From Experiments to Expertise: Scientific Knowledge Consolidation for AI-Driven Computational Physics

Haonan Huang [1]

Abstract

While large language models (LLMs) have transformed AI agents into proficient executors of computational materials science, performing a hundred simulations does not make a researcher. What distinguishes research from routine execution is the progressive accumulation of knowledge — learning which approaches fail, recognizing patterns across systems, and applying understanding to new problems. However, the prevailing paradigm in AI-driven computational science treats each execution in isolation, largely discarding hard-won insights between runs. Here we present QMatSuite, an open-source platform closing this gap. Agents record findings with full provenance, retrieve knowledge before new calculations, and in dedicated reflection sessions correct erroneous findings and synthesize observations into cross-compound patterns. In benchmarks on a six-step quantum-mechanical simulation workflow, accumulated knowledge reduces reasoning overhead by 67% and improves accuracy from 47% to 3% deviation from literature — and when transferred to an unfamiliar material, achieves 1% deviation with zero pipeline failures.

1. Introduction

Large language model agents can now autonomously plan, execute, and interpret computational physics simulations, in many cases matching the proficiency of trained researchers on simulation tasks of various complexity (Boiko et al., 2023; Bran et al., 2024; Ghafarollahi & Buehler, 2025; Zou et al., 2025; Gustin et al., 2025; Pérez-Sánchez et al., 2026; Kumar et al., 2026; Bai et al., 2026; Pham et al., 2026; Wang et al., 2025; Liu et al., 2025b;a). The execution problem — can AI run quantum-mechanical simulations correctly? — is largely solved. Meanwhile, the broader AI community has seen landmark systems like the 'AI Scientist' (Lu et al., 2026) successfully automate entire research lifecycles within machine learning environments.

Yet realizing this vision in the physical sciences exposes a fundamental barrier: physical execution is not a frictionless, rapid-feedback loop. A researcher who performs a hundred calculations over six months does not simply accumulate a hundred results. They internalize technical nuances, distinguish physical reality from artifacts, and distill cumulative observations into general principles. This progressive transformation of experience into expertise is what current AI systems lack. Existing agents start each session without access to findings from previous sessions: cross-session knowledge is either architecturally absent, limited to session-scoped scratchpads that reset between runs, or consists of static rules curated by human experts (Zou et al., 2025; Pham et al., 2026; Wang et al., 2025). RAG over calculation logs provides a partial remedy but lacks mechanisms for quality validation, knowledge abstraction, and traceable provenance (Lewis et al., 2020). Recent surveys have identified experiential accumulation and consolidation as a critical frontier for AI agents broadly (Hu et al., 2025); in computational physics specifically, structured memory for multi-step derivation and verification remains an open problem.

This gap will not close by building smarter models. The issue is structural: an agent session lasts minutes to hours, a physics research program unfolds over months to years. Bridging this requires infrastructure — a platform where findings from one session are available to the next, where those findings can be reviewed and corrected against physical reasoning, and where accumulated observations can be synthesized into higher-order understanding.

We present QMatSuite, an open-source computational physics platform built to provide this infrastructure. We validate it across 135 autonomous solid-state calculations spanning six material categories and 98 molecular geometry optimizations — using two different AI models and three simulation engines — and through a controlled learning-curve experiment on the anomalous Hall conductivity (AHC) of bcc iron, a transport property that requires a six-step pipeline

[1]Department of Physics, Princeton University, Princeton, NJ 08540, USA. Correspondence to: Haonan Huang <hnhuang@princeton.edu>.

Accepted by the ICML 2026 Workshop on AI for Physics (AI4Physics@ICML 2026).

coupling two simulation codes and physical judgment at multiple steps. Accumulated knowledge transforms both the efficiency and the character of the agent's work, and transfers cleanly to an unfamiliar ferromagnet. We further show that knowledge consolidation — synthesis of individual findings into cross-compound patterns — requires dedicated reflection sessions separate from task execution, paralleling the cognitive rhythm of human physics research. The platform supports 15 simulation engines and connects to any AI model via the Model Context Protocol (Anthropic, 2024), decoupling accumulated physics knowledge from both the computational engine that produced it and the AI model that will use it.

2. Platform design and scale validation

QMatSuite is built on three core pillars. First, it abstracts simulation complexity through engine-agnostic structured tools (`set_parameters`, `run_calculation`, `get_results_summary`) that translate high-level agent calls into engine-specific inputs for 15 simulation codes (Appendix Table 1); both AI agents (via MCP) and human researchers (via a desktop GUI) access the same core through a unified public API (Fig. 1a). Second, end-to-end provenance tracks the entire history from raw inputs to final insights, creating an auditable chain in the spirit of reproducible workflow infrastructures (Huber et al., 2020; Jain et al., 2015; Wilkinson et al., 2016). Third — and central to this work — the platform provides a persistent scientific memory system. Knowledge entries are organized in a graded hierarchy: *findings* record observations from individual calculations ("PBE overestimates the GaAs lattice constant by 1.6%"), *patterns* synthesize regularities across multiple findings ("PBE overestimation scales with atomic mass across III-V compounds"), and *principles* encode general rules. The knowledge base comprises curated best practices (read-only), agent-generated insights that accumulate across sessions (read-write), and community knowledge packs; this work focuses on the agent-generated pipeline.

Because calculation execution consumes the agent's full attention, the platform does not rely on the agent spontaneously pausing to manage knowledge. Instead, it embeds reminders at natural workflow junctures — tool preambles prompt the agent to search prior knowledge before configuring a new calculation; post-execution returns prompt error-recovery recording; and results summaries prompt logging of numerical outcomes. These lightweight nudges make knowledge bookkeeping a natural byproduct of the calculation workflow rather than a separate task (Appendix A); their efficiency is tested directly below. All experiments are driven by short natural-language prompts specifying only scientific intent (reproduced verbatim in each figure banner); agents receive no guidance on computational parameters, software syntax, or workflow structure.

Scale validation. To validate the platform at scale, we deployed an autonomous agent (Claude Sonnet 4.6) via QMatSuite's MCP interface to drive structural relaxations and band structure calculations in Quantum ESPRESSO (QE) (Giannozzi et al., 2009; 2017). Testing across 135 diverse materials spanning six categories (elemental metals and semiconductors, binary semiconductors, simple oxides, halides, layered transition metal compounds, and topological insulators) (Fig. 1b,c) yielded an 85.2% autonomous completion rate (115 materials). Success rates scaled with structural complexity, from 100% in simple metals to 62–65% in topological insulators and layered transition metal compounds. Notably, the 20 incomplete cases resulted exclusively from timeouts rather than crashes or erroneous outputs.

Lattice constants for 114 materials agree with experiment at 1.02% mean absolute error (MAE), consistent with established PBE benchmarks (Haas et al., 2009). Band gaps for 68 non-metallic compounds show 1.76 eV MAE, dominated by the expected PBE underestimation. All 42 metals were correctly identified; five strongly correlated insulators (α-Sn, CdO, VO_2, $LaMnO_3$, CoO) were predicted metallic — known semilocal-DFT failures, not agent errors (Mori-Sánchez et al., 2008; Cohen et al., 2008). The agent recorded a median of 3 insights per material, searching prior knowledge 4.2 times and recording 3.2 times on average — confirming that the platform's nudge-based knowledge bookkeeping integrates naturally into the calculation workflow without explicit user instructions.

Cross-engine and cross-model validation. To verify engine and model agnosticism, we deployed GPT 5.4 with ORCA 6.1.1 (Neese et al., 2020) for 98 molecular geometry optimizations (Appendix Fig. 6), of which 91 completed successfully. Bond lengths for the reference-backed raw-ORCA-converged subset agree with spectroscopic references at MAE = 0.0069 Å (0.52%), and bond angles at MAE = 0.51°, consistent with established hybrid-DFT geometry benchmarks (Bühl & Kabrede, 2006; Mardirossian & Head-Gordon, 2017). The seven final failures comprise two agent-declared failures despite converged raw outputs, three missing or non-terminated last-output cases, and two explicit last-attempt error terminations.

3. Knowledge transforms a complex workflow

3.1. The learning curve

The 135-material validation demonstrates competence on routine computational tasks. The more demanding test is whether accumulated knowledge helps when task complexity exceeds what pre-training alone can handle — when

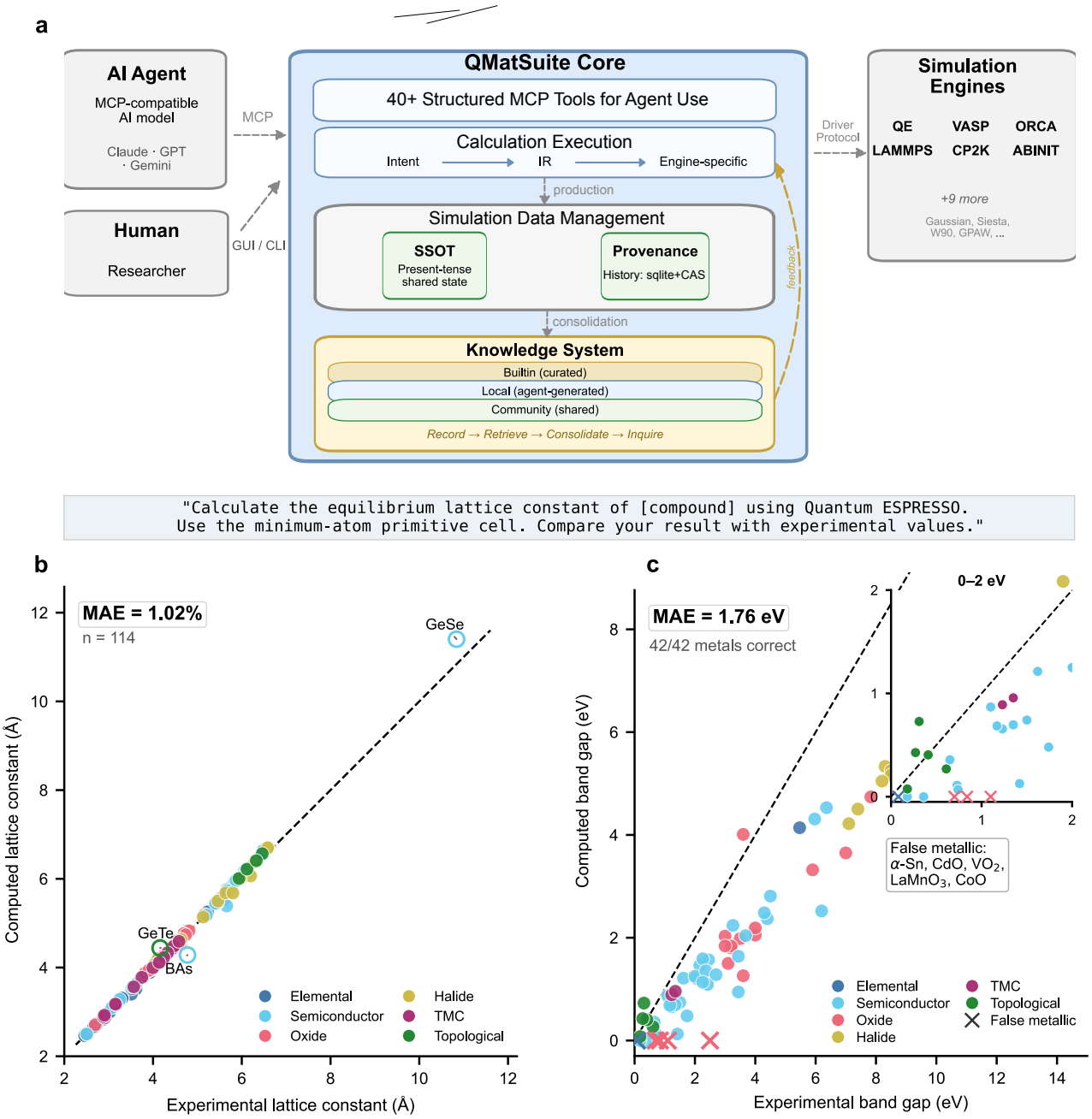

Figure 1. **Platform architecture and scale validation. a**, QMatSuite architecture showing symmetric access by AI agents (via MCP) and human researchers (via GUI) to the shared core, which contains the knowledge system, provenance tracking, present-tense shared state, and engine dispatch layer. **b**, Computed versus experimental lattice constants for 114 materials (MAE = 1.02%). **c**, Computed versus experimental band gaps for 68 non-metallic compounds (MAE = 1.76 eV); 42/42 metals correctly identified; 5 correlated insulators predicted metallic (known PBE limitation, marked ×). Inset: 0–2 eV region.

parameters must be coordinated across multiple codes, failure modes are undocumented, and physical judgment is required at multiple steps. We chose the anomalous Hall conductivity (AHC) of bcc iron as the testbed: a quantum-mechanical transport property (Nagaosa et al., 2010) whose computation requires a six-step pipeline coupling two simulation codes (QE and Wannier90 (Pizzi et al., 2020)), with a large parameter space where many settings require non-trivial physical reasoning. Even experienced practitioners require significant trial and error; the literature value is approximately -751 S/cm (Yao et al., 2004).

We ran the identical task three times in sequence (Fig. 2), each as a fresh session with a fresh project directory and the same model (Claude Opus 4.6). The sole variable carried between runs was the knowledge database: the baseline started empty, the second run inherited 6 insights from the first, and the third inherited 9.

Every metric improves monotonically (Fig. 2a). API reasoning time — which isolates cognitive effort from simulation compute — dropped from 42.8 to 16.1 minutes (62% reduction), total tool calls from 251 to 143, and pipeline execution attempts from 23 to 10. AHC accuracy improved from -1100 S/cm (46.5% error) to -846.2 S/cm (12.7%) to -771.5 S/cm (2.7%).

The dominant time saving traces to a single insight. In the baseline, the agent spent around 3 hours and 14 pipeline attempts diagnosing why the computed AHC was zero — ultimately discovering that QE's non-self-consistent (NSCF) step requires explicit non-zero `starting_magnetization`; otherwise, it silently defaults to a non-magnetic state, producing zero Berry curvature and zero AHC. This behavior — physically counter-intuitive because NSCF reuses converged charge densities — is neither clearly documented in standard manuals nor represented in training data. The finding was recorded as insight #2. In both subsequent runs, the agent retrieved this insight before its first calculation and applied the fix proactively, entirely eliminating the debugging episode (Fig. 2b).

Three observations establish the knowledge database, rather than pre-training knowledge, as the causal variable (Appendix Table 2). First, the identical model without the database required 14 attempts and 3.5 hours to discover the spin initialization requirement. Second, the third run failed to retrieve an available insight and consequently re-encountered the corresponding error — demonstrating that retrieval, not latent model knowledge, mediates transfer. Third, the monotonic improvement across three runs with identical model weights but increasing database content is consistent only with the database as the independent variable.

3.2. Cognitive liberation

Total wall time does not monotonically decrease — and this tells the most interesting story. The baseline agent finished at 6.2 hours, most of it consumed by infrastructure debugging; it never had the opportunity to question whether its result was converged. The second run completed in under two hours but also did not attempt convergence optimization. The third run reached its first valid result (-921 S/cm, 23% error) in 58 minutes — then voluntarily spent an additional 19 hours on a systematic seven-iteration convergence study, varying disentanglement parameters and comparing uniform versus adaptive mesh refinement strategies. It discovered that adaptive refinement outperforms brute-force mesh densification at one-seventh the computational cost — a genuine methodology insight that no previous run had the cognitive bandwidth to pursue.

The three runs thus exhibit a qualitative behavioral progression (Fig. 2c and Appendix Fig. 4): from debugger (the baseline, spending 70% of tool calls on infrastructure), through solver (the second run, resolving known and novel errors to reach a valid result), to optimizer (the third run, freed from debugging to pursue systematic physical exploration). The accuracy improvement from 47% to 3% is a consequence of this liberation, not of the knowledge system directly optimizing parameters.

3.3. Knowledge quality and self-correction

Across three runs, the agent accumulated 15 insights (Appendix Table 3): 9 error-recovery findings (60%), 3 result recordings (20%), and 3 methodology insights (20%) — the last category appearing only in the third run, reflecting the behavioral shift toward physics exploration. Of the 15, 13 are fully correct, one is partially correct, and one contains correct data but an incorrect conclusion. This last case involves Wannier90's disentanglement parameter `dis_froz_max`, which sets the upper energy boundary of the "frozen window" — the range within which Bloch states are included exactly in the Wannier representation (Souza et al., 2001). For transport properties that depend on states near the Fermi level, this window should extend well above the Fermi energy to ensure accuracy of relevant bands. The agent's third run found that setting `dis_froz_max` to 17 eV (slightly below the Fermi energy of $\sim$17.5 eV) gave the best AHC value and recommended keeping it low — prioritizing an accidental fit to literature over rigorous convergence validation (Fig. 3a).

In a dedicated review session, the same model operating in reflection mode identified this error (Fig. 3a). Applying convergence analysis to the agent's own parameter sweep data, it recognized that the recommended value was an unconverged outlier while higher values produced stable results. It further verified this against official tutorial input

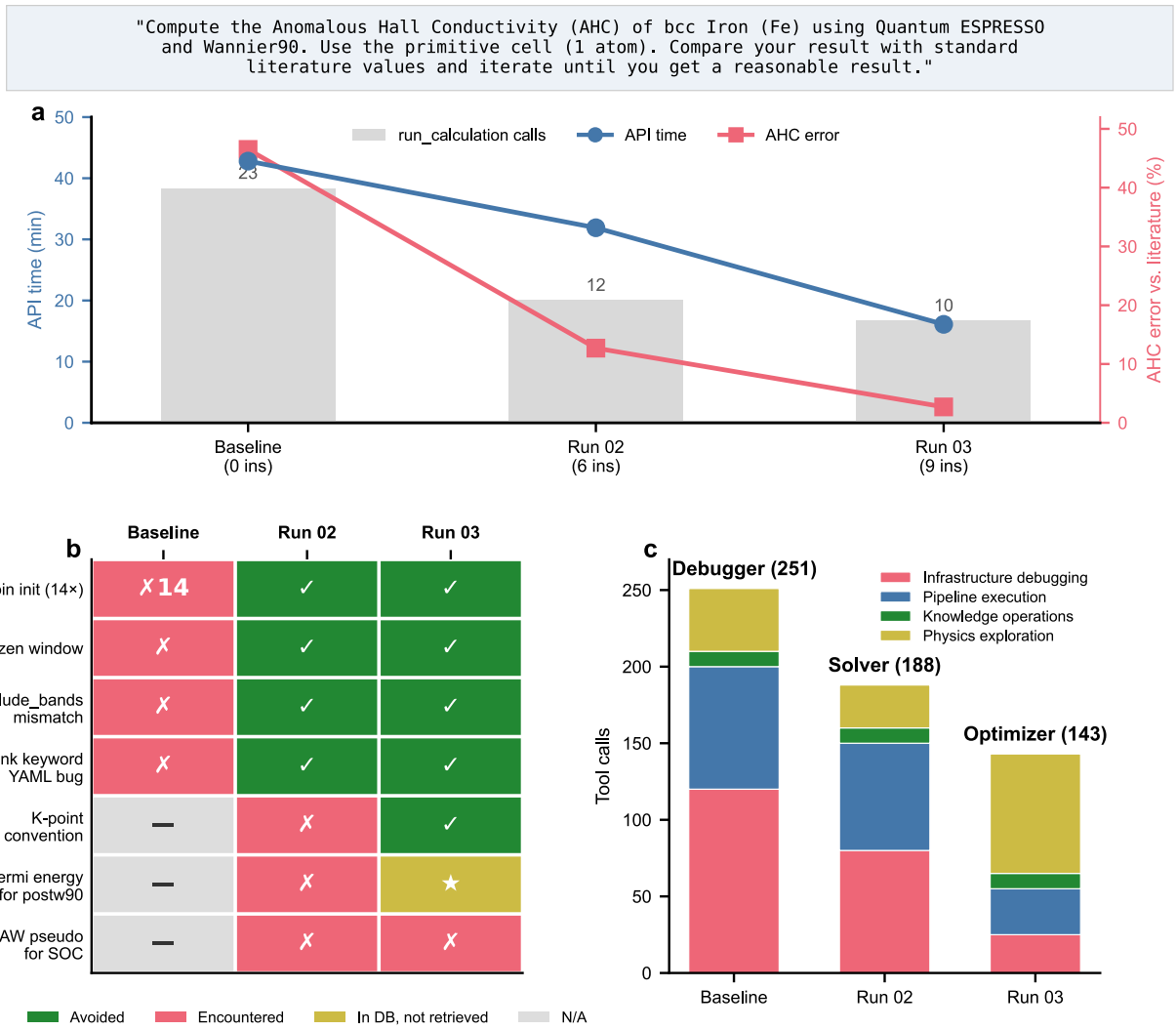

Figure 2. **Knowledge transforms a complex workflow. a**, Fe AHC learning curve across three runs with 0, 6, and 9 accumulated insights. Gray bars: `run_calculation` calls; blue line: API reasoning time; red line: AHC error versus literature. **b**, Episode avoidance matrix showing which pitfalls each run encountered (red) versus avoided via knowledge (green). Yellow: insight in database but not retrieved; gray: not applicable. **c**, Tool call composition shifting from infrastructure debugging ("Debugger") to physics exploration ("Optimizer").

files retrieved from raw source repositories, which directly contradicted the calculation agent's recommendation. The insight was deprecated and replaced with a corrected version citing both the convergence data and the tutorial evidence. This self-correction required structured general guidance, but the agent finds answers itself (Appendix E).

3.4. Cross-material transfer

To test whether knowledge generalizes beyond the material that produced it, we applied the reviewed iron knowledge database (21 insights) to a nickel AHC calculation — a system with no prior entries in the database. The prompt was identical except for the element name (Fig. 3b,c).

The agent completed the Ni AHC in 66 tool calls with only 3 pipeline executions — all successful — achieving $|\sigma_z| =$ 2079 S/cm (within 1.0% deviation from the literature value of $\sim$2073 S/cm (Wang et al., 2007)). API reasoning time was 14.1 minutes. The agent proactively applied correct pseudopotentials, spin settings, and convergence strategies from iron knowledge.

A controlled three-way Ni comparison shed light on the effect of knowledge quality (Fig. 3b,c). The naive baseline agent (0 insights) required 14 pipeline executions — 7 failures — and 272 tool calls to reach $\sim$10% error. The agent with 15 unreviewed Fe insights reduced this to 9 executions but inherited an incorrect parameter recommendation, costing three additional iterations, but the agent can still recover from such knowledge poisoning. The reviewed 21-insight agent completed the task in 3 executions with zero failures.

The most revealing finding concerns what did *not* transfer. Without a prior nickel calculation record, the 21-insight agent could not engage in recipe replication — the behavior observed in all six same-material iron sessions (Appendix E), where a key parameter was copied verbatim from the highest-performing prior result despite being flagged as incorrect. For nickel, it instead reasoned from general principles, arriving at a physically correct parameter choice. Strikingly, the same agent achieved better accuracy on an unfamiliar material (1.0%) than on the familiar one (16–23% in same-material reruns). While the Ni result may partly reflect favorable error cancellation in the absence of a systematic convergence study, the contrast suggests that recipe availability can actively hinder principle-based reasoning.

4. Knowledge consolidates into understanding

4.1. From findings to patterns

The AHC experiments demonstrate that accumulated findings improve efficiency and accuracy on a single complex workflow. A natural next question is whether agents can move beyond individual findings to synthesize higher-order understanding — recognizing regularities that span multiple calculations and materials. To test this, we ran a chain of 24 calculation sessions across 19 zinc-blende semiconductors, during which the agent accumulated 25 individual findings — each a structured record of one calculation's outcome (e.g., "PBE overestimates the GaAs lattice constant by 1.6%"). No higher-order synthesis occurred during any task-execution session: across 398 execution sessions, 84.9% received the platform's recording nudges, yet 95.3% of all knowledge recordings occurred in end-of-session batches and zero pattern-grade entries — cross-compound generalizations that synthesize multiple findings — were produced (Appendix Fig. 5b,c). Trace analysis reveals that agents informally recognized cross-compound trends during execution, yet never deposited these as higher-grade knowledge entries, because attention was fully consumed by the current calculation.

A dedicated reflection session with a 12-word prompt ("Review accumulated findings and summarize any patterns you see across compounds") produced three quantitative patterns from 25 individual findings in under three minutes (Appendix Fig. 5a). Pattern 1 identified that PBE lattice constant overestimation scales systematically with atomic mass across all three compound families (Group IV, III-V, II-VI), with quantitative error ranges tabulated per family — a synthesis spanning 19 compounds that no single session could have produced. Pattern 2 distinguished quantitative band gap underestimation from qualitative topological failures in near-zero-gap systems. Pattern 3 established that variable-cell relaxation with semicore pseudopotentials requires kinetic energy cutoffs above 80 Ry, supported by 13 successful calculations with Pulay stress identified as the causal mechanism (Francis & Payne, 1990). Each pattern references its supporting findings by unique identifier, maintaining full provenance to source calculations. These are not trivial restatements — they contain structured quantitative data, causal reasoning, and explicit scope boundaries.

An independent session using a less capable model (Claude Sonnet 4.6 instead of Opus 4.6) produced three patterns with identical themes from the same 25 findings, confirming that pattern structure is latent in the data rather than an artifact of a specific model.

This parallels a well-established observation in human cognition: synthesis does not occur during task execution but requires dedicated reflective activity (Sio & Ormerod, 2009; Baird et al., 2012). Independently, recent work on experiential reinforcement learning has identified an analogous reflection-consolidation loop as necessary for durable improvement in AI agents (Shi et al., 2026). Our findings provide empirical evidence for this principle in a scientific computing context (Appendix Fig. 5).

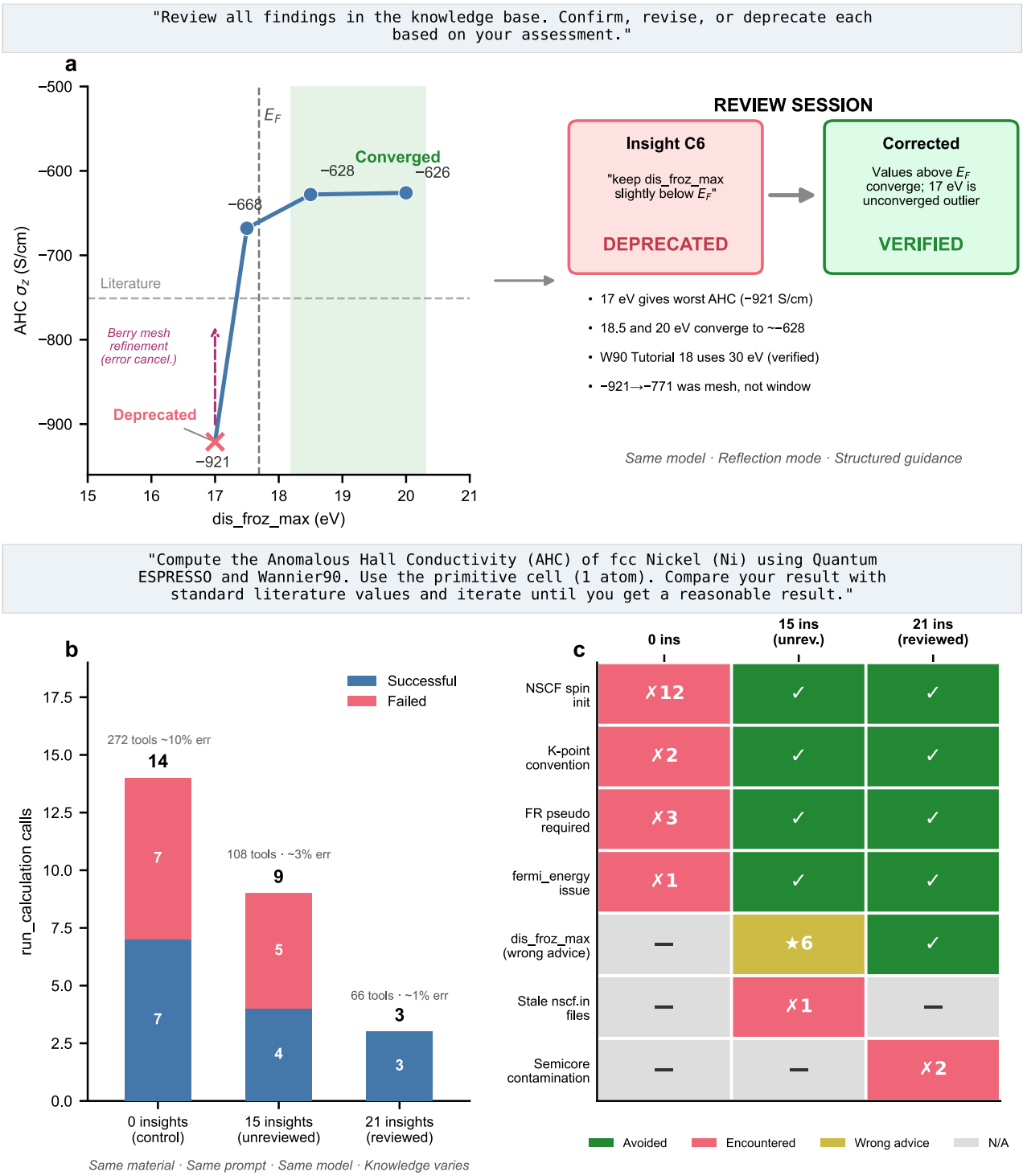

Figure 3. **Knowledge self-corrects and transfers across materials. a**, Left: convergence analysis showing the deprecated parameter recommendation (`dis_froz_max` = 17 eV, red ×) as an unconverged outlier, with the upward arrow indicating that Berry mesh refinement coincidentally pulled the AHC toward the literature value (error cancellation). Right: the review session's deprecation reasoning and corrected replacement. **b**, Ni AHC three-way comparison: 0 insights (14 calculations, 7 failures), 15 unreviewed insights (9 calculations, 5 failures), 21 reviewed insights (3 calculations, 0 failures). **c**, Ni pitfall avoidance matrix showing progressive transfer; yellow star (★6) marks the `dis_froz_max` row where unreviewed knowledge provided incorrect advice requiring 3 extra iterations.

4.2. From understanding to new inquiry

As a proof of principle, we prompted the agent to test one of its patterns. It selected InP, predicted a PBE band gap of $\sim$0.7 eV by interpolating Pattern 2's underestimation trend, and computed 0.696 eV — confirming the prediction. While this amounts to interpolation along a monotonic curve, it demonstrates that the full knowledge cycle — from observation through consolidation to hypothesis-driven inquiry — can close autonomously. Whether agents can design more physically creative experiments from richer knowledge bases remains an intriguing open question for future work.

5. Discussion

The experiments presented here demonstrate that persistent scientific memory transforms AI agent behavior beyond simple acceleration. The learning curve shows the most visible effect — a 67% reduction in reasoning overhead from baseline to cross-material transfer — but the deeper transformation is qualitative. The runs exhibit a behavioral progression from debugger to solver to optimizer to transferable expert: the baseline agent spent its entire session on infrastructure; the knowledge-equipped agent on an unfamiliar material completed the task with zero pipeline failures.

Agent-generated knowledge is predominantly correct, demonstrating that the recording system produces reliable scientific documentation. Like any research output, it occasionally captures incorrect conclusions: the `dis_froz_max` recommendation was based on correct numerical data but drew the wrong inference from under-converged calculations. Critically, unreviewed incorrect knowledge has measurable cost: the 15-insight Ni agent required 3 extra iterations to override inherited bad advice, while the reviewed 21-insight agent encountered zero failures. This establishes knowledge review not as an academic exercise but as a practical necessity that measurably reduces wasted computation.

The knowledge system directly addresses the RAG limitations (Lewis et al., 2020; Liu et al., 2024; Packer et al., 2023) identified in the Introduction, with our experiments providing direct evidence for each. Regarding correctness, unreviewed knowledge containing flawed recommendations results in performance penalty; the reviewed database eliminated this cost entirely. For compression, 25 individual findings were consolidated into 3 patterns in under three minutes; such abstraction renders retrieval and community sharing trivial due to the reduced data footprint. Finally, concerning provenance, when the review session deprecated the `dis_froz_max` recommendation, the system preserved the complete lineage from the erroneous conclusion back to the source data, ensuring the correction remained fully auditable and verifiable.

Cross-session memory for LLM agents is by now an established line of work: reflective and episodic memory (Shinn et al., 2023; Park et al., 2023), memory-augmented language models (Packer et al., 2023), skill and workflow memory (Wang et al., 2023; 2024), and a recent wave of dedicated agent-memory systems (Xu et al., 2025; Chhikara et al., 2025; Rasmussen et al., 2025; Fang et al., 2025). What none of these combine, and QMatSuite does, is three things at once: a graded, *reviewable*, provenance-linked findings→patterns→principles hierarchy with explicit deprecation, in which unreviewed knowledge carries a *measured* cost rather than an assumed one; a *quantified* separation of execution from reflection, with consolidation empirically confined to dedicated review sessions; and a domain — computational physics — in which both the process and the outcome of a calculation have *independent* ground truth. That independent ground truth is exactly what makes the error-cancellation and recipe-replication phenomena reported here observable at all, and it is out of reach for memory studied only on chat or question-answering benchmarks. The typed knowledge graph proposed in review — with deprecation status as a first-class node property — is a natural next step that could make recipe replication detectable structurally rather than behaviorally, and we see it as promising future work.

Agents exhibit distinct cognitive modes that have direct implications for system design. In execution mode, agents are goal-directed and resistant to mid-task reflection — only 4.0% of sessions show any mid-session knowledge recording despite 84.9% receiving explicit nudges, and zero pattern-grade entries were produced across 398 sessions. In reflection mode, the same model performs convergence analysis, tutorial verification, and self-correction (Park et al., 2023; Shinn et al., 2023). Effective AI research programs should alternate between execution and reflection sessions, with tool design adapted to each mode.

The cross-material experiment reveals the essential role of knowledge quality, and that recipe replication — copying parameters from the highest-performing prior result rather than reasoning about their applicability — is an anti-pattern triggered by recipe availability. In all six same-material iron sessions, the agent selected the deprecated `dis_froz_max` value by extracting it from the best-performing result record, ignoring deprecation warnings. On unfamiliar nickel, the absence of a material-specific recipe forced the agent to reason from principles — and it achieved comparable or better accuracy. Knowledge systems should encourage principle-level entries over numerical recipes.

QMatSuite demonstrates robust generality across engines (Quantum ESPRESSO, Wannier90, ORCA) and models (Claude and GPT). These results are directly relevant to

physics-centric reasoning with LLM agents: structured, provenance-linked, reviewable memory is what converts a tool-augmented agent — which can already run quantum-mechanical simulations correctly — into one that performs multi-step derivation, convergence verification, and physical-principle reasoning across sessions.

Sufficiently capable models will not obviate this infrastructure: the bottleneck is session isolation, not reasoning capability; smarter models will produce higher-quality insights and sharper reviews, making the infrastructure more valuable, not less. The platform supports community-contributed, domain-specific knowledge packs (Wilkinson et al., 2016); persistent scientific memory, demonstrated here for condensed-matter calculations, represents a general requirement for AI-driven research across computational physics.

The reflection loop is not infallible. A review session could in principle consolidate a false pattern, promoting a spurious regularity to higher-grade knowledge. Because review verdicts are themselves recorded as reviewable entries and every entry retains its provenance chain back to the source calculations, such errors remain auditable and reversible, and that same provenance supports human-in-the-loop pruning of poisoned entries. How robustly the loop behaves when a weaker or open-weights model occupies the reflection role — rather than the frontier model used here — is untested, and we leave it to future work.

Several limitations bound these claims. The learning-curve evidence rests on a single sequential three-run chain, without repeated trials or variance estimates; the Fe→Ni transfer stays within the $3d$-ferromagnet family, a relatively mild out-of-distribution test; and we do not benchmark against long-context replay or raw-log RAG baselines, which we leave as direct comparisons for future work. The efficiency result is likewise best read as a gain in cognitive reallocation — from debugging toward physical exploration — rather than as reduced wall time.

Code and data availability

QMatSuite is open-source and available at https://github.com/QMatSuite/QMatSuite. Complete experimental artifacts — full prompts, agent traces, knowledge-database snapshots, and reviewed-insight tables — are archived on Zenodo at https://doi.org/10.5281/zenodo.20588420 under CC-BY 4.0.

Impact Statement

This paper presents work whose goal is to advance AI-driven computational physics. Persistent scientific memory systems may accelerate physics research, with corresponding risks for the propagation of incorrect inherited knowledge when review practices are weak — risks we discuss through the knowledge-quality and recipe-replication analyses above. The manuscript itself was written by human authors describing an AI-agent platform; we do not advocate for autonomous publication of physics results without human review.

References

Anthropic. Model context protocol, 2024. URL https://modelcontextprotocol.io. Open standard for AI model–tool integration.

Bai, J., Aldossary, A., Swanick, T., Müller, M., Kang, Y., Zhang, Z., Lee, J. W., Ko, T. W., Vakili, M. G., Bernales, V., et al. El agente gr\'afico: Structured execution graphs for scientific agents. *arXiv preprint arXiv:2602.17902*, 2026.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., and Schooler, J. W. Inspired by distraction: mind wandering facilitates creative incubation. *Psychol. Sci.*, 23(10):1117–1122, 2012. doi: 10.1177/0956797612446024.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Bühl, M. and Kabrede, H. Geometries of transition-metal complexes from density-functional theory. *J. Chem. Theory Comput.*, 2(5):1282–1290, 2006. doi: 10.1021/ct600118l.

Chhikara, P., Khant, D., Aryan, S., Singh, T., and Yadav, D. Mem0: Building production-ready AI agents with scalable long-term memory. 2025.

Cohen, A. J., Mori-Sánchez, P., and Yang, W. Insights into current limitations of density functional theory. *Science*, 321(5890):792–794, 2008. doi: 10.1126/science.1158722.

Fang, R., Liang, Y., Wang, X., Wu, J., Qiao, S., Xie, P., Huang, F., Chen, H., and Zhang, N. Memp: Exploring agent procedural memory. 2025.

Francis, G. and Payne, M. C. Finite basis set corrections to total energy pseudopotential calculations. *Journal of Physics: Condensed Matter*, 2(19):4395–4404, 1990.

Fuh, H.-R. and Guo, G.-Y. Intrinsic anomalous hall effect in nickel: A gga+ u study. *Physical Review B—Condensed Matter and Materials Physics*, 84(14):144427, 2011.

Ghafarollahi, A. and Buehler, M. J. Automating alloy design and discovery with physics-aware multimodal multiagent AI. *Proc. Natl Acad. Sci. USA*, 122(4):e2414074122, 2025. doi: 10.1073/pnas.2414074122.

Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M., Dabo, I., Dal Corso, A., de Gironcoli, S., Fabris, S., Fratesi, G., Gebauer, R., Gerstmann, U., Gougoussis, C., Kokalj, A., Lazzeri, M., Martin-Samos, L., Marzari, N., Mauri, F., Mazzarello, R., Paolini, S., Pasquarello, A., Paulatto, L., Sbraccia, C., Scandolo, S., Sclauzero, G., Seitsonen, A. P., Smogunov, A., Umari, P., and Wentzcovitch, R. M. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter*, 21:395502, 2009. doi: 10.1088/0953-8984/21/39/395502.

Giannozzi, P., Andreussi, O., Brumme, T., Bunau, O., Buongiorno Nardelli, M., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Cococcioni, M., Colonna, N., Carnimeo, I., Dal Corso, A., de Gironcoli, S., Delugas, P., DiStasio, R. A., Ferretti, A., Floris, A., Fratesi, G., Fugallo, G., Gebauer, R., Gerstmann, U., Giustino, F., Gorni, T., Jia, J., Kawamura, M., Ko, H.-Y., Kokalj, A., Küçükbenli, E., Lazzeri, M., Marsili, M., Marzari, N., Mauri, F., Nguyen, N. L., Nguyen, H.-V., Otero-de-la Roza, A., Paulatto, L., Poncé, S., Rocca, D., Sabatini, R., Santra, B., Schlipf, M., Seitsonen, A. P., Smogunov, A., Timrov, I., Thonhauser, T., Umari, P., Vast, N., Wu, X., and Baroni, S. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter*, 29:465901, 2017. doi: 10.1088/1361-648X/aa8f79.

Gustin, I., Calderón, L. M., Pérez-Sánchez, J. B., Gonthier, J. F., Nakamura, Y., Panicker, K., Ramprasad, M., Zhang, Z., Zou, Y., Bernales, V., et al. El agente cu\'antico: Automating quantum simulations. *arXiv preprint arXiv:2512.18847*, 2025.

Haas, P., Tran, F., and Blaha, P. Calculation of the lattice constant of solids with semilocal functionals. *Phys. Rev. B*, 79:085104, Feb 2009. doi: 10.1103/PhysRevB.79.085104. URL https://link.aps.org/doi/10.1103/PhysRevB.79.085104.

Hu, Y., Liu, S., Yue, Y., Zhang, G., Liu, B., Zhu, F., Lin, J., Guo, H., Dou, S., Xi, Z., Jin, S., Tan, J., Yin, Y., Liu, J., Zhang, Z., Sun, Z., Zhu, Y., Sun, H., Peng, B., Cheng, Z., Fan, X., Guo, J., Yu, X., Zhou, Z., Hu, Z., Huo, J., Wang, J., Niu, Y., Wang, Y., Yin, Z., Hu, X., Liao, Y., Li, Q., Wang, K., Zhou, W., Liu, Y., Cheng, D., Zhang, Q., Gui, T., Pan, S., Zhang, Y., Torr, P., Dou, Z., Wen, J.-R., Huang, X., Jiang, Y.-G., and Yan, S. Memory in the age of AI agents. 2025.

Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A. V., Andersen, C. W., Ramirez, F. F., Adorf, C. S., Gargiulo, F., Kumbhar, S., Passaro, E., Johnston, C., Merkys, A., Cepellotti, A., Mounet, N., Marzari, N., Kozinsky, B., and Pizzi, G. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data*, 7:300, 2020. doi: 10.1038/s41597-020-00638-4.

Jain, A. et al. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency Computat.: Pract. Exper.*, 27(17):5037–5059, 2015. doi: 10.1002/cpe.3505.

Kumar, S. G. H., Zou, Y., Wang, A., Valdés-Hernández, J., Ko, T. W., Yue, N., Leng, O., Xu, H., Crebolder, C., Aspuru-Guzik, A., et al. El agente s\'olido: A new age (nt) for solid state simulations. *arXiv preprint arXiv:2602.17886*, 2026.

Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020.

Liu, G., Yang, S., and Zhong, Y. Masgent: An AI-assisted materials simulation agent. 2025a.

Liu, J., Zhu, T., Ye, C., Fang, Z., Weng, H., and Wu, Q. VASPilot: MCP-facilitated multi-agent intelligence for autonomous VASP simulations. *Chinese Phys. B*, 34(11): 117106, 2025b. doi: 10.1088/1674-1056/ae0681.

Liu, N. F. et al. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.*, 12: 157–173, 2024. doi: 10.1162/tacl_a_00638.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., and Clune, J. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Mardirossian, N. and Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.*, 115(19):2315–2372, 2017. doi: 10.1080/00268976.2017.1333644.

Mori-Sánchez, P., Cohen, A. J., and Yang, W. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Phys. Rev. Lett.*, 100:146401, 2008. doi: 10.1103/PhysRevLett.100.146401.

Nagaosa, N., Sinova, J., Onoda, S., MacDonald, A. H., and Ong, N. P. Anomalous hall effect. *Rev. Mod. Phys.*, 82:1539–1592, May 2010. doi: 10.1103/RevModPhys.82.1539. URL https://link.aps.org/doi/10.1103/RevModPhys.82.1539.

Neese, F. et al. The ORCA quantum chemistry program package. *WIREs Comput. Mol. Sci.*, 10(6):e1459, 2020. doi: 10.1002/wcms.1459.

Packer, C. et al. MemGPT: Towards LLMs as operating systems. 2023.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.

Pérez-Sánchez, J. B., Zou, Y., Campos-Gonzalez-Angulo, J. A., Müller, M., Gustin, I., Wang, A., Hao, H., Ko, T. W., Choi, C., Isbrandt, E. S., et al. El agente quntur: A research collaborator agent for quantum chemistry. *arXiv preprint arXiv:2602.04850*, 2026.

Pham, T. D., Tanikanti, A., and Keçeli, M. ChemGraph as an agentic framework for computational chemistry workflows. *Commun. Chem.*, 9:33, 2026. doi: 10.1038/s42004-025-01776-9.

Pizzi, G., Vitale, V., Arita, R., Blügel, S., Freimuth, F., Géranton, G., Gibertini, M., Gresch, D., Johnson, C., Koretsune, T., Ibañez-Azpiroz, J., Lee, H., Lihm, J.-M., Marchand, D., Marrazzo, A., Mokrousov, Y., Mustafa, J. I., Nohara, Y., Nomura, Y., Paulatto, L., Poncé, S., Ponweiser, T., Qiao, J., Thöle, F., Tsirkin, S. S., Wierzbowska, M., Marzari, N., Vanderbilt, D., Souza, I., Mostofi, A. A., and Yates, J. R. Wannier90 as a community code: new features and applications. *J. Phys.: Condens. Matter*, 32:165902, 2020. doi: 10.1088/1361-648X/ab51ff.

Prandini, G., Marrazzo, A., Castelli, I. E., Mounet, N., and Marzari, N. Precision and efficiency in solid-state pseudopotential calculations. *npj Comput. Mater.*, 4:72, 2018. doi: 10.1038/s41524-018-0127-2.

Rasmussen, P., Paliychuk, P., Beauvais, T., Ryan, J., and Chalef, D. Zep: A temporal knowledge graph architecture for agent memory. 2025.

Shi, T., Chen, S., Jiang, B., Song, L., Yang, L., and Zhao, J. Experiential reinforcement learning. 2026.

Shinn, N., Cassano, F., Berman, A., Gopinath, K., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. 2023.

Sio, U. N. and Ormerod, T. C. Does incubation enhance problem solving? a meta-analytic review. *Psychol. Bull.*, 135(1):94–120, 2009. doi: 10.1037/a0014212.

Souza, I., Marzari, N., and Vanderbilt, D. Maximally localized wannier functions for entangled energy bands. *Phys. Rev. B*, 65:035109, Dec 2001. doi: 10.1103/PhysRevB.65.035109. URL https://link.aps.org/doi/10.1103/PhysRevB.65.035109.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. 2023.

Wang, X., Vanderbilt, D., Yates, J. R., and Souza, I. Fermi-surface calculation of the anomalous hall conductivity. *Physical Review B—Condensed Matter and Materials Physics*, 76(19):195109, 2007.

Wang, Z., Huang, H., Zhao, H., Xu, C., Zhu, S., Janssen, J., and Viswanathan, V. DREAMS: Density functional theory based research engine for agentic materials simulation. 2025.

Wang, Z. Z., Mao, J., Fried, D., and Neubig, G. Agent workflow memory. 2024.

Weigend, F. and Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7:3297–3305, 2005. doi: 10.1039/B508541A.

Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., and Zhang, Y. A-MEM: Agentic memory for LLM agents. 2025.

Yao, Y., Kleinman, L., MacDonald, A. H., Sinova, J., Jungwirth, T., Wang, D.-s., Wang, E., and Niu, Q. First principles calculation of anomalous Hall conductivity in ferromagnetic bcc Fe. *Phys. Rev. Lett.*, 92:037204, 2004. doi: 10.1103/PhysRevLett.92.037204.

Železnỳ, J., Yahagi, Y., Gomez-Olivella, C., Zhang, Y., and Sun, Y. High-throughput study of the anomalous hall effect. *npj Computational Materials*, 9(1):151, 2023.

Zou, Y., Cheng, A. H., Aldossary, A., Bai, J., Leong, S. X., Campos-Gonzalez-Angulo, J. A., Choi, C., Ser, C. T., Tom, G., Wang, A., Zhang, Z., Yakavets, I., Hao, H., Crebolder, C., Bernales, V., and Aspuru-Guzik, A. El Agente: An autonomous agent for quantum chemistry. *Matter*, 8(7):102263, 2025. doi: 10.1016/j.matt.2025.102263.

A. Methods

Platform architecture. QMatSuite provides 40+ structured tools via the Model Context Protocol (MCP), organized into five categories: Project (initialization, structure management), Config (parameter setting, presets), Execute (calculation launch, monitoring), Analysis (results extraction, structure promotion), and Knowledge (search, record, review). Engine drivers translate structured tool calls into engine-specific input files. The platform uses a three-layer execution model: user/agent intent (through structured operations such as `set_parameters`, `apply_preset`, `run_calculation`), an engine-agnostic intermediate representation, and engine-specific materialization into QE namelists, VASP INCAR tags, ORCA keyword blocks, or other engine-native inputs. Fifteen simulation engines are supported at varying integration depth: Quantum ESPRESSO (Giannozzi et al., 2009; 2017), VASP, ORCA, LAMMPS, CP2K, ABINIT, Gaussian, Siesta, Wannier90, GPAW, Psi4, PySCF, xTB, QMCPACK, and Yambo. All experiments in this work used Quantum ESPRESSO 7.5 (Giannozzi et al., 2009; 2017) with SSSP precision pseudopotentials (Prandini et al., 2018) for solid-state calculations and ORCA 6.1.1 (Neese et al., 2020) for molecular calculations.

Persistent scientific memory. The knowledge system comprises three tiers stored in SQLite databases with FTS5 full-text search indexing. The builtin tier contains curated best practices (read-only). The local tier stores agent-generated insights (read-write) in three grades: findings (individual observations), patterns (cross-system regularities), and principles (general rules). Each entry includes scope tags, source calculation references, and free-text reasoning. The community tier supports sharable domain-specific insight packs. All tiers are searched via a unified `search_knowledge` tool with relevance ranking. Recording uses `record_insight` with mandatory grade, scope, and reasoning fields. Review uses `review_insight` with verdicts (confirmed, deprecated, promoted) and provenance links. Insights may reference other insights by ULID, enabling hierarchical knowledge structures.

Prompt design and nudge mechanism. All experiments used short natural-language task prompts specifying scientific intent only (e.g., "Calculate the equilibrium lattice constant of [material]"); they contained no computational parameters, software syntax, or workflow structure. Experiment-level system preambles imposed workflow discipline (engine execution through structured tool calls, one active run at a time, explicit error-recovery recording) without prescribing chemistry-specific parameter choices. The nudge mechanism embedded reminders in tool response text at natural workflow junctures: a front-loaded reminder in `create_calculation` to search prior knowledge before configuration; a post-recovery reminder in `run_calculation` to record the error and its fix; a results-linked reminder in `get_results_summary` to record numerical findings; and a synthesis reminder in `list_insights` (e.g., "N findings pending synthesis. Consider reviewing with `list_insights(grade='finding')` when your current task is complete"). These nudges increased opportunity for knowledge capture but did not guarantee compliance.

AHC learning curve experiment. The anomalous Hall conductivity of bcc iron was computed using the Quantum ESPRESSO + Wannier90 six-step pipeline: SCF → NSCF → wannierprep → pw2wannier90 → wannier90 → postw90. The agent was Claude Opus 4.6 with no timeout. The prompt was identical across all three runs. Each run used a fresh Claude session, a fresh project directory, and the same model. The sole variable was the knowledge database: Condition A started empty (0 insights), Run 02 inherited 6 insights, and Run 03 inherited 9. The builtin knowledge database was disabled (`QMS_KNOWLEDGE_BUILTIN=0`) to isolate local knowledge effects. To ensure the knowledge database was the sole information channel between sessions, web search tools were blocked, engine installation tutorial and example files were removed from disk, and session persistence was disabled. All agent traces were audited post-hoc to verify that no external information sources — web, filesystem tutorials, or cached content from prior runs — were accessed in any experimental session. The model's pre-training knowledge constitutes an uncontrolled stochastic baseline; the learning curve design isolates the database contribution from this baseline.

Cross-material transfer experiment (Ni). Three Ni AHC runs used the same prompt (substituting "fcc Nickel (Ni)" for the Fe element) and the same model (Claude Opus 4.6, no timeout). The baseline (0 insights) used an empty knowledge database with all leakage channels blocked; the trace was audited post-hoc to confirm no external information access. Run 26 (15 insights) used the unreviewed Fe knowledge database containing two incorrect parameter recommendations. Run 24 (21 insights) used the reviewed database with deprecated entries corrected. The literature reference for fcc Ni AHC is approximately 2000–2400 S/cm (Yao et al., 2004; Wang et al., 2007; Fuh & Guo, 2011; Železný et al., 2023).

Meta session and knowledge consolidation experiment. 19 zinc-blende semiconductors underwent sequential relaxation and band structure calculations (24 sessions, Claude Opus 4.6) accumulating 25 findings in the local knowledge database. Two independent reflection sessions — one using Claude Sonnet 4.6, one using Claude Opus 4.6 — were then run with an identical 12-word prompt. Both models produced 3 patterns with identical themes, confirming that pattern structure is latent in the data rather than model-specific. A subsequent design experiment session (Claude Opus 4.6) used a 29-word prompt with no preamble, attempt limits, or tool instructions.

Nudge compliance analysis. Nudge compliance was measured across 398 execution sessions (270 from the 135-material benchmark, 128 from AHC and reflection experiments) by parsing all `trace.jsonl` files programmatically. Recording nudge delivery was identified by matching nudge phrases in tool response text. Each `record_insight` call was located by event index within its session and classified by position (early <30%, mid 30–70%, late 70–90%, end-of-session >90%). Of 398 sessions, 338 (84.9%) received at least one recording nudge; 16 (4.0%) contained any mid-session `record_insight` call; 95.3% of all recordings occurred in the final 10% of session events; and zero sessions produced pattern-grade entries. All 41 patterns in the knowledge base were produced exclusively in dedicated review sessions.

Cross-engine molecular validation. 98 molecules spanning diatomics, triatomics, small polyatomics, organics, radicals, and transition metal complexes were computed using ORCA 6.1.1 with agent-selected methods (predominantly B3LYP/def2-TZVP-level hybrid DFT (Weigend & Ahlrichs, 2005; Mardirossian & Head-Gordon, 2017)). The agent was GPT 5.4. Experimental reference bond lengths and angles were compiled from NIST CCCBDB, the NIST Chemistry WebBook, Huber & Herzberg spectroscopic constants, and primary spectroscopic literature.

B. Platform comparison and retrieval evidence

Table 1. Comparison of AI agent systems for computational science. Eight systems evaluated across capability dimensions relevant to knowledge-driven research. "Intent-level tools" means the agent expresses scientific purpose (e.g. `run_calculation`) rather than writing engine-specific input files. Knowledge persistence refers to whether agent-generated findings survive across independent sessions.

System	Venue	Input gen.	Param select.	Error recovery	Cross-session knowledge	Knowledge review	Engines	Open src
El Agente Q	*Matter* 2025	ORCA files	✓	✓	Static rules*	—	1	No
ChemGraph	*Commun. Chem.* 2026	ASE API	Partial	Limited	None	—	6	Yes
DREAMS	arXiv 2025	QE via ASE	✓	✓	Session-scoped	—	1	Yes
AtomAgents	*PNAS* 2025	Pre-built fn	Partial	Limited	None	—	1	Yes
VASPilot	*Chin. Phys. B* 2025	VASP files	Hybrid	✓	RAG (wiki)	—	1	Yes
Masgent	arXiv 2025	VASP files	Defaults	Preventive	None	—	1	Yes
Aitomia	arXiv 2025	✓	✓	✓	None	—	ML	Partial
QMatSuite	**This work**	**Intent tools**	✓	✓	**Persistent, graded**	✓	**15**	**Yes**

*El Agente Q's "hierarchical memory" consists of static rules manually curated by human experts; episodic memory is explicitly disabled in the published implementation.

C. Knowledge inventory for the Fe AHC learning curve

Across three runs, the agent accumulated 15 insights. In summary: 13 are fully correct, 1 is partially correct (A5: correct calculation, incomplete convergence attribution), and 1 contains correct data but an incorrect conclusion (C6: `dis_froz_max` recommendation based on unconverged outlier). The load-bearing deprecated entry is reproduced in full below as the worked example for the self-correction claim; the complete insight table is available in the archived artifacts (https://doi.org/10.5281/zenodo.20588420).

Counts by run. Baseline (Run A): 6 entries (A1 frozen-window fix, A2 NSCF `starting_magnetization` critical requirement, A3 semicore exclusion, A4 `exclude_bands` trap, A5 -1100 S/cm result, A6 blank-keyword crash). Run 02: 3 entries (B1 k-point convention, B2 `fermi_energy` required, B3 -846 S/cm result). Run 03: 6 entries (C1 projections string form, C2 redundant `fermi_energy`, C3 FR pseudopotential required for SOC, C4 -772 S/cm adaptive-mesh result, C5 adaptive-refinement vs brute-force cost comparison, C6 deprecated method entry above). Three entries are methodology insights — all appearing in Run 03, reflecting the behavioral shift toward physics exploration.

Table 2. Insight retrieval and application across Fe AHC runs. Matrix showing which insights from the knowledge database were retrieved and applied in each run. ★ marks an insight present in the database but not retrieved, causing the agent to re-encounter the corresponding error. This case (insight B2 in Run 03) provides causal evidence that retrieval — not latent model knowledge — mediates transfer: the same model with the same pre-training failed to avoid an error whose solution was available but not found by BM25 search.

		Run 02		**Run 03**	
ID	**Content (brief)**	**Retrieved**	**Applied**	**Retrieved**	**Applied**
A1	Frozen window fix	✓	✓	—	—
A2	`starting_mag` (CRITICAL)	✓	✓	✓	✓
A3	Semicore exclusion	—	Set indep.	✓	✓
A4	`exclude_bands` trap	✓	Avoided	—	—
A5	AHC -1100 (prior result)	✓	Planning	✓	Planning
A6	Blank keyword bug	✓	Avoided	—	—
B1	K-point convention	*(Recorded this run)*		✓	✓
B2	`fermi_energy` required	*(Recorded this run)*		★ **Not retr.**	Hit error
B3	AHC -846 (improved)	*(Recorded this run)*		✓	Planning

★ Insight B2 was in the database but none of Run 03's four search queries matched it. The agent independently resolved the error and recorded a redundant insight (C2). This demonstrates that BM25 retrieval is imperfect — having knowledge in the database does not guarantee it will be found.

Table 3. Deprecated insight C6 (full content), the load-bearing example for the knowledge self-correction claim. Entry contains correct sensitivity data but draws an incorrect conclusion: `dis_froz_max = 17` produced the best AHC (-921 S/cm) through error cancellation (Berry mesh refinement compensated for unconverged Wannier functions); higher values (18.5, 20) converge to ~ -628 S/cm. Deprecated and replaced in a dedicated review session via convergence analysis and tutorial verification.

Field	**Value**
ID	C6
Source run	Run 03 (iron AHC, Claude Opus 4.6)
Grade	Finding (method)
Scope tags	wannier90, AHC, disentanglement
Original content	"`dis_froz_max`=17 eV (slightly below $E_F = 17.5$ eV) gave the best AHC result (-921 S/cm baseline, -771.5 S/cm with adaptive mesh); keep this value low."
Supporting data	Sweep: `dis_froz_max`=17→-921; 17.5→-668; 18.5→-628; 20→-626 S/cm
Correctness	Wrong conclusion (data correct)
Review verdict	Deprecated
Replacement	"`dis_froz_max` must be well above E_F to ensure convergence of states near the Fermi level; the best literature agreement at 17 eV arose from error cancellation with Berry-mesh refinement. Recommended: $\geq$20 eV, verified against Wannier90 Tutorial 18 (`dis_froz_max=30`)."

D. AHC trace analyses

Cross-run summary for the Fe learning curve.

Metric	Cond. A (0)	Run 02 (6)	Run 03 (9)
Final AHC (S/cm)	-1100	-846.2	-771.5
Error vs. literature	46.5%	12.7%	2.7%
API time (min)	42.8	31.9	16.1
Wall time (min)	372.4	113.3	1211.3
Total tool calls	251	188	143
`run_calculation` calls	23	12	10
`search_knowledge` calls	4	3	4
`record_insight` calls	6	3	6

Every cognitive-effort metric improved monotonically as the database grew. Wall time did not, because Run 03 spent the majority of its time on a deliberate convergence study rather than infrastructure debugging.

Condition A: discovery from a blank database. Condition A lasted 21,761 s and terminated by timeout after 251 tool calls. The session began with empty `search_knowledge` and `list_insights` results. The first major challenge was

pseudopotential selection: the auto-resolution routine returned a PAW file inappropriate for spin–orbit coupling, forcing the agent to locate and stage a fully relativistic pseudopotential manually. The next obstacle arose at Wannierization: with `num_wann = 18`, a frozen window placed too high included too many states; the agent diagnosed this from the eigenvalue spectrum and lowered the frozen-window boundary. The decisive diagnostic came from comparing SCF and NSCF outputs: the SCF calculation was magnetic, but the NSCF output reported a non-magnetic spin–orbit calculation. The agent eventually inferred that `starting_magnetization(1)` must also be present in the NSCF input. This single discovery consumed roughly 3.5 h and 14 pipeline attempts in the baseline, and it became the most important transferable error-recovery finding in the learning curve. The final AHC value was -1100 S/cm, overestimated by 46.5%.

Run 02: inherited knowledge removes the zero-AHC trap. Run 02 began with six findings from Condition A in the local database. Before the first calculation was launched, the agent retrieved and explicitly planned around four of them. This inheritance removed the entire zero-AHC episode from the session. The dominant bottleneck instead became a new interoperability bug: a k-point convention mismatch between Quantum ESPRESSO and Wannier90. Roughly 40% of the trace was spent exploring explicit crystal-format k-point lists, stale input-file collisions, and shifted mesh conventions. The eventual fix was to supply explicit crystal k-points folded into the $[-0.5, 0.5)$ convention for the NSCF step. After a further correction to supply the Fermi energy to the transport post-processing stage, the pipeline completed cleanly at -846.2 S/cm. Run 02 thus provides the first direct causal demonstration of persistent knowledge: a previously discovered failure mode disappeared entirely, while a new and previously unseen failure mode became the dominant source of effort.

Run 03: knowledge liberates the agent for convergence analysis. Run 03 inherited nine findings and retrieved five unique insights before the first calculation. Those retrieved findings directly shaped the initial setup: explicit NSCF k-points, explicit NSCF magnetization, a lower frozen-window floor, a denser SCF mesh. The run still encountered three problems (FR pseudopotential bug in the automatic resolver; incorrect serialization of the Wannier90 projection field; omission of the Fermi energy in the final transport stage), all corrected and recorded. Once the first successful pipeline returned -921 S/cm, the character of the session changed — the agent was no longer infrastructure-bound and used the rest of the session for a seven-iteration convergence study:

Iter.	Key change	σ_z (S/cm)	Interpretation
1	Baseline, uniform Berry mesh 100^3	-921	First end-to-end converged pipeline, still overestimated
2	`dis_froz_max = 20` eV	-626	Upper frozen window moved into the converged regime
3	`dis_froz_max = 18.5` eV	-628	Same converged branch as 20 eV
4	`dis_froz_max = 17.5` eV	-668	Intermediate regime
5	No frozen window	-3.4	Berry curvature collapses
6	Uniform Berry mesh 200^3	-864	Brute-force refinement helps, at very high cost
7	Adaptive refinement + `use_ws_distance`	-771.5	Best result, 2.7% error

The database did not directly encode the final adaptive-refinement recipe. Rather, it removed enough repeated debugging that the agent could spend the remainder of the session on systematic physical exploration — the mechanistic basis for the "cognitive liberation" claim.

Cross-material transfer to Ni. The three-way Ni comparison isolates the effects of knowledge quantity and quality:

Run	Start insights	Final AHC (S/cm)	Error	`run_calc` calls	Failed runs	Tool calls
0 insights	0	+2289	∼10%	14	7	272
15 insights (unreviewed Fe)	15	+2040	∼3%	9	5	108
21 insights (reviewed Fe)	21	+2079	1.0%	3	0	66

The reviewed 21-insight agent reached $|\sigma_z| = 2078.6$ S/cm in three successful launches and 14.1 min of API time. Importantly, it did not copy the Fe-specific `dis_froz_max = 17` eV recipe. Because the database contained Fe-specific findings but no Ni-specific "best-result" record, the agent instead reasoned from general principles: use a fully relativistic pseudopotential, keep NSCF explicitly magnetic, use explicit crystal k-points, and inspect the eigenvalue spectrum to choose the frozen-window placement. It therefore began at `dis_froz_max = 20` eV, diagnosed residual semicore contamination after the second run, widened the outer window, and converged in the third iteration.

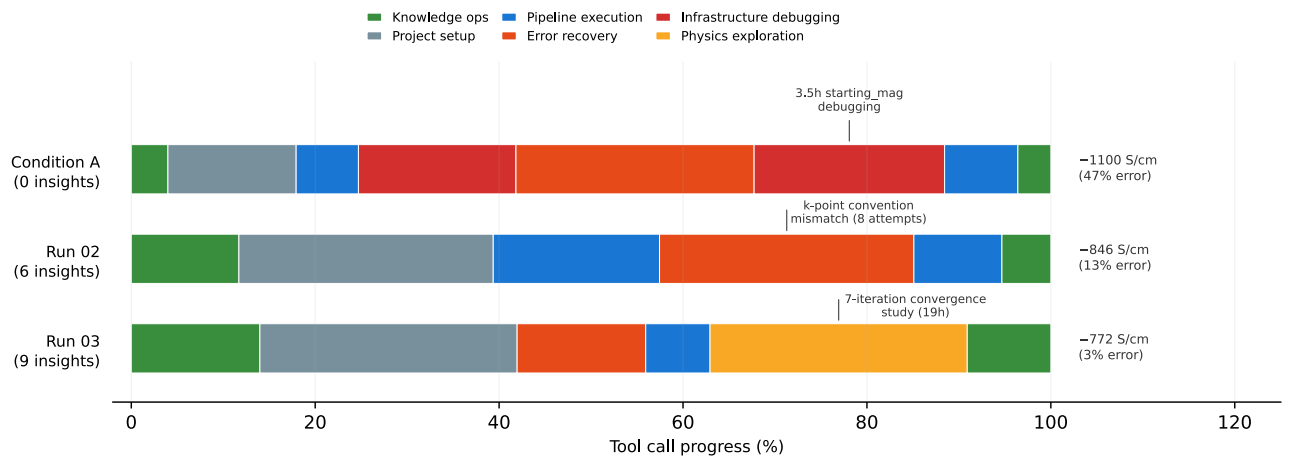

Figure 4. **Temporal structure of agent activity across three Fe AHC runs.** Gantt-style visualization of tool call composition, showing the shift from infrastructure debugging to physics exploration as knowledge accumulates. **Condition A** (0 insights): dominated by debugging and error recovery, with 3.5 h spent diagnosing why the computed AHC was zero (the `starting_magnetization` issue). **Run 02** (6 insights): proactively avoided the `starting_mag` pitfall but encountered a novel k-point convention mismatch requiring 8 attempts. **Run 03** (9 insights): resolved all setup issues within the first 80 tool calls, then devoted the remainder to a voluntary 7-iteration convergence study — a qualitative shift from "getting it to work" to "understanding the physics."

E. Review-session journey and the knowledge-delivery gap

The target finding. The review experiment was designed around a subtle but scientifically important failure mode: a finding containing correct numerical data but an incorrect conclusion. Insight C6 correctly recorded the Fe AHC sensitivity sweep but concluded that `dis_froz_max` should be kept slightly below the Fermi level because the final adaptive-refinement run at 17 eV happened to give -771.5 S/cm, close to the literature value. In reality, that agreement arose from error cancellation between an unconverged frozen-window choice and later adaptive Berry-mesh refinement.

Layered defenses. The review experiment did not converge on a single perfect prompt. Different preamble layers resolved different cognitive failure modes, and the resulting progression required a layered defense stack. The scientific lesson is not prompt sensitivity in the abstract but that claim-testing review required composing multiple distinct mechanisms:

Preamble layer	Failure mode addressed	Mechanism
Citation URL + verbatim excerpt	Hallucinated verification	Forces genuine external retrieval rather than memory-only verification
"Separate data from conclusions"	Correct data, wrong inference	Makes the model test the conclusion rather than only the data record
"Search recommendations directly"	Topic-level instead of claim-level retrieval	Encourages dedicated search for the recommendation itself
Step 0 PREPARATION (fetch raw input files)	Evidence-discovery gap	Moves the search target from prose pages to raw calculation inputs
"No middle ground"	Confirmed-with-caveat rationalization	Removes hedged verdicts for partially wrong findings

The decisive transition occurred with the addition of the last layer. One reviewing run downloaded the raw `Fe.win` file from Tutorial 18 and found `dis_froz_max = 30` but rationalized the contradiction away as pseudopotential-specific and confirmed the original finding with caveats. Adding the instruction "A finding is either completely correct or it is not. There is no middle ground" removed that escape hatch. Subsequent runs then independently deprecated the target finding.

Delivery gap: recipe replication anti-pattern. The delivery experiment asked whether the corrected 21-insight Fe database would alter the behaviour of a fresh Fe calculation agent. It did not. Across six delivery runs, every agent still chose `dis_froz_max = 17` eV. By runs 19–23 the corrected finding was visible; what failed was the comparative, claim-testing cognitive mode required to interpret it. One run provided the clearest evidence:

> *"I'm reconsidering the* `dis_froz_max` *value — the knowledge base shows that 17 eV gives* -921 *S/cm without adaptive refinement and* -771.5 *S/cm with it, while the official tutorial uses 30 eV... Based on the adaptive refinement results,* `dis_froz_max = 17` *eV is the better choice since it produced the strongest conductivity value of* -771.5 *S/cm."*

This is the same error-cancellation trap the review agent learned to catch. The calculation agent compressed the retrieved evidence into a recipe-oriented summary and anchored on the best prior numerical outcome rather than on parameter convergence. The delivery failure is therefore not merely a search-ranking problem; it is evidence that in execution mode, the claim-testing cognitive mode required to interpret contradictory knowledge entries fails even when the corrected entries are visibly retrieved. This recipe-replication behaviour is the same anti-pattern observed in the Ni cross-material experiment in the main text.

F. Benchmark statistics

Category-resolved statistics for the 135-material matrix.

Category	Materials	Complete	Success (%)	Lattice MAE (%)	Band-gap MAE (%)	Mean insights
Elemental	31	31	100.0	0.77	24.3	2.6
Semiconductor	35	32	91.4	1.43	51.1	2.7
Oxide	25	19	76.0	0.66	41.5	3.5
Halide	13	13	100.0	1.23	38.8	3.1
TMC	23	15	65.2	0.44	28.4	2.8
Topological	8	5	62.5	2.50	61.7	2.6
All	135	115	85.2	1.02	46.4	2.9

All 20 failures were timeouts under a uniform end-to-end session budget; no wrong-answer completions and no unrecovered engine crashes occurred in the final statistics. The ORCA benchmark yielded 91/98 completed under the conservative publication criterion; the reference-backed raw-ORCA-converged subset achieved bond-length MAE = 0.0069 Å (0.52%) across 78 comparisons and bond-angle MAE = 0.51° across 33 comparisons.

Tool usage across the 135-material benchmark.

Tool	Total calls	Per material
`inspect_calculation`	603	5.2
`set_parameters`	485	4.2
`search_knowledge`	480	4.2
`get_results_summary`	400	3.5
`run_calculation`	386	3.4
`search_demos`	382	3.3
`record_insight`	371	3.2

Two behavioural features follow directly. Knowledge retrieval was systematic rather than exceptional (`search_knowledge` used 4.2 times per material). Knowledge recording was likewise routine at the finding level: the median completed material generated three recorded insights.

Nudge compliance and the bandwidth problem. Across 398 execution sessions, 338 (84.9%) received at least one recording nudge, yet only 16 sessions (4.0%) contained any mid-session `record_insight` call, 95.3% of all recordings occurred in the final 10% of session events, and zero execution-mode sessions produced pattern-grade entries. All 41 patterns in the knowledge base were created in dedicated reflection sessions rather than during active calculations. Agents often recognized cross-compound trends informally during calculations, but execution mode did not provide enough attentional slack to convert those observations into higher-grade, provenance-linked knowledge objects.

G. Dedicated reflection distills findings into predictive knowledge

Two independent reflection sessions converged on the same three patterns. The dedicated distillation experiment asked whether a short reflection session could extract higher-order structure from 25 findings accumulated across 24 execution

sessions over 19 zinc-blende semiconductors. Two independent sessions were run on the same starting database with the same 12-word prompt: one used Claude Sonnet 4.6 (2 min 46 s), one used Claude Opus 4.6 (2 min 28 s). Both produced exactly three pattern-grade entries with identical scientific themes:

- PBE lattice-constant overestimation increases systematically with constituent atomic mass across the tetrahedral semiconductor families.
- PBE band-gap underestimation worsens as the true gap narrows and becomes qualitatively unreliable in the near-zero-gap regime.
- Variable-cell relaxations with semicore pseudopotentials require higher cutoffs and smearing to avoid Pulay-stress artifacts.

The wording differed, but the numerical backbone was the same (e.g., the monotonic Group IV sequence C $+0.14\%$, Si $+0.71\%$, Ge $+1.80\%$, α-Sn $+2.58\%$). The reflection sessions were not prompted with target answers or suggested categories. Pattern structure appears to have been latent in the findings themselves; dedicated review simply exposed it.

Pattern-guided design produced an accurate quantitative prediction for InP. The follow-up session tested whether distilled patterns could drive new scientific reasoning. Starting from the Opus-generated database of 25 findings plus 3 patterns, the agent was asked to propose and execute one calculation testing a specific prediction. It selected InP (relaxed structure existed; no band structure yet), used the distilled band-gap pattern to predict a PBE direct gap of $\sim$0.7 eV (corresponding to $\sim$45–48% underestimation relative to the 1.344 eV experimental direct gap), and used the workflow pattern to choose `ecutwfc = 80` Ry, SSSP precision pseudopotentials, and a $12 \times 12 \times 12$ mesh. The calculation returned a direct band gap of 0.696 eV at Γ — magnitude prediction accurate to 0.6%, with correct direct-gap character. This is the strongest evidence that pattern-grade knowledge functions as a predictive scientific model rather than a descriptive summary.

Implications. The execution sessions generated the empirical substrate. Dedicated review sessions compressed that substrate into higher-order, provenance-linked patterns in under three minutes. The design session then used one of those patterns to guide an efficient new calculation. The full progression from observation to pattern to prediction was reproducible from recorded data and robust across two different models.

Availability. Complete experimental artifacts, including full prompts, agent traces, knowledge database snapshots, and reviewed-insight tables, are archived on Zenodo at https://doi.org/10.5281/zenodo.20588420 under CC-BY 4.0.

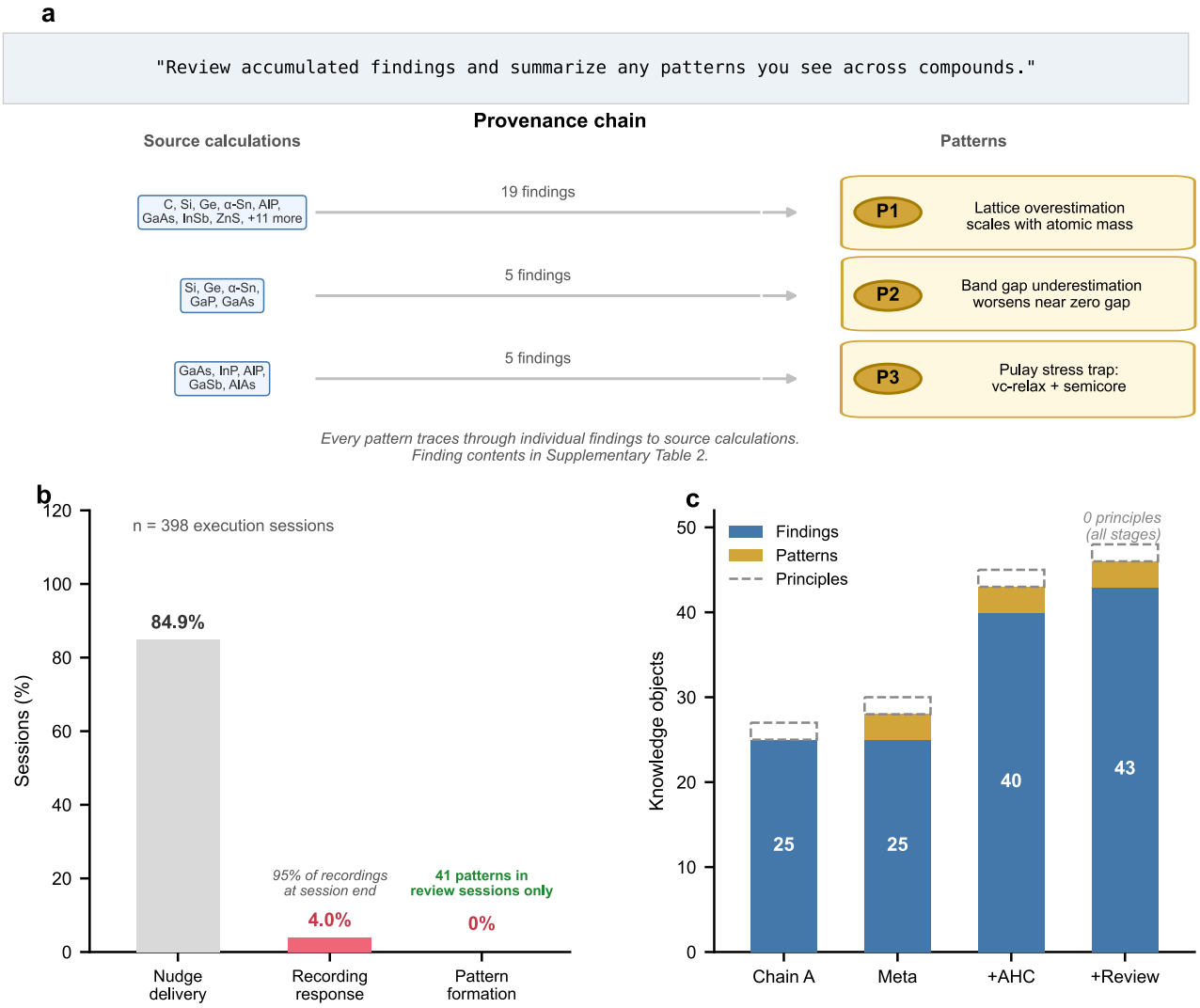

Figure 5. **Knowledge consolidates through dedicated reflection. a**, Three patterns distilled from 25 findings by a dedicated reflection session (Pattern 1: lattice overestimation scaling; Pattern 2: band gap underestimation; Pattern 3: Pulay stress trap), with provenance chains linking each pattern to its source calculations. **b**, Nudge compliance across 398 execution sessions: 84.9% received recording nudges, 4.0% showed any mid-session recording, 0% produced patterns during execution. All 41 patterns were produced exclusively in dedicated review sessions. **c**, Knowledge grade evolution across experiments: findings accumulate during execution (25→43), patterns appear only after dedicated reflection (0→3), principles remain an open frontier.

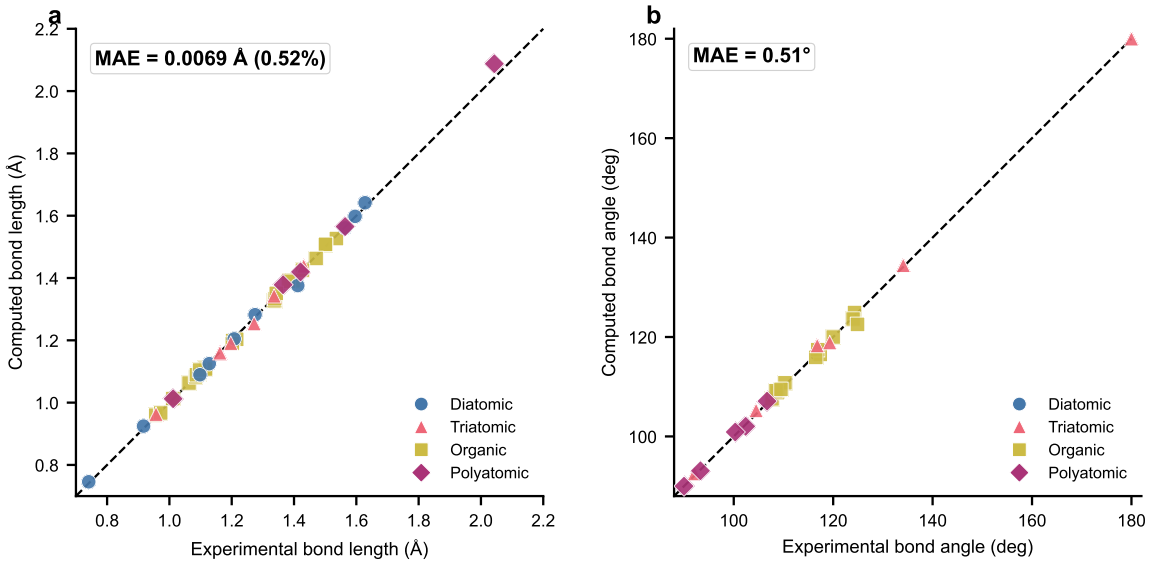

Figure 6. **Cross-engine, cross-model validation: molecular geometry optimization with ORCA 6.1.1 and GPT 5.4. a**, Computed versus experimental bond lengths for 78 comparisons from the reference-backed raw-ORCA-converged subset (MAE = 0.0069 Å, 0.52%), consistent with established hybrid-DFT benchmarks. **b**, Computed versus experimental bond angles for 33 comparisons (MAE = 0.51°). This experiment uses a different simulation engine (ORCA vs. Quantum ESPRESSO), a different AI model (GPT 5.4 vs. Claude), and a different chemical domain (molecular vs. solid-state), confirming that the platform generalizes across all three axes. Of 98 molecules, 91 completed under the conservative publication criterion; the seven final failures comprise two agent-declared failures despite converged raw outputs, three missing or non-terminated last-output cases, and two explicit last-attempt error terminations.