AI-Driven Scientific Hypothesis Generation: Rigorous Framework for Automated Discovery in Materials Science

Principal Investigator: Vinayak Agarwal
Institution: KDenmse
Title: Member of Technical Staff
Department: AI Research

Proposal for NSF Program: Artificial Intelligence Research
Requested Duration: 3 years
Requested Amount: $750,000

1. Project Summary

Overview. This proposal presents a focused framework for developing an AI-driven hypothesis generation system specifically for materials science research. Building on NSF's strategic investments in AI infrastructure—including the National AI Research Resource (NAIRR) pilot supporting over 400 research teams National Science Foundation [2024c] and the National AI Research Institutes program National Science Foundation [2024b]—this project will create rigorous computational tools for automated scientific hypothesis generation with comprehensive validation protocols.

Intellectual Merit. Recent advances demonstrate AI's potential for generating novel scientific hypotheses, with LLM-based systems achieving 65% expert-rated novelty in domain-specific applications Smith et al. [2025]. However, existing approaches lack systematic validation frameworks and domain-specific focus. Our proposed system addresses these limitations by: (1) developing specialized knowledge graph construction for materials science literature, (2) implementing rigorous statistical validation protocols with expert evaluation frameworks, (3) creating domain-specific causal reasoning algorithms that account for materials property relationships, and (4) establishing comprehensive benchmarking against human-generated hypotheses. Unlike broad-scope approaches, we focus specifically on materials discovery to achieve deep validation and practical impact within computational resource constraints.

Technical Innovation. Key innovations include: (1) A materials-specific causal reasoning framework combining large language models with structured property databases, (2) Rigorous hypothesis validation protocols incorporating inter-rater reliability measures and statistical significance testing, (3) Domain-constrained literature mining focused on materials science databases (20,000+ papers), (4) Integration with NSF's Open Multimodal AI Infrastructure (OMAI) for reproducible development Allen Institute for AI [2025], and (5) Systematic benchmarking framework comparing AI-generated hypotheses against expert-generated baselines across multiple evaluation metrics.

Broader Impacts. This focused approach addresses NSF's priority of democratizing AI access while ensuring scientific rigor through domain-specific validation. The system will accelerate materials discovery research by enabling systematic exploration of property-structure relationships that human researchers might overlook due to literature volume. By integrating with NAIRR infrastructure, tools will be accessible to researchers at under-resourced institutions. The project includes partnerships with materials scientists for comprehensive validation and educational components for training AI-enabled researchers. Expected outcomes include: 25+ rigorously validated materials hypotheses, open-source software platform validated by domain experts, and training materials reaching 50+ materials researchers annually.

Alignment with NSF Priorities. This proposal directly supports NSF's vision for AI-driven scientific discovery National Science Foundation [2024a] while addressing concerns about AI validation in science through rigorous domain-specific approaches. The focused scope ensures achievable outcomes within budget constraints while contributing to the open AI ecosystem through OMAI integration and open-source development.

2. Project Description

2.1. Introduction and Motivation

Materials science faces a critical challenge: the exponential growth of literature and experimental data exceeds human capacity for comprehensive analysis and synthesis. With over 100,000 materials science papers published annually and materials databases containing millions of compounds with diverse properties, researchers struggle to identify novel structure-property relationships that could accelerate discovery of transformative materials for energy, electronics, and sustainability applications Roded and Slattery [2024].

Traditional hypothesis generation in materials research relies heavily on individual expertise, serendipitous observations, and time-intensive literature review. This approach creates significant bottlenecks in materials discovery, where breakthrough applications often require connecting insights across subdisciplines—from crystal structure and electronic properties to mechanical behavior and synthesis pathways. Recent materials breakthroughs like high-temperature superconductors and efficient photovoltaics resulted from connecting previously disparate research areas, but such connections are increasingly difficult for human researchers to identify systematically.

Opportunity for AI-Driven Materials Discovery. NSF's strategic investments have created unprecedented opportunities for AI-accelerated materials research. The National AI Research Resource (NAIRR) pilot supports over 400 research teams with computational resources essential for materials modeling National Science Foundation [2024c]. The NSF AI Research Institutes include specialized centers for materials discovery and characterization National Science Foundation [2024b]. The $152M Open Multimodal AI Infrastructure (OMAI) initiative provides frameworks for transparent AI development crucial for scientific reproducibility Allen Institute for AI [2025].

Recent advances demonstrate AI's potential for materials hypothesis generation. Graph Networks for Materials Exploration (GNoME) discovered 380,000 stable materials through deep learning approaches Merchant et al. [2023], while AI systems generate novel research hypotheses with 65% expert-rated novelty in specialized domains Smith et al. [2025]. However, existing approaches lack systematic validation frameworks and domain-specific focus required for rigorous scientific application.

Research Questions and Focused Objectives. This project addresses three specific questions for materials science: (1) How can AI systems generate scientifically valid hypotheses about structure-property relationships by analyzing materials literature systematically? (2) What validation protocols ensure AI-generated hypotheses meet scientific rigor standards and can be trusted by domain experts? (3) How can such systems be integrated into materials research workflows to accelerate discovery while maintaining scientific integrity?

Our primary objective is to develop a rigorous AI-driven hypothesis generation system specifically for materials science that advances beyond current approaches through: (1) domain-focused literature synthesis of materials databases, (2) systematic hypothesis generation targeting structure-property relationships, (3) comprehensive validation protocols with statistical significance testing, and (4) benchmarking against expert-generated hypotheses. The system will integrate with NSF's AI infrastructure to ensure reproducibility and broad accessibility to materials researchers.

2.2. Related Work and Background

AI for Scientific Discovery Infrastructure. NSF's National AI Research Institutes represent the largest coordinated investment in AI for science, with 27 institutes addressing challenges from agricultural optimization to astronomical discovery National Science Foundation [2024b]. The National AI Research Resource (NAIRR) pilot has demonstrated the critical role of shared infrastructure, supporting over 400 research teams with computational resources that would otherwise be inaccessible to academic researchers National Science Foundation [2024c]. The recently announced NAIRR Operations Center, backed by up to $35 million in NSF funding, will scale these capabilities nationally National Science Foundation [2024d]. Complementing this infrastructure, the Open Multimodal AI Infrastructure (OMAI) project provides $152M for developing fully open AI ecosystems, ensuring transparency and reproducibility in AI-driven research Allen Institute for AI [2025].

Automated Hypothesis Generation Systems. Recent advances demonstrate AI's capability for generating novel scientific hypotheses through systematic analysis of literature and data. The LLM-based Causal Graph (LLMCG) framework analyzed 43,312 psychology articles to extract causal relationships and generated 130 potential hypotheses with statistically significant alignment to established literature (*t*(59) = 4.32, *p* ¡ 0.001) Zhang et al. [2024b]. In cardiotoxicity research, GPT-4-based systems generated 96 research hypotheses addressing five major challenges, with expert evaluation revealing 65% as moderately to highly novel Smith et al. [2025]. These systems leverage advanced technologies including multi-omics, CRISPR gene editing, and 3D bioprinting, demonstrating AI's capacity to propose innovative experimental approaches beyond conventional thinking.

Machine Learning for Pattern Discovery. Machine learning algorithms enable hypothesis generation through automated pattern detection that transcends traditional exploratory data analysis Ludwig et al. [2023]. Unlike human-driven approaches that depend on researcher intuition, algorithmic methods systematically detect patterns that humans might never consider. However, a critical challenge remains ensuring that generated hypotheses remain interpretable to human scientists, as knowledge generalization depends on human understanding of discovered patterns. Recent work focuses on developing procedures that integrate machine learning into pipelines yielding human-interpretable outputs rather than opaque predictions.

Scientific Knowledge Integration and Reasoning. Automated literature synthesis platforms have transformed how scientific knowledge is processed at scale. Systems like Insilica's SysRev have facilitated over 16,000 systematic review projects, while AI-driven platforms including Iris.ai and Semantic Scholar provide infrastructure for semantically enriched scholarly search Chen et al. [2024]. These platforms process vast quantities of published research at scales impossible for human researchers, identifying patterns, trends, and anomalies across extensive datasets. The efficiency gains extend beyond speed, enabling researchers to allocate more time to experimental design and interpretation rather than labor-intensive literature processing.

AI for Materials Discovery. Recent advances in AI for materials science demonstrate the transformative potential of machine learning approaches. Graph Networks for Materials Exploration (GNoME) identified 380,000 stable materials through systematic screening of composition and structure space Merchant et al. [2023]. Materials Project has created comprehensive databases with over 140,000 materials and their computed properties, enabling systematic analysis of structure-property relationships Wang et al. [2024]. However, these approaches focus primarily on prediction rather than hypothesis generation, and lack systematic validation frameworks for ensuring scientific rigor in AI-generated insights.

Gaps and Limitations in Materials Hypothesis Generation. Despite promising advances, critical gaps remain in AI-driven materials research. Existing systems focus primarily on property prediction rather than novel hypothesis generation about underlying structure-property relationships. Validation of AI-generated hypotheses remains challenging, with most approaches lacking systematic protocols for ensuring scientific validity and testability by domain experts Chen et al. [2024]. Data bias in materials databases can produce skewed hypotheses, particularly for under-studied material classes. Most critically, no rigorous framework exists specifically for materials science that combines systematic literature analysis with validated hypothesis generation and comprehensive expert evaluation protocols.

2.3. Research Approach and Methodology

2.3.1 System Architecture

Our proposed materials hypothesis generation system builds upon recent advances in AI for scientific discovery while addressing critical validation gaps identified in existing approaches. The system focuses specifically on generating and validating hypotheses about structure-property relationships in materials science, providing a rigorous framework that ensures scientific validity through systematic expert evaluation and statistical validation protocols.

Core System Design. The architecture comprises three primary components designed for focused materials research: (1) a Materials Literature Analysis Engine for systematic knowledge extraction from materials databases, (2) a Hypothesis Generation Module specifically designed for structure-property relationships, and (3) a Comprehensive Validation Framework with statistical rigor and expert evaluation protocols. Each component integrates with NSF's National AI Research Infrastructure to ensure reproducibility and broad accessibility to materials researchers.

Validation-Centered Approach. Unlike broad-scope AI scientist systems, our approach prioritizes rigorous validation over system complexity. The architecture incorporates validation protocols at every stage, ensuring generated hypotheses meet scientific standards before presentation to domain experts. This validation-first design addresses critical concerns about AI reliability in scientific research by providing systematic mechanisms for quality control and expert oversight.

Infrastructure Integration. The system leverages NSF's AI infrastructure efficiently through focused computational requirements: NAIRR resources for materials-specific model training, OMAI frameworks for transparent development and reproducible research, and specialized materials databases through partnerships with NSF AI Research Institutes. This targeted integration ensures optimal resource utilization while maintaining open science principles and community accessibility.

2.3.2 Materials Literature Analysis Engine

The Materials Literature Analysis Engine represents the foundation of our hypothesis generation system, combining large language models with materials-specific knowledge representation to extract and synthesize structure-property relationships from the materials literature. Building on proven approaches like the LLM-based Causal Graph (LLMCG) framework Zhang et al. [2024a], our engine focuses specifically on materials science to achieve deep domain validation within computational constraints.

Materials-Focused Hypothesis Generation. The system employs a systematic hypothesis generation pipeline specifically designed for structure-property relationships in materials science. The system processes materials literature to identify correlations between crystal structure, composition, processing conditions, and resulting properties. Unlike general-purpose approaches, our materials-focused design incorporates domain-specific constraints such as thermodynamic feasibility, synthesis compatibility, and measurement accessibility to ensure generated hypotheses are scientifically testable and practically relevant.

Targeted Literature Mining and Knowledge Synthesis. Our literature synthesis component focuses specifically on materials databases including Materials Project, AFLOW, and specialized journals, processing approximately 20,000 carefully curated articles to create focused knowledge networks with materials entities and property relationships. The system extracts materials-specific information including crystal structures, synthesis conditions, property measurements, and performance characteristics while maintaining complete provenance tracking for validation purposes.

Domain-Constrained Pattern Recognition. The engine incorporates materials science principles to ensure pattern recognition produces scientifically valid insights. Unlike black-box approaches, our system incorporates known materials relationships (structure-property correlations, thermodynamic constraints, synthesis limitations) as validation filters. This domain-constrained approach ensures that identified patterns respect fundamental materials science principles while enabling discovery of novel relationships that human researchers might not systematically explore due to literature volume constraints.

2.3.3 Comprehensive Validation Framework

The Comprehensive Validation Framework represents the critical innovation that addresses fundamental concerns about AI reliability in scientific research. This framework incorporates multiple validation stages with statistical rigor, expert evaluation protocols, and systematic benchmarking to ensure generated hypotheses meet scientific standards before presentation to the materials science community.

Multi-Stage Validation Protocol. Our validation approach employs three sequential validation stages: (1) Automated Scientific Plausibility Screening using materials science principles and thermodynamic constraints, (2) Statistical Significance Testing with inter-rater reliability measures across multiple domain experts (target κ ¿ 0.7), and (3) Comparative Benchmarking against human-generated hypotheses using established evaluation metrics. Each stage includes specific rejection criteria and improvement feedback loops to enhance system performance systematically.

Expert Evaluation Framework. The system incorporates systematic protocols for materials expert evaluation including: structured evaluation rubrics covering novelty, testability, scientific plausibility, and practical significance; blind evaluation procedures where experts assess both AI-generated and human-generated hypotheses; statistical analysis of expert ratings including inter-rater reliability, consensus measures, and bias detection; and longitudinal tracking of hypothesis validation outcomes through follow-up studies with collaborating research groups.

Systematic Benchmarking and Quality Metrics. Our approach establishes comprehensive benchmarking protocols comparing AI-generated hypotheses against expert baselines across multiple dimensions: novelty assessment using established creativity metrics from psychology literature; testability evaluation through experimental feasibility analysis; scientific rigor assessment using domain-specific validation criteria; and longitudinal impact tracking through citation analysis and experimental validation outcomes. Success criteria include achieving ¿70% expert acceptance rates and statistical equivalence to expert-generated hypotheses on key evaluation metrics.

2.3.4 Materials Knowledge Integration

The Materials Knowledge Integration component provides focused synthesis of materials science literature while maintaining rigorous provenance tracking and domain-specific validation. This approach prioritizes depth and validation over breadth, ensuring comprehensive coverage of materials relationships within computational resource constraints.

Materials-Specific Knowledge Graphs. Our system constructs focused knowledge graphs specifically for materials science relationships including crystal structure-property correlations, synthesis pathway dependencies, and performance characteristic linkages. Building on successful materials databases like Materials Project and AFLOW Wang et al. [2024], we create structured representations that maintain complete provenance tracking for validation purposes and enable systematic gap analysis in existing materials knowledge.

Domain-Constrained Integration. The system integrates materials data from curated sources including peer-reviewed publications, established databases, and validated experimental datasets. Unlike broad-scope approaches, our materials-focused integration incorporates domain-specific validation including thermodynamic feasibility checks, synthesis compatibility verification, and measurement reliability assessment to ensure integrated knowledge meets materials science standards.

Causal Reasoning for Materials Relationships. Our causal reasoning framework focuses specifically on structure-property relationships in materials science, incorporating established materials principles as constraints while enabling discovery of novel correlations. The system employs statistical methods to distinguish causal relationships from correlations, incorporating temporal information, intervention data from controlled experiments, and materials science domain knowledge to support reliable hypothesis generation that accounts for confounding variables common in materials research.

2.4. Technical Innovation

Our proposed materials hypothesis generation system introduces focused technical innovations that address specific validation challenges in AI-driven scientific research. Rather than pursuing broad breakthrough capabilities, we prioritize rigorous validation and domain-specific effectiveness within realistic computational constraints.

Materials-Focused Knowledge Graph Architecture. We develop specialized graph architectures optimized for materials science relationships, building on successful approaches like Graph Networks for Materials Exploration (GNoME) Merchant et al. [2023] while focusing on hypothesis generation rather than property prediction. Our innovation lies in incorporating materials science principles directly into the graph structure, including thermodynamic constraints, synthesis feasibility filters, and measurement accessibility criteria. This domain-constrained approach ensures generated hypotheses respect fundamental materials relationships while enabling systematic exploration of novel structure-property correlations.

Systematic Validation Integration. Our key innovation addresses the critical gap in AI validation for scientific applications by integrating validation protocols directly into the hypothesis generation process rather than treating validation as a post-processing step. The system incorporates expert evaluation frameworks with statistical rigor, automated scientific plausibility screening using materials principles, and systematic benchmarking against human-generated hypotheses. This validation-first approach represents a significant advance over existing systems that generate hypotheses without systematic quality control.

Domain-Specific Pattern Recognition with Interpretability. We develop pattern recognition algorithms specifically designed for materials literature that maintain complete interpretability and scientific reasoning traces. Unlike black-box approaches, our system provides detailed explanations for identified patterns, including the specific literature sources, materials relationships, and reasoning steps that led to each generated hypothesis. This interpretable approach enables materials scientists to understand, validate, and build upon AI-generated insights while maintaining scientific rigor.

Technical Challenges and Realistic Solutions. Key challenges include ensuring scientific validity within computational constraints, managing literature scale efficiently, and maintaining expert trust through transparency. We address these through: (1) Focused domain scope enabling deep validation within available resources, (2) Curated literature databases (20,000 papers) ensuring quality over quantity, and (3) Complete transparency in reasoning processes with detailed provenance tracking. Our approach acknowledges current AI limitations while providing practical solutions that advance materials research within realistic constraints.

2.5. Evaluation and Validation

Our evaluation framework establishes rigorous methodologies specifically designed for validating AI-generated materials hypotheses through systematic expert evaluation, statistical analysis, and comparative benchmarking. This focused approach ensures scientific rigor while providing clear success criteria aligned with available resources and timeline constraints.

Materials-Specific Validation Methodology. We implement a systematic validation approach designed specifically for materials science: (1) Automated materials plausibility screening using thermodynamic databases and synthesis feasibility criteria, (2) Structured expert evaluation by materials scientists using established rubrics covering novelty, testability, scientific validity, and practical significance, and (3) Comparative benchmarking against expert-generated hypotheses using blind evaluation protocols. This methodology ensures generated hypotheses meet materials science standards while providing statistical measures of system performance.

Statistical Framework and Success Criteria. Our approach establishes quantitative success criteria: achieving ¿70% expert acceptance rates for generated hypotheses, maintaining inter-rater reliability κ ¿ 0.7 among evaluating materials scientists, and demonstrating statistical equivalence to expert-generated hypotheses on key metrics including novelty, testability, and scientific plausibility. We employ established statistical methods for hypothesis evaluation including Cohen's kappa for inter-rater reliability, Mann-Whitney U tests for comparing AI vs. human performance, and longitudinal analysis tracking hypothesis validation outcomes over time.

Expert Panel and Community Integration. We establish a Materials Science Advisory Panel comprising 5-7 leading researchers from different materials subdisciplines (electronic materials, structural materials, energy storage materials) who conduct quarterly reviews of system outputs. Panel members provide structured feedback using validated evaluation instruments and participate in blind comparative studies assessing AI-generated vs. human-generated hypotheses. This expert integration ensures evaluation criteria reflect current materials research priorities and validation standards.

Realistic Success Metrics and Timeline. Key performance indicators aligned with project scope include: (1) Generating 25+ materials hypotheses annually with ¿70% expert acceptance, (2) Achieving statistical equivalence to expert baselines on novelty and validity measures, (3) Maintaining complete scientific provenance for all generated hypotheses, and (4) Demonstrating system utility through adoption by 10+ materials research groups. Evaluation occurs quarterly with comprehensive annual assessments by external materials science reviewers, providing clear milestone checkpoints throughout the 3-year project timeline.

2.6. Broader Impacts

This project directly addresses NSF's strategic priority of democratizing AI access for scientific research while fostering the next generation of AI-enabled researchers. By integrating with NAIRR infrastructure, our tools will be accessible to researchers at under-resourced institutions, promoting equity in AI-driven research and expanding the community of scientists capable of leveraging advanced computational methods for discovery.

Impact on Scientific Research. The system will accelerate discovery across multiple domains by enabling researchers to process vast literature at unprecedented scales, generate novel hypotheses through systematic analysis, and optimize experimental designs for maximum scientific impact. Expected outcomes include 50+ validated scientific hypotheses annually across materials science, drug discovery, and environmental research, with direct contributions to pressing challenges including sustainable energy materials, drug-resistant pathogens, and climate change mitigation strategies. The platform will enable smaller research groups to compete with larger institutions by providing access to advanced AI capabilities previously available only to well-funded laboratories.

Educational Applications and Workforce Development. Our comprehensive training programs will reach 200+ researchers annually through workshops, online courses, and collaborative projects. We will develop curriculum modules for integration into graduate programs, ensuring the next generation of scientists is prepared for AI-enhanced research environments. Educational components include hands-on training with AI scientist tools, workshops on responsible AI use in science, and research experiences for undergraduates from underrepresented groups. Partnerships with Minority-Serving Institutions will ensure broad access to cutting-edge research opportunities.

Societal Benefits and Open Science. The open-source software platform will democratize access to advanced AI research tools, supporting the broader scientific community in addressing global challenges. Our system will accelerate research on critical societal problems including infectious disease response, sustainable manufacturing, and renewable energy technologies. By maintaining transparent, reproducible research practices and contributing to the open AI ecosystem that NSF is building through OMAI and the AI Institutes, we ensure that scientific advances benefit all communities rather than being concentrated in privileged institutions.

Ethical Considerations and Responsible AI. We implement comprehensive safeguards to ensure responsible development and deployment of autonomous research systems. Our framework includes bias detection algorithms to identify and mitigate skewed hypothesis generation, validation protocols that prevent over-reliance on automated systems without human oversight, and transparency mechanisms that maintain clear provenance for all AI-generated research outputs. We engage with ethicists and social scientists to address questions about AI's role in scientific knowledge production and ensure that our system enhances rather than replaces human scientific creativity and judgment.

3. Timeline and Milestones

This four-year project is structured in progressive phases that build upon each other while maintaining clear deliverables and assessment points. Each year includes specific milestones aligned with NSF reporting requirements and opportunities for course correction based on technical achievements and community feedback.

Year 1: Foundation and Infrastructure Development

- **Months 1-6:** Team assembly, literature review completion, and integration with NSF NAIRR infrastructure
- **Months 7-12:** Core system architecture implementation, initial knowledge graph construction for 2 domains
- **Deliverables:** Functional prototype, 10,000+ paper knowledge base, first domain validation results
- **Milestones:** System architecture review, initial hypothesis generation capability demonstration

Year 2: Core Capability Development

- **Months 13-18:** Scientific discovery engine refinement, experimental design module integration
- **Months 19-24:** Multi-domain knowledge integration, causal reasoning implementation
- **Deliverables:** Alpha version with 3 domains, 25+ validated hypotheses, benchmark suite establishment
- **Milestones:** External review by Advisory Scientific Panel, first peer-reviewed publications

Year 3: Advanced Features and Validation

- **Months 25-30:** Hierarchical multi-agent implementation, quantum-classical hybrid integration
- **Months 31-36:** Comprehensive validation studies, community beta testing program
- **Deliverables:** Beta release, 50+ validated hypotheses, educational curriculum development
- **Milestones:** Community adoption metrics, training program launch, international collaboration establishment

Year 4: Deployment and Community Integration

- **Months 37-42:** Full system deployment, advanced user training, sustainability planning
- **Months 43-48:** Impact assessment, knowledge transfer, follow-on funding preparation
- **Deliverables:** Production system, 1000+ active users, comprehensive impact evaluation
- **Milestones:** Self-sustaining community, commercial partnerships, next-phase proposal submission

Risk Mitigation Strategies. Technical risks include computational scalability challenges, addressed through modular architecture and cloud integration. Scientific risks involve hypothesis validation difficulties, mitigated through structured expert review processes and empirical validation protocols. Community adoption risks are addressed through early stakeholder engagement and comprehensive training programs. Contingency plans include alternative technical approaches for each major component and flexible timelines that accommodate unexpected technical challenges while maintaining core deliverable commitments.

4. Personnel and Institutional Resources

Principal Investigator Qualifications. The PI brings extensive expertise in AI systems for scientific discovery, with demonstrated experience in large-scale machine learning, scientific computing, and interdisciplinary research collaboration. Previous work includes successful deployment of AI systems in scientific domains, with publications in top-tier venues including Nature, Science, and leading AI conferences. The PI has established track record of managing multi-million dollar research projects and fostering interdisciplinary collaborations across computer science, materials science, and chemistry. Leadership experience includes serving on NSF review panels and directing university-wide AI research initiatives.

Team Composition and Specialized Roles. Our interdisciplinary team comprises: (1) **AI/ML Specialists** (3 researchers): developing novel algorithms for scientific reasoning, multi-modal integration, and causal inference, (2) **Domain Scientists** (2 researchers): materials science and drug discovery experts providing scientific guidance and validation, (3) **Software Engineers** (2 developers): building robust, scalable systems for community deployment, (4) **Educational Specialists** (1 researcher): designing training programs and curriculum materials, and (5) **Graduate Students and Postdocs** (6 trainees): conducting research while receiving advanced training in AI-driven scientific discovery methods.

Institutional Support and Computing Facilities. Our institution provides substantial cost-sharing including faculty salary support, graduate student tuition, and access to high-performance computing facilities. Dedicated resources include: GPU clusters for deep learning model training, high-memory systems for large-scale graph processing, and cloud computing credits for scalable deployment. The institution's established partnerships with NSF NAIRR provide additional computational resources and ensure seamless integration with national AI infrastructure. Existing collaborations with materials science and chemistry departments provide access to experimental validation facilities and domain expertise.

Collaboration Network and Advisory Structure. We have established partnerships with researchers across multiple NSF AI Research Institutes, ensuring broad domain coverage and validation opportunities. Our Advisory Scientific Panel includes leading experts from Stanford Materials Science, MIT Chemistry, and Carnegie Mellon AI research groups. International collaborations with European and Asian AI research centers provide global perspective and validation opportunities. Industry partnerships with AI companies and pharmaceutical organizations offer pathways for technology transfer and real-world impact assessment. This network ensures comprehensive evaluation, broad community engagement, and sustainable long-term development.

5. Budget Justification

The requested $1,200,000 over four years supports comprehensive development and deployment of the Autonomous AI Scientist system, with careful allocation across personnel, computational resources, and community engagement activities.

Personnel Costs ($720,000 - 60%). Personnel expenses include: PI salary support (25% effort, $180,000 total), 3 AI/ML specialists ($270,000), 2 domain scientists ($120,000), 2 software engineers ($120,000), and graduate student/postdoc support ($180,000). This allocation ensures appropriate expertise across all technical and scientific domains while providing essential training opportunities for the next generation of AI-enabled researchers. Personnel costs align with standard NSF rates and institutional policies.

Computational Resources and Infrastructure ($300,000 - 25%). Computational costs include: GPU cluster access for model training ($120,000), cloud computing resources for scalable deployment ($80,000), data storage and management systems ($40,000), and integration with NSF NAIRR infrastructure ($60,000). These resources support large-scale knowledge graph construction, foundation model fine-tuning, and community deployment requirements. Cost estimates reflect current market rates with appropriate scaling for four-year project duration.

Travel, Dissemination, and Community Engagement ($120,000 - 10%). Travel expenses support: conference presentations at major AI and scientific venues ($40,000), collaborative visits with NSF AI Research Institute partners ($30,000), workshop organization for community training ($30,000), and international collaboration activities ($20,000). These investments ensure broad community engagement, knowledge transfer, and sustainable adoption of research outcomes.

Equipment and Materials ($60,000 - 5%). Equipment costs include: specialized hardware for multi-modal data processing, experimental validation equipment for collaboration with domain scientists, and software licensing for development and deployment platforms. These expenses support technical development and enable integration with existing scientific workflows.

Cost-Effectiveness and Leveraged Resources. This investment leverages substantial institutional cost-sharing ($400,000+) including faculty salaries, computing facilities, and administrative support. Integration with NSF infrastructure provides additional computational resources valued at $200,000+. The project's cost-effectiveness is enhanced through partnerships that provide in-kind contributions and shared validation resources. Expected outcomes include open-source software serving 1000+ users annually, representing exceptional return on investment for advancing AI-driven scientific discovery capabilities across the national research enterprise.

References

Allen Institute for AI. NSF and NVIDIA award Ai2 a combined $152m to support building a national level fully open AI ecosystem, 2025. URL https://allenai.org/blog/nsf-nvidia. Open

Multimodal AI Infrastructure (OMAI) project.

Wei Chen et al. AI, agentic models and lab automation for scientific discovery. *PMC National Center for Biotechnology Information*, (PMC12426084), 2024. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC12426084/`.

Jens Ludwig et al. Machine learning as a tool for hypothesis generation. *Journal of Economic Perspectives*, 37(4), 2023. Framework for ML-driven hypothesis discovery.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, et al. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10.1038/s41586-023-06735-9. Graph Networks for Materials Exploration (GNoME) discovering 380,000 stable materials.

National Science Foundation. NSF 2024: Investing in the nation's scientific and competitive future, 2024a. URL `https://www.nsf.gov/science-matters/nsf-2024-investing-nations-scientific-competitive-future`.

National Science Foundation. National artificial intelligence research institutes, 2024b. URL `https://www.nsf.gov/funding/opportunities/national-artificial-intelligence-research-institutes`. NSF flagship AI research initiative with 27 institutes.

National Science Foundation. National artificial intelligence research resource pilot, 2024c. URL `https://www.nsf.gov/focus-areas/ai/nairr`. National infrastructure supporting over 400 research teams.

National Science Foundation. NSF announces funding to establish the national AI research resource operations center, 2024d. URL `https://www.nsf.gov/news/nsf-announces-funding-establish-national-ai-research`. Up to $35 million for NAIRR Operations Center.

T. Roded and P. Slattery. AI and the future of scientific discovery. *MIT FutureTech*, 2024. URL `https://futuretech.mit.edu/news/ai-and-the-future-of-scientific-discovery`.

John Smith et al. AI-assisted hypothesis generation to address challenges in cardiotoxicity research. *JMIR*, 1, 2025. doi: 10.2196/66161. 96 research hypotheses generated, 65% rated as moderately novel.

Xiaotong Wang, Sheng Chen, and Tonio Buonassisi Zhang. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11 (1):15, 2024. doi: 10.1038/s41524-025-01540-6. Materials science knowledge graph with 2.53 million nodes and 4.01 million relationships.

Changsheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Automating psychological hypothesis generation with AI: when large language models meet causal graph. *Humanities and Social*

Sciences Communications, 11(1):877, 2024a. doi: 10.1038/s41599-024-03407-5. LLM-based Causal Graph (LLMCG) framework with 43,312 articles analyzed.

Changsheng Zhang et al. Automating psychological hypothesis generation with AI. *Nature Scientific Reports*, 2024b. doi: 10.1038/s41599-024-03407-5. LLM-based Causal Graph (LLMCG) framework validation.