Supplementary Materials

From large language models to AI agents in energy materials research: enabling discovery, design, and automation

Tongao Yao[1,2], Junming Huang[1,2], Yujie Yan[1,2], Yang Yang[3], Ziye Wang[1,2], Xuqiang Shao[3], Zhengyang Gao[1,2], Weijie Yang[1,2]

[1]Department of Power Engineering, North China Electric Power University, Baoding 071003, Hebei, China.

[2]Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding 071003, Hebei, China.

[3]Department of Computer Science, North China Electric Power University, Baoding 071003, Hebei, China.

Correspondence to: Prof. Weijie Yang, Department of Power Engineering, North China Electric Power University, Baoding 071003, Hebei, China. E-mail: yangwj@ncepu.edu.cn

Supplementary Tables

Supplementary Table 1. Exemplary AI Agent Systems and their Key Contributions to Autonomous Discovery

System Name	Primary Contribution	Key Technology/Approach	Application Domain	Reference
ChemCrow	Autonomous planning and execution of chemical synthesis using a comprehensive toolset.	LLM (GPT-4), Tool integration, Safety checks	Organic Chemistry	[15]
SciToolAgent	Intelligent planning of complex, multi-step computational toolchains via a knowledge graph.	Scientific Tool Knowledge Graph (SciToolKG), Graph reasoning	Computational Science	[57]
Perovskite-R1	Domain-specific hypothesis generation for new materials after fine-tuning on literature.	Fine-tuning, Instruction datasets	Perovskite Solar Cells	[1]
aLLoyM	Prediction of alloy phase diagrams from fine-tuning on	Fine-tuning, Large-scale computational data	Alloy Design	[63]

computational
(CALPHAD)
data.

Supplementary Table 2. Extended Quantitative Comparison of AI Systems and Baselines Across the Materials Discovery Pipeline.

Pipeline Stage	Task	Representative AI System	Key Metric(s)	Reported Performance / Outcome	Representative Baseline(s)
Literature Mining	Structured data extraction from legacy papers	LMExt [24]	Extraction accuracy (per field); OCR robustness	84.2 % on modern texts; 43.8 % on legacy PDFs	Manual curation by experts
	Table / figure comprehension	ChartQA / DIVE[29, 30]	Figure parsing F1 score / chart QA accuracy	≈ 78 – 82 % (domain-dependent)	Traditional OCR + template methods
Design & Prediction	Band-gap prediction from text	LLM-Prop [23]	MAE (eV) for band gap	0.23 eV	CGCNN (0.29); ALIGNN (0.25)
	Formation-energy prediction (graph)	ALIGNN / MEGNet [10, 11]	MAE (eV / atom)	ALIGNN 0.026; MEGNet 0.028	DFT reference calculations
	Synthesizability prediction	ChemCrow[15]	Retrosynthesis success rate (%)	~76 % routes successfully planned	Rule-based synthesis planners
Simulation & Automation	Code generation for DFT / MD scripts	LLaMP, ChemGraph [49, 50]	Script success / execution accuracy	≈ 85 – 90 % valid execution	Manual template generation
	Tool use / task orchestration	SciToolAgent[57]	Tool-use accuracy; multi-tool success rate	~88 % average tool call success	Single-tool scripts / manual workflow
Optimization / Automation	Closed-loop electrolyte optimization	Robson[21]	Sample efficiency; cycle time	Novel electrolytes in ≈ 4 weeks	Bayesian Optimization (BO); human DoE

	Autonomous cathode discovery	ChatBattery [66]	Candidate generation / filtering rate	~100 candidates screened in 2 months	Manual trial-and-error (years)
Agentic Systems	Hypothesis generation and refinement	SciAgents[1]	Long-horizon task success; diversity metrics	Generated novel multi-step hypotheses (qualitative)	Single LLM prompt / RAG baseline
	Error recovery and multi-agent debate	ChatGPT multi-agent prototype [N/R]	Recovery rate / debate consensus accuracy	Not reported (N/R)	Single-agent prompting

Notes

1. *N/R = Not Reported* in the cited literature.
2. Values are drawn from peer-reviewed or officially documented sources.
3. Baselines correspond to established materials-informatics methods (e.g., CGCNN/ALIGNN, BO).
4. Reported numbers are indicative of relative performance; experimental setups vary.

Supplementary Table 3. Representative Benchmarks and Reproducibility Resources for AI-Driven Materials Discovery

Category	Example Benchmark / Dataset	Task Type	Evaluation Metric	Reference
Literature Mining	MatScholar NER corpus	Materials NER (materials, phase, property, method)	F1, entity accuracy	[2]
	MatSciBERT downstream suite	NER / relation classification / abstract classification (materials domain)	F1 / accuracy	[3]
	SciREX (for scientific IE; cross-domain)	Document-level IE / relation extraction	Macro/Micro-F1	[2]
Property Prediction	Matbench v0.1 (Materials Project tasks)	Band gap, formation energy, elastic moduli, etc.	MAE, RMSE; leaderboard	[4]
	Matbench-Discovery	Discovery-centric evaluation with in-/out-of-distribution and prospective settings	Stability hit-rate, precision@k, AUROC (task-dependent)	[4]
	JARVIS / AFLOW (curated tasks)	Formation energy, band gap, dielectric/elastic properties	MAE, RMSE; OOD splits (as defined per study)	[5]
Planning & Reasoning	ScienceAgentBench	Multi-step tool use & reasoning across 100+ science tasks	Task success, tool-call accuracy, reasoning faithfulness	[1]
	DiscoveryBench	Multi-step data-driven discovery tasks	End-to-end task success,	[6]

		(hypothesis→ evidence→ decision)	evidence grounding	
	PaRoutes (retrosynthesis)	Multi-step route planning on curated patent routes	Route quality/diversity, success rate, steps	[7]
	USPTO-50k (one-step)	Single-step retrosynthesis	Top-n accuracy	[8]
Automation / Closed-Loop Optimization	Olympus	Benchmarking experiment-planning under noise; mixed-parameter, multi-objective	Regret, sample efficiency, hypervolume (MO)	[9]
	ScienceAgentBench / DiscoveryBench (automation subsets)	Tool-enabled execution sub-tasks; success under constraints	Completion rate, tool reliability	[1]

Reference

[1] Chen, Z.; Chen, S.; Ning, Y.; Zhang, Q.; Wang, B.; Yu, B.; Li, Y.; Liao, Z.; Wei, C.; Lu, Z.; Dey, V.; Xue, M.; Baker, F. N.; Burns, B.; Adu-Ampratwum, D.; Huang, X.; Ning, X.; Gao, S.; Su, Y.; Sun, H. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *arXiv* **2024**, arXiv:2410.05080. Available online: https://arxiv.org/abs/2410.05080 (accessed 07 October 2024).

[2] Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J. Chem. Inf. Model.* **2019**, *59*, 3692-3702. DOI: 10.1021/acs.jcim.9b00470

[3] Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8*, 102. DOI: 10.1038/s41524-022-00784-w

[4] Riebesell, J.; Goodall, R. E. A.; Benner, P.; Chiang, Y.; Deng, B.; Ceder, G.; Asta, M.; Lee, A. A.; Jain, A.; Persson, K. A. Matbench Discovery -- A framework to evaluate machine learning crystal stability predictions. *arXiv* **2023**, arXiv:2308.14920. Available online: https://arxiv.org/abs/2308.14920v3 (accessed 01 August 2023).

[5] Omee, S. S.; Fu, N.; Dong, R.; Hu, M.; Hu, J. Structure-based out-of-distribution (OOD) materials property prediction: a benchmark study. *npj Comput. Mater.* **2024**, *10*, 144. DOI: 10.1038/s41524-024-01316-4

[6] Prasad Majumder, B.; Surana, H.; Agarwal, D.; Dalvi Mishra, B.; Meena, A.; Prakhar, A.; Vora, T.; Khot, T.; Sabharwal, A.; Clark, P. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. *arXiv* **2024**, arXiv:2407.01725. Available online: https://arxiv.org/abs/2407.01725 (accessed 01 July 2024).

[7] Genheden, S.; Bjerrum, E. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. *ChemRxiv* **2022**, *1*, 527-539. DOI: 10.1039/D2DD00015F

[8] Wei, Y.; Shan, L.; Qiu, T.; Lu, D.; Liu, Z. Machine learning-assisted retrosynthesis planning: Current status and future prospects. *Chinese Journal of Chemical Engineering* **2025**, *77*, 273-292. DOI: https://doi.org/10.1016/j.cjche.2024.10.014

[9] Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. Olympus: a benchmarking framework for noisy optimization and experiment planning. *arXiv* **2020**, arXiv:2010.04153. Available online: https://arxiv.org/abs/2010.04153 (accessed 08 October 2020).