Grounded autonomous scrutiny at scale: emergent critique from reproduction of published computational physics papers

Anonymous Authors[1]

Abstract

Autonomous LLM agents now produce complete research artifacts in machine-learning sandboxes, but real computational physics is harder: experiments are first-principles calculations against re-runnable physical ground truth, and meaningful new work almost always builds on a key existing paper. We ask whether such an agent can perform *grounded scrutiny* of published computational physics — reading a paper, reproducing it from scratch, and surfacing methodological concerns from execution. We deploy a single Claude Opus 4.6 configuration at two complementary scopes. At scale, across 111 open-access Quantum ESPRESSO papers, an autonomous agent runs the read–plan–compute–compare loop and, although never asked to critique, raises substantive methodological concerns on $\sim$42% of papers; **86 of 88 of these critiques (97.7%) surface only after the agent has actually run a calculation, with a reading-only ceiling of 0.9%**. Critique emerges from reproduction, not from reading. In depth, on one *Nature Communications* paper on multiscale device simulation of a 2D-material MOSFET, a fresh agent inheriting a verified reproduction pipeline autonomously produces a 14-concern physics inventory and a six-page publishable Comment that revises the paper's $L_G = 5$ nm headline. Two of its $L_G = 5$ nm headline-challenging attacks — a source-degeneration contact-resistance bound and a Sb-doping degradation ratio — are absent from the published 21-reviewer peer review.

1. Introduction

Recent work has shown that an autonomous LLM agent can navigate the entire research life cycle of a machine learning project — ideating, coding, running the training experiment, analyzing data, and writing the manuscript (Lu et al., 2026). The natural next question is whether the same kind of system can take on real-world physical science, where experiments are not training loops but first-principles calculations of real physical systems. **A fundamental difference between sandboxed automated research and real physical science is that the latter is computationally and intellectually demanding — physics reasoning that cannot be reduced to interpolation, multi-scale calculations on decades-mature scientific software, and verifiability against re-runnable physical ground truth — so meaningful new work almost always begins from a key existing paper.** A researcher reads the relevant literature, reproduces the central calculations in their own setup, critically evaluates what those calculations show, conceives what is missing or wrong, runs follow-up calculations to test the new idea, and possibly writes a publishable document. We refer to this stance — audit by execution against the same physics the literature describes — as *grounded scrutiny*.

In this work we ask whether an autonomous LLM agent can perform grounded scrutiny on its own, in computational physics — the natural testbed because numerical claims can be independently re-computed against the same physics, and community effort has established reproducibility standards for the simulations themselves (Lejaeghere et al., 2016; Bosoni et al., 2024). We focus on density functional theory (DFT) and the Quantum ESPRESSO ecosystem (Giannozzi et al., 2009; 2017), and deploy a single Claude Opus 4.6 configuration at two complementary scopes: **at scale**, where a fresh agent reproduces an arbitrary published paper end-to-end; and **in depth**, where a single agent is unleashed on one carefully chosen paper to push reproduction-driven scrutiny as far as it will go.

At scale, across 111 open-access Quantum ESPRESSO papers, the agent autonomously runs the read–plan–compute–compare loop and reproduces roughly three-quarters of in-scope claims within 5% of the published value. Unprompted, it raises substantive methodological concerns on $\sim$42% of papers — **97.7% of which require execution to surface, and only 0.9% are catchable by reading alone**. *Critique emerges from reproduction, not from reading*: critical sci-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

entific scrutiny is execution-bound, not a property of prior knowledge.

In depth, for one *Nature Communications* paper on multi-scale device simulation of a 2D-material MOSFET (Pizzi et al., 2016), a single agent goes well beyond reproduction. Given a verified multi-code reproduction pipeline, in one unsupervised session it inventories its physics concerns about the paper, runs three classes of new calculation the original work did not perform, and produces a six-page publishable Comment — composed, figured, typeset, and PDF-iterated entirely by itself — whose two main findings revise the paper's $L_G = 5$ nm headline conclusion. Neither finding appears in the published peer review, a complementarity we examine in §4.2. **Unlike prior work in which an LLM reads a paper and flags errors by reading alone (Son et al., 2025), every numerical claim in our Comment is backed by a first-principles calculation the agent itself runs against the same physics the original paper used.** Scrutiny here is computational, not textual.

The capabilities exercised — multi-scale physics reasoning, autonomous execution against re-runnable ground truth, and synthesis into a publishable artifact — are precisely those a full real-world research loop demands; grounded scrutiny is the natural first stage of that loop and the central object of this paper.

2. Grounded scrutiny at scale

2.1. Setup

The corpus is **111 unique Quantum ESPRESSO papers** in which QE is the primary computational tool, filtered from all open-access QE literature (Giannozzi et al., 2009; 2017) published 2010–2024 across 12 journal families (Appendix A M1); restricting to open-access literature and open-source software lets us release every input, output, and trace. Across both the scale and depth regimes, the harness is the same: the **Claude Code CLI** as agentic orchestrator with **Claude Opus 4.6** as the underlying model, shelling out to bash for QE, Wannier90 (Pizzi et al., 2020), and any analysis the agent writes itself in Python. There is no central tool layer — no MCP server, no library wrapper — by deliberate choice, to keep the harness honest about what the model does unaided. This differs from contemporary DFT-agent systems (Wang et al., 2025; Kumar et al., 2026; Zou et al., 2025) that serve as execution layers for human-specified tasks; none reads a published paper. The scale-mode pipeline (Fig. 1a) wraps this harness in a Python outer loop that iterates over the corpus, handing each paper to a fresh agent under a boilerplate prompt that converged in development and is then held fixed for production (Appendix A M2, Appendix H). The agent receives the paper plus a small knowledge envelope covering core QE and the Wannier90 ecosystem.

Each agent is told to do four things in order: load the **full paper into context** — information is interleaved across a paper in ways that require coherent understanding rather than discrete retrieval; write a structured reading summary with its plan and targets; execute calculations serially while updating a worklog; and emit a structured verdict. A wall-clock soft cap of **2–4 h** per paper is set as a flexibility budget rather than a hard timeout (Appendix A M2); agents declare their own scope.

2.2. Aggregate reproduction quality

Within the agent's capability envelope, reproduction is uniformly strong. Papers are graded on a four-tier scale (T4 exact, T3 qualitatively correct, T2 partial, T1 failure); *in-scope* is what fits the time budget, and agents declare scope in the verdict (Appendix A M4). Across **571 deduplicated quantitative claims**, the agent matches **75.8% within 5%** and **83.2% within 10%** of the published value, with a **median deviation of 0.9%**. At the paper level, 6 of 111 reach T4 and **90.9% of the 110 scorable papers** reach T3 or T4. Figure 1b shows lattice constants across a 24-paper subset (63 claims): median |dev| 0.25%, 62/63 within 5% — the cleanest single-quantity calibration.

Two small knowledge files were nevertheless operationally important. Early runs frequently produced false refusals — agents prematurely concluding "QE cannot do this" and skipping calculations they should not have. We tested this with a controlled ablation on 15 open-access papers under otherwise identical conditions, adding two compact text files (QE command idioms; pseudopotential-selection heuristics (Prandini et al., 2018)) and a rule requiring the agent to consult them before declaring a capability absent. The false-refusal class was eliminated and attempted-workflow breadth expanded materially: phonon workflows attempted on **15% of papers rose to 29%** on the same set. The agent did not become more capable; it stopped declining capabilities it already had. **Two small text files unlock cognitive scope the model already has: the binding constraint is the knowledge harness, not model capability.**

2.3. Agent behaviour during a session

Full trace analysis (Appendix G) reveals a uniform **load-then-execute-serially** pattern: agents ingest paper and knowledge files first, write the reading summary, then interleave QE-input creation with execution, closing with the verdict. Three observations matter for §2.4. Once the paper is in context, agents essentially never re-read it (nearly all sessions show zero late reads), treating the initial load and reading summary as sufficient working state. Under context compaction (16% of sessions), agents rely on their on-disk QE inputs, outputs, and worklog rather than re-reading —

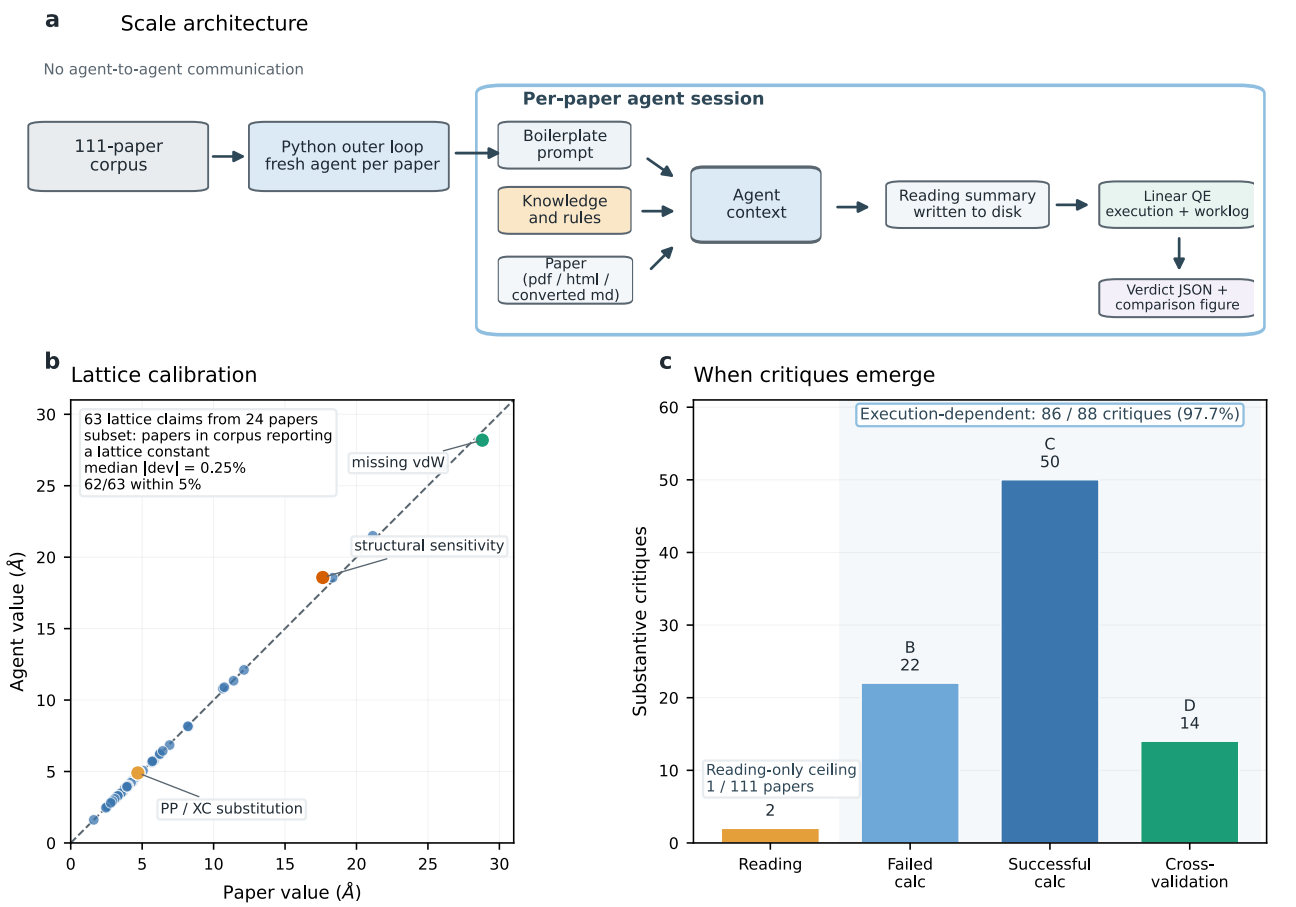

Figure 1. **Grounded scrutiny at scale: architecture, calibration, and execution-dependence. (a)** Two-level pipeline: a Python outer loop iterates over the 111-paper corpus, handing each paper to a fresh Claude Opus 4.6 agent running inside the Claude Code CLI; there is no agent-to-agent communication. Within each per-paper session, three fixed inputs — a boilerplate task prompt, a required-reading envelope of knowledge and house rules, and the paper itself — feed the agent's working context, from which the agent writes a reading summary to disk, executes QE serially alongside a worklog, and emits a verdict JSON and comparison figures. **(b)** Lattice-constant calibration on the 24-paper subset reporting a lattice constant (63 claims), agent vs paper on a 45° line; median $|\text{dev}| = 0.25\%$, 62/63 within 5%. Outliers are labelled by dominant systematic: PP/XC substitution, missing vdW, structural sensitivity. **(c)** Phase classification of the 88 substantive critiques: A = 2 (during reading), B = 22 (after a failed calculation), C = 50 (after a successful calculation compared to the paper), D = 14 (after cross-validation). Execution-dependent: 86/88 = 97.7%. Reading-only ceiling: 1/111 papers = 0.9%.

the filesystem becomes external memory, an emergent strategy we did not prompt for; this is consistent with long-standing observations that LLM compliance is strongest on verifiable file-level artefacts (Zhou et al., 2023). And agents are **open-loop on visual output**: in zero of 61 inspected sessions does an agent open a figure it just wrote, notice a problem, and regenerate. We return to this in §5.

2.4. Emergent scrutiny and the execution requirement

Nothing in the production prompt asks the agent to critique the paper; the instruction is to reproduce, and deliverables are reproduction artifacts. With that caveat, the spontaneous critique rate is striking: $\sim$42% **of papers contain at least one substantive methodological concern** raised by the agent unprompted, across nine categories (Appendix F). The total number of substantive critiques across the corpus — a paper can contribute more than one — is **88**.

The more interesting question is *when* these critiques surface. Figure 1c classifies each post hoc by phase: during reading (A), after a failed calculation (B), after a successful calculation compared to the paper (C), or after cross-validation (D). Counts: **A = 2, B = 22, C = 50, D = 14**.

Eighty-six of eighty-eight critiques (97.7%) emerged only after the agent had actually run a calculation. The two Phase-A critiques are one file-corruption artifact and one genuine reading catch (a phosgene adsorption energy reported at 142 Ry, three-to-five orders beyond physical), giving a reading-only ceiling of **1 in 111 pa-**

pers (0.9%). Phase C is the modal discovery point — the agent runs, compares, notices — and Phase B is the error-recovery path where diagnosing a failed run reveals a paper-side concern. The 97.7% execution requirement complements passive-reading verification baselines (∼21% recall on SPOT's 91 retraction/errata-level errors (Son et al., 2025)) and code-provided reproduction baselines in the 19–27% range (CORE-Bench (Siegel et al., 2024), PaperBench (Starace et al., 2025), ReplicationBench (Ye et al., 2025)). To our knowledge this is the **first quantitative evidence that critical scientific scrutiny is execution-bound in an autonomous-agent corpus of this size**. On a 12-paper two-machine cross-check with no shared state, 7/12 produced overlapping critique category sets (5 substantive, 2 convergence-only) — qualitatively consistent scrutiny patterns across independent runs.

2.5. Inside the envelope

Within the envelope, agents execute multi-code workflows autonomously across DFT+U, anomalous Hall conductivity, spin–orbit coupling, LDA+U correlated magnetism, `epsilon.x` optics, and DFPT dielectrics — a six-panel agent-vs-paper diversity gallery is in Figure 4 (Appendix B). Representative autonomous depth within one session includes full SCF → NSCF → Wannierization → MLWF hopping pipelines (Marzari et al., 2012) on infinite-layer nickelates (nearest-neighbour hopping to 1.1%, bandwidths to 0.3%), and 44-MLWF Wannierizations with Berry-curvature integration (Wang et al., 2006) recovering Mn_3Al σ_{xy} within 4%. **Scale mode is not running `pw.x` and reading off a gap** — it autonomously executes the QE ecosystem as the paper demands.

Spontaneous discoveries have a concrete texture (four per-phase vignettes in Appendix F): on a WS_2 monolayer paper (Phase C), the agent's fully-relativistic calculation gave spin–orbit splitting 429 meV against the paper's 571 meV, with MoS_2 reproducing the paper within 2% as an internal control and experiment (400–410 meV, supplied post hoc) agreeing with the agent. The envelope has limits scale alone cannot cross — the 2–4 h budget and the QE + Wannier90 knowledge scope, both prices of a uniform pipeline across a hundred-plus papers. Can an agent go deep on a single paper, all the way to a physically meaningful verdict and on to a publication-shaped scientific artifact?

3. From scale to depth: a test case on the boundary

One paper in the corpus sits exactly on that boundary: Pizzi et al., arsenene and antimonene double-gate MOSFETs (*Nat. Commun.*, 2016) (Pizzi et al., 2016), a first-principles device study arguing that 2D As and Sb form ultra-scaled sub-10 nm MOSFETs whose performance meets the ITRS industry roadmap. The paper combines four codes — QE (Giannozzi et al., 2009; 2017), Wannier90 (Pizzi et al., 2020), NanoTCAD ViDES (Marian et al., 2023; Fiori & Iannaccone, 2005), and custom post-processing — across DFT structure, hybrid-functional bandgaps (Heyd et al., 2003; 2006), ballistic NEGF transport, phonon scattering, and device figures of merit. Its central claim, on which its scientific interest rests, is the sub-10 nm ITRS-compliance of these devices. In scale mode the agent reproduces the QE-accessible claims at strong fidelity and stops at Wannier90, because NanoTCAD ViDES sits outside the knowledge envelope. Pizzi 2016 is close enough to the envelope that the agent gets most of the way autonomously, and far enough beyond it that completing the chain exercises depth-mode grounded scrutiny end-to-end.

4. Reproduce–Review–Reflect: depth-mode grounded scrutiny

The depth pipeline (Fig. 2a) is a three-stage sequence on one paper: **Reproduce** once, build a verified end-to-end reproduction across all four codes; then **Review** and **Reflect**, two fresh autonomous agent sessions that, respectively, audit the paper against the verified pipeline and upgrade the audit into a scientific Comment.

4.1. Reproduce: building a verified reference pipeline

The Reproduce stage on Pizzi 2016 was carried out as human–agent collaboration, almost entirely because of the legacy state of the transport solver the paper relies on. NanoTCAD ViDES (Marian et al., 2023; Fiori & Iannaccone, 2005) was last updated in 2016 and carries silent failure modes that require instrumentation to surface (engineering-challenges summary in Appendix I). The human work was mainly tool repair; the downstream multi-code reproduction across QE → Wannier90 → instrumented NanoTCAD — pseudopotential selection, convergence parameters, and calibration against the paper's published figures of merit — was the agent's. This is a one-time tool-engineering cost, not a recurring per-paper cost; the resulting verified pipeline is what the autonomous Review and Reflect stages inherit.

4.2. Review: a 14-concern inventory, four attacks, and comparison with human peer review

The Review agent received the paper plus the verified reproduction pipeline under a "physics-first, tools-second" prompt (Appendix A M3, Appendix H): write a structured concerns inventory before any tool use, then pursue attacks. This ordering is itself the result of a prompt ablation — a tools-first inventory, obtained by reading paper and tooling knowledge together as in scale mode, produces a narrower list of attacks anchored on what the tools easily compute

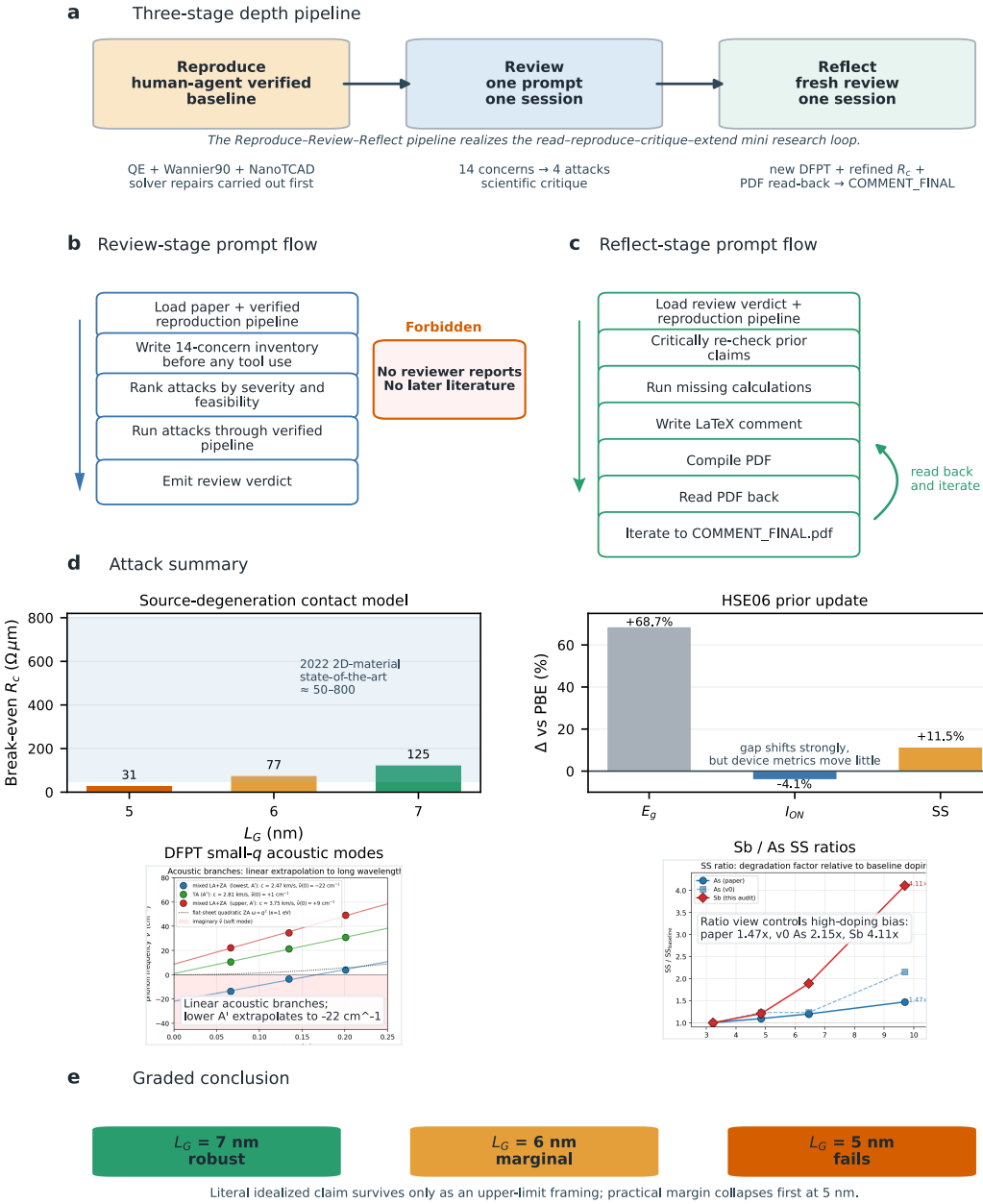

Figure 2. **The Reproduce–Review–Reflect pipeline applied to Pizzi 2016. (a)** Three-stage flow: Reproduce (human–agent verified baseline across QE + Wannier90 + NanoTCAD) → Review (one session; 14-concern inventory, four attacks) → Reflect (fresh session; new DFPT, refined R_c, PDF read-back loop → COMMENT_FINAL). **(b)** Review-stage prompt flow: load paper + verified pipeline → write 14-concern inventory before any tool use → rank attacks → run attacks → emit review verdict. Forbidden inputs: reviewer reports, subsequent literature. **(c)** Reflect-stage prompt flow: load verdict → re-check claims → run missing calculations → write LaTeX → compile → read PDF back → iterate to COMMENT_FINAL.pdf. **(d)** Attack summary: source-degeneration R_c = 31/77/125 $\Omega\cdot\mu$m at L_G = 5/6/7 nm vs 2022 2D-material state-of-the-art ≈ 50–800 $\Omega\cdot\mu$m (Shen et al., 2021); HSE06 prior update (E_g +68.7%, I_{ON} −4.1%, SS +11.5%); DFPT small-q acoustic (linear branches; lower A′ → -22 cm^{-1}); Sb/As SS ratios (paper 1.47×, pipeline As 2.15×, Sb 4.11×; Sb-vs-As acceleration 1.91×). **(e)** Graded conclusion: $L_G = 7$ nm robust, 6 nm marginal, 5 nm fails.

(Appendix A M3). The Review agent ran 2 h 11 min and issued 239 tool calls. It produced a **14-concern physics inventory** (up from 5 under the tools-first ablation), ranked by severity and feasibility, attempted **four computational attacks** end-to-end (contact resistance, HSE+SOC bandgap, an analytical flexural-phonon argument, a Sb doping sweep), and ran **nine further cross-checks** validating individual paper claims. We summarize two attacks — one headline-

challenging (contact resistance), one null-result control where the agent's own initial intuition is overturned by the calculation (HSE prior update).

Attack A — contact resistance. The paper sets source/drain contact resistance to zero in one sentence and never quantifies it. The agent post-processed the paper's own f_T and τ with two analytical series-resistance models, finding break-even contact resistances at $L_G = 5$ nm well below the state-of-the-art for any 2D contact ($\sim$50–800 $\Omega{\cdot}\mu$m) (Shen et al., 2021); the Reflect-stage refinement (§4.3) later replaced these toy bounds with a source-degeneration model yielding **31/77/125 $\Omega{\cdot}\mu$m at $L_G = 5/6/7$ nm**. Under any realistic contact resistance, the $L_G = 5$ nm headline collapses.

Attack B — HSE+SOC bandgap and the prior-update story. Unprompted by paper or prompt, the agent proposed HSE itself from the inventory, on the standard observation that PBE (Perdew et al., 1996) underestimates gaps by 30–50% and the naïve expectation — it stated explicitly — that a larger gap should give a steeper subthreshold slope through reduced band-to-band tunneling. It ran a full HSE06+SOC pipeline through Wannier and back into the device simulation. The gap rose by **+68.7%**, as expected, but the device-level result inverted the prior: I_{ON} shifted **$-$4.1%**, local SS shifted **+11.5%** (PBE $\rightarrow$ HSE), because the gate work-function offset absorbs approximately $\Delta E_g/2$ and the carrier-relevant conduction-band alignment is approximately conserved (Appendix A M5). The agent recognised the inversion in its worklog and scoped the falsification to this attack rather than over-generalizing. **The agent originated the physics question, held an incorrect prior, and updated itself in writing once the calculation refuted it — the failure mode that grounded execution is structurally protective against (§5).**

The other two Review attacks bore on the headline indirectly. On flexural-phonon scattering, Review made an analytical argument that the paper's Takagi-formula treatment was inadequate but did not have time to back it computationally. On Sb doping, Review ran the sweep Pizzi 2016 omitted and obtained a qualitative scaling, but the final bracket did not fully enclose the I_{OFF} target. Both open threads were picked up in Reflect (§4.3).

Comparison with the published peer review. *Nat. Commun.* publishes its peer review files; Pizzi 2016 received 21 reviewer concerns across two rounds. The Review agent was forbidden from reading reviewer reports or subsequent literature, creating an orthogonal information asymmetry: the agent has computational reach without literature fluency; reviewers have the converse. Figure 3 shows the overlap as a sparse 14×21 matrix. Under our coding scheme, two Review concerns map directly to referee concerns (SAME), two are loose family overlaps (LOOSE), and the remaining ten are new — the exact counts depending on a small number of judgment calls (Appendix D). The more load-bearing observation is qualitative and robust: **both $L_G = 5$ nm headline-challenging attacks — contact resistance (P9) and Sb doping (P12) — are Review-only.** To our knowledge this is the first end-to-end existence proof that an autonomous agent, starting from a published paper, produces grounded findings on its central claim that human peer review did not. **Agent and human peer review have orthogonal attack surfaces, and their union is strictly larger than either alone**; humans do not run calculations during review.

4.3. Reflect: from Review verdict to a grounded follow-up artifact

Review left two scientific threads open — the analytical phonon argument and the incomplete Sb doping bracket — alongside the usual marks of a single audit run: tone problems ("FALSIFIED everything"), an arithmetic error in the phonon mean-free-path scaling, and a binary verdict where a graded one is more honest. We handed the Review corpus to a fresh agent under a boilerplate prompt (Appendix H): review the audit, verify what is verifiable, run the calculations Review did not, produce a scientific Comment. The agent converged end-to-end on a publication-shaped artifact without human intervention, closing the audit by reading its own compiled PDF back into context and iterating (Fig. 2c; Appendix E).

The output, **COMMENT_FINAL.pdf**, is a six-page scientific Comment with figures, references, methods, and an explicit list of open questions; substantive improvements are in Fig. 2d,e and summarised as $L_G = 7$ nm robust / 6 nm marginal / 5 nm fails. None of the three pieces of work Reflect added is interpolation from the paper's parameters. The **source-degeneration contact model** is a different physical model from the Review-stage toy bound: it identifies that the paper's own f_T and τ already constrain the maximum extractable transconductance, derives the resulting break-even contact resistance from first principles, and gives a ceiling independent of any external 2D-contact-resistance benchmark. The **DFPT phonon calculation** replaces Review's analytical Takagi argument with first-principles DFPT on freestanding arsenene, recovering an in-plane acoustic branch linear in q with a soft long-wavelength mode (lower A′ extrapolating to -22 cm^{-1}) the analytical treatment was structurally incapable of capturing. The **Sb high-doping closing point** completes Review's sweep at the doping level that actually brackets the I_{OFF} target, converting the qualitative scaling into a quantitative degradation ratio (Sb degrades $1.91\times$ faster than As under matched doping). **The Comment is the form, but the content is grounded follow-up scientific work**: three classes of calculation that go beyond what Pizzi 2016 reported, packaged in the shape of an object that

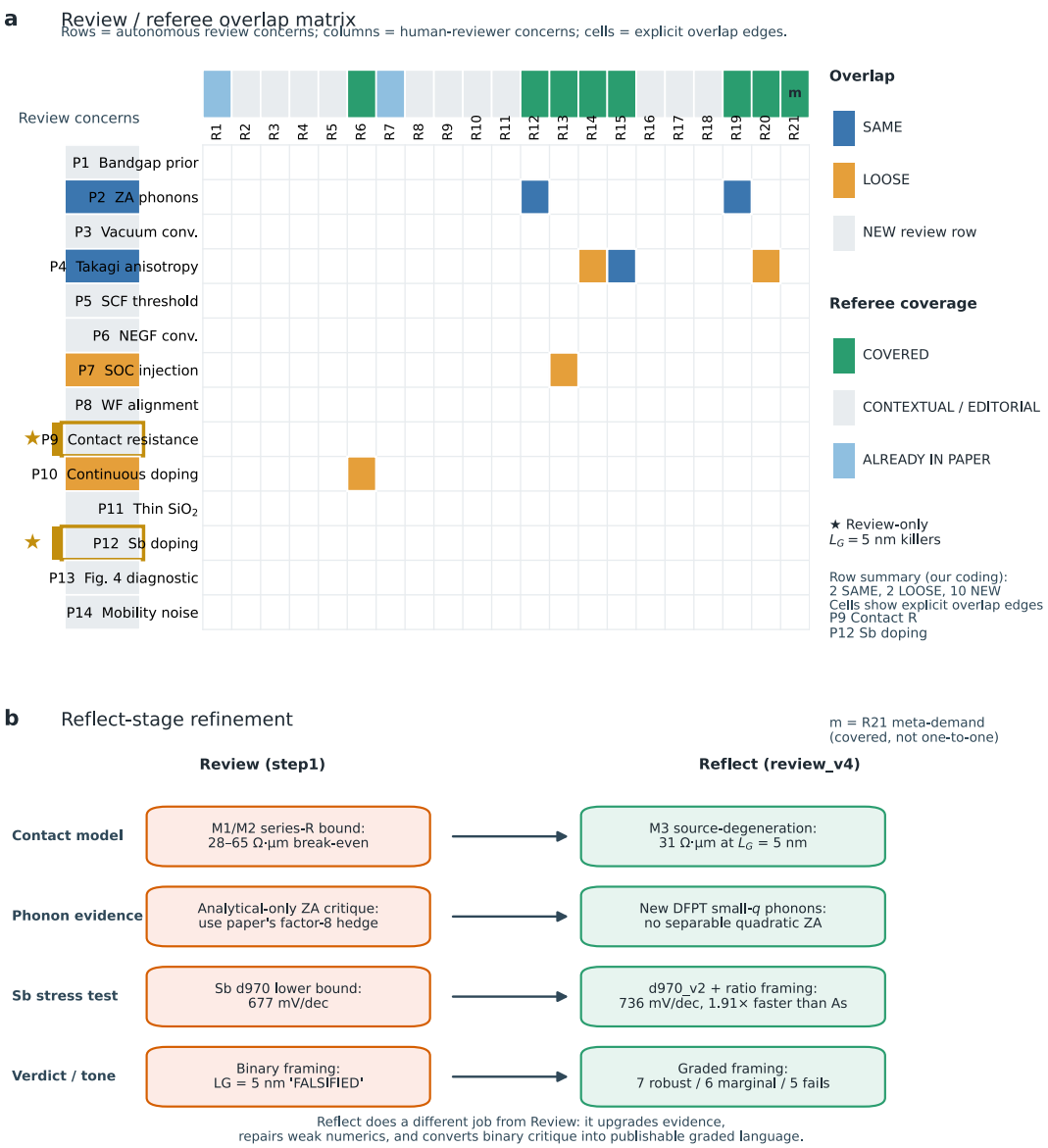

Figure 3. **Review ↔ referee overlap and Reflect-stage refinement.** *Top:* sparse 14 × 21 overlap matrix, rows = Review concerns (P1–P14), columns = referee concerns (R1–R21). Row-level coding (Review side, SAME/LOOSE/NEW) summarises each Review concern's strongest overlap class. Under our coding scheme: **2 SAME, 2 LOOSE, 10 NEW** (Review side); 8 COVERED / 11 EDITORIAL-or-CONTEXTUAL / 2 ALREADY-IN-PAPER (referee side); Appendix D gives the full rule-set. Gold outline and star mark the two Review-only $L_G = 5$ nm headline-challenging attacks, P9 (contact resistance) and P12 (Sb doping). *Bottom:* four paired Review → Reflect refinement rows: contact modelling M1/M2 bound (28–65 Ω·μm) → M3 source-degeneration (31 Ω·μm at $L_G = 5$ nm); analytical ZA-phonon critique → first-principles DFPT; Sb stress test (d970 lower bound → d970_v2 ratio framing, 1.91× faster than As); binary "FALSIFIED" verdict → graded (7 robust / 6 marginal / 5 fails).

could be submitted.

5. Discussion

Computation as the central epistemic act. The failure modes documented in the AI Scientist literature (Lu et al., 2026; Beel et al., 2025) — hallucination (Sui et al., 2024; Huang et al., 2024), judge–writer loops, naïve idea generation — are failures of the **blank-slate paradigm**, where the system generates content with no physical reference. Our results are for a different paradigm, *grounded scrutiny*, where every step is anchored against an external physical reference: the paper's numbers (themselves validated against experiment or independent first-principles work) and the

agent's own re-runnable simulations of the same physics. No hallucinated number can survive — running the calculation either reproduces the target or does not, and the judge is the physics. *Grounding in published physical science is structurally protective against hallucination modes any blank-slate generative system must contend with — not as a feature of our implementation, but as a property of grounded scrutiny itself.* The 97.7% execution requirement at scale and the HSE prior-update at depth are the same finding at two scales: physical reasoning is not the bottleneck; *running the thing* is. The thesis of this paper is the same sentence read in either direction — *critique emerges from reproduction, not from reading*.

What grounded scrutiny demands. The capabilities exercised end-to-end here are exactly the three the introduction identifies as discriminators of real physical science: physics reasoning that cannot be reduced to interpolation (pseudopotential, functional, k-mesh, convergence criterion); multi-scale execution on decades-mature scientific software (QE $\rightarrow$ Wannier90 $\rightarrow$ NEGF, with the paper's prose and figures as the only spec); and verifiability against re-runnable physical ground truth at every step. Every step is physics-level reasoning, not text-level pattern matching.

Harness, not model. Every limitation we encountered traces to the harness, along four engineering-addressable axes. (i) **Knowledge.** Two text files unlock cognitive scope the model already has (§2); required-reading envelopes are a primitive solution, and a richer structured tool layer is the natural next step. (ii) **Tools.** The Reproduce-stage cost on Pizzi 2016 was itself a tool-maturity problem; a tool layer with well-engineered solvers amortises that cost across future depth cases. (iii) **Compute resource management and planning.** Agents over-subscribe cores, leak subprocesses on long sessions, and even when told to spend unlimited time bias toward narrowing scope and finishing — both reflect using a short-task coding CLI as a long-horizon research orchestrator. (iv) **Visual capability.** Agents miss topological differences between band-structure plots, extract values from plots with large error, and never look back at their own figures — a multimodal-reading problem the harness can solve through programmatic figure checks and structured plot data. Scaling grounded scrutiny means scaling the harness along all four axes, not waiting for a new model.

Outlook. Grounded scrutiny is the natural first stage of a full real-world research loop — read, reproduce, critique, extend. The depth case here already does the third and a measure of the fourth; an agent that originates its own scientific question and writes a full paper beyond an incremental Comment is the natural extension. Separately, a direct practical use already exists: deployed alongside human peer review, pre-verified reproduction pipelines plus autonomous Review and Reflect stages give the literature a second epistemic mode it does not currently have — not "was this paper read carefully" but "was this paper *run*" — and could function as a fully automated complement to the existing process.

References

Agapito, L. A., Curtarolo, S., and Buongiorno Nardelli, M. Reformulation of DFT+U as a pseudohybrid Hubbard density functional for accelerated materials discovery. *Physical Review X*, 5:011006, 2015.

Beel, J., Kan, M.-Y., and Baumgart, M. Evaluating Sakana's AI Scientist: Bold claims, mixed results, and a promising future? *ACM SIGIR Forum*, 59:1–20, 2025.

Bosoni, E. et al. How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. *Nature Reviews Physics*, 6:45–58, 2024.

Fiori, G. and Iannaccone, G. NanoTCAD ViDES. *Journal of Computational Electronics*, 4:63–66, 2005.

Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21: 395502, 2009.

Giannozzi, P. et al. Advanced capabilities for materials modelling with QUANTUM ESPRESSO. *Journal of Physics: Condensed Matter*, 29:465901, 2017.

Hasan, T. et al. Strain-dependent electronic and optical properties of boron-phosphide and germanium-carbide hetero-bilayer: A first-principles study. *AIP Advances*, 10:085128, 2020.

Heyd, J., Scuseria, G. E., and Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *Journal of Chemical Physics*, 118:8207–8215, 2003.

Heyd, J., Scuseria, G. E., and Ernzerhof, M. Erratum: "Hybrid functionals based on a screened Coulomb potential". *Journal of Chemical Physics*, 124:219906, 2006.

Huang, J. et al. Large language models cannot self-correct reasoning yet. In *Proceedings of ICLR*, 2024.

Kumar, S. G. H. et al. El Agente Sólido: A new age(nt) for solid state simulations. Preprint at arXiv:2602.17886, 2026.

Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science*, 351:aad3000, 2016.

Lu, C. et al. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026.

Marian, D., Marin, E. G., Perucchini, M., Iannaccone, G., and Fiori, G. Multi-scale simulations of two dimensional material based devices: the NanoTCAD ViDES suite. *Journal of Computational Electronics*, 22:1327–1337, 2023.

Marzari, N., Mostofi, A. A., Yates, J. R., Souza, I., and Vanderbilt, D. Maximally localized Wannier functions: Theory and applications. *Reviews of Modern Physics*, 84: 1419–1475, 2012.

Nomura, Y., Nomoto, T., Hirayama, M., and Arita, R. Magnetic exchange coupling in cuprate-analog d^9 nickelates. *Physical Review Research*, 2:043144, 2020.

Park, M., Han, G., and Rhim, S. H. Anomalous Hall effect in a compensated ferrimagnet: Symmetry analysis for Mn_3Al. *Physical Review Research*, 4:013215, 2022.

Perdew, J. P., Burke, K., and Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters*, 77:3865–3868, 1996.

Pizzi, G. et al. Performance of arsenene and antimonene double-gate MOSFETs from first principles. *Nature Communications*, 7:12585, 2016.

Pizzi, G. et al. Wannier90 as a community code: new features and applications. *Journal of Physics: Condensed Matter*, 32:165902, 2020.

Prandini, G., Marrazzo, A., Castelli, I. E., Mounet, N., and Marzari, N. Precision and efficiency in solid-state pseudopotential calculations. *npj Computational Materials*, 4: 72, 2018.

Priem, J., Piwowar, H., and Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint at arXiv:2205.01833, 2022.

Reyes-Retana, J. A. and Cervantes-Sodi, F. Spin-orbital effects in metal-dichalcogenide semiconducting monolayers. *Scientific Reports*, 6:24093, 2016.

Shen, P.-C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature*, 593: 211–217, 2021.

Siegel, N., Kapoor, S., Nagdir, N., Stroebl, B., and Narayanan, A. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2024.

Son, G. et al. When AI co-scientists fail: SPOT—a benchmark for automated verification of scientific research. Preprint at arXiv:2505.11855, 2025.

Starace, J. et al. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of ICML*, volume 267 of *PMLR*, pp. 56843–56873, 2025.

Sui, X. et al. The surprising value of the confabulated: how LLM hallucinations support a creative revision process. In *Proceedings of ACL*, 2024.

Wang, J. et al. Layers dependent dielectric properties of two dimensional hexagonal boron nitride nanosheets. *AIP Advances*, 6:125126, 2016.

Wang, X., Yates, J. R., Souza, I., and Vanderbilt, D. Ab initio calculation of the anomalous Hall conductivity by Wannier interpolation. *Physical Review B*, 74:195118, 2006.

Wang, Z. et al. DREAMS: Density functional theory based research engine for agentic materials simulation. Preprint at arXiv:2507.14267, 2025.

Ye, C. et al. ReplicationBench: Can AI agents replicate astrophysics research papers? Preprint at arXiv:2510.24591, 2025.

Zhou, J. et al. Instruction-following evaluation for large language models. Preprint at arXiv:2311.07911, 2023.

Zou, Z. et al. El Agente: An autonomous agent for quantum chemistry. *Matter*, 8:102263, 2025.

A. Methods

M1. Corpus construction. The scale-mode corpus was assembled from an OpenAlex (Priem et al., 2022) snapshot (March 2026) of all papers citing Quantum ESPRESSO, yielding **30,779 candidates**. We filtered in three passes. (i) Asset availability: **10,926 entries** retaining both full-text PDF and supplementary material downloadable without institutional authentication. (ii) Open-access verification: **4,708 entries** with a publicly licensed route, of which **2,641** were verified by direct download. (iii) Primary-tool filter: an LLM classifier (Claude Sonnet 4.6) read each paper and identified **708 entries** in which Quantum ESPRESSO is the primary computational tool. From this pool we drew a stratified random sample plus earlier development batches used for prompt and knowledge-envelope iteration. Work proceeded in five batches: B1–B4 were development runs used to converge the prompt, the knowledge envelope, and the harness; B5 is the single-machine controlled production run. Corpus-wide reproduction-quality and critique-phase statistics are computed on the full deduplicated 111-paper OA corpus; the $\sim 42\%$ paper-level critique rate is 47/111. The controlled knowledge-ablation referenced in §2 is the B3→B4 comparison on a fixed 15-paper open-access subset (same model, same compute, only the knowledge files and the consult-before-refuse rule added). The cross-machine reproducibility figure uses a 12-paper subset run on a second workstation with no shared state, filesystem, or prompt history.

M2. Harness, prompt, and required-reading envelope. Both scale and depth regimes use the same harness: Claude Code CLI + Claude Opus 4.6, on a single workstation with ~ 12 CPU cores and ~ 64 GB RAM. The agent interacts with the system only through bash shell calls. **No central tool layer** is exposed: no MCP server, no library wrapper. This is a deliberate choice to make trace analysis tractable and to keep the harness honest about what the model can do with shell access alone.

Scale mode wraps this harness in a Python outer loop that iterates over the corpus, spinning up a fresh agent per paper with a 2–4 h wall-clock soft cap framed as a flexibility budget; scope reductions are declared in the verdict. Each agent receives five required-reading files that define the envelope: `HOUSE_RULES.md` (resource-management discipline, time-budget discipline, honesty conventions); `READING_GUIDE.md` (structure of the reading summary and the reproduction-target list); `VERDICT_FORMAT.md` (JSON schema for the per-paper verdict); `INDEX.md` (Quantum ESPRESSO command idioms — `pw.x`, `ph.x`, `epw.x`, `epsilon.x`, `open_grid.x`, `pp.x`); and `PSEUDOPOTENTIALS.md` (pseudopotential-selection heuristics: SSSP vs PseudoDojo, NLCC compatibility with hybrid functionals, SOC-capable families). Core Wannier90 ecosystem documentation is also included; NanoTCAD ViDES and other transport solvers are not. The converged production prompt emerged across B1–B4 and is stable for B5. The B3→B4 ablation on a fixed 15-paper OA subset (same model, same compute, only the knowledge files and consult-before-refuse rule added) is the controlled ablation reported in the main text.

M3. Depth-mode pipeline. Depth mode uses the same CLI + Opus 4.6 harness but without a Python outer loop: three sequential agent sessions on one paper (Reproduce, Review, Reflect), each under its own prompt. The Review and Reflect prompts enforce a "physics-first, tools-second" structure: inventory of concerns written before any tool use, attack ranking by severity and feasibility, explicit forbidden-inputs clauses excluding reviewer reports and subsequent literature. The physics-first ordering was converged on after a tools-first ablation in which paper and tooling knowledge were read together before planning (the scale-mode ordering): under that ordering the concerns inventory anchored on what the available tools easily computed and stayed narrow, yielding 5 concerns against the 14 of the physics-first run on the same paper.

The Reproduce stage on Pizzi 2016 was carried out over roughly one week of intense human–agent debugging cycles. A summary of the tooling arc is in Appendix I: the verified QE + Wannier90 + instrumented NanoTCAD toolchain produced by this stage is what Review and Reflect inherit.

M4. Grading, in-scope definition, and verdict schema. Each paper is graded on a four-tier scale: **T4** = all in-scope claims the agent attempted reproduce within its self-declared tolerance; **T3** = all in-scope claims reproduce qualitatively with small numerical offsets and the underlying mechanism preserved; **T2** = some in-scope claims reproduce while others fail or are abandoned; **T1** = no in-scope claim reproduces. *In-scope* is relative to the agent's own declared scope at the start of the session: what its QE installation, knowledge envelope, time budget, and available compute allow it to attempt. Claims the agent declares out-of-scope are excluded from the grade. The verdict JSON schema includes, for each claim, the paper-side value, the agent-side value, a numerical deviation, a scope classification, and a free-text reasoning field.

M5. Gate work-function mechanism behind the HSE prior update. The naïve "larger gap → steeper SS" expectation is borrowed from band-to-band-tunneling (TFET) device physics and does not transfer to a thermionic MOSFET, whose subthreshold slope is governed by gate electrostatics and band-edge effective mass rather than by the bandgap. Operationally, under a functional change PBE → HSE the gate work-function offset required to hold the same I_{OFF} target shifts by

approximately $\Delta E_g/2$, the carrier-relevant OFF-state conduction-band alignment is approximately conserved, and the observed (I_{ON} −4.1%, SS +11.5%) shift is in the direction predicted by conduction-band curvature changes.

M6. Phase A/B/C/D classification. Each of the 88 substantive critiques was classified post hoc by manual inspection of the agent's reading summary, worklog, and tool-call sequence against the explicit rule-set in Appendix C. The rule-set and four worked examples (one per phase) are released there.

M7. Data availability. All production inputs, outputs, traces, required-reading envelopes, prompts, verdict JSONs, the verified Pizzi 2016 reproduction pipeline, COMMENT_FINAL.pdf, the cross-machine reproducibility subset, development-batch artifacts (marked as excluded from quantitative results), figure-generation scripts, and the Pizzi 2016 Reproduce-stage debug journal will be released upon publication under CC-BY 4.0.

B. Workflow diversity gallery

Figure 4 gives six autonomous agent-vs-paper comparisons, each produced in a single agent session with no human intervention, spanning DFT+U, Wannier90+postw90 anomalous Hall, spin–orbit coupling, LDA+U correlated magnetism, `epsilon.x` optics, and DFPT dielectrics.

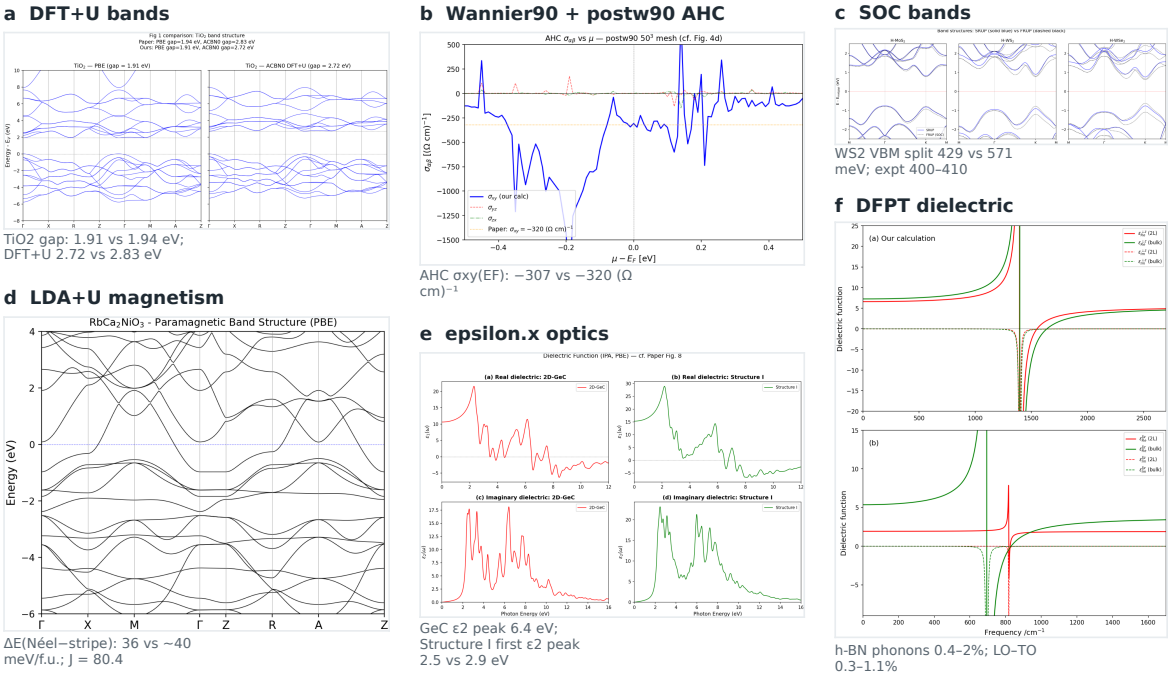

Figure 4. **Workflow diversity gallery.** Six autonomous agent-vs-paper comparisons in a three-column mosaic: **(a)** DFT+U bands, TiO_2 gap 1.91 vs 1.94 eV; DFT+U 2.72 vs 2.83 eV (Agapito et al., 2015). **(b)** Wannier90 + postw90 AHC, $\sigma_{xy}(E_F) = -307$ vs -320 $(\Omega\ \mathrm{cm})^{-1}$ (Park et al., 2022). **(c)** SOC bands, WS_2 VBM split 429 vs 571 meV; experiment 400–410 meV (Reyes-Retana & Cervantes-Sodi, 2016). **(d)** LDA+U magnetism, $\Delta E(\text{Néel} - \text{stripe}) = 36$ vs $\sim$40 meV/f.u.; $J = 80.4$ vs 80–90 meV (Nomura et al., 2020). **(e)** `epsilon.x` optics, GeC ε_2 peak 6.4 eV; Structure I first ε_2 peak 2.5 vs 2.9 eV (Hasan et al., 2020). **(f)** DFPT dielectric, h-BN phonons 0.4–2%, LO–TO 0.3–1.1% agreement (Wang et al., 2016). Each panel was produced in a single agent session with no human intervention.

C. Phase A/B/C/D classification rule-set

Each substantive critique was classified against the first stage at which it became a paper-side concern in the trace.

- **Phase A:** the concern is present during reading, before execution beyond paper/guide inspection.
- **Phase B:** the concern surfaces while diagnosing a failed or misbehaving calculation.

- **Phase C:** the concern emerges after a successful calculation is compared to the paper.
- **Phase D:** the concern becomes substantive only after a second independent calculation or cross-check changes the confidence state.

The phase totals in the audited substantive ledger are $A = 2$, $B = 22$, $C = 50$, $D = 14$, for 88 substantive critiques. Of those, 86/88 (97.7%) are execution-dependent. The reading-only ceiling is 1/111 = 0.9%, because the second Phase-A case was a corrupted-source artifact rather than a genuine scientific catch.

Table 1. Worked examples of the phase-coding rule-set.

Phase	PID	Short label	Why it belongs there
A	QEL-2020-01088	Phosgene +142 Ry	Adsorption-energy number is already physically impossible at reading time; no calculation is needed to know that 142 Ry is not a plausible molecular adsorption energy.
B	QEL-2012-00155	Hybrid-NLCC failure mode	The critique emerges while trying to make the HSE workflow behave: PP/NLCC incompatibility and resulting gap collapse are discovered in failure diagnosis.
C	QEL-2016-00265	WS_2 SOC splitting	The concern appears only after a successful fully relativistic calculation yields 429 meV against the paper's 571 meV.
D	QEL-2019-00174	Fe_2TaAl phonons	The stronger paper-side concern only appears after cross-checking the Γ-point optical frequencies across two PP families, both of which refute the paper's quoted scale.

Boundary rule. Boundary cases were resolved by the *first decisive surfacing* rule. If a reading-stage intuition later became a genuine critique only after execution, the phase is execution-bound, not reading-bound. Likewise, if an initial comparison-stage discrepancy later received a stronger independent confirmation, the critique remains Phase C unless the second computation is what made it credible in the first place. This conservative coding biases against overstating reading-only detection.

D. Review ↔ referee overlap coding rule

The overlap coding uses the following rule on each Review-side concern (rows P1–P14) against each human referee concern (columns R1–R21).

- **SAME:** same physical objection family, same load-bearing mechanism, and close enough in scope that a referee could reasonably regard them as the same criticism.
- **LOOSE:** same broad family or device-nonideality direction, but materially different mechanism or narrower/wider scope.
- **NEW:** no meaningful overlap on the review side.

On the referee axis, **COVERED** means directly addressed by a SAME or LOOSE Review concern, **EDITORIAL/CONTEXTUAL** means outside the Review session's load-bearing physics-audit frame, and **ALREADY-IN-PAPER** means a reviewer-raised issue that had already been incorporated into the published version later read by Review. Under this coding scheme the row summary is **2 SAME, 2 LOOSE, 10 NEW**, and the referee-side summary is **8 COVERED, 11 EDITORIAL/CONTEXTUAL, 2 ALREADY-IN-PAPER**. The two SAME mappings are P2↔R12/R19 (ZA-phonon / ballistic-regime vulnerability) and P4↔R15 (Takagi oversimplification). P7↔R13 is deliberately only LOOSE: the referee asked whether SOC was present at all, whereas Review questioned the transport-pipeline injection. P9 (contact resistance) and P12 (Sb doping sensitivity) remain the two Review-only $L_G = 5$ nm headline-challenging attacks. R21 (a reviewer sensitivity-analysis meta-demand across the whole inventory) is kept as a covered *meta* column rather than a one-to-one cell.

Review-side inventory P1–P14. P1 PBE gap / effective-mass bias; P2 ZA-phonon exclusion undermining ballistic justification; P3 20 Å vacuum for charged 2D supercell; P4 Takagi isotropic/parabolic mismatch with anisotropic valleys; P5 SCF-threshold misreport; P6 NEGF k-mesh/energy-step convergence not demonstrated; P7 SOC injection into ViDES; P8 work-function-shifted I_{OFF} normalisation; **P9 contact resistance set to zero**; P10 continuous-sheet doping approximation;

P11 sub-1 nm physical SiO_2; **P12 doping sensitivity of the gate barrier**; P13 Fig. 4 "no tunneling charges" diagnostic conflation; P14 $\sim 5\%$ mobility-tensor numerical noise.

Referee-side concerns R1–R21. Opening texts (abridged): R1 tight-binding model construction; R2 phosphorene/As/Sb detail; R3 puckered As allotrope; R4 indirect-gap discussion; R5 why large mobilities; R6 realistic disorder sources; R7 finite-T effects; R8 As/Sb look too similar; R9 Table I effective-mass interpretation; R10 why stop at $L_G = 7$; R11 Table II explanation; R12 ZA flexural phonons in buckled crystals; R13 SOC in DFT/TB pipeline; R14 electron–phonon anisotropy; R15 Takagi oversimplified; R16 Wannier-TB related-work crediting; R17 experimental realization; R18 shorten long paragraphs; R19 unresolved ZA concern (round 2); R20 mobility method remains unconvincing (round 2); R21 sensitivity-analysis meta-demand (round 2).

E. Reflect-stage iteration summary

The depth pipeline separates into a long Review session and a shorter Reflect session. Session metrics: Review (v15) = 239 tool calls, 131.4 min; Reflect (review_v4) = 234 tool calls, 53.7 min. The reflect-stage document loop itself used four compile invocations and two explicit PDF read-backs before `COMMENT_FINAL.pdf` was frozen. The iteration history is four drafts: (1) Review verdict (14-concern inventory, four attacks, nine verified claims); (2) Reflect v1/v2 recalibration; (3) Reflect v3 locks in graded conclusion and the DFPT replacement plan; (4) Reflect v4 adds the source-degeneration contact model, the arsenene DFPT calculation, the clean d970_v2 Sb point, and the final six-page Comment with PDF read-back.

Table 2. Key depth-stage numerical anchors referenced in the main text.

Quantity / attack	Review stage	Reflect stage (final)
HSE06+SOC gap E_g (arsenene, $L_G = 5$ nm)	1.4681 eV (PBE+SOC) $\rightarrow$ 2.4767 eV	same; $+68.7\%$ relative shift
I_{ON} PBE $\rightarrow$ HSE propagation	2903 A/m $\rightarrow$ 2783 A/m (-4.1%)	confirmed
Local SS PBE $\rightarrow$ HSE	$+11.5\%$	confirmed
Contact resistance model M1 / M2 / M3 at $L_G = 5$ nm	27.7 / 65.2 / — $\Omega\cdot\mu$m	M3 source degeneration: 31 $\Omega\cdot\mu$m
R_c break-even at $L_G = 5/6/7$ nm	—	30.95 / 76.75 / 125.30 $\Omega\cdot\mu$m
Sb degradation (d970 / baseline, verified pipeline)	677 mV/dec lower bound	736 / 179, ratio 4.11
As degradation (d970 / baseline, verified pipeline)	—	272 / 126, ratio 2.15
Sb-vs-As SS acceleration	qualitative	$4.11/2.15 = 1.91\times$
DFPT arsenene lower A′ extrapolation	analytical only	-21.89 cm^{-1} at $q \rightarrow 0$

The HSE result is the depth-side null-result control: it validates a severe upstream bandgap correction while showing that the specific device-level headline remains comparatively robust to that correction once the work-function alignment is re-calibrated. The source-degeneration R_c lands close to Review's harsher M1 bound, not its more forgiving M2 bound. The DFPT result recovers acoustic branches linear (not quadratic) in q with a soft long-wavelength mixed-A′ mode, inconsistent with the LA-only Takagi picture the original paper uses. The Sb ratio converts the qualitative Review-stage scaling into a quantitative degradation ratio.

F. Per-phase representative vignettes

We include one vignette per phase. PIDs are our internal identifiers.

Phase A: phosgene adsorption energy (+142 Ry), QEL-2020-01088. The reading summary listed adsorption energies of 142.09 Ry for phosgene on a [5,0] CNT and 136.96 Ry on a BN nanotube. That is orders of magnitude beyond plausible molecular adsorption; typical adsorption energies are on the order of 10^{-3}–10^{-1} Ry. This is the rare genuine reading-only physics catch in the corpus.

Phase B: hybrid-NLCC failure mode in GaAs/InAs defects, QEL-2012-00155. The paper's main headline depends on hybrid-functional bulk gaps and downstream defect energetics. In reproduction, the critique surfaced while debugging the HSE setup itself: PseudoDojo PPs with NLCC produced a near-zero hybrid gap, while SG15 no-NLCC PPs restored the expected qualitative behaviour. The final verdict confirms the paper's qualitative need for hybrids (PBE gives a zero GaAs

gap, as the paper states) while showing that the narrow-gap side of the setup is extremely PP-sensitive. Textbook Phase-B: the critique is born inside failure diagnosis.

Phase C: WS_2 SOC splitting, QEL-2016-00265. The fully relativistic reproduction gave a WS_2 valence-band splitting of 429 meV against the paper's 571 meV, a 24.9% discrepancy. The key feature is not that the calculation failed — it succeeded, reproduced the qualitative physics, and then disagreed with the paper in a scientifically informative way. The worklog attributes the discrepancy to the paper's custom RRKJ ultrasoft fully-relativistic pseudopotentials overestimating W-based SOC splittings; the reproduced value is closer to experiment (400–410 meV). Phase C is the modal discovery point in the corpus.

Phase D: Fe_2TaAl/Fe_2TaGa phonons, QEL-2019-00174. The initial QE DFPT result already suggested a large discrepancy in the optical-mode scale, but the stronger critique only became credible after a second PP-family cross-check. GBRV ultrasoft and PSLibrary PAW calculations both gave Γ-point optical frequencies in the $\sim$180–330 cm^{-1} range, whereas the paper quotes $\sim$510–680 cm^{-1} for Fe_2TaAl and $\sim$470–591 cm^{-1} for Fe_2TaGa. The paper's qualitative dynamical-stability claim survives, but the quantitative optical-mode scale looks systematically wrong by roughly a factor of two. The second independent computation is what converts a suspicious mismatch into a substantive critique.

Critique taxonomy. The OA-v2 audit maps the 88 substantive critiques onto nine categories: `convergence_issue` (60, non-substantive operational subset), `paper_error` (21), `paper_bias` (20), `agent_outperforms` (14), `missing_correction` (11), `method_limitation` (9), `pp_artifact` (5), `paper_omission` (5), `experimental_benchmark` (3). The `convergence_issue` category is tracked operationally but excluded from the 88-row substantive ledger, because a convergence failure is not by itself a paper-side methodological concern.

Denominator conventions. Paper-level substantive-critique rates across denominators: full deduplicated OA corpus 47/111 (42.3%); deduplicated production papers only 47/90 (52.2%); clean single-machine B5 Mac run 30/50 (60.0%). The value 88 is not a paper-level rate; it is the critique-instance count used for phase and category analyses.

G. Trace-level behavioural observations

The trace-behaviour audit used 61 production-trace sessions from B3–B5 and three compact findings that anchor the main-text claims.

Load-then-execute is the aggregate signature. Across the first 20 tool calls of 61 sessions, 586/1020 calls (57.5%) were direct file reads and 664/1020 (65.1%) were read-like actions once closely related lookup actions were included. No session begins by launching QE. Reading summaries are written in 50/61 sessions and the trace explicitly references the summary in 51/61. Among sessions with scorable late behaviour, 49/50 show zero paper rereads after execution starts. Comparison-figure compliance in the audited B5 figure total is 107/107. The safe claim is procedural: once the summary became mandatory, production traces show a stable read $\rightarrow$ plan $\rightarrow$ execute discipline with durable on-disk planning state. The "load-then-execute" framing is a tool-call statement, not a cognition statement: after the initial load, agents may still be attending to earlier paper text internally even when no explicit reread appears in the trace.

Filesystem as external memory under compaction. Compaction affected 10/61 sessions (16.4%), with 11 total compaction events. Zero sessions re-read the paper after the last compaction; only 2 sessions re-opened the reading summary. The emergent strategy is not "reconstruct the paper state from scratch" but "continue from on-disk artifacts" — generated inputs, outputs, worklogs, and the reading summary already written to disk. We did not prompt for this behaviour.

Open-loop visual behaviour. In 0/61 sessions does an agent generate a figure, open it, notice a problem, and regenerate. Agents reliably produce *requested* visual artifacts but do not yet treat their own visual output as a self-auditing feedback channel. This is the basis for treating visual verification as a harness limitation rather than a reasoning limitation.

H. Prompt and knowledge-envelope structural overview

The production prompts and the required-reading envelope are structured artifacts rather than monolithic instructions. We give a one-paragraph structural description of each rather than verbatim text; complete files will be released upon publication under CC-BY 4.0.

Scale prompt (B5, converged production form). A short orchestration layer ($\sim$30 lines) that references the five required-reading files by name, sets the four-step workflow (load paper $\rightarrow$ write reading summary $\rightarrow$ execute calculations with

worklog → emit verdict), defines a single output directory per paper, and states the flexibility time budget. It does not ask for critique; it asks for reproduction artifacts, comparison figures, and a structured verdict.

`HOUSE_RULES.md` ($\sim$30 lines). Resource-management and process discipline: one-calculation-at-a-time, core-count discipline, full-paper-load requirement, comparison-figure-deliverable requirement, honesty conventions (when-in-doubt-declare-out-of-scope rather than silently drop a claim), and time-awareness scaffolding.

`READING_GUIDE.md` ($\sim$40 lines). Specifies the structured pre-compute planning output: section-by-section summary, a reproduction-target list with per-target scope/feasibility annotations, and explicit scope declarations for items the agent will not attempt. Forces a complete paper understanding before execution.

`VERDICT_FORMAT.md` ($\sim$50 lines). JSON schema for the per-paper verdict, including paper-side value, agent-side value, deviation, scope classification, and free-text reasoning for each attempted claim. Standardizes outputs for post hoc aggregation.

`INDEX.md` and `qe_executables.md` ($\sim$600 lines combined). Verified QE and Wannier90 capability map: executable idioms (`pw.x`, `ph.x`, `epw.x`, `epsilon.x`, `open_grid.x`, `pp.x`), known gotchas for `pw2wannier90.x` k-point format, `postw90.x` Fermi-energy handling, adaptive Berry meshes, and `open_grid.x` use in hybrid+SOC workflows. Encodes working idioms beyond the small subset agents spontaneously recall.

`PSEUDOPOTENTIALS.md` ($\sim$200 lines). PP decision tree covering XC-matching, library choice (SSSP, PseudoDojo, SG15, GBRV, PSLibrary), NLCC compatibility with hybrid functionals, SOC-capable families, and workflow-specific PP constraints.

Depth-mode Review and Reflect prompts ($\sim$150 and $\sim$120 lines). Both enforce a physics-first, tools-second structure: inventory of concerns written before any tool use, attack ranking by severity and feasibility, explicit forbidden-inputs clauses excluding reviewer reports and subsequent literature. The Reflect prompt additionally requires the compile-PDF-read-back-and-iterate loop and specifies the Comment-shaped output.

B3→B4 knowledge ablation. On the same 15 OA papers, same model family, same 12-core compute budget, adding `INDEX.md` + `PSEUDOPOTENTIALS.md` + the consult-before-refuse rule. Aggregate: attempted quantitative claims 44 → 47 (+7%); within-5% claims 40/44 (91%) → 44/47 (94%); mean absolute deviation 2.2% → 1.3% (-44%); phonon workflows attempted on 15% → 29% of papers; false "QE cannot do this" refusals eliminated on paired cases. The point is mechanism, not a large-N performance claim: the agent did not become more intelligent; it stopped declining workflows already available in its local toolchain.

I. Engineering challenges summary for the Reproduce stage

The Reproduce stage on Pizzi 2016 was dominated by tool maturity differences across the QE / Wannier90 / NanoTCAD ViDES stack. We summarize the recurrent challenge classes without reproducing the day-by-day journal; the full day-by-day notes will be released upon publication under CC-BY 4.0.

1. **Legacy-runtime port.** The 2016-era transport-solver wrappers expected a legacy Python 2 interpreter and project-local conventions that did not survive unchanged into the 2026 environment. A dedicated runtime, wrapper repair, and an explicit rebuild path were prerequisites for any device result to be trustable.

2. **Spinless-TB versus SOC-spinor factor-of-two mismatch.** The SOC spinor Wannier Hamiltonian traces both spin channels, while legacy transport code paths assumed a spinless tight-binding Hamiltonian and applied an extra factor of two. The correction had to be propagated consistently through the Landauer current path and the charge feed to the Poisson loop; the latter was the more consequential of the two because it distorted every SCF iteration. Several of these bugs had numerically compensating effects on the paper's published geometries, making disentanglement a systematic alternate-geometry exercise rather than a direct diagnostic.

3. **Hard-coded lead-context assumptions.** The lead self-energy code implicitly assumed a minimum contact-pad headroom. In the principal-layer wrapper the effective lead context occupied four super-slices; at some short-pad settings this consumed essentially the full pad length, leaving no bulk-like lead region and distorting the device behaviour. The assumption was not visible from the Python interface; it was surfaced by reading the C source and then checked numerically by a pad-size diagnostic.

4. **Silent-failure traps in the legacy core.** Memory-management issues on long k-point sweeps and a 600-iteration SCF ceiling could return numerically untrustable results without a clean exception. Source-level patches and a tracked-SCF wrapper converted silent solver behaviour into logged diagnostics.

5. **Transport-truncation ambiguity.** The paper's "58 neighbours" language left genuine ambiguity about how much long-range Wannier hopping the device model retained, which triggered a principal-layer detour and a bit-identity unit-test suite verifying equivalence to the atomic reference implementation.

Methodological lessons. Two lessons survive into the final paper story. First, multiple bugs were simultaneously present and partly cancelled on the publication geometry; the pipeline therefore passed through a deceptive intermediate regime where some observables looked nearly correct for the wrong reasons. Alternate geometries and continuation paths were not redundant checks; they were necessary to distinguish genuine agreement from accidental cancellation. Second, "read everything from scratch" repeatedly acted as a recovery mechanism: several decisive breakthroughs came from forcing a fresh read of the code path or the paper after the prior explanatory frame had become too sticky. The verified QE + Wannier90 + instrumented NanoTCAD toolchain produced at the end of this stage is what Review and Reflect inherit.

Provenance of COMMENT_FINAL.pdf. The six-page Comment produced autonomously by Reflect has the following provenance chain, audited end-to-end: Review-stage verdict and worklog $\rightarrow$ Reflect-stage new calculations (source-degeneration contact model, arsenene DFPT, d970_v2 Sb point) $\rightarrow$ LaTeX draft $\rightarrow$ four compile invocations $\rightarrow$ two explicit PDF read-back events $\rightarrow$ final six-page PDF. No human editing was applied to the PDF body itself.

Release note. Complete experimental artifacts, including full prompts, required-reading envelope files, per-paper traces, the NanoTCAD ViDES debug journal, and the complete 9-category critique catalog, are available upon publication under CC-BY 4.0.