `TARGETS`, 1,841 `RELATES_TO`, 431 `TREATS_TRADITIONALLY`, 257 `MAPS_TO`, 301 `KNOWN_TREATS`, 32 `PREPARED_AS`). A visual example of nodes originating from a single extracted source is presented in Figure S1.

3. Results

3.1. Knowledge-graph ablation: edge perturbation tests structural dependence

First, to verify that DeepRoot Discovery genuinely relies on graph structure, we progressively shuffled the graph edges and tasked the critic agent with evaluating 30 extracted closed-loop source–malady claims. As expected, the Critic's self-confidence in the therapeutic plausibility of the source based on the text decreases as edge perturbation increases, demonstrating responsiveness to the KG's integrity (Figure 2A). Around 50% perturbation, the critic's confidence converges with the raw LLM baseline, suggesting that the KG signal has been degraded enough that the critic behaves similarly to an LLM without structured graph support. Furthermore, past 50%, the score continues to decrease to 0.30, reflecting KG-dependent scoring.

3.2. KG-guided recovery of mechanistically supported candidates

Next, we tested whether DeepRoot can use the KG to recover mechanistically grounded candidates from noisy historical text. For this, we synthesized evaluation cases by selecting sets of 3 closed-loop and 7 non-closed-loop distractor sources. The associated paragraphs of those sources were then interweaved into a mini-corpus and fed to different models to rank the sources and candidate chemical compounds (Figure 2B). We report source recall@3, compound recall@10 (the fraction of closed-loop compounds recovered within the top-10 candidates), and mean self-confidence (0–1) related to the therapeutic plausibility of candidate compounds. Because each passage may contain ~ 500 compounds, compound recall@10 directly tests whether KG-grounded scoring concentrates the likely leads.

Over 30 mini-corpora, DeepRoot Discovery outperforms the LLM baseline, achieving $1.95\times$ higher source recall and $6.11\times$ higher compound recall (Figure 2B). Surprisingly, despite DeepRoot Discovery being theoretically capable of fully traversing the KG, recovery was not perfect. This is likely due to two factors: framing evaluation as an inference

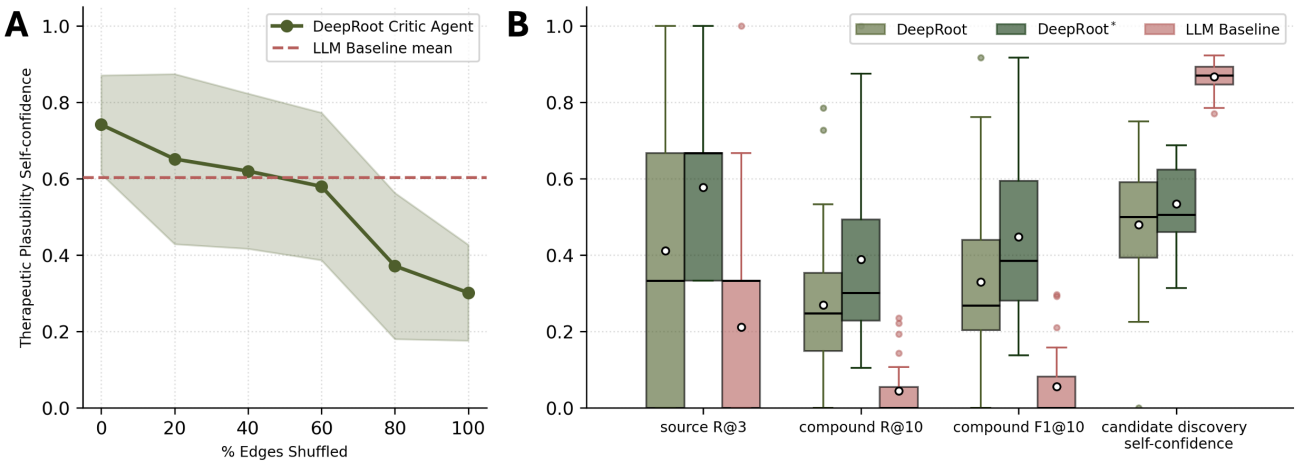

Figure 2. **(A)** Critic agent self-reported confidence for the therapeutic plausibility of each source-text pair vs. KG edge-shuffle fraction (n=30 source-text pairs). **(B)** Source-and-compound recovery of DeepRoot Discovery, LLM baseline (both using Gemini3.1 Flash Lite). *Batch evaluation by processing all 30 mini-corpora in a single invocation. Candidate discovery self-confidence refers to the mean self-reported confidence in each model's proposed compound candidates per mini-corpus (n=30 mini-corpora).

task (Edwards & Camacho-Collados, 2024) and incomplete subgraph traversal, which could explain why batch evaluation across all mini-corpora improved all metrics since it could indirectly surface shared relevant paths for the Critic.

Nevertheless, the recall and F1 gains validate that KG augmentation meaningfully enhances parsing and ranking. Notably, the LLM baseline overstates therapeutic relevance, with a self-confidence of 0.87, versus DeepRoot Discovery's 0.48, which closely aligns with the latter's source recall@3 of 0.41 (Figure 2A). This alignment suggests that self-confidence in a KG-augmented system is effectively bounded by retrieval accuracy. In contrast, the other modalities demonstrated high self-confidence hallucinations, which is a phenomenon previously reported for both LLMs and agents (Lin et al., 2022; 2025). Together with the KG ablation study, we establish that KG contributes meaningfully to the reasoning capabilities of an underlying LLM.

3.3. Blind rediscovery of held-out validated treatments

Whereas Section 3.2 tested whether DeepRoot surfaces *mechanistically grounded* candidates—compounds for which the graph itself closes a compound→target→disease loop—this experiment tests whether the system can blind-rediscover *empirically validated* compound–disease treatments after we hide them. Concretely, for each held-out pair (a `KNOWN_TREATS` edge sourced from ChEMBL clinical indications) we delete the edge and all stereochemical siblings (planar-InChIKey prefix) from the validator, then ask DeepRoot Discovery to re-rank candidates for the disease. We evaluate on a 21-pair historical set, and compare against a raw-corpus LLM given the full *Shen Nong Ben Cao Jing* and asked to rank the same top-K (Table 1).

DeepRoot Discovery recovers 10 of 21 held-out pairs (R@20 = 47.6%), compared with the LLM baseline (1 of 21, R@20 = 4.8%). Per-disease candidate pools span 87–1,954 compounds (median 835), so random $R@20 \approx 2.4\%$, suggesting that the result is far above random retrieval.

Table 1. Held-out `KNOWN_TREATS` recovery on 21 historically reachable ChEMBL indication pairs. R@k in %, MRR unitless.

Method	R@1	R@5	R@10	R@20	MRR
DeepRoot Discovery	9.5	28.6	33.3	**47.6**	**0.161**
Raw-corpus LLM	0.0	4.8	4.8	4.8	0.012

3.4. Benchmarking DeepRoot's therapeutic reasoning against diverse baselines

We audit critic-agent outputs with an independent LLM judge (Claude Sonnet 4.6, cross-family from the graded systems) on 30 stratified source–malady claims across seven conditions (Table 2): the DeepRoot Discovery at three LLM tiers (Gemini 3.1 Pro / 2.5 Flash / 3.1 Flash Lite), a graph-only baseline (no LLM), an LLM-only baseline given corpus passages, and a tool-call LLM with direct access to the same APIs (ChEMBL, Open Targets, PubMed, MeSH) that DeepRoot Assembly itself queries. The judge scores six dimensions on $[1, 5]$ and flags hallucinated evidence per claim.

All three KG-augmented configurations outperform every baseline on overall score. Even DeepRoot–Lite (3.70) exceeds both the graph-only condition (3.55) and the tool-calling LLM (2.47). This contrast highlights a tradeoff between grounding and synthesis. The tool-calling agent triggers the judge's hallucinated-evidence flag on 87% of

Table 2. Reasoning-quality evaluation: seven conditions graded by Claude Sonnet 4.6 over a stratified sample of 30 source→malady claims. Scores are means on $[1, 5]$; Hallu. is the rate of the judge's `hallucinated_evidence` flag in $[0, 1]$. **Bold** = best per column.

System	Components	Overall↑	EF↑	VA↑	RC↑	CM↑	UC↑	AC↑	Hallu.↓
DeepRoot — Gemini 3.1 Pro	graph + LLM	**3.83**	4.53	4.47	**3.97**	**4.07**	**3.73**	**3.67**	0.10
DeepRoot — Gemini 2.5 Flash	graph + LLM	3.77	**4.67**	4.37	3.83	3.63	3.57	3.63	0.07
DeepRoot — Gemini 3.1 Flash Lite	graph + LLM	3.70	4.60	4.27	3.73	3.70	3.60	**3.67**	0.07
Graph-only	graph, no LLM	3.55	4.55	**4.55**	2.69	3.21	3.31	2.93	**0.00**
Text + LLM (G3.1 FL)	corpus + LLM	3.17	3.10	2.80	3.47	3.67	3.27	3.17	0.13
Tool-call + LLM (G3.1 FL)	ChEMBL/OT/PubMed/MeSH	2.47	2.30	2.97	2.80	3.30	2.63	2.70	0.87

EF: evidence fidelity; VA: verdict alignment; RC: reasoning coherence; CM: clinical mapping; UC: uncertainty calibration; AC: actionability; Hallu.: hallucination rate.

claims, despite having on-the-fly access to the same set of APIs. By contrast, the graph-only condition produces no hallucinated evidence by construction, but exhibits the weakest reasoning coherence (2.69). KG-augmented LLMs therefore occupy a favorable middle ground: they maintain low hallucination rates (7–10%) while preserving the reasoning and synthesis capacity absent from graph-only scoring.

3.5. Human-expert evaluation of DeepRoot reasoning as a qualitative case study

To evaluate the reasoning quality of DeepRoot Discovery as a traditional medicine knowledge assistant, we constructed a mini-corpus of 50 randomly sampled source–malady pairs from *Shen Nong Ben Cao Jing*, each paired with its associated textual evidence. This setting reflects a potential use case in which a scientist seeks to assess whether observed historical claims about a source's therapeutic potential are grounded in modern biological evidence.

Table 3. Classified modern disease agreement and verdict agreement with reference to DeepRoot. Verdict refers to the system-interpreted therapeutic plausibility of a source–malady pair.

System	DA	VA	PR	PB	W	U
DeepRoot — (G3.1 FL)	–	–	3	39	4	4
Biomni (w/ full agent env.)	42	30	0	32	12	6
Text + LLM (G3.1 FL)	33	13	1	11	19	19

DA: disease agreement (% of source–malady pairs mapped to the same modern disease as DeepRoot); VA: verdict agreement (% of pairs assigned the same verdict as DeepRoot); PR: Previously Reported; PB: Plausible or Better (Previously Reported, Very Plausible, or Plausible); W: Weak; U: Unsupported.

Comparing DeepRoot with Biomni (a biomedical reasoning agent environment) (Huang et al., 2025), we find good alignment of disease classification and verdicts (Table 3). We provide full responses from DeepRoot and Biomni Appendix A.10. Overall, DeepRoot Discovery clearly leverages the KG for reasoning. Biomni also shows strong reasoning and tool-calling capabilities. Notably, both DeepRoot and Biomni cite the same biological targets for their reasoning in example 1. But example 8 highlights a key limitation of KG over-reliance. Although DeepRoot's underlying LLM identified that the hydrolyzed version of certain compounds were bioactive, those downstream products were absent from the KG, preventing KG path completion and leading to low confidence despite biological plausibility.

4. Discussion and Conclusion

DeepRoot shows that historical materia medica can be converted from pre-ontological prose into an auditable biomedical knowledge graph that supports mechanistic therapeutic reasoning. On the *Shen Nong Ben Cao Jing*, this construction pass enables held-out treatment recovery and substantially lower hallucinated-evidence rates than LLM-only, tool-calling, and biomedical-agent baselines; with broader historical corpora, the same framework could support larger-scale drug repurposing, de novo therapeutic candidate nomination, and prioritization of experimentally testable natural-product hypotheses.

It is also important to highlight the gap between DeepRoot (LLM + KG) and LLMs given direct access to the same biomedical APIs. Our results suggest that building a verified knowledge graph suppresses hallucination in a way that querying those resources at inference time does not. We show that building a verified knowledge graph suppresses hallucination in a way where querying those same APIs at inference time does not. For corpora that predate modern ontologies, retrieval-augmented and tool-using agents need a construction pass first, rather than on-demand calling. The same pattern may transfer to other historical materia medica, including Ayurvedic and broader ethnopharmacological archives, as well as structured ranking problems beyond traditional medicine. Furthermore, DeepRoot Assembly is a one-time invocation costing $\sim$ \$0.25/corpus and avoiding the recurring expense of on-demand database retrieval. In these settings, an agentically constructed KG offers the additional advantage that new curated claims, user submissions, and external evidence can be incorporated over time, expanding coverage while preserving traceability for future candidate ranking.

Limitations. *(i) Corpus:* a single 71-chunk materia medica; transfer to other historical corpora is unverified. *(ii) Sample size:* the held-out slice is $N = 21$ pairs, single seed, no bootstrap CIs. *(iii) Priors:* flat, face-validity, uncalibrated (Appendix A.5). *(iv) Coverage:* Open Targets is human-disease-only, leaving non-modern indications unscored. *(v) DeepRoot Discovery reasoning:* rationale quality is bounded by the underlying LLM. *(vi) Comparasion evalutions with other LLM modalities:* While we reported the best result for LLM or biomni from prompt engineering, limited analysis was placed in this area, but previous studies have demonstrated that such interventions seldom provide substantial improvements (Qian et al., 2024; Wu et al., 2024). Furthermore, we tried evaluating Biomni/Phylo within the LLM judge suite, but its access to recently updated papers and external validation evidence may exceed the judges' closed-book biomedical knowledge, particularly for claims grounded in post-cutoff literature.

Impact Statement

DeepRoot is a research tool for hypothesis generation, not medical advice. By converting historical materia medica into auditable source–compound–target–disease chains, it may help researchers prioritize natural-product candidates for experimental follow-up and drug repurposing.

The main risks are overinterpretation, unsafe self-medication, and misuse of traditional knowledge. Historical claims may be ineffective, toxic, or culturally specific, and graph-supported plausibility does not establish safety or efficacy. Any downstream use requires expert review, provenance tracking, toxicity assessment, and experimental validation.

Data Availability

The *Shen Nong BenCao Jing* corpus and code for generating the KG and traversing it are provided at `github.com/CarlisleMa/deeprootv1`. Provided within are also the scripts and raw data generated for the evaluations.

References

Brook, K., Bennett, J., and Desai, S. P. The chemical history of morphine: An 8000-year journey, from resin to de-novo synthesis. *Journal of Anesthesia History*, 3(2):50–55, 2017. doi: 10.1016/j.janh.2017.02.001. URL https://www.sciencedirect.com/science/article/pii/S2352452916301293.

Chandrasekhar, V., Rajan, K., Kanakam, S. R. S., Sharma, N., Weißenborn, V., Schaub, J., and Steinbeck, C. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Research*, 53(D1):D634–D643, 2025. doi: 10.1093/nar/gkae1063. URL https://academic.oup.com/nar/article/53/D1/D634/7908792. Published online 2024.

Dai, Y., Shao, X., Zhang, J., Chen, Y., Chen, Q., Liao, J., Chi, F., Zhang, J., and Fan, X. TCMChat: A generative large language model for traditional Chinese medicine. *Pharmacological Research*, 210: 107530, 2024. doi: 10.1016/j.phrs.2024.107530. URL https://www.sciencedirect.com/science/article/pii/S1043661824004754.

Edwards, A. and Camacho-Collados, J. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10058–10072, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.879/.

Ghafarollahi, A. and Buehler, M. J. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: 10.1002/adma.202413523. URL https://doi.org/10.1002/adma.202413523. Published online 2024.

He, J., Guo, Y., Lam, L. K., Leung, W., He, L., Jiang, Y., Wang, C. C., Xing, G., and Chen, H. Opentcm: A graphrag-empowered llm-based system for traditional chinese medicine knowledge retrieval and diagnosis, 2025. URL https://arxiv.org/abs/2504.20118.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., Yin, D., Marwaha, S., Carter, J. N., Zhou, X., Wheeler, M., Bernstein, J. A., Wang, M., He, P., Zhou, J., Snyder, M., Cong, L., Regev, A., and Leskovec, J. Biomni: A general-purpose biomedical AI agent. *bioRxiv*, June 2025. doi: 10.1101/2025.05.30.656746. URL https://www.biorxiv.org/content/10.1101/2025.05.30.656746v1. Preprint.

Hui, Y., Du, L., Lin, S., Qu, Y., and Cao, D. Extraction and classification of TCM medical records based on BERT and Bi-LSTM with attention mechanism. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1626–1631, 2020. doi: 10.1109/BIBM49941.2020.9313359. URL https://ieeexplore.ieee.org/document/9313359.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton,

E. E. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956. URL https://academic.oup.com/nar/article/51/D1/D1373/6777787.

Koehn, F. E. Biosynthetic medicinal chemistry of natural product drugs. *MedChemComm*, 3(8): 854–865, 2012. doi: 10.1039/C2MD00316C. URL https://pubs.rsc.org/en/content/articlelanding/2012/md/c2md00316c.

Li, Y., Luan, Z., Liu, Y., Liu, H., Qi, J., and Han, D. Automated information extraction model enhancing traditional Chinese medicine RCT evidence extraction (Evi-BERT): Algorithm development and validation. *Frontiers in Artificial Intelligence*, 7: 1454945, 2024. doi: 10.3389/frai.2024.1454945. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1454945/full.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Lin, X., Ning, Y., Zhang, J., Dong, Y., Liu, Y., Wu, Y., Qi, X., Sun, N., Shang, Y., Wang, K., Cao, P., Wang, Q., Zou, L., Chen, X., Zhou, C., Wu, J., Zhang, P., Wen, Q., Pan, S., Wang, B., Cao, Y., Chen, K., Hu, S., and Guo, L. LLM-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*, 2025. doi: 10.48550/arXiv.2509.18970. URL https://arxiv.org/abs/2509.18970.

Liu, Y., Yuan, Y., Yan, K., Li, Y., Sacca, V., Hodges, S., Cannistra, M., Jeong, P., Wu, J., and Kong, J. Evaluating the role of large language models in traditional Chinese medicine diagnosis and treatment recommendations. *npj Digital Medicine*, 8:466, 2025. doi: 10.1038/s41746-025-01845-2. URL https://www.nature.com/articles/s41746-025-01845-2.

McLaughlin, J., Lagrimas, J., Iqbal, H., Parkinson, H., and Harmse, H. OLS4: A new ontology lookup service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics*, 41(5):btaf279, 2025. doi: 10.1093/bioinformatics/btaf279. URL https://academic.oup.com/bioinformatics/article/41/5/btaf279/8125017.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., Hersey, A., and Leach, A. R. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019. doi: 10.1093/nar/gky1075. URL https://academic.oup.com/nar/article/47/D1/D930/5162468.

Newman, D. J. and Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83(3):770–803, 2020. doi: 10.1021/acs.jnatprod.9b01285. URL https://pubs.acs.org/doi/10.1021/acs.jnatprod.9b01285.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., Baker, J., Ferrer, J., Raies, A., Razuvayevskaya, O., Faulconbridge, A., Petsalaki, E., Mutowo, P., Machlitt-Northen, S., Peat, G., McAuley, E., Ong, C. K., Mountjoy, E., Ghoussaini, M., Pierleoni, A., Papa, E., Pignatelli, M., Koscielny, G., Karim, M., Schwartzentruber, J., Hulcoop, D. G., Dunham, I., and McDonagh, E. M. Open Targets platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*, 49(D1):D1302–D1310, 2021. doi: 10.1093/nar/gkaa1027. URL https://academic.oup.com/nar/article/49/D1/D1302/6024045.

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., and Sun, M. Chatdev: Communicative agents for software development, 2024. URL https://arxiv.org/abs/2307.07924.

Rasmussen, P., Paliychuk, P., Beauvais, T., Ryan, J., and Chalef, D. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. doi: 10.48550/arXiv.2501.13956. URL https://arxiv.org/abs/2501.13956.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I. NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020. doi: 10.1093/database/baaa062. URL https://academic.oup.com/database/article/doi/10.1093/database/baaa062/5881509.

Tu, Y. The discovery of artemisinin (Qinghaosu) and gifts from Chinese medicine. *Nature Medicine*, 17(10):1217–

1220, 2011. doi: 10.1038/nm.2471. URL https://www.nature.com/articles/nm.2471.

Wu, Y., Yue, T., Zhang, S., Wang, C., and Wu, Q. Stateflow: Enhancing llm task-solving through state-driven workflows, 2024. URL https://arxiv.org/abs/2403.11322.

A. Technical appendices and supplementary material

A.1. Knowledge graph schema

Edge types. Seven typed edges, each carrying a numeric `confidence_score` (flat prior, see Appendix A.5), an `evidence_type` tag (see Appendix A.6), and a `source_db` provenance field where applicable.

- `TREATS_TRADITIONALLY` (Source → Traditional_Malady): evidence span quoted from the source chunk text.
- `MAPS_TO` (Traditional_Malady → Modern_Disease): `is_primary`, `mapping_role` ∈ {primary, syndrome_component}, `mapping_source` ∈ {gemini+icd10_exact, gemini+mesh_exact, gemini+snomed_exact, gemini_unverified}, `mapping_alternatives` (JSON).
- `IS_EXTRACTED_FROM` (Chemical_Compound → Source): evidence type encodes COCONUT/PubChem provenance and resolution level (canonical vs. alias vs. formula).
- `TARGETS` (Chemical_Compound → Biological_Target): `pchembl_score`, `assay_id`, `assay_type` ∈ {B, F}, `assay_description`, `mechanism_action`.
- `RELATES_TO` (Biological_Target → Modern_Disease): `ot_overall_score`, `match_tier` ∈ {efo_id, mondo_id, doid_id, mesh_id, norm_name}.
- `KNOWN_TREATS` (Chemical_Compound → Modern_Disease): `clinical_phase` ∈ $\{1, 2, 3, 4\}$, materialized from ChEMBL drug indications; held-out evaluation slice (§3.3).
- `PREPARED_AS` (Source → Preparation_Method).

Identity and write semantics. All writes are idempotent `MERGE`-on-identity. Compound identity is the RDKit-computed InChIKey, which is invariant to canonical-SMILES variants and to naming differences across COCONUT and PubChem. Target identity is the ChEMBL ID, which unifies SINGLE PROTEIN, PROTEIN COMPLEX, PROTEIN FAMILY, and ORGANISM target types under one key. Modern disease identity is the canonical name, with ontology codes coalesce-backfilled as later agents verify them against additional services.

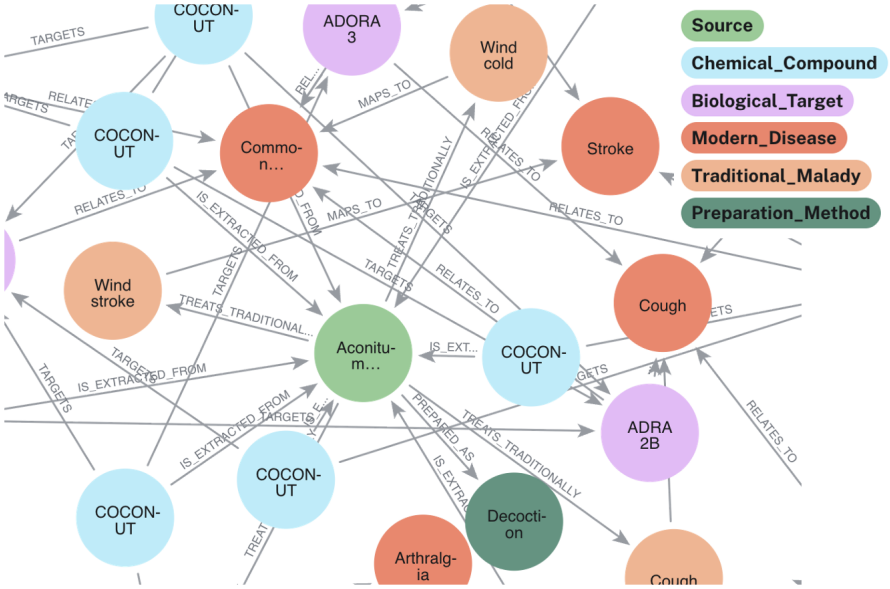

Figure S1. Example of a node cluster in Neo4j.

Table S1. Complete node schema for the DeepRoot knowledge graph.

Node	Property	Description
Source	name	Canonical Latin binomial or common name (primary key)
	aliases	Alternative names from source text
	evidence_span	Verbatim passage from which the node was extracted
	source_document	Origin corpus file identifier
	canonical_name	Auditor-resolved canonical name
	canonical_type	Taxonomic category (herb, mineral, animal, fungus, ...)
	canonical_part	Plant/animal part used (root, bark, seed, whole, ...)
	canonical_source	Database used to resolve canonical form
	canonical_raw_response	Raw LLM response from canonicalization step
	linker_status	Compound-linker outcome (ok, skipped, failed)
	linker_attempted_at	ISO timestamp of last linker run
	linker_compound_count	Number of compounds linked from this source
	linker_evidence_type	Evidence type used (coconut, chembl)
Chemical_Compound	name	IUPAC or common compound name (primary key)
	smiles	Canonical SMILES string
	inchikey	Standard InChIKey identifier
	molecular_formula	Molecular formula (e.g. C21H23NO4)
	np_likeness	Natural-product likeness score from COCONUT (-5 to $+5$)
	annotation_level	Structural confidence tier (1 = MS2 confirmed, 5 = predicted)
	source_db	Source database (COCONUT, ChEMBL, ...)
	coconut_row	COCONUT row index for traceability
	pubchem_cid	PubChem Compound ID
	target_linker_status	ChEMBL target-linker outcome
	linker_attempted_at	ISO timestamp of target-linker run
	linker_chembl_id	ChEMBL molecule ID used for target lookup
	linker_lookup_method	Match method (inchikey, smiles, name)
	linker_target_count	Targets written after filtering
	linker_dropped_count	Targets dropped below pChEMBL floor
	linker_pchembl_floor	pChEMBL activity threshold applied
	linker_max_targets	Cap on targets written per compound
	kt_linker_status	KNOWN_TREATS linker outcome
	kt_linker_attempted_at	ISO timestamp of KNOWN_TREATS linker run
	kt_linker_indication_count	Drug indications written
	kt_linker_dropped_count	Indications dropped below phase threshold
	kt_linker_min_phase	Minimum clinical trial phase accepted
Biological_Target	name	Target protein name (primary key)
	target_pref_name	ChEMBL preferred target name
	gene_symbol	HGNC gene symbol
	uniprot_id	UniProt accession
	target_chembl_id	ChEMBL target identifier
	target_type	Target class (SINGLE PROTEIN, PROTEIN COMPLEX, ...)
	ncbi_tax_id	NCBI taxonomy ID of the target organism
	td_linker_status	Target-disease linker outcome
	td_linker_attempted_at	ISO timestamp of target-disease linker run
	td_linker_association_count	Disease associations written
	td_linker_dropped_count	Associations dropped below score threshold
	td_linker_min_score	Open Targets association score floor applied

Continued on next page

Table S1 *(continued)*

Node	Property	Description
Modern_Disease	name	Disease name (primary key)
	doid_id	Disease Ontology identifier
	mondo_id	MONDO disease ontology identifier
	mesh_id	MeSH descriptor
	efo_id	Experimental Factor Ontology (EFO) identifier
	icd10_code	ICD-10 classification code
	snomed_id	SNOMED CT concept identifier
	verified_by	Agent or curator that confirmed the mapping
Traditional_Malady	name	TCM ailment name (primary key)
	description	Classical definition from source text
	evidence_span	Verbatim passage supporting the malady
	source_document	Origin corpus file identifier
	mapper_status	Malady-to-disease mapper outcome
	mapper_classification	Ontology mapping confidence class
	mapper_attempted_at	ISO timestamp of mapper run
	mapper_raw_response	Raw LLM response from mapping step
	archived	Reviewer flag: duplicate or low-quality node
	archive_reason	Free-text reason for archiving
	reviewed_by	Reviewer agent identifier
Preparation_Method	name	Preparation name (decoction, pill, powder, . . .)
	route	Administration route (oral, topical, inhaled, . . .)
	evidence_span	Verbatim passage describing the preparation

A.2. Assembly agent protocols

Table S2. The seven DeepRoot Assembly agents in dependency order. Each agent combines LLM proposal with deterministic verification against the listed grounding source.

#	Agent	Role	Grounding source
i	Extraction	emit Source / Malady / Preparation; TREATS, PREPARED_AS	—
ii	Auditor	canonicalize; verify evidence spans; merge duplicates	COCONUT, PubChem
iii	Malady→Disease	generate-then-verify ontology mapping (MAPS_TO)	MeSH, ICD-10, SNOMED
iv	Source→Compound	natural-product / chemical lookup (IS_EXTRACTED_FROM)	COCONUT, PubChem
v	Compound→Target	molecular targets and bioactivities (TARGETS)	ChEMBL
vi	Target→Disease	mechanism-to-disease, dispatched on target type (RELATES_TO)	Open Targets, NCBI Tax, OLS4
vii	Reviewer	rules + LLM archival pass (orphans, OCR artifacts, off-domain entities)	—

Extraction. A Gemini call per text chunk emits `Source` nodes with aliases and species hints, `Traditional_Malady` nodes with descriptions, `Preparation_Method` nodes, and the `TREATS_TRADITIONALLY` and `PREPARED_AS` edges among them. Each edge records the literal evidence span from the source chunk. Confidence is the LLM's self-assessed score, used only as a soft signal for downstream auditing (the auditor verifies span and identity independently).

Auditor. Three deterministic post-extraction jobs. (1) *Canonicalization.* Gemini Flash-Lite (temperature 0, batched 20 sources/call, structured-JSON schema) labels each source `organism` (Latin binomial + part), `chemical` (name or formula), or `uncanonicalized`. Per-type external lookup verifies: organisms hit a local COCONUT inverted index over 62,792 species keys, chemicals hit PubChem REST. Canonical labels with external-DB hits are tagged `gemini+coconut` / `gemini+pubchem`; the rest are tagged `gemini_unverified_*`, `uncanonicalized`, or `error`. (2) *Source merge.* Sources sharing the same (`canonical_name`, `canonical_part`) collapse onto a keeper (highest-degree, alphabetical tiebreak). Merged-from nodes are soft-archived with reason `merged_into:<keeper>`; outgoing edges are re-targeted (parallel edges take maximum confidence) and aliases are unioned. (3) *Evidence-span verification.* Substring check (whitespace-normalized) of every `TREATS_TRADITIONALLY` evidence span against its source chunk. Hallucinated spans (typically LLM-introduced ellipsis) trigger soft-archival with reason `hallucinated_evidence`.

Malady→Disease (generate-then-verify). One Gemini call per malady (temperature 0, six-shot system prompt) emits a typed exit: `disease`, `symptom`, `syndrome`, `ambiguous`, or `tcm_no_equivalent`. Crucially, the LLM emits canonical *names only*—never codes—together with an ontology hint. Each proposed name is then verified in parallel against ICD-10 (NLM Clinical Tables), MeSH (NLM RDF Lookup), and SNOMED (EBI OLS4). Verification accepts only *tolerant exact match* (case- and punctuation-insensitive); fuzzy hits, even from the API itself, are rejected. Syndromes can produce one primary plus up to two `syndrome_component` edges, but only components that pass exact-match verification are written (unverified components are dropped and logged in the primary edge's `mapping_alternatives`). Default mode rejects unverified mappings entirely; an `--allow-unverified` flag stores them with `requires_review=true` and a degraded prior.

Source→Compound. Routes by `canonical_type`. Organisms hit a local COCONUT inverted index (in-memory, $\sim$700,000 structures with species provenance) by exact normalized species name and the "first two tokens" (genus species) prefix. Chemicals hit a disk-cached PubChem REST client (4 RPS), with formula fallback via `compound/fastformula` for formula-shaped canonicals. Compound identity is the RDKit-computed InChIKey from canonical SMILES, which collapses canonicalization variants and unifies COCONUT/PubChem name splits onto a single node. Edges are written with `evidence_type` encoding the resolution path (`coconut_organism_canonical`, `pubchem_chemical_canonical`, alias, formula, unverified). Per-source `linker_status` stamping makes the agent fully resumable across network interruptions.

Compound→Target. ChEMBL queries by InChIKey (91.4% resolution rate), falling back to canonical SMILES then preferred name. For each resolved compound, the agent retrieves mechanism-of-action records and bioactivity records. Activity filtering: `pchembl_value` ≥ 5.0 (10 μM floor) for quantitative tier, `assay_type` $\in \{B, F\}$, `standard_relation` $\in \{=, \sim\}$; data validity flags must be empty. Target type is *not* restricted to SINGLE PROTEIN: PROTEIN COMPLEX (subunit fan-out), PROTEIN FAMILY (broad-spectrum inhibitors), and ORGANISM (anti-pathogen evidence, e.g. *Plasmodium falciparum* for antimalarials) are all admitted. A `--include-phenotypic` flag additionally retrieves phenotypic activities (`pchembl_value` null, `assay_type` F/B) at a lower confidence prior; for terpenes, sterols, and other natural products tested phenotypically rather than against named molecular targets, this opt-in is necessary to avoid silent loss of $\sim$2,000 compounds. Salt and tautomer parents are aggregated via the ChEMBL molecule hierarchy; without hierarchy expansion, 30–50$\%$ of activities are missed for multi-form compounds.

Target→Disease. Four-way dispatch on `target_type`. SINGLE PROTEIN: Open Targets GraphQL via UniProt→Ensembl, returning disease associations with overall scores binned to confidence tiers. PROTEIN COMPLEX: subunit fan-out via ChEMBL, then per-subunit Open Targets, with per-disease max-score deduplication. PROTEIN FAMILY: intentionally skipped (`td_linker_status = skipped_protein_family`), as family-level evidence is too coarse for clinical association. ORGANISM: an EFO/DOID walk over OLS4 starting from the NCBI `tax_id`, emitting both the specific disease class and its ancestors, falling back to a parallelized LLM safety net (Gemini proposes candidate disease names, NLM MeSH verifies via tolerant exact match) when the ontology walk returns empty. Three-tier disease matching: exact ontology-ID match first, then normalized-name exact match, then MeSH-synonym expansion. A plan-then-apply

phase performs all writes via UNWIND-batched Cypher transactions ($\sim$12 transactions for $\sim$5,500 rows), with DELETE restricted to terminal-status rows so transient API failures cannot wipe live edges.

Reviewer. Two-pass deterministic-then-LLM archival. Pass 1 catches OCR artifacts (single-character entities, mojibake), orphans (degree 0), generic categories ("herb", "compound"), and metaphysical concepts that escaped extraction. Pass 2 batches the residual ambiguous nodes ($\sim$20) to Gemini for biomedical-relevance filtering. Cascade archival propagates to incident edges. Archival is soft (archived=true with reason); no records are deleted.

A.3. Discovery agent protocols

DeepRoot Discovery comprises two agent roles operating over the typed graph: a *critic agent* that scores existing Source→Malady claims, and a *discovery agent* that nominates novel compound candidates for a target Modern_Disease. Both consume the tier-bucketed path-scoring layer described in Appendix A.4.

Critic agent. For each (Source, Malady) claim, the agent receives a structured payload assembled from the KG: the claim itself (Source name + aliases, Malady description, primary mapped Modern_Disease, mapping rationale), deterministic Pass-1 signals (path bucket distribution, loop-closure counts, top-bucket score), the top-K mechanistic chains (Source→Compound→Target→Disease) with edge metadata, and four cross-cutting enrichments—compound profiles (KNOWN_TREATS for other diseases, target spectrum), target genericity (number of associated diseases per target), source-level target convergence (multi-compound hits on the same target), and sibling verdicts (Pass-1 verdicts of other claims on the same source). The model returns a structured JSON CriticVerdict with: a verdict on the four-tier ladder (VALIDATED / PLAUSIBLE / WEAK / UNSUPPORTED); biological_plausibility and evidence_coherence scores in $[0, 1]$, defensively clamped to that range; a key_evidence list of cited compound–target– disease triples; a concerns list with typed enum values (generic_target, weak_evidence_only, indirect_mechanism, wrong_disease_mapping, syndrome_underutilized, promiscuous_compound, unverified_evidence); a free-form rationale; and a requires_human_review flag (auto-set when the LLM and deterministic Pass-1 verdicts disagree by ≥ 2 rungs). The prompt instructs the model to quote specific input fields and never speculate beyond the provided evidence; numeric ranges are enforced via post-hoc clamping rather than relying on the model to obey them.

Discovery (nominator) agent. Given a target Modern_Disease query d^*, the agent walks the KG backward (disease $\leftarrow$ malady $\leftarrow$ source $\leftarrow$ compound) to enumerate all corpus-supported candidate compounds, then walks forward (compound $\rightarrow$ target $\rightarrow$ disease) to score each candidate's mechanistic plausibility. A novelty filter drops compounds whose KNOWN_TREATS edge already reaches d^* (supporting an in-memory mask for held-out evaluation without mutating the graph). The remaining candidates are ranked lexicographically by (i) has_loop_closure (does at least one forward chain reach d^*), (ii) forward_bucket (T1 $>$ T2 $>$ T3 $>$ T4; weakest-link tier of the strongest loop-closing path), (iii) unique_sources_count, (iv) unique_maladies_count, and (v) forward_max_score (multiplicative product of edge confidences along the strongest path). The output is an ordered list of CompoundCard entries containing top historical paths (source, malady, evidence span), top forward paths (target, assay description, OT score), and KNOWN_TREATS for other diseases as polypharmacology context. The discovery agent is fully deterministic—no LLM is in the loop—making the ranking auditable and stable across re-runs.

A.4. Tier-bucket path scoring

A *path* is a sequence of typed edges connecting a source node to a disease node through compound and target intermediaries. Each edge carries an evidence_type tag (e.g., chembl_mechanism, ot_association_strong, coconut_organism_canonical) which maps to one of four tiers $T \in \{\text{T1}, \text{T2}, \text{T3}, \text{T4}\}$ (Appendix A.5, Table S3) and a flat numeric prior $c \in [0, 1]$.

For a path p with edges $e_1, \ldots, e_n$, the *path bucket* $B(p)$ and *path score* $S(p)$ are

$$B(p) \;=\; \min_{i=1\ldots n} T(e_i), \qquad S(p) \;=\; \prod_{i=1\ldots n} c(e_i).$$

Paths are ordered lexicographically by $(B(p), S(p))$ with the bucket as the primary key (highest-tier bucket wins) and the multiplicative score as tiebreak within a bucket. The bucket captures the qualitative claim "a chain is only as strong as its weakest edge" (weakest-link), while the score gives a continuous ordering inside each tier.

The same scoring layer is consumed by both Discovery agents (Appendix A.3) and by the deterministic Pass-1 signals fed to the critic. Because priors are flat (Table S3) rather than learned, raw external scores (`ot_overall_score`, `pchembl_value`, `np_likeness`) are preserved as edge attributes so downstream consumers can recalibrate without re-running Assembly.

A.5. Confidence priors

Tier ladder (used for path scoring). T1 > T2 > T3 > T4. Path bucket is the minimum tier across edges (weakest-link); within a bucket, ranking uses the multiplicative product of edge confidences as tiebreak.

Table S3. Per-edge-type confidence priors. Priors are flat (not learned), chosen on biomedical face validity, and never replaced by self-reported LLM confidence. Raw external scores (e.g., `ot_overall_score`, `pchembl_value`, `np_likeness`) are preserved as edge properties so downstream consumers can recalibrate without re-running Assembly.

Edge	Evidence type	Tier (prior)
`IS_EXTRACTED_FROM`	coconut_organism_canonical / pubchem_chemical_canonical	T1 (0.70–0.80)
	coconut_organism_alias	T2 (0.55)
	coconut_organism_unverified / pubchem_chemical_unverified	T3 (0.50–0.55)
	pubchem_chemical_formula	T4 (0.50)
`TARGETS`	chembl_mechanism	T1 (0.95)
	chembl_activity_strong (pchembl ≥ 7)	T2 (0.75)
	chembl_activity_moderate (pchembl ≥ 6)	T3 (0.60)
	chembl_activity_weak (pchembl ≥ 5) / chembl_phenotypic	T4 (0.40)
`RELATES_TO`	ncbi_pathogen_consensus	T1 (0.92)
	ot_association_strong (OT ≥ 0.7)	T1 (0.85)
	ot_association_moderate (OT ≥ 0.4) / complex_aggregate / pathogen_llm_verified	T2 (0.65–0.75)
	ot_association_weak (OT ≥ 0.2)	T3 (0.45)
`MAPS_TO`	icd10/mesh/snomed exact (primary)	T1 (0.80–0.85)
	syndrome_component / symptom	T2 (0.65–0.75)
`KNOWN_TREATS`	clinical_phase = 4 (approved)	T1 (0.95)
	clinical_phase = 3	T1 (0.85)
	clinical_phase = 2	T2 (0.65)
	clinical_phase = 1	T3 (0.45)

A.6. Graph statistics

Node and edge totals (active, post-Assembly). 21,111 nodes active, 94 archived. By type: 415 Source, 294 Traditional_Malady, 129 Modern_Disease, 18,012 Chemical_Compound, 2,211 Biological_Target, 50 Preparation_Method. Edges: 52,467 active. By type: 32,909 `IS_EXTRACTED_FROM`, 16,696 `TARGETS`, 1,841 `RELATES_TO`, 431 `TREATS_TRADITIONALLY`, 301 `KNOWN_TREATS`, 257 `MAPS_TO` (208 primary +49 syndrome_component), 32 `PREPARED_AS`.

Per-evidence-type breakdown. IS_EXTRACTED_FROM: 32,885 organism_canonical, 21 chemical_canonical, 3 formula. TARGETS: 60 mechanism, 1,148 strong, 1,185 moderate, 2,264 weak, 12,039 phenotypic. RELATES_TO: 936 ot_weak, 666 ot_moderate, 14 ot_strong, 203 complex_aggregate, 17 llm_verified, 3 pathogen_consensus, 2 efo. KNOWN_TREATS: 60 phase 4, 92 phase 3, 84 phase 2, 65 phase 1. The phenotypic-heavy distribution of TARGETS reflects the natural-product corpus: terpenes, sterols, and flavonoids are predominantly characterized by phenotypic bioassays rather than named molecular targets.

Convergence. 4,605 compounds appear in ≥ 2 sources (classic phytomedicine pattern: β-sitosterol $112\times$, quercetin $87\times$, kaempferol $66\times$). Across the 129 Modern_Disease nodes, 257 MAPS_TO edges resolve to an average of 1.99 maladies per disease, indicating strong canonical convergence rather than fragmentation.

Coverage. 504 of 2,211 targets (22.8%) link to at least one disease; 88 of 129 disease nodes (68%) are reached by at least one target. 3,221 of 18,012 compounds (17.9%) have at least one TARGETS edge; the remainder either lack ChEMBL records or have no admissible target-class data, a documented limitation of the underlying databases rather than of the pipeline.

A.7. Evaluation protocols

Eval 1: edge-perturbation sensitivity. A fixed test set of closed-loop Source–Malady claims is sampled from the KG. For each perturbation level $p \in \{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$, a fraction p of edges across four mechanistic edge types (TARGETS, RELATES_TO, KNOWN_TREATS, MAPS_TO) is selected uniformly at random and their target endpoints are shuffled among themselves. A single perturbation is applied per level; all test claims are then evaluated by the Critic against the same perturbed graph. Self-confidence (mean biological plausibility over all claims) is reported per level.

Eval 2: source and compound recovery on mini-corpora. To build each mini-corpus, build_recovery_eval_corpus.py first queries the knowledge graph for two disjoint source pools: **closed-loop sources** (those with at least one complete Source → Compound → Target → Disease chain where the source also TREATS_TRADITIONALLY → Malady → MAPS_TO the same disease) and **distractor sources** (all other non-archived KG sources). The full *Shen Nong Ben Cao Jing* text is split into paragraphs, each tagged by keyword regex against every source name in the KG. Gemini Flash then verifies which tagged sources a paragraph actually *describes therapeutically* (rather than merely cross-referencing in a compatibility list). Verified single-source paragraphs are banked into the two pools. Each synthetic eval case is assembled by a diversity-maximising greedy algorithm: it picks K=3 least-used closed-loop source paragraphs and N=7 least-used distractor source paragraphs, ensures no overlap between the two sets, then deterministically shuffles all 10 paragraphs into an interleaved mini-corpus. The label set for compound recovery includes both **closed-loop compounds** (retrieved via a Cypher walk confirming Compound → TARGETS → Target → RELATES_TO → Disease for a disease the source already treats; typically 1–5 per source) and **distractor compounds** (IS_EXTRACTED_FROM compounds of the distractor sources). Compound recall@k is reported separately against each label set so the closed-loop and broad-coverage signals can be distinguished. Thirty such mini-corpora are generated, each with a distinct closed-loop source signature enforced by deduplication.

Eval 3: positive-control recovery of hidden known treatments. The 301 KNOWN_TREATS edges are filtered to those whose Modern_Disease has a backward chain $d \leftarrow$ malady $\leftarrow$ source $\leftarrow$ compound in the KG (the *historical-reachability* subset), yielding 21 (compound, disease) pairs across 10 diseases. For each test pair (c^*, d^*), we compute c^*'s planar InChIKey prefix (first 14 characters, dropping stereochemistry) and mask every KNOWN_TREATS edge from any compound sharing that prefix to d^* (in-memory mask only; the graph is not mutated). The discovery agent is then run on d^* with top-$K = 20$. A trial succeeds at rank r if any nominee within the top r shares c^*'s planar prefix. The candidate pool—all compounds reachable via the backward chain from d^*—is recorded per-disease (range 87–1,954, median 835), giving a uniform random recall@20 baseline of $\approx 2.4\%$.

Eval 4: LLM-as-judge reasoning quality. A stratified sample of 30 closed-loop Source–Malady claims is drawn from the 431 candidate claims, with the strata chosen to exercise distinct verdict regimes (representative balance of *unsupported*, *strong_support*, *T1-bucket-without-loop-closure*, *mechanistic-only*, and *traditional-only*). The same 30 claims are scored by six conditions: DeepRoot Discovery at three LLM tiers (Gemini 3.1 Pro, 2.5 Flash, 3.1 Flash-Lite), a graph-only deterministic baseline (Pass-1 verdict only, no LLM), an LLM baseline given just the corpus pas-

sages, and a tool-call LLM baseline given direct API access to ChEMBL, Open Targets, PubMed, and MeSH. Outputs are graded by Claude Sonnet 4.6 (cross-family from the graded systems) on six dimensions in $[1, 5]$: Evidence Fidelity (does the critic cite evidence present in the payload?), Verdict Alignment (does the verdict follow from the visible evidence?), Reasoning Coherence (does the rationale explain *this* claim's chain?), Clinical Mapping (does the critic responsibly handle the malady→disease mapping?), Uncertainty Calibration (are the scores, concerns, and review flag calibrated?), and Actionability (would a curator know what to inspect next?). The judge additionally returns six binary flags (`hallucinated_evidence`, `unsupported_verdict_jump`, `ignored_loop_closure_status`, `overclaims_strength`, `contradictory_scores`, `needs_human_review`) and a recommended status in {`pass`, `weak_pass`, `fail`, `human_review`}. The judge sees only the critic's visible artifacts (verdict, scores, key_evidence, concerns, rationale) plus, where applicable, the structured payload that the critic was given; it does not have access to ground truth and grades the quality of the critic's argument rather than its absolute correctness.

A.8. Prompt templates

This section summarizes the four critical prompts in DeepRoot. Each template is described as a tuple of *(role, input schema, output schema, key instructions)*; the verbatim text is in the released codebase.

Extraction prompt. *Role:* extract typed entities from a single text chunk of the source corpus. *Input:* the chunk text plus a `source_document` identifier. *Output schema (structured JSON):* `sources[]`, `maladies[]`, `preparations[]`, `treats_edges[]`, `prepared_as_edges[]`; each entity carries `name`, `aliases`, `evidence_span` (verbatim quote from the chunk), and a self-assessed `confidence`. *Key instructions:* evidence spans must be substring-matchable to the chunk text; identifiers (binomials, ChEMBL IDs) are never to be invented; the LLM emits names, not codes.

Malady→Disease mapping prompt (generate-then-verify). *Role:* classify each Traditional_Malady into one of {`disease`, `symptom`, `syndrome`, `ambiguous`, `tcm_no_equivalent`} and propose canonical English name(s). *Input:* the malady's `name`, `description`, `evidence_span`, and source classical context. *Output schema:* a top-level `classification` field plus a `mappings[]` list, where each mapping carries `name` (canonical English), an `ontology_hint` ∈ {`icd10`, `mesh`, `snomed`}, a `role` ∈ {`primary`, `syndrome_component`}, and a one-sentence `rationale`. *Key instructions:* the model emits canonical *names only, never codes*; for `syndrome` maladies, up to two `syndrome_component` entries may be returned in addition to the primary; six in-context examples cover the typical TCM patterns (*wind heat*, *gu toxin*, *counterflow*, etc.). Codes (ICD-10, MeSH, SNOMED) are recovered downstream by deterministic exact-match verification against the corresponding ontology services—never trusted from the LLM.

Critic agent prompt. *Role:* biomedical reasoning expert evaluating whether a historical Source plausibly treats a Modern_Disease via known mechanisms. *Input:* the structured payload described in Appendix A.3 (claim, Pass-1 signals, evidence paths, compound profiles, target profiles, source-level target convergence, sibling verdicts), serialized as a single JSON object. *Output schema:* the `CriticVerdict` object described in Appendix A.3. *Key instructions (highlighted in the system prompt):* (i) assess TARGET QUALITY via the per-target disease count (pleiotropy of ≥ 50 associated diseases is flagged `generic_target`); (ii) ground COMPOUND PHARMACOLOGY in the compound's `KNOWN_TREATS` record and target spectrum; (iii) scrutinize the DISEASE MAPPING for clinical plausibility against the historical evidence span; (iv) upweight POLYPHARMACOLOGY / CONVERGENCE when multiple compounds from the same source hit the same target; (v) discount SPECIFICITY when the source has many unrelated claims (kitchen-sink remedy). The prompt instructs the model to quote specific input fields and never speculate beyond the provided payload.

Baseline LLM prompt. *Role:* Candidate nominator. You are a pharmaceutical and natural products expert with broad knowledge of traditional Chinese medicine, pharmacognosy, and bioactive plant compounds. You will be given a passage from the Shen Nong Ben Cao Jing and will be asked to discover and rank plausibility of therapeutic compounds. Read this passage from a historical Chinese herbal text. List each medicinal source discussed; for each, propose up to 10 therapeutic chemical compounds it likely contains, with a plausibility score 0.0-1.0 (0.0 = incoherent or no known mechanism, 1.0 = biologically obvious) and a one-sentence reasoning grounded in known biochemistry or pharmacology.

Biomni prompt. *Role:* Biomedical reasoning expert. *Input:* You are a pharmaceutical and natural-products expert with broad knowledge of traditional Chinese medicine, pharmacognosy, bioactive plant/mineral/animal compounds, and modern clinical pharmacology. You will be given an extracted corpus of historical medicinal entries. Each block is one claim: a

SOURCE (a medicinal substance) and a MALADY (the traditional ailment the historical text says it treats), followed by the original passage. You must judge to your fullest capability, using any databases and subagents available necessitating @ChemBL, @PubChem, @PubMed, @OpenTargets as places for you to find relevant compounds or links to mechanistically evaluate the plausibility that the source provided can treat the malady listed. Rules: Score each claim independently; do not let one claim bias another. Do not invent citations or database IDs. Reason from known pharmacology. If the source is an inert mineral or has no plausible bioactive route to the disease, say so and mark Unsupported.

LLM-judge prompt. *Role:* independent biomedical evaluation judge grading the visible reasoning of an automated critic, not deciding whether the underlying therapeutic claim is true. *Input:* `condition` under judgment (one of the six in Eval 4), the claim, the Pass-1 deterministic signals, the visible payload the critic received (or empty for graph-free conditions), and the critic's full structured output. *Output schema:* per-dimension scores in $[1, 5]$ for the six rubric dimensions (Appendix A.7), six binary flags, a recommended status, and a free-text justification referencing specific input fields. *Key instructions:* treat Pass-1 as an input signal, not ground truth (a critic can be good even when it agrees with an imperfect Pass-1 verdict, if it explains the limitation correctly); penalize hallucinated citations (every compound, target, and disease named by the critic must trace to a payload field, except for widely accepted biomedical facts in text-only conditions); penalize rationales that confuse global top-bucket evidence with loop-closing disease support (a claim with 76 T1-bucket paths but zero of those paths reaching the mapped disease is *not* well-supported).

A.9. Implementation notes

Models. Extraction, Auditor canonicalization, Malady mapper, and Reviewer Pass 2 use Gemini 3.1-Flash-Lite at temperature 0 with structured-JSON schemas (rate-limited 1 request/sec, retried up to 4 times on transient failure). The Discovery LLM stages use Gemini Pro under the same rate-limit envelope.

Storage. Neo4j AuraDB. All writes are idempotent `MERGE` statements with status fields (e.g., `linker_status`, `td_linker_status`, `mapper_status`) written *last* so transient failures retry safely. Default re-run filters select only nodes with `status IS NULL`; `--retry-misses` reprocesses `error` and `no_*` branches; `--force-relink` deletes prior edges from terminal-status branches only, never from in-flight nodes. This makes the pipeline crash-safe end-to-end.

Caching. PubChem responses are disk-cached by query; COCONUT is loaded once into an in-memory inverted index. Re-runs after network interruptions complete in seconds.

End-to-end cost. Full Assembly over the 71-chunk corpus runs in $\sim$30–40 minutes wall-clock, dominated by per-edge Neo4j MERGE latency on AuraDB. Total LLM spend is $\sim\$0.25$ (Flash-Lite); ChEMBL, COCONUT, PubChem, Open Targets, NLM, and EBI are all free.

A.10. Qualitative scoring of representative critic agent responses

We assembled fifty source–malady pairs, constrained so that every pair has both a traditional `TREATS_TRADITIONALLY` edge and a primary mapping to a modern disease. For each pair we extracted the corresponding historical passage from the *Shen Nong Ben Cao Jing* to form a mini-corpus. The mini-corpus was given to DeepRoot Discovery, Biomni, or baseline LLM and tasked to score each pair based on the plausibility that the particular source can treat the traditional malady it is claimed to treat. Shown below are the first 10 examples for DeepRoot and Biomni.

Table S4. DeepRoot versus Biomni baseline): malady→modern-disease mappings and plausibility verdicts across ten (Source, Malady) pairs.

		DeepRoot		Biomni	
#	Source → Malady	Mapped disease	Verdict	Mapped disease	Verdict
1	*Epimedium* → impotence	Erectile Dysfunction	Very Plausible	Erectile Dysfunction	Very Plausible
2	*Trichosanthes cucumeroides* → menstrual block	Amenorrhea	Unsupported	Amenorrhea	Plausible
3	*Rhizoma Arisaematis* → heart pain	Angina Pectoris	Previously Reported	Angina Pectoris	Weak
4	*Calcareous Spar* → generalized fever	Fever	Unsupported	Pyrexia (Fever)	Unsupported
5	*Aster tataricus* → cough	Cough	Plausible	Bronchitis; Asthma	Very Plausible
6	*Fructus Ailanthi Altissimi* → impotence	Erectile Dysfunction	Plausible	Erectile Dysfunction	Weak
7	*Ge Gen* → toxicity	Poisoning	Plausible	Alcohol Poisoning	Plausible
8	*Sanguisorba officinalis* → vaginal discharge	Leukorrhea	Unsupported	Leukorrhea; Candidiasis	Plausible
9	*Hedgehog Pelt* → hemorrhoids	Hemorrhoids	Plausible	Hemorrhoidal Disease	Unsupported
10	*Calcareous Spar* → cough	Cough	Unsupported	Cough; Bronchitis	Unsupported

Each box below shows one verdict for a (Source, Malady) pair. VERY PLAUSIBLE PREVIOUSLY REPORTED PLAUSIBLE WEAK UNSUPPORTED badges reflect the reported plausibility tier (green = strongest; PREVIOUSLY REPORTED denotes a source compound with an already-known therapeutic edge to the mapped disease).

Example 1: *Epimedium* (Source) and *Impotence* (Malady)

DeepRoot Verdict: VERY PLAUSIBLE

Weakest Link: The KG contains several non-closing paths for icariin that distract from the primary, well-established PDE5 inhibitory mechanism.

Reasoning: Epimedium is well-documented for its primary active constituent, icariin, a known phosphodiesterase-5 (PDE5) inhibitor. The KG confirms the icariin→PDE5A→erectile dysfunction path, which is the canonical mechanism for treating impotence by enhancing nitric oxide-mediated vasodilation. Although other paths in the graph do not close the loop, the primary mechanism is highly robust and clinically validated.

Biomni Verdict: VERY PLAUSIBLE

Weakest Link: Icariin's PDE5A IC_{50} (∼5,900 nM) is ∼1,000-fold weaker than sildenafil (∼3.5 nM); oral bioavailability of the native glycoside in humans remains the key uncertainty.

Reasoning: Icariin (CID 5318997; CHEMBL553204) is a confirmed PDE5A inhibitor (IC_{50} = 5,900 nM, ChEMBL; active in PubChem AID 362785 — human recombinant PDE5A1 assay), sharing the same mechanism as sildenafil: cGMP preservation → corpus cavernosum smooth muscle relaxation. In vivo rat models confirm efficacy via GRK2

downregulation in penile tissue (PMID 39905744) and protection of cavernous tissue under hyperglycemia (PMID 40161546). The metabolite icariside II is significantly more potent. Icariin has reached ChEMBL Max Phase 3, making this the best-validated claim in the dataset.

Example 2: ***Trichosanthes cucumeroides*** **(Source) and** ***Menstrual block*** **(Malady)**

DeepRoot Verdict: UNSUPPORTED

Weakest Link: The lack of any mechanistic path connecting cucurbitacins to reproductive hormonal regulation or uterine function.

Reasoning: Trichosanthes cucumeroides contains cucurbitacins, which are potent cytotoxic and anti-inflammatory triterpenoids. While these compounds modulate pathways like IGF2BP1 and ITGAL, there is no clear mechanistic link to the regulation of the menstrual cycle or the treatment of amenorrhea. The KG paths converge on unrelated conditions like diabetes and malaria, failing to provide a coherent basis for this traditional claim.

Biomni Verdict: PLAUSIBLE

Weakest Link: The primary active agent, β-trichosanthin, is a large protein ($\sim$28 kDa) with an abortifacient rather than emmenagogue mechanism — it does not address functional amenorrhea unrelated to pregnancy.

Reasoning: Trichosanthes root contains β-trichosanthin, a Type I ribosome-inactivating protein (RIP) that depurinates 28S rRNA, arresting protein synthesis in trophoblast cells and inducing abortion in vivo (PMID 3596896; 198 PubMed results for 'trichosanthin ribosome inactivating'). This coherently explains the TCM use for pregnancy-related amenorrhea. Cucurbitacin B (CID 5281316; CHEMBL2180736) contributes cytotoxic activity (IDH1 IC_{50} = 640 nM; 79/122 active PubChem assays). The mechanism is abortifacient rather than broadly emmenagogue, limiting applicability to non-pregnancy amenorrhea.

Example 3: ***Rhizoma Arisaematis*** **(Source) and** ***heart pain*** **(Malady)**

DeepRoot Verdict: PREVIOUSLY REPORTED

Weakest Link: The evidence tier for the adenosine-receptor interactions is limited to bronze, suggesting a need for higher-confidence binding data.

Reasoning: Rhizoma Arisaematis contains adenosine, which acts on adenosine receptors (ADORA1, ADORA2A, ADORA2B, ADORA3) to modulate cardiac rhythm and coronary blood flow. These paths close the loop to angina pectoris, providing a plausible mechanism for alleviating heart pain through vasodilation and metabolic protection. The evidence is consistent across multiple receptor subtypes, though the overall tier remains bronze.

Biomni Verdict: WEAK

Weakest Link: Arisaema species (the actual botanical source) have no documented cardiac pharmacology in any modern database; the claim's plausibility depends entirely on resolving 'Hu Zhang' as Polygonum cuspidatum (resveratrol), which is a different plant.

Reasoning: Rhizoma Arisaematis (Tian Nan Xing) derives from Arisaema species; PubMed and ChEMBL searches returned no cardiac or anti-anginal activity for any Arisaema compound. The classical text's reference to 'Hu Zhang' maps to Polygonum cuspidatum in modern pharmacopoeia — a rich source of resveratrol (CID 445154; CHEMBL165, Max Phase 3) with COX-1 (IC_{50} = 380–2,840 nM), COX-2 (IC_{50} = 3,490 nM), and SIRT1 inhibition relevant to anti-platelet and cardioprotective activity (PMID 41978708). If the source is confirmed as knotweed, the rating rises to Plausible; as Arisaema, it remains Weak.

Example 4: ***Calcareous Spar*** **(Source) and** ***Generalized fever*** **(Malady)**

DeepRoot Verdict: UNSUPPORTED

Weakest Link: The substance is an inert mineral with no pharmacological activity relevant to systemic fever.

Reasoning: Calcareous Spar is primarily composed of calcium carbonate (CaCO3), an inert mineral in the context of systemic fever reduction. It lacks bioactive organic compounds capable of modulating thermoregulatory centers or inflammatory cytokines. The KG correctly shows no paths, as there is no plausible pharmacological mechanism for this substance to treat fever.

Biomni Verdict: UNSUPPORTED

Weakest Link: CaCO3 has no antipyretic target in any database; the antipyretic TCM mineral is Gypsum (CaSO4), a distinct compound with documented TLR4/NF-κB inhibition.

Reasoning: Calcareous Spar is calcium carbonate (CaCO3; CID 10112; CHEMBL1200539, Approved). Its 31 ChEMBL and 22 OpenTargets indications cover hypocalcemia, renal osteodystrophy, GERD, and osteoporosis — none include fever. No antipyretic bioactivity was found in ChEMBL or PubMed. The TCM antipyretic mineral is Gypsum (CaSO4, Shi Gao), which inhibits TLR4/NF-κB signaling, reduces IL-6/PGE2/TNF-α, and upregulates AVP in LPS-fever models (PMID 37769495). The TCM 'cold nature clears heat' rationale has no pharmacological correlate for CaCO3.

Example 5: ***Aster tataricus*** **(Source) and** ***Cough*** **(Malady)**

DeepRoot Verdict: PLAUSIBLE

Weakest Link: The low-confidence wood-tier evidence for the strobopinin-PTGS interaction limits the certainty of this specific mechanism.

Reasoning: Aster tataricus is traditionally used as an antitussive, and its constituents like strobopinin target PTGS1/2 (cyclooxygenases). Inhibition of these enzymes can reduce airway inflammation and prostaglandin-mediated cough reflexes. While the KG confirms the loop to cough, the evidence is wood-tier, and other paths for quercetin and kaempferol do not close the loop to respiratory conditions.

Biomni Verdict: VERY PLAUSIBLE

Weakest Link: Shionone (CID 12315507) has zero PubChem bioassay data and is absent from ChEMBL; specific antitussive IC_{50} values for any Aster compound against a defined respiratory target have not been reported.

Reasoning: Aster tataricus (Zi Wan) root contains shionone, epifriedelanol (CID 119242; 2 active PubChem assays), and triterpenoid saponins. Multiple in vivo mouse studies confirm mechanistically specific anti-inflammatory activity in respiratory disease: PI3K/AKT/NF-κB inhibition in asthma (PMID 38081397), HIF-1α/VEGFA pathway suppression in chronic bronchitis, and iNOS/LPS-mediated mediator suppression in acute bronchitis. Saponins additionally provide expectorant activity via mucociliary stimulation. Comprehensive reviews (PMIDs 40271061, 36080154) confirm the traditional antitussive use is well-corroborated by modern pharmacology.

Example 6: ***Fructus Ailanthi Altissimi*** **(Source) and** ***impotence*** **(Malady)**

DeepRoot Verdict: PLAUSIBLE

Weakest Link: The reliance on wood-tier evidence for the identified targets makes the specific pharmacological contribution of these compounds to erectile function speculative.

Reasoning: Fructus Ailanthi Altissimi contains flavonoids like quercetin and rutin, which exhibit antioxidant and vasorelaxant properties. The KG identifies paths such as quercetin→AKR1B1 and rutin→ADRA2A/C, which are linked to erectile dysfunction. While these targets are relevant to vascular health and smooth muscle tone, the evidence

is primarily from wood-tier sources, limiting the strength of the mechanistic link.

Biomni Verdict: WEAK

Weakest Link: Ailanthone and related quassinoids have no demonstrated activity at PDE5, androgen receptors, eNOS, or any pro-erectile target; the pharmacological profile is anticancer and antimalarial only.

Reasoning: Ailanthus altissima fruit contains quassinoids (ailanthone, CID 72965; CHEMBL161015) with documented anticancer (HepG2 IC_{50} = 240 nM), antiparasitic (P. falciparum IC_{50} = 0.003 μg/mL), and NF-κB/PKM2 inhibitory activity. PubMed returned zero results for 'Ailanthus altissima erectile dysfunction.' No PDE5A, androgen receptor, or nitric oxide pathway activity was found in ChEMBL (8 activities) or PubChem (11 assays, 1 active). The TCM bark use is for spermatorrhea — a distinct condition — and even that lacks modern mechanistic validation.

Example 7: *Ge Gen* (Source) and *Toxicity* (Malady)

DeepRoot Verdict: PLAUSIBLE

Weakest Link: The term 'toxicity' is too non-specific to definitively map to the ACHE-inhibition mechanism identified in the KG.

Reasoning: Ge Gen (Pueraria lobata) contains isoflavones like daidzin and flavonoids like rutin and apigenin. The KG identifies paths where these compounds target ACHE (acetylcholinesterase), which is a critical enzyme in managing certain types of neurotoxicity and organophosphate poisoning. While the loop closes, the clinical application of Ge Gen for general 'toxicity' is broad and requires more specific evidence regarding the type of toxin.

Biomni Verdict: PLAUSIBLE

Weakest Link: The detoxification evidence is specific to alcohol-related hepatotoxicity and does not extend to the broader traditional claim of resolving 'various toxins.'

Reasoning: Ge Gen (Radix Puerariae) contains puerarin (CID 5281807; CHEMBL486386, Max Phase 2) and daidzein (CID 5281708; CHEMBL8145). Puerarin modulates macrophage polarization (M1$\rightarrow$M2) and activates PI3K/AKT signaling to suppress alcohol-induced hepatocellular cytotoxicity (PMID 42010989; 14 PubMed papers). Daidzein inhibits ALDH2 (IC_{50} = 9,000 nM) and acts as a phytoestrogen via ERβ (IC_{50} = 303 nM). OpenTargets confirms Phase 2 clinical registration for alcohol abuse/dependence. The mechanism is coherent for alcohol detoxification specifically, but the broader 'various toxins' claim lacks support.

Example 8: *Sanguisorba officinalis* (Source) and *Vaginal discharge* (Malady)

DeepRoot Verdict: UNSUPPORTED

Weakest Link: The absence of any KG paths linking the known astringent tannins of Sanguisorba to the treatment of vaginal discharge.

Reasoning: Sanguisorba officinalis is rich in hydrolyzable tannins and ellagic acid, which provide astringent and antimicrobial properties useful for treating inflammatory exudates. However, the KG paths for its constituents like astragalin and genistein do not close the loop to vaginal discharge or related gynecological infections. The current graph evidence focuses on unrelated targets like ESR2 and CA12, which do not explain the traditional use.

Biomni Verdict: PLAUSIBLE

Weakest Link: The most direct anti-Candida evidence uses the charred/processed form (not raw root), and no human clinical trial for vaginal discharge specifically has been identified.

Reasoning: Sanguisorba officinalis contains ellagic acid (CID 5281855; CHEMBL6246; Src IC_{50} = 300 nM, DNA

gyrase IC_{50} = 3.3–47 μg/mL), sanguiin H-6 (CID 16130897), and ziyuglycoside I/II (CID 71609288). Direct anti-Candida albicans activity is confirmed for the whole extract (PMID 12169406). Ziyuglycoside II synergizes with tetracycline against MDR S. aureus (MIC = 32 μg/mL; FICI = 0.25–0.375; PMID 41385864) via membrane disruption and ROS generation. Astringent ellagitannins reduce mucosal secretion and inflammation, providing a dual antimicrobial-astringent mechanism supporting the traditional use.

Example 9: *Hedgehog Pelt* (Source) and *Hemorrhoids* (Malady)

DeepRoot Verdict: PLAUSIBLE

Weakest Link: The reliance on wood-tier evidence for the interaction between the pelt's constituents and the identified receptors.

Reasoning: Hedgehog Pelt contains various organic acids and signaling molecules that the KG links to ADRA1A and NR3C1 (glucocorticoid receptor). These targets are involved in vasoconstriction and anti-inflammatory responses, which are relevant to reducing the swelling and pain associated with hemorrhoids. While the mechanism is plausible, the source material is complex and the evidence tier is low.

Biomni Verdict: UNSUPPORTED

Weakest Link: No bioactive compounds from hedgehog pelt (Erinaceus europaeus skin) have been characterized in any database; the entire mechanistic chain is unestablished.

Reasoning: Hedgehog pelt (Wei Pi, Pellis Erinacei) from Erinaceus europaeus returned zero PubMed results for 'hedgehog pelt hemorrhoids' or 'Ci Wei Pi TCM hemorrhoids.' No ChEMBL entries exist for E. europaeus skin compounds. Compounds found in searches (erinacines, hericenones) derive from the unrelated hedgehog mushroom (Hericium erinaceus). The traditional use likely reflects sympathetic magic or empirical use of charred animal material as a non-specific hemostatic, neither of which constitutes a credible pharmacological mechanism.

Example 10: *Calcareous Spar* (Source) and *Cough* (Malady)

DeepRoot Verdict: UNSUPPORTED

Weakest Link: The substance is an inert mineral with no pharmacological activity relevant to cough.

Reasoning: As a mineral composed of calcium carbonate, Calcareous Spar possesses no antitussive properties or mechanisms to modulate the cough reflex. It does not interact with respiratory receptors or inflammatory pathways involved in cough. The lack of KG paths reflects the biological reality that this substance is inert for this indication.

Biomni Verdict: UNSUPPORTED

Weakest Link: CaCO3 has no antitussive, bronchodilatory, or anti-inflammatory mechanism in any database; PubMed returned zero results for 'calcium carbonate antitussive cough.'

Reasoning: Calcareous Spar is the same CaCO3 (CID 10112) as Example 4, with no antitussive indication in ChEMBL or OpenTargets. While calcium signaling is mechanistically relevant to cough — Ca2+-activated Cl-channels (TMEM16A) regulate airway smooth muscle tone and cough reflex sensitivity (PMID 39608849) — these are endogenous calcium-signaling mechanisms unrelated to exogenous CaCO3 supplementation. The TCM rationale of minerals 'descending rebellious qi' to suppress cough has no pharmacological correlate for this compound.