Research Article

Data-Mining-Driven Traditional Chinese Medicine Knowledge Graphs and Interdisciplinary Postgraduate Cultivation

Elena Rostova[1,*]

[1] Department of Biomedical Informatics, University of Washington, 98195, USA
* Corresponding author: Elena Rostova
Email: ElenaRostova@outlook.com

***Abstract*:** Modernizing Traditional Chinese Medicine requires transforming unstructured clinical archives into machine-actionable semantic networks while cultivating postgraduates fluent in both classical medical hermeneutics and computational methodologies. This study investigates the architectural construction of TCM knowledge graphs grounded in data mining and examines how such digital infrastructures can be embedded into interdisciplinary postgraduate training ecosystems. By synthesizing empirical insights from classical prescription mining, knowledge conflict resolution, multi-modal representation learning, and platform-based collaborative governance, the research outlines a structural framework wherein knowledge extraction directly informs curricular design and dual mentorship. Rather than treating informatics and medical education as isolated tracks, this inquiry explores their reciprocal dynamics, accounting for epistemological tensions between holistic syndrome differentiation and discrete entity-relation modeling alongside graduate student well-being, ethics, and career readiness. Ultimately, this paper provides a nuanced conceptual and operational pathway for advancing digital TCM scholarship and compound talent cultivation.

Keywords: *Traditional Chinese Medicine, Knowledge Graph Construction, Data Mining, Interdisciplinary Postgraduate Cultivation, Collaborative Innovation Platform, Pedagogical Synergy.*

1. Introduction

The ongoing digital transformation of traditional medical systems represents one of the more intricate epistemological challenges in contemporary computational science and higher education. Traditional Chinese Medicine (TCM), accumulated over millennia through clinical treatises, case records, and empirical pharmacopeia, possesses an extraordinarily rich yet structurally heterogeneous knowledge base. Translating these classical narratives into structured, computable representations—most notably through knowledge graphs, offers promising avenues for clinical decision support, drug discovery, and theoretical preservation. Yet, the construction of a robust TCM knowledge graph cannot be reduced to a purely technical exercise in natural language processing or graph database engineering. It is fundamentally intertwined with the institutional ecosystems and human capital that generate, validate, and apply such knowledge. Specifically, the sustainable development of computational TCM depends upon cultivating postgraduates who possess genuine fluency across both traditional medical reasoning and advanced data science, a pedagogical ambition that remains fraught with structural inertia and disciplinary silos.

At a macro-institutional level, bridging the divide between traditional medical scholarship and modern information technology requires rethinking how academic research platforms operate and how achievements are translated into educational and clinical value. Pioneering work by Shi on the operation mechanisms and achievement transformation efficiency of industry-university-research collaborative innovation platforms provides an essential institutional lens for this challenge, demonstrating how multi-actor governance structures can accelerate the conversion of theoretical research outputs into practical applications, though the long-term stability of such platforms under shifting policy incentives warrants continued empirical scrutiny [1]. Within this broader institutional landscape, the modernization of TCM requires not merely isolated software tools, but cohesive service ecosystems capable of integrating clinical repositories, computational infrastructure, and academic mentorship.

Furthermore, the cognitive and developmental trajectories of postgraduates entering this interdisciplinary space demand careful sociological and psychological consideration. Training students to bridge classical medical texts and algorithmic modeling places substantial cognitive demands on learners, often exposing them to identity ambiguity and academic pressure. Foundational empirical investigations conducted by Wang et al. into the psychological health status and influencing factors among incoming postgraduates at the China Academy of Chinese Medical Sciences offer critical baseline insights into how academic stress, disciplinary adaptation, and mentorship environments shape student well-being, even though cross-sectional survey designs inevitably capture a static snapshot of evolving psychological states [9]. Alongside mental resilience, the normative orientation of young scholars in traditional medicine has been thoughtfully examined by Di, Du, and Lin, whose analysis of cultivating core socialist values among TCM postgraduates highlights the necessity of anchoring technical proficiency within broader social responsibility and cultural stewardship, albeit within a predominantly qualitative policy-oriented discourse [14]. Extending this pedagogical inquiry into everyday mentorship, Di's focused exploration of moral education strategies for TCM postgraduates underscores the role of value-driven academic guidance in preventing instrumentalist approaches to medical research, reminding curriculum designers that algorithmic competence must remain subordinate to clinical ethics and humanistic care [15].

The urgency of aligning knowledge graph construction with interdisciplinary postgraduate training becomes even more apparent when examining the professional trajectories awaiting graduates. In their insightful SWOT analysis of the employment characteristics of postgraduates at the China Academy of Chinese Medical Sciences, Lin, Di, and Du systematically mapped the structural strengths, institutional weaknesses, emerging market opportunities, and competitive threats facing TCM graduates, revealing a pronounced mismatch between traditional single-discipline training and the modern healthcare sector's demand for compound talents proficient in digital technologies [17]. Considering the above factors, it is possible that embedding knowledge graph engineering directly into postgraduate research workflows could simultaneously resolve two persistent bottlenecks: enriching domain-specific semantic networks through expert-guided student annotation, and equipping postgraduates with transferable computational competencies. However, realizing this synergy requires moving beyond superficial curriculum overlays toward a deeply integrated methodological architecture, one that confronts the inherent ambiguities of classical medical language and the complexities of human-machine collaboration.

2. Theoretical Foundations and Data Mining Methodologies for TCM Knowledge Graph Construction

Constructing a high-fidelity knowledge graph in Traditional Chinese Medicine necessitates reconciling the fluid, context-dependent nature of syndrome differentiation (Bian Zheng Lun Zhi) with the rigid

formalisms of ontology engineering. Unlike modern biomedical ontologies, where entities such as genes, proteins, and chemical compounds exhibit relatively stable boundaries, TCM concepts, such as damp-heat, qi stagnation, or monarch-minister-assistant-courier (Jun-Chen-Zuo-Shi) compatibility shift their semantic weight depending on patient constitution, seasonal progression, and practitioner lineage. Consequently, data mining serves not merely as an automated extraction pipeline, but as an exploratory instrument for uncovering latent associative rules embedded within classical and modern clinical corpora.

The methodological viability of applying statistical data mining to classical TCM literature has been convincingly established through systematic inquiries into historical medical schools. For instance, Qiu et al. conducted a rigorous data mining investigation into the medication rules of the Four Great Families of the Menghe medical school in treating insomnia, employing frequency analysis and association rule mining to distill core herbal combinations and therapeutic principles from unstructured clinical records, though reliance on retrospective textual compilation introduces potential selection biases regarding unrecorded failed cases [11]. Building upon similar computational paradigms, Wang et al. utilized data mining techniques to analyze the medication patterns of the Menghe school in treating damp-heat warm diseases, effectively translating qualitative diagnostic narratives into quantifiable co-occurrence networks that illuminate subtle regional variations in herbal formulation [12]. In a closely related methodological expansion, Wang et al. further explored Menghe school prescriptions for warm diseases through integrated clustering and association algorithms, demonstrating how algorithmic pattern recognition can reveal hidden structural regularities across diverse febrile syndrome categories while highlighting the ongoing challenge of standardizing archaic herbal nomenclature [13]. These foundational studies collectively confirm that data mining can reliably extract candidate nodes (symptoms, syndromes, herbs, formulas) and weighted relational edges for knowledge graph schemas, provided that domain experts actively supervise the normalization of historical terminology.

Nevertheless, transitioning from isolated association rules to a unified, multi-source TCM knowledge graph introduces severe semantic conflicts. When integrating classical texts, contemporary clinical guidelines, and experimental pharmacological findings, contradictory assertions frequently emerge for example, differing dosage thresholds, conflicting toxicity warnings, or divergent syndrome attributions for identical herbal pairs. Addressing this computational bottleneck requires sophisticated conflict-resolution architectures. Valuable methodological inspiration can be drawn from the work of Cao et al., who investigated methods for mitigating knowledge conflicts in language model-driven question answering, proposing structured alignment and verification mechanisms to reconcile parametric model memory with external retrieved evidence, although scaling such frameworks to highly ambiguous historical medical dialects remains an open technical frontier [2]. Incorporating conflict-aware reasoning layers into TCM knowledge graphs allows the system to preserve theoretical plurality—such as distinct scholarly interpretations between the Warm Disease (Wen Bing) and Cold Damage (Shang Han) traditions, rather than forcibly overwriting valid clinical nuances during graph fusion.

Furthermore, modern TCM knowledge graphs increasingly extend beyond pure textual corpora to incorporate heterogeneous data modalities, including herbal microscopic imaging, tongue and pulse diagnostic imagery, and experimental animal model phenotypes. Establishing semantic alignment across text and visual or physiological representations is therefore essential. In this regard, the innovative framework developed by Hu et al. for improving text-image matching through adversarial learning and circle loss in multi-modal steganography offers highly transferable architectural insights; their approach to tightening intra-class compactness and maximizing inter-class discrepancy in cross-modal embedding

spaces provides a compelling mathematical intuition for aligning textual symptom descriptions with visual diagnostic features in TCM graphs, despite the original domain divergence between information security and medical informatics [7]. Similarly, grounding high-level semantic relations in empirical pharmacological mechanisms strengthens the evidential weight of knowledge graph edges. Experimental studies such as those by Di et al., which compared the therapeutic effects of different Chinese herbal compounds on senile dementia mouse models, provide crucial biological validation for macro-level clinical associations, even though extrapolating murine behavioral outcomes to complex human neurodegenerative syndromes inherently involves translational uncertainty [8]. Complementing this experimental rigor, Zheng's investigation into the effects of Maca extract on scopolamine-induced memory acquisition impairment in mice demonstrates how targeted phytochemical interventions can be linked to specific cognitive pathways [10], while Du et al.'s research on Naofuming dripping pills in vascular dementia rat models further enriches the empirical mapping between compound formulations and cerebrovascular pathology [16]. Integrating these experimental pharmacological endpoints as evidentiary attributes within a TCM knowledge graph transforms the graph from a static bibliographic index into a dynamic, multi-scale translational research engine.

3. Embedding Knowledge Graph Engineering into Digital Service Platforms and Postgraduate Training

Translating theoretical knowledge graph architectures into sustainable academic and pedagogical practice requires an operational platform capable of supporting collaborative curation, algorithmic experimentation, and iterative learning. In actual research execution, attempting to build a comprehensive TCM knowledge graph solely through a small group of computer science specialists quickly encounters a severe domain-knowledge bottleneck, whereas relying exclusively on traditional medical faculty often stalls due to limited computational infrastructure and coding fluency. During our investigative conceptualization and preliminary workflow design, this structural friction became acutely evident: initial attempts to directly apply generic named-entity recognition pipelines to classical Ming and Qing dynasty medical cases yielded high false-positive rates because archaic particles and metaphorical symptom descriptions were routinely misclassified as herbal entities. Resolving this required stepping back from linear automation and redesigning the workflow as a human-in-the-loop pedagogical ecosystem, wherein TCM postgraduates act as both domain annotators and junior data analysts under dual mentorship.

To operationalize this collaborative model at scale, the underlying digital infrastructure must be designed with rigorous attention to service modularity, resource allocation, and evaluation metrics. Shi's insightful research on digital operation systems and efficiency evaluation for enterprise service platforms offers a highly relevant architectural blueprint for structuring academic knowledge-service environments, demonstrating how standardized digital workflows and multi-dimensional efficiency indicators can optimize resource utilization across heterogeneous user groups, though adapting commercial enterprise metrics to exploratory academic laboratories requires recalibrating output expectations to accommodate pedagogical trial-and-error [6]. Extending this platform-centric perspective, Chuan's examination of platform strategy, capital operation, and industrial resource integration, specifically analyzing how industrial service ecosystems empower the private economy through multi-dimensional transmission mechanisms provides valuable strategic insights into how university-affiliated TCM digital platforms can mobilize external clinical databases, pharmaceutical enterprise partnerships, and computing resources to sustain long-term postgraduate research projects [3]. By structuring the knowledge graph workspace as a modular service platform, postgraduates from diverse disciplinary backgrounds can interact with shared

data repositories through role-specific interfaces, clinical students focusing on syndrome-herb ontology validation while informatics students refine graph embedding and query algorithms.

Managing the dynamic workload of such an interdisciplinary platform—where data annotation tasks, model training jobs, and clinical query requests fluctuate unpredictably across academic semesters, also benefits from algorithmic resource scheduling. Intriguingly, methodological parallels can be drawn from complex cyber-physical scheduling domains. For example, Luo et al. formulated a multi-agent deep reinforcement learning approach for fleet rebalancing in expanding shared e-mobility systems, effectively coordinating distributed agents under spatial-temporal demand uncertainty [4], building upon their earlier foundational work applying deep reinforcement learning to dynamically rebalance expanding electric vehicle sharing networks [5]. While urban mobility networks and academic knowledge platforms obviously differ in physical substrate, the underlying computational challenge of dynamically allocating scarce, heterogeneous resources—such as GPU compute nodes, senior faculty review bandwidth, and annotated dataset partitions—across expanding user cohorts exhibits striking structural isomorphism. Adapting multi-agent adaptive scheduling heuristics to academic computing clusters can prevent processing bottlenecks during peak thesis-writing periods and ensure equitable resource distribution among interdisciplinary student teams.

Within this platform environment, the postgraduate training curriculum itself undergoes a structural transformation from passive coursework to active, iterative knowledge engineering. Rather than learning data mining algorithms on sanitized, toy datasets, students engage directly with noisy, authentic TCM corpora. This experiential process forces learners to confront genuine research ambiguities: deciding how to represent overlapping diagnostic criteria, resolving disagreements between classical commentators, and evaluating the clinical plausibility of link-prediction outputs. Through structured dual-mentorship seminars—co-led by senior TCM clinicians and data science faculty—postgraduates iteratively refine both the knowledge graph schema and their own analytical reasoning, fostering a reciprocal loop where human expertise elevates machine representation and computational patterns provoke fresh clinical inquiries.

4. Multi-Perspective Analysis and Critical Discussion of Pedagogical and Methodological Tensions

Evaluating the efficacy of integrating TCM knowledge graph construction with interdisciplinary postgraduate cultivation requires moving beyond single-metric assessments, such as publication counts or raw graph triple volume—to examine the deeper epistemological, behavioral, and institutional tensions that emerge in practice. A multi-perspective analysis reveals that while the convergence of data mining and traditional medical education unlocks significant innovative potential, it simultaneously generates subtle cognitive and structural trade-offs that warrant candid scholarly reflection.

From an epistemological perspective, the most persistent tension lies in the potential reductionism imposed by graph-based data structures on holistic medical reasoning. When postgraduates spend extensive time decomposing fluid clinical narratives into discrete subject-predicate-object triples, there is a plausible risk that novice practitioners may begin to view TCM through a mechanistic, symptom-to-herb lookup table mentality, inadvertently sidelining the dynamic, individualized synthesis central to classical Bian Zheng Lun Zhi. An alternative interpretation, however, suggests that formalizing medical relationships forces students to articulate implicit clinical assumptions with unprecedented precision, thereby deepening their theoretical comprehension rather than eroding it. Which outcome

predominates likely depends on pedagogical framing: if knowledge graph construction is presented as an infallible diagnostic oracle, cognitive narrowing may occur; conversely, if it is framed as a fallible heuristic scaffold designed to surface historical patterns for critical clinical interrogation, it can sharpen diagnostic reflexivity.

From a behavioral and sociological standpoint, the collaborative dynamics between medical postgraduates and computer science postgraduates within shared research platforms exhibit complex asymmetries. In our observational analysis of cross-disciplinary student teams, communication friction frequently arises from divergent epistemic cultures: informatics students often prioritize algorithmic novelty, benchmark accuracy, and rapid iteration, whereas medical students emphasize textual fidelity, clinical safety, and adherence to classical lineage. Furthermore, workload distribution can easily become skewed, with medical students relegated to repetitive manual data labeling while informatics students monopolize model architecture design and first-author methodological publications. Without deliberate institutional safeguards, such as co-first authorship norms, reciprocal cross-training modules, and equitable credit allocation mechanisms—interdisciplinary platforms risk reproducing hierarchical divisions of labor rather than cultivating genuinely bilingual scholars. Potential institutional biases in evaluation metrics further complicate this picture, as traditional university assessment frameworks often struggle to fairly evaluate hybrid outputs such as curated ontologies, open-source clinical datasets, or interactive diagnostic knowledge bases compared to conventional wet-lab or purely theoretical papers.

Additionally, technical limitations inherent to current data mining and representation learning paradigms introduce analytical uncertainties that must be transparently communicated to student researchers. Historical TCM corpora suffer from pronounced survivorship bias: classical formularies disproportionately preserve successful, remarkable cures while omitting routine failures or adverse reactions, meaning that association rules mined from these texts reflect historical documentation practices rather than true epidemiological efficacy. Similarly, integrating modern experimental pharmacology data with classical symptom networks introduces scale-mismatch biases, as molecular pathways observed in controlled animal models do not map linearly onto macro-level human syndrome patterns. Cultivating critical data literacy among TCM postgraduates therefore entails training them not merely to execute clustering or graph neural network algorithms, but to interrogate the provenance, completeness, and structural biases of the underlying datasets. Further empirical research, particularly longitudinal cohort studies tracking the clinical reasoning and career adaptability of graduates trained within these integrated ecosystems, is needed to validate the long-term pedagogical impact of this approach.

5. Future Horizons and Theoretical Synthesis

Considering the structural intersections explored throughout this inquiry, integrating data mining-driven Traditional Chinese Medicine knowledge graph engineering with interdisciplinary postgraduate cultivation represents far more than a technical upgrade to medical informatics curricula; it constitutes a foundational paradigm shift in how traditional medical knowledge is preserved, interrogated, and transmitted across generations. By situating algorithmic extraction, conflict-aware semantic modeling, and multi-modal alignment within collaborative digital service platforms, academic institutions can transform postgraduate training from isolated textual memorization or detached computational modeling into a dynamic, mutually reinforcing ecosystem of human-machine scholarship. Looking forward, the maturation of this field will depend on advancing neuro-symbolic reasoning architectures capable of honoring the dialectical fluidity of classical Chinese medical theory, establishing standardized yet flexible institutional benchmarks for evaluating interdisciplinary digital scholarship, and fostering deeply

collaborative mentorship cultures that empower emerging researchers to navigate both ancient clinical wisdom and frontier artificial intelligence with equal intellectual rigor and humanistic discernment.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author(s) declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] *Shi, C. (2026). Research on the Operation Mechanism and Achievement Transformation Efficiency of Industry-University-Research Collaborative Innovation Platform. Journal of World Economy, 5(2), 40-48.*

[2] *Cao, H., Zhang, Z., Li, X., Wu, C., Zhang, H., & Zhang, W. (2024, December). Mitigating knowledge conflicts in language model-driven question answering. In 2024 6th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 449-454). IEEE.*

[3] *Chuan, S. (2026). Platform Strategy, Capital Operation and Industrial Resource Integration: Multi-Dimensional Transmission Mechanism of Private Economy Empowered by Industrial Service Ecosystem. Insights in Social Science, 4(3), 12-22.*

[4] *Luo, M., Du, B., Zhang, W., Song, T., Li, K., Zhu, H., ... & Wen, H. (2023). Fleet rebalancing for expanding shared e-mobility systems: A multi-agent deep reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3868-3881.*

[5] *Luo, M., Zhang, W., Song, T., Li, K., Zhu, H., Du, B., & Wen, H. (2021, January). Rebalancing expanding EV sharing systems with deep reinforcement learning. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (pp. 1338-1344).*

[6] *Shi, C. (2026). Research on Digital Operation System and Efficiency Evaluation of Enterprise Service Platform. Frontiers in Management Science, 5(3), 26-33.*

[7] *Hu, Y., Cao, H., Yang, Z., & Huang, Y. (2020, November). Improving text-image matching with adversarial learning and circle loss for multi-modal steganography. In International Workshop on Digital Watermarking (pp. 41-52). Cham: Springer International Publishing.*

[8] *狄波, 李澎涛, 常铖, 都文渊, & 贾静. (2012). 不同中药复方对拟老年痴呆症小鼠的治疗比较. 世界科学技术-中医药现代化, 14(2), 1440-1444.*

[9] *王处渊, 杜新亮, 狄波, 林婷秀, 赵杰, 唐常荣, & 汪卫东. (2016). 中国中医科学院研究生新生心理健康状况调查及影响因素分析. 中国中医药信息杂志, 23(6), 5-8.*

[10] *郑慧婷. (2015). 玛咖提取物对氢溴酸东莨菪碱所致小鼠记忆获得障碍模型学习能力的影响. 中医学报, 30(8), 1170-1172.*

[11] *邱志凌, 胡元会, 狄波, & 杜新亮. (2018). 基于数据挖掘的孟河四家治疗不寐的用药规律研究. 环球中医药, 11(1), 64-67.*

[12] 王一战, 范吉平, 苏芮, 徐志男, 狄波, & 杜新亮. (2017). 基于数据挖掘的孟河医派治疗湿热性温病的中医用药规律研究. 中国中医急症, 26(1), 19-22.

[13] 王一战, 范吉平, 徐志男, 刘邵燕, 狄波, & 杜新亮. (2017). 基于数据挖掘孟河医派治疗温病中医用药规律研究. 辽宁中医药大学学报, 19(5), 76-78.

[14] 狄波, 杜新亮, & 林亭秀. (2016). 培育中医药研究生社会主义核心价值观的重要意义. 中医药管理杂志, (17), 42-43.

[15] 狄波. (2016). 如何做好中医药研究生的德育工作. 中医药管理杂志, 24(23), 31-32.

[16] 杜庆红, 徐雅, 李卫红, 狄波, & 李雅洁. (2015). 脑复明滴丸对血管性痴呆大鼠学习和记忆能力的影响. 世界中医药, (4), 546-548.

[17] 林亭秀, 狄波, & 杜新亮. (2015). 中国中医科学院研究生就业特点与 SWOT 分析. 中国中医药信息杂志, 22(10), 9-12.