A Universal Autonomous Agent for Atomistic Simulation and Benchmarking Its Capabilities

Fengxu Yang [1] **Jack D. Evans** [1]

[1] *School of Physics, Chemistry and Earth Sciences, Adelaide University, South Australia 5000, Australia*. Correspondence to: Jack D. Evans j.evans@adelaide.edu.au.

1. Introduction

The rapid advancement of agentic capabilities in large language models (LLMs) presents a compelling opportunity to transform computational chemistry for materials science research [1]. As a discipline fundamentally driven by complex computer-mediated workflows, computational chemistry tasks are naturally suited for integration with LLM-based control and can serve as a testing ground for approaches that manage self-driving labs [2, 3]. However, despite this natural alignment, most computational chemistry agentic systems remain constrained by rigid tool-calling paradigms [4, 5] and overly fine-grained agent-role decompositions [6, 7, 8], which force models to operate within narrowly predefined functions and significantly limit agent autonomy and system-level flexibility. These design choices inherently limit the capacity of LLMs to navigate complex scenarios and engage in self-correction, effectively reducing capable reasoning engines to mere instruction followers. As a result, these systems struggle with tasks outside the curated demonstrations and require extensive manual intervention to expand capabilities.

We introduce QUASAR, a universal atomistic computation system designed to overcome these limitations through robust and open-source architecture that optimally orchestrates LLM reasoning, delivering enhanced usability and flexibility for truly autonomous scientific research.

2. System Architecture and Methods

QUASAR employs a three-agent architecture (Figure 1): the Strategist interprets user-defined research objectives and decomposes them into scientifically grounded subtasks; the Operator executes these subtasks by interfacing with atomistic simulation software, the file system, and external resources, handling input preparation, job execution, and analysis; and the Evaluator assesses task completion, returning unsatisfactory results to the Operator for autonomous refinement. The system also supports plan-level iteration to further improve overall outcomes.

The included software stack bridges multiple levels of theory, integrating Quantum ESPRESSO [9] for ab initio density functional theory (DFT) calculations, MACE [10] for machine-learned potentials (MLP), LAMMPS [11] for classical molecular dynamics (MD), and RASPA3 [12] for grand canonical Monte Carlo simulations, unified through ASE [13] and pymatgen [14] for structure manipulation and analysis. This multiscale integration enables QUASAR to autonomously chain simulations across length and time scales. For example, the system can parameterize force fields from quantum calculations, then deploy them for large-scale dynamics or adsorption studies that would be intractable at the DFT level.

QUASAR addresses key limitations of other LLM-based agentic systems through intelligent context and memory management. The Evaluator acts as a checkpoint mechanism, summarizing key actions and results to distill each task to its essential information. This approach keeps the context window manageable and reduces API costs. Task-level and complete run summaries are stored locally, enabling long-term memory that supports iterative improvement across runs and provides an accessible entry point for human review. A double-pass planning mechanism ensures high-quality execution: initial plans from the Strategist undergo a second-pass review to identify missing elements before execution. Planning is further regulated by two user-adjustable parameters: Granularity (controls how finely tasks are decomposed and therefore how frequently evaluation and compression occur) and Accuracy (balances computational effort against accuracy). These features allow QUASAR to dynamically adapt its planning and memory architecture to diverse research objectives.

The system is designed for robust operation in real-world high-performance computing (HPC) environments where interruptions from API exhaustion, wall-time limits, or unexpected shutdowns are common. QUASAR maintains checkpoints that enable seamless state recovery and continuation, while simulation restart awareness allows the Operator to resume interrupted calculations. A time-monitoring mechanism tracks whether simulations are actively progressing, enabling early termination when anomalies or stalls are detected. Simulation tool knowledge is accessed through a hybrid retrieval system that combines semantic retrieval-augmented generation (RAG) over documentation with direct access to example input files, allowing the system to reason over sparse, non-annotated simulation inputs that embedding-only approaches handle poorly. Furthermore, the system is delivered as a containerized solution, ensuring high portability across diverse computing environments.

3. Results

We evaluated QUASAR using a three-tiered benchmark suite (Appendix A) with the `gemini-3-flash-preview` LLM. The three tiers progressively challenge the system from routine calculations to novel research scenarios, assessing

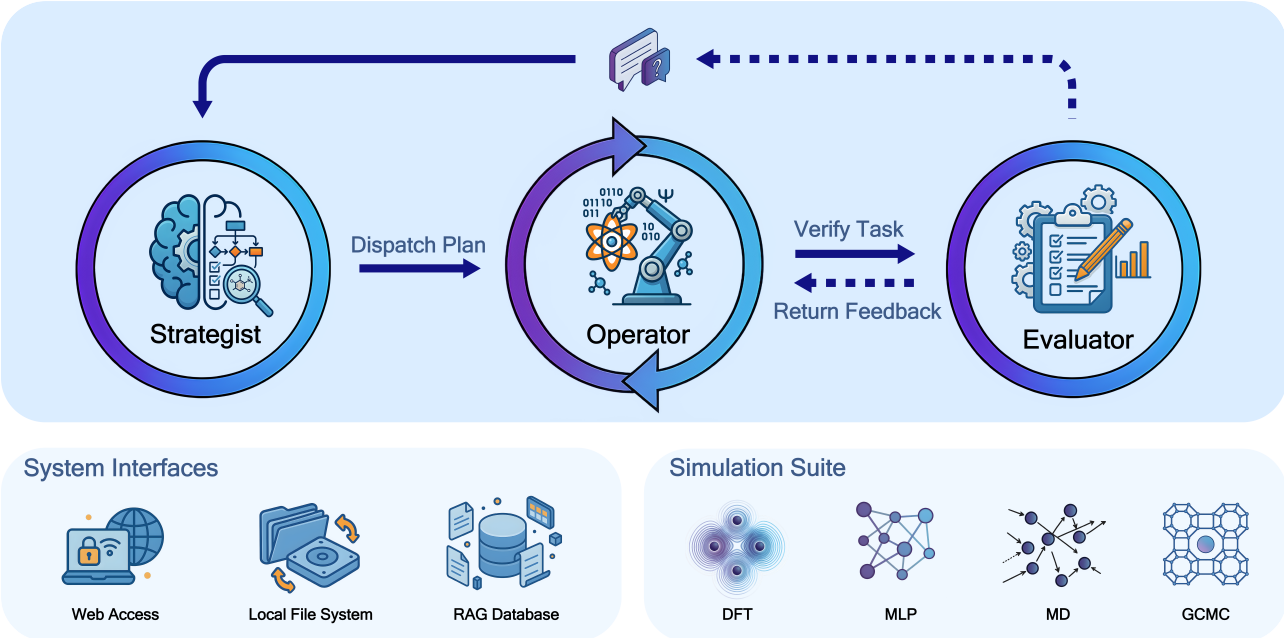

Fig. 1: Overview of the QUASAR architecture, where dashed lines represent optional feedback paths for iterative refinement.

not only domain knowledge but also planning capability and robustness under realistic conditions. We propose this benchmark suite as a standardized evaluation framework for autonomous computational chemistry systems, enabling systematic comparison of different LLM backends and agent architectures while providing reproducible metrics for tracking progress in the field.

3.1 Tier I. Task Execution

To verify baseline usability, we selected three representative tasks spanning distinct methodological archetypes: DFT k-point convergence, which ensures reliable Brillouin-zone sampling for electronic structure simulations; *NPT* ensemble equilibration in molecular dynamics, which is essential for obtaining thermodynamically consistent densities before production simulations; and helium void-fraction analysis, used to characterize accessible porosity in materials for gas storage and separation applications.

3.2 Tier II. Workflow Orchestration

Tier II evaluates the ability of QUASAR to decompose high-level scientific questions into coherent, multi-step computational workflows. Tasks include calculating the electronic band gap of bulk nickel oxide, generating CO_2 adsorption isotherms for UiO-66 at 298 K, and determining the melting point of aluminum. Each of these tests required the system to identify and sequence prerequisite calculations autonomously. For example, the band gap calculation requires structure optimization, appropriate DFT+U setup for strongly correlated electrons, and post-processing of electronic structure outputs, while the adsorption isotherm demands force field selection, framework preparation, and systematic pressure-sweep simulations.

3.3 Tier III. Frontier Research

While Tier I and II tasks are well-documented in the literature and likely represented in LLM training data, Tier III targets research problems that lack established protocols. We tasked QUASAR with screening 5% La-doped $ATaO_3$ perovskites (A = Li, Na, K) for photocatalytic methyl orange degradation under UV irradiation, and identifying which covalent-organic framework (COF) maximizes Xe/Kr adsorption selectivity at 298 K. Finally, a new material generated by latent diffusion [15] was provided and simulations of its mechanical properties and CO_2 adsorption requested. These problems require comparative screening across multiple candidate materials, property predictions that combine electronic structure and thermodynamic sampling, and synthesis of results into actionable scientific conclusions.

Across all tiers, QUASAR powered by a fast, low-cost LLM (`gemini-3-flash-preview`) demonstrated robust performance, successfully completing each benchmark with scientifically valid results. The Tier I tasks were completed with minimal iteration, validating both domain knowledge and software integration. Tier II showcased effective workflow decomposition, with the Strategist correctly identifying prerequisite calculations (e.g., structure optimization before property calculation) and the Evaluator catching intermediate errors for correction. Importantly, Tier III results demonstrate the ability for QUASAR to handle complex, multi-step research campaigns requiring long-term planning and execution over extended computational experiments.

Acknowledgments

J.D.E is the recipient of an Australian Research Council Discovery Early Career Award (project number DE220100163) funded by the Australian Government. The Phoenix HPC service at the Adelaide University is acknowledged for providing high-performance computing resources. This research was supported by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS), with access to computational resources provided by Pawsey Supercomputing Research Centre through the National Computational Merit Allocation Scheme.

References

[1] Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, Shuangjia Zheng, Ning Ding, Wanli Ouyang, Nanqing Dong, Yu Cheng, Siqi Sun, Lei Bai, and Bowen Zhou. From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery, October 2025. arXiv:2508.14111.

[2] Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. Self-Driving Laboratories for Chemistry and Materials Science. *Chemical Reviews*, 124(16):9633–9732, August 2024.

[3] Theo Jaffrelot Inizan, Sherry Yang, Aaron Kaplan, Yen-hsu Lin, Jian Yin, Saber Mirzaei, Mona Abdelgaid, Ali H. Alawadhi, KwangHwan Cho, Zhiling Zheng, Ekin Dogus Cubuk, Christian Borgs, Jennifer T. Chayes, Kristin A. Persson, and Omar M. Yaghi. System of Agentic AI for the Discovery of Metal-Organic Frameworks, April 2025. arXiv:2504.14110.

[4] Orlando A. Mendible-Barreto, Misael Díaz-Maldonado, Fernando J. Carmona Esteva, J. Emmanuel Torres, Ubaldo M. Córdova-Figueroa, and Yamil J. Colón. DynaMate: leveraging AI-agents for customized research workflows. *Molecular Systems Design & Engineering*, 10(7):585–598, June 2025.

[5] Thang D. Pham, Aditya Tanikanti, and Murat Keçeli. ChemGraph: An Agentic Framework for Computational Chemistry Workflows, June 2025. arXiv:2506.06363.

[6] Alireza Ghafarollahi and Markus J. Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent AI. *Proceedings of the National Academy of Sciences*, 122(4):e2414074122, January 2025.

[7] Yunheng Zou, Austin H. Cheng, Abdulrahman Aldossary, Jiaru Bai, Shi Xuan Leong, Jorge Arturo Campos-Gonzalez-Angulo, Changhyeok Choi, Cher Tian Ser, Gary Tom, Andrew Wang, Zijian Zhang, Ilya Yakavets, Han Hao, Chris Crebolder, Varinia Bernales, and Alán Aspuru-Guzik. El Agente: An autonomous agent for quantum chemistry. *Matter*, 8(7):102263, July 2025.

[8] Ziqi Wang, Hongshuo Huang, Hancheng Zhao, Changwen Xu, Shang Zhu, Jan Janssen, and Venkatasubramanian Viswanathan. DREAMS: Density Functional Theory Based Research Engine for Agentic Materials Simulation, July 2025. arXiv:2507.14267.

[9] P Giannozzi, O Andreussi, T Brumme, O Bunau, M Buongiorno Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A Dal Corso, S De Gironcoli, P Delugas, R A DiStasio, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H-Y Ko, A Kokalj, E Küçükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N L Nguyen, H-V Nguyen, A Otero-de-la Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A P Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni. Advanced capabilities for materials modelling with Quantum ESPRESSO. *Journal of Physics: Condensed Matter*, 29(46):465901, November 2017.

[10] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, January 2023. arXiv:2206.07697.

[11] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, February 2022.

[12] Y. A. Ran, S. Sharma, S. R. G. Balestra, Z. Li, S. Calero, T. J. H. Vlugt, R. Q. Snurr, and D. Dubbeldam. RASPA3: A Monte Carlo code for computing adsorption and diffusion in nanoporous materials and thermodynamics properties of fluids. *The Journal of Chemical Physics*, 161(11):114106, September 2024.

[13] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, July 2017.

[14] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013.

[15] Vaidotas Simkus, Anders Christensen, Steven Bennett, Ian Johnson, Mark Neumann, James Gin, Jonathan Godwin, and Benjamin Rhodes. Mofasa: A Step Change in Metal-Organic Framework Generation, December 2025. arXiv:2512.01756.

Appendix A. Benchmark

Table A1: Overview of the three-tiered benchmarking suite used to validate QUASAR. All test cases were executed using `gemini-3-flash-preview`.

Tier	Test Case	Prompt	Result
I. Task Execution	K-point Convergence	Calculate the k-point density to converge bulk Cu energy to 1 meV/atom.	$\Delta E < 0.6$ meV
	NPT Equilibration	Calculate the density of water at 298 K and 1 bar.	$\bar{\rho} = 0.99786$ g/cm^3
	Helium Void Fraction	Calculate the helium-accessible void fraction and pore volume for IRMOF-1.cif at 298 K.	$\phi_{He} = 0.7988$
II. Workflow Orchestration	Band Gap	Calculate the electronic band gap for bulk nickel oxide.	$E_g = 4.18$ eV
	Adsorption Isotherm	Calculate the adsorption isotherm for CO_2 in UiO-66 at 298 K.	$q_{1\,\mathrm{bar}} = 3.46$ mmol/g
	Melting Point	Calculate the melting point of aluminum	$T_m = 934.47 \pm 0.61$ K
III. Frontier Research	Photocatalyst Screening	Determine which 5% La-doped $ATaO_3$ perovskite (A = Li, Na, K) exhibits the best photocatalytic degradation performance against methyl orange under UV irradiation.	$NaTaO_3$: 5% La
	Gas Separation	Determine which COF maximizes the Xe/Kr adsorption selectivity at 298 K. The experimental CIF structures are provided.	COF-X-Br
	Virtual MOF Assessment	Determine the CO_2 adsorption performance and mechanical properties of the provided MOF.	Non-porous ($q \approx 0$); Stiff ($K = 165.8$ GPa)