Graph of Thoughts Nanobiomaterials Assistants: Towards Logical, Tool-Augmented, and Multi-Agent Reasoning in Scientific LLMs

Xianghang Peng and David Scott Lewis
AIXC
research@aiexecutiveconsulting.com

Abstract

The integration of Large Language Models (LLMs) into materials science has accelerated literature review, hypothesis generation and testing, property prediction, and code generation. However, the safety-critical domain of nanobiomaterials—defined by complex multi-attribute constraints involving surface chemistry, cytotoxicity, and immune response—exposes the fundamental limitations of current autoregressive models. Standard approaches, such as Chain-of-Thought (CoT) prompting, frequently suffer from logical inconsistencies, hallucinated constraints, and a lack of formal verifiability. This **position paper** argues that the next generation of AI assistants for nanobiomaterials must transition from linear, probabilistic generation to structured, rigorous reasoning architectures. We propose the **Graph of Thought Nanobiomaterials Assistant (GoT-Nano)**, a framework that integrates Graph of Thoughts (GoT) reasoning, symbolic solver augmentation (SatLM), and multi-agent debate systems. By restructuring inference as an arbitrary graph, we enable non-linear exploration of design spaces, backtracking from toxic candidates, and the aggregation of multi-modal insights. Furthermore, we advocate for the integration of formal logical layers to enforce consistency (e.g., negation and transitivity) and the use of external symbolic tools for verifiable calculation. We contend that this shift from pattern matching to deliberate, verifiable reasoning is essential for the trustworthy deployment of AI in the high-stakes design of therapeutic nanostructures.

Introduction

The rapid proliferation of Large Language Models (LLMs) has fundamentally altered the landscape of scientific discovery. In materials science, models pretrained on vast corpora of scientific text have demonstrated surprising utility, from extracting synthesis recipes (Odobesku et al. 2025) to predicting material properties via text regression (Lei, Docherty, and Cooper 2024). However, as these models migrate from auxiliary tools for literature review to active participants in experimental design ("copilots"), a critical gap has emerged between their linguistic fluency and their logical reliability.

This gap is particularly acute in the field of nanobiomaterials. Unlike bulk materials, nanobiomaterials require the precise orchestration of physicochemical properties at the nanoscale to achieve specific biological outcomes. Designing a nanoparticle for drug delivery involves a combinatorial optimization problem: the core must be kinetically stable, the surface chemistry must prevent opsonization (immune recognition), and the hydrodynamic radius must fall within a narrow window to ensure cellular uptake. A single design flaw—such as a miscalculation in ligand density or an overlooked toxicity pathway—can render a therapeutic candidate dangerous.

Current LLM approaches, dominated by standard Input-Output (IO) or linear Chain-of-Thought (CoT) prompting, struggle with the complexity of these multi-objective optimization problems. They frequently suffer from *logical inconsistencies*, where a model might correctly identify a polymer as hydrophobic in one context but recommend it for a hydrophilic application in the next (Cheng et al. 2025). Furthermore, standard LLMs lack the intrinsic ability to perform the precise arithmetic or symbolic constraint satisfaction required for stoichiometry and kinetic modeling. They operate as "System 1" thinkers—fast and intuitive—but lack the "System 2" capabilities of deliberation and verification (Yao et al. 2023).

In this paper, we posit that the future of AI-driven nanobiomaterials design lies not in larger models, but in structured reasoning architectures. We argue for a transition from linear generation to *Graph of Thoughts* (GoT) processing (Besta et al. 2024), supported by external symbolic solvers (Ye et al. 2023) and verified by multi-agent debate systems (Zhou and Chen 2025). By mapping the design process to a graph structure, we allow the AI to explore multiple synthesis pathways simultaneously, backtrack when constraints are violated, and merge insights from disparate subthoughts.

Our contributions are threefold:

- We analyze the limitations of current reasoning paradigms (CoT, ToT) in the context of nanobiomaterials, highlighting failure modes regarding logical consistency and planning.
- We propose a conceptual architecture for a "Graph of Thoughts Nanobiomaterials Assistant" (GoT-Nano) that integrates solver-based reasoning and formal logical layers to ensure design validity.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We outline a research agenda for the community to bridge the gap between formal methods and scientific LLMs, emphasizing the need for domain-specific logical benchmarks.

The Landscape of Reasoning Paradigms

To understand the necessary evolution of scientific AI, we must first categorize the current and emerging reasoning paradigms available to LLMs. Recent literature has moved rapidly beyond simple prompt engineering toward architectures that mimic human metacognition and formal logic.

From Chains to Graphs: Structured Inference

The standard Chain-of-Thought (CoT) approach (Wei et al. 2022) improved LLM performance on math and logic tasks by inducing intermediate reasoning steps. However, CoT is linear and irreversible; once the model heads down an incorrect reasoning path (e.g., selecting a toxic precursor), it rarely self-corrects within the same generation. This "error cascading" is fatal in scientific workflows.

Tree of Thoughts (ToT). To address this, Yao et al. (2023) introduced Tree of Thoughts, which enables the model to explore multiple reasoning branches and backtrack when a branch is evaluated as unpromising. In materials design, this is analogous to considering three different synthesis methods (e.g., precipitation, sol-gel, hydrothermal), evaluating the theoretical yield and purity of each, and discarding the low-yield paths before generating the final protocol. Janshagen (2024) have further explored how ToT can be integrated with symbolic solvers, though they note the challenge of maintaining thought diversity.

Graph of Thoughts (GoT). Besta et al. (2024) generalized this further to arbitrary graphs. GoT allows for the *aggregation* of thoughts, where information from multiple distinct reasoning paths can be merged. This is crucial for nanobiomaterials, where a final design might require merging a core synthesis strategy from Path A with a surface functionalization strategy from Path B. The graph structure supports complex topologies like feedback loops and recursive refinement, mirroring the iterative nature of scientific research.

Solver-Augmented Reasoning

LLMs are probabilistic token predictors, not calculators. Consequently, they struggle with tasks requiring strict adherence to constraints or precise arithmetic.

Program-Aided and Symbolic Solvers. Paradigms such as Program-Aided Language Models (PAL) (Gao et al. 2023) and SatLM (Ye et al. 2023) decouple reasoning from computation. SatLM, for instance, uses the LLM to parse a natural language problem into a declarative logical specification, which is then solved by an external theorem prover or SAT solver. Pan et al. (2023) demonstrated similar gains by translating problems into symbolic logic (Logic-LM). In the context of nanobiomaterials, this allows the LLM to define the *constraints* of a design (e.g., "size < 50nm AND surface_charge > 0mV") while a solver ensures the proposed components mathematically and physically satisfy these conditions.

Logical Consistency and Formal Layers

A pervasive issue in LLMs is logical inconsistency. Cheng et al. (2025) categorize these failures into negation consistency (answering "yes" to P and "yes" to $\neg P$), implication consistency, and transitivity consistency. For a scientific assistant, such inconsistencies are unacceptable. If a model asserts that "Silver nanoparticles are antimicrobial" and "Agent X is a silver nanoparticle," it must logically conclude that "Agent X is antimicrobial." Failure to do so represents a breakdown in reasoning that undermines trust.

Frameworks like *Maieutic Prompting* (Jung et al. 2022) and *BeliefBank* (Kassner, Oyama, and Schütze 2021) attempt to maintain a consistent graph of beliefs, updating connected nodes when a new fact is introduced. We argue that nanobiomaterials assistants require a "Formal Layer"—a module dedicated solely to checking the logical satisfiability of the generated scientific claims against a knowledge base of axioms (e.g., chemistry rules) (Yang et al. 2024).

Multi-Agent Systems and Debate

Finally, single-model hallucinations can be mitigated through multi-agent architectures. Li et al. (2024) review how systems of agents can decompose complex tasks. Approaches like *Adaptive Heterogeneous Multi-Agent Debate* (A-HMAD) (Zhou and Chen 2025) utilize agents with different personas or specializations to critique each other's reasoning. In a materials context, a "Toxicology Agent" could debate a "Synthesis Agent," forcing the system to resolve conflicts between efficacy and safety before presenting a solution to the human user. This aligns with the *Self-Refine* (Madaan et al. 2023) paradigm, where models iteratively improve outputs based on feedback.

The Nanobiomaterials Challenge

Nanobiomaterials science presents a unique set of challenges that make it an ideal testbed for these advanced reasoning paradigms.

Complexity and Constraints

Designing a nanobiomaterial is a constraint satisfaction problem with high dimensionality. Parameters include:

- **Core Material:** Gold, Iron Oxide, Silica, Liposomes, etc.
- **Size and Shape:** Affects circulation time and cellular uptake.
- **Surface Chemistry:** Ligands, polymers (PEG), antibodies.
- **Biological Interaction:** Protein corona formation, immune clearance, toxicity.

These parameters are interdependent. Increasing size might increase drug loading (positive) but accelerate clearance by the reticuloendothelial system (negative). An LLM must navigate this trade-off space intelligently.

The Status Quo: Extraction vs. Reasoning

Current state-of-the-art systems like *nanoMINER* (Odobesku et al. 2025) and *LLaMat* (Mishra et al. 2024) focus heavily on information extraction. They excel at mining values (e.g., zeta potential, diameter) from unstructured text to build databases. While essential, extraction is retrospective; it organizes what is already known. The gap lies in *prospective design*. When an LLM is asked, "Design a nanoparticle to cross the blood-brain barrier with minimal liver accumulation," it typically retrieves a generic prototype from its training data. It rarely performs first-principles reasoning to deduce that a specific combination of a transferrin-targeting ligand and a specific zwitterionic coating might solve the problem, unless that exact combination appears in the training set. Zaki et al. (2024) highlight that while LLMs contain vast materials knowledge, their ability to apply it to novel logical deduction remains limited.

Proposed Architecture: GoT-Nano

We propose a modular architecture for a Nanobiomaterials Design Assistant that moves beyond simple prompting to a structured, neuro-symbolic workflow.

System Overview

The system is composed of a central **Reasoning Controller** (an LLM) that orchestrates a dynamic **Graph of Thoughts**. This controller interacts with three specialized modules: the **Tool Interface**, the **Formal Verifier**, and the **Multi-Agent Critic**.

1. The Graph of Thoughts (GoT) Engine Instead of generating a linear protocol, the Controller initializes a graph where the root node is the User Query (e.g., "Design a DOX-loaded carrier").

- **Expansion:** The Controller generates multiple initial nodes representing different material classes (e.g., Node A: Liposome, Node B: Polymeric Micelle, Node C: Gold Nanosphere).
- **Transformation:** Each node undergoes transformation steps. For Node A, edges might branch to different lipid ratios or hydration methods.
- **Aggregation:** Crucially, the system can merge nodes. If Node B has excellent stability and Node A has high loading, the GoT engine can generate a Node D (Lipid-Polymer Hybrid) that attempts to combine these attributes. This aggregation is a unique strength of graph-based reasoning over tree-based search.

2. Tool-Augmented Verification Material design involves hard constraints. A "Thought" in the graph containing a proposed formulation is passed to the Tool Interface.

- **Stoichiometry Check:** A symbolic Python calculator validates that the molar ratios are physically possible and balanced (Chen et al. 2022).
- **Property Prediction:** Calls to ML models (e.g., Random Forests trained on MatSciBERT embeddings (Lei, Docherty, and Cooper 2024)) predict the zeta potential or toxicity of the structure described in the node.
- **SatLM Integration:** Complex constraints are parsed into boolean logic (e.g., `Assert(Toxicity < Threshold AND Stability > 24h)`) and checked by a SAT solver (Ye et al. 2023). Nodes violating these constraints are pruned immediately, preventing the model from pursuing physically impossible designs.

3. The Logic and Consistency Layer To prevent the "hallucinated properties" problem, a Formal Verifier module checks logical consistency. Using the taxonomy from Cheng et al. (2025), this module ensures:

- **Factuality Consistency:** The properties attributed to a material in a Thought node must match retrieved knowledge base facts (e.g., if the model claims PEGylation increases circulation time, this is verified against an axiomatic knowledge base).
- **Transitivity:** If the system deduces $A \to B$ and $B \to C$, it must uphold $A \to C$ in the final report.

4. Multi-Agent Refinement Before a final path in the graph is selected, a Multi-Agent Debate is triggered (Zhou and Chen 2025).

- **Agent 1 (The Toxicologist):** Scrutinizes the design solely for safety risks and biocompatibility.
- **Agent 2 (The Engineer):** Scrutinizes for synthesis scalability and cost.
- **Agent 3 (The Clinician):** Scrutinizes for therapeutic efficacy and targeting.

Using methods similar to *Reflexion* (Shinn et al. 2023), these agents critique the leading graph nodes. The Controller aggregates this feedback to refine the design or spawn new, improved nodes.

Workflow Example

Consider a request to design a contrast agent for MRI.

1. **GoT Initialization:** The model proposes Gadolinium-chelates and Iron Oxide nanoparticles (SPIONs).
2. **Evaluation:** The Tool Interface (Solvers) flags free Gadolinium as toxic (nephrogenic systemic fibrosis risk).
3. **Backtracking/Branching:** The graph prunes free Gadolinium paths. It branches the SPION path into "Dextran coated" and "PEG coated."
4. **Aggregation:** The system identifies a separate reasoning chain discussing tumor targeting via folic acid. It creates a new node: "Folic-Acid conjugated PEG-SPION."
5. **Debate:** The Toxicologist agent argues that high iron load might cause oxidative stress. The Logic Layer checks if the concentration proposed violates known safety thresholds using SAT solvers.
6. **Output:** The system presents the optimized design with a trace of the verification steps.

Discussion: The Formal Turn in AI for Materials

The proposed architecture represents a shift towards what has been termed "Formal Mathematical Reasoning" in AI (Yang et al. 2024). By grounding LLM generations in formal systems—whether they are SAT solvers for logic or simulators for physics—we address the trustworthiness crisis in generative AI.

Implications for Discovery

This approach moves AI from a retrieval engine to a reasoning engine. It allows for the discovery of counter-intuitive designs that satisfy logical constraints but do not appear explicitly in the training data (out-of-distribution generalization). Qu et al. (2024) noted that language representations are powerful for exploration; adding graph-based topology allows that exploration to be systematic rather than random.

Challenges and Limitations

Implementing this architecture faces hurdles. **Latency:** Graph search and multi-agent debate are computationally expensive compared to single-turn inference. **Formalization Gap:** Translating vague biological constraints ("biocompatible") into precise symbolic logic for a SAT solver is non-trivial (Ye et al. 2023). We need better "auto-formalization" models that can bridge natural language science and formal logic. **Data Scarcity:** While papers exist, structured logical datasets for materials science (like MaScQA (Zaki et al. 2024)) are rare compared to general math benchmarks.

Research Agenda

To realize GoT-Nano, we outline a 3-5 year research agenda for the community.

1. Reasoning-Aware Benchmarks

Current benchmarks measure factual recall or simple entity extraction. We need *reasoning* benchmarks for materials science, akin to *ProofWriter* or *FOLIO* but for chemical logic (Tafjord, Dalvi, and Clark 2021; Han et al. 2024). These benchmarks should present design problems where success depends on multi-step deduction and constraint satisfaction, not just memory.

2. Formalizing Domain Constraints

We must build libraries of formal constraints for nanobiomaterials. Just as software verification uses formal specifications, material design needs formal rules (e.g., "Zeta potential < -30mV implies stability" as a probabilistic rule). Integrating these with LLMs via Logic-LM approaches (Pan et al. 2023) is a key frontier.

3. Tool-LLM Interfaces

Improving the reliability of tool use is essential. *Toolformer* (Schick et al. 2023) showed the way, but scientific tools are complex. We need LLMs fine-tuned to write input files for standard simulation packages (e.g., LAMMPS, Gaussian) and robustly parse their outputs (Bran et al. 2023).

Conclusion

The intersection of Nanobiomaterials and Large Language Models is ripe for transformation. However, reliance on the statistical probabilities of next-token prediction is insufficient for the engineering of life-saving technologies. We strongly advocate for the adoption of *Graph of Thoughts* architectures augmented by symbolic verification and multi-agent critique. Only by imposing logical rigor on our generative models can we trust them to design the medicines of the future. This roadmap aligns directly with the AAAI call for systems that marry the flexibility of neural networks with the reliability of formal reasoning.

Acknowledgments

This work was conceptualized to align with the AAAI 2026 Bridge Program themes. We acknowledge the foundational insights provided by the authors cited herein, whose work on reasoning paradigms and materials informatics made this synthesis possible.

References

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *arXiv preprint arXiv:2308.09687*.

Bran, A. M.; Cox, S.; White, A. D.; and Schwaller, P. 2023. Chemcrow: Augmenting large-language models with chemistry tools. arXiv:2304.05376.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *arXiv preprint arXiv:2211.12588*.

Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 10400–10408.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2023. PAL: Program-aided Language Models. In *International Conference on Machine Learning*, 10764–10799.

Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; Benson, L.; Sun, L.; Zuberi, E.; Qiao, Y.; Burtell, M.; et al. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. *arXiv preprint arXiv:2209.00840*.

Janshagen, A. 2024. Reasoning with Large Language Models Using Tree-of-Thought and Symbolic Solvers. *arXiv preprint arXiv:2402.09123*.

Jung, J.; Qin, L.; Welleck, S.; Brahman, F.; Bhagavatula, C.; Le Bras, R.; and Choi, Y. 2022. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1266–1279.

Kassner, N.; Oyama, O.; and Schütze, H. 2021. BeliefBank: Adding Memory to a Pre-Trained Language Model

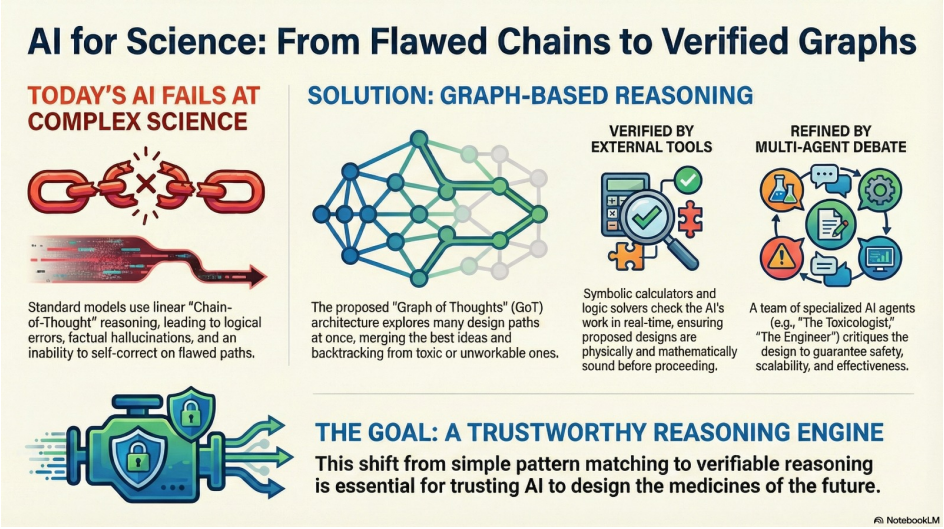

Figure 1: Comparison of linear chain-of-thought failures in current AI versus a proposed graph-based reasoning architecture for trustworthy scientific design.

for a Systematic Notion of Belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8849–8861.

Lei, G.; Docherty, R.; and Cooper, S. J. 2024. Materials science in the era of large language models: a perspective. *Digital Discovery*, 3: 1257.

Li, Z.; Zhang, A.; Wei, C.; Shi, Y.; Kim, D.; Liu, L.; Liu, M.; Su, Y.; and Liu, Y. 2024. Multi-Agent Systems Meet Large Language Models: Architectures, Synergies, and Future Directions. *arXiv preprint arXiv:2402.01680*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint arXiv:2303.17651*.

Mishra, V.; Singh, S.; Zaki, M.; Ahlawat, D.; Grover, H.; Mishra, B.; Miret, S.; Mausam; and Krishnan, N. M. A. 2024. LLaMat: Large Language Models for Materials Science Information Extraction. In *NeurIPS 2024 Workshop on AI for Accelerated Materials Design (AI4Mat)*.

Odobesku, R.; Romanova, K.; Mirzaeva, S.; Zagorulko, O.; Sim, R.; Khakimullin, R.; Razlivina, J.; et al. 2025. Agent-based multimodal information extraction for nanomaterials. *npj Computational Materials*, 11(1): 194.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. Y. 2023. Logic-LM: Empowering LLMs with Symbolic Solvers for Faithful Logical Reasoning. *arXiv preprint arXiv:2305.12295*.

Qu, J.; Xie, Y. R.; Ciesielski, K. M.; Porter, C. E.; Toberer, E. S.; and Ertekin, E. 2024. Leveraging language representation for materials exploration and discovery. *npj Computational Materials*, 10: 58.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint arXiv:2303.11366*.

Tafjord, O.; Dalvi, B.; and Clark, P. 2021. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3621–3634.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Yang, K.; Poesia, G.; He, J.; Li, W.; Lauter, K.; Chaudhuri, S.; and Song, D. 2024. Formal Mathematical Reasoning: A New Frontier in AI. *arXiv preprint arXiv:2412.16075*.

Yao, S.; Griffiths, T. L.; Yu, D.; Zhao, J.; Cao, Y.; Shafran, I.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, 36.

Ye, X.; Chen, Q.; Dillig, I.; and Durrett, G. 2023. SatLM: Satisfiability-Aided Language Models Using Declarative Prompting. *Advances in Neural Information Processing Systems*, 36.

Zaki, M.; Jayadeva; Mausam; and Krishnan, N. M. A. 2024. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2): 313–327.

Zhou, Y.; and Chen, Y. 2025. Adaptive heterogeneous multi-agent debate for enhanced educational and factual reasoning in large language models. *Journal of King Saud University – Computer and Information Sciences*, 37: 330.