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

**Haonan Huang** [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; **85 of 88 of these critiques (96.6%) surface only after the agent has actually run a calculation, with a reading-only ceiling of 1.8%**. 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 complete, submission-form six-page Comment that revises the paper's $L_{\mathrm{G}} = 5$ nm headline. Two of its $L_{\mathrm{G}} = 5$ 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 — **96.6% of which require execution to surface, and only 1.8% are catchable by reading alone**. *Critique emerges from reproduction, not from reading*: critical scientific scrutiny is execution-bound, not a property of prior

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

Accepted at ICML 2026 AI for Science Workshop.

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 complete, submission-form six-page Comment — composed, figured, typeset, and PDF-iterated entirely by itself — whose two main findings revise the paper's $L_{\mathrm{G}} = 5$ 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 complete, submission-form 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 **rose from 3/15 to 5/15 papers (20% → 33%)** 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 for this failure class 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 —

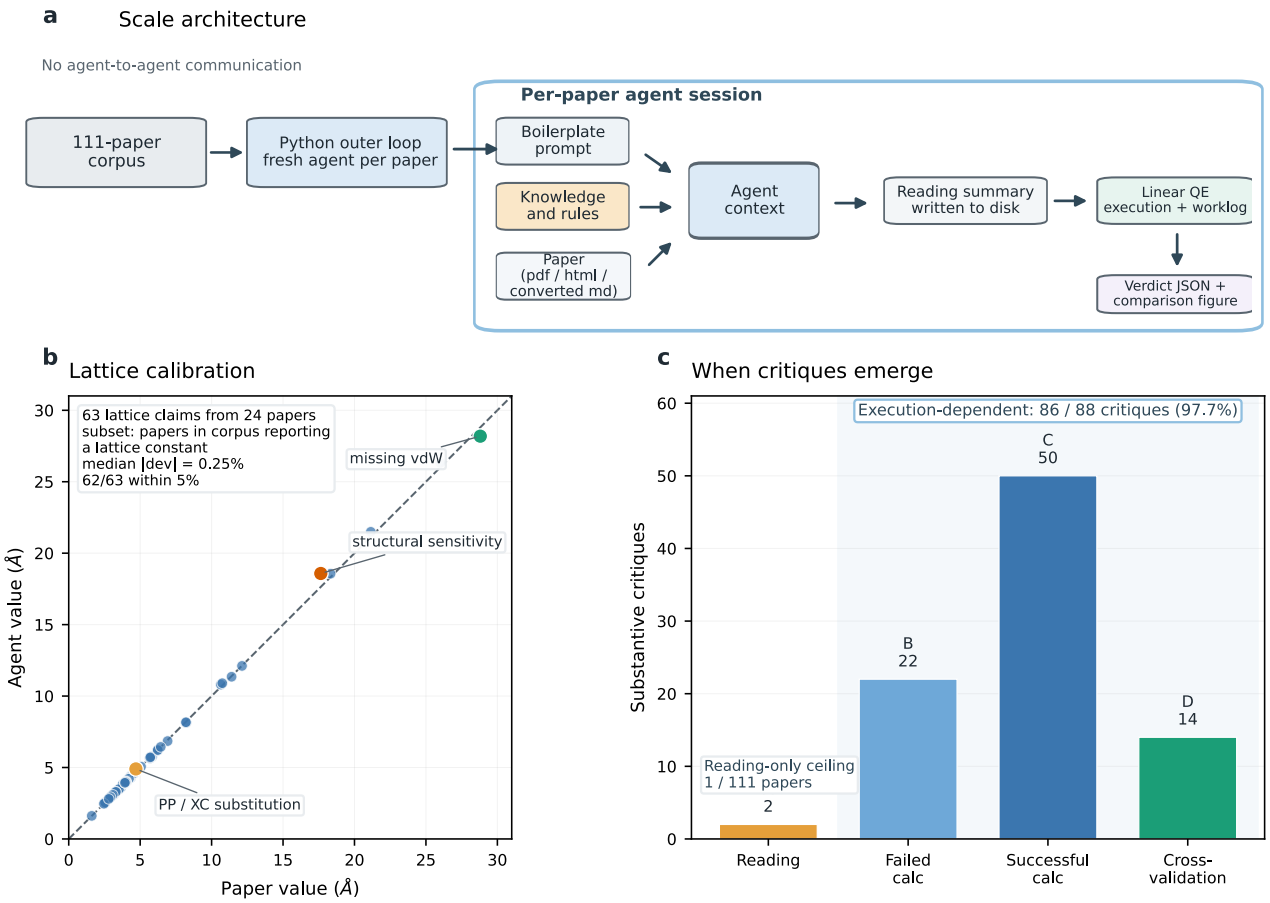


*Figure 1.* **Grounded scrutiny at scale: architecture, calibration, and execution-dependence. (a)** Two-level pipeline: a Python outer loop iterates over the 111-paper corpus, handing each paper to a fresh Claude Opus 4.6 agent running inside the Claude Code CLI; there is no agent-to-agent communication. 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 = 3 (during reading), B = 22 (after a failed calculation), C = 49 (after a successful calculation compared to the paper), D = 14 (after cross-validation). Execution-dependent: 85/88 = 96.6%. Reading-only ceiling: 2/111 papers = 1.8%. Caption counts are the corrected ledger (one row reclassified per the no-calculation clause; Appendix A M6); the panel is drawn from the pre-correction ledger (A = 2, C = 50, 86/88, 1/111) and is superseded by these counts.

**the filesystem becomes external memory**, an emergent strategy we did not prompt for. 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 (47/111) contain at least one substantive methodological concern** raised by the agent unprompted, across eight substantive categories (a ninth tracked category, agent-side convergence issues, is excluded; Appendix F). A *substantive methodological concern* is a paper-side concern under the Appendix F taxonomy; agent-side convergence issues do not count. Denominators, defined once: 111 = the full deduplicated open-access corpus; 90 = the deduplicated union of unique papers across converged-harness batches B3–B5 (47/90 = 52.2%); 50 = the controlled single-machine B5 run (30/50 = 60.0%; 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 success-

ful calculation compared to the paper (C), or after cross-validation (D). Counts: **A = 3, B = 22, C = 49, D = 14** (corrected ledger; Appendix A M6).

**Eighty-five of eighty-eight critiques (96.6%) emerged only after the agent had actually run a calculation** (coding-sensitivity analysis: Appendix A M6). The three Phase-A critiques are one file-corruption artifact and two genuine reading catches: a phosgene adsorption energy reported at 142 Ry, three-to-five orders beyond physical, and a reproducibility-omission catch — a paper whose structure coordinates are omitted from both text and SI, the structure being available only behind a subscription — which is not a physics catch like the phosgene case. The reading-only ceiling is **2 in 111 papers (1.8%)**. 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 96.6% execution requirement complements passive-reading verification baselines ($\sim$21% recall on SPOT's 91 retraction/errata-level errors (Son et al., 2025)) and code-provided reproduction baselines in the 19–21% 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 $\rightarrow$ NSCF $\rightarrow$ Wannierization $\rightarrow$ 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 $Mn_3Al$ $\sigma_{xy}$ 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 and band structure, 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 $\rightarrow$ Wannier90 $\rightarrow$ 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

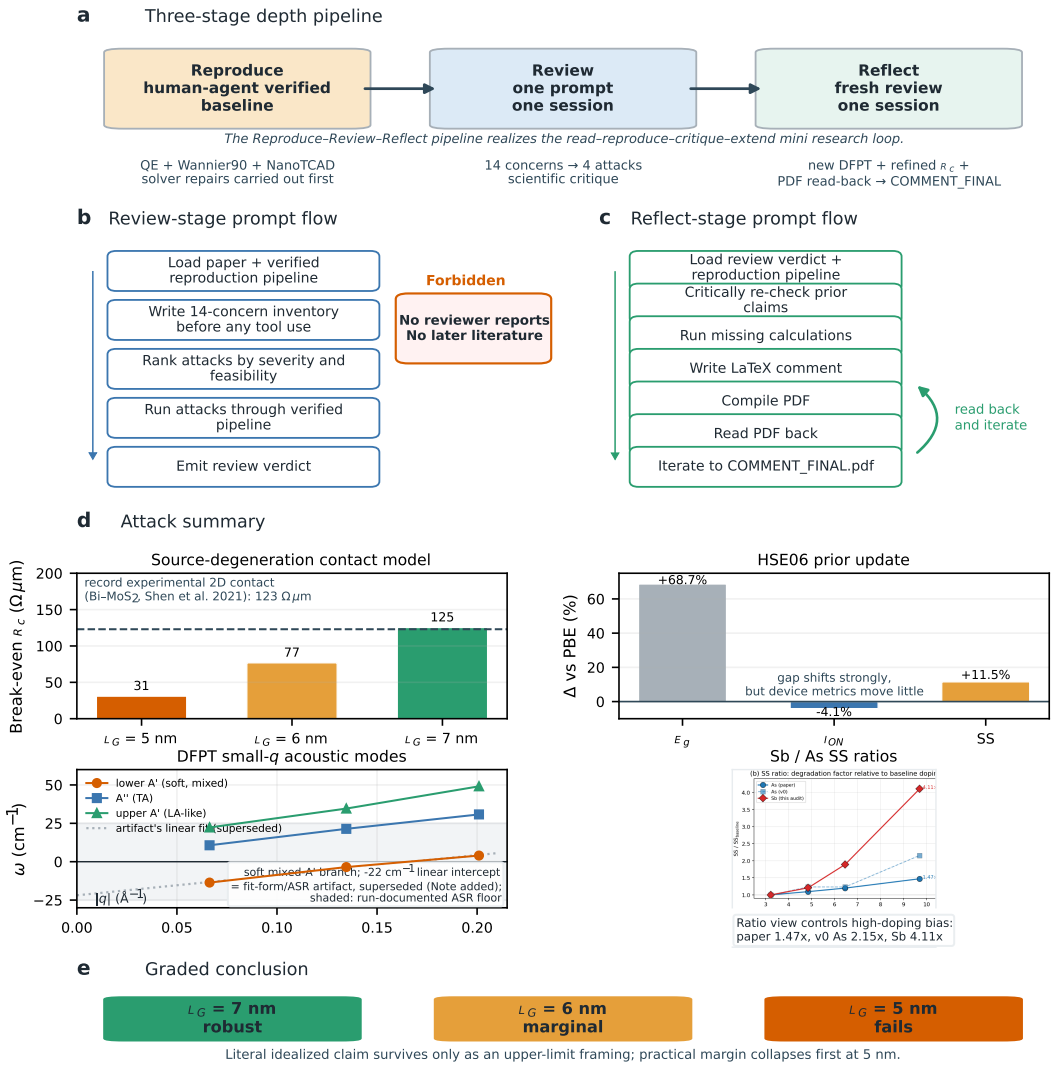


*Figure 2.* **The Reproduce–Review–Reflect pipeline applied to Pizzi 2016. (a)** Three-stage flow: Reproduce (human–agent verified baseline across QE + Wannier90 + NanoTCAD) → Review (one session; 14-concern inventory, four attacks) → Reflect (fresh session; new DFPT, refined $R_c$, PDF read-back loop → 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. the record experimental 2D contact, 123 $\Omega{\cdot}\mu$m for semimetal–$MoS_2$ (Shen et al., 2021); HSE06 prior update ($E_g$ +68.7%, $I_{ON}$ −4.1%, SS +11.5%); DFPT small-$q$ acoustic (soft mixed-A$'$ flexural-derived branch; the artifact's $-22$ cm$^{-1}$ linear intercept is an ASR/fit-form artifact, superseded per the Note added in §4.3; panel redrawn from the run's raw frequencies); 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.

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 (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 $\tau$ with two analytical series-resistance models, finding break-even contact resistances at $L_G = 5$ nm below even the experimental record for a 2D contact, 123 $\Omega{\cdot}\mu$m for semimetal–$MoS_2$ (Shen et al., 2021) — an author-supplied benchmark; the agent's own figures ($\sim$100–1000 $\Omega{\cdot}\mu$m, Bi–$MoS_2$ $\sim$120) came from parametric memory in the blind inventory phase, and the source-degeneration bound is benchmark-independent; 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 (Heyd et al., 2003; 2006) 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 → 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 $14 \times 21$ 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 (reproduced as-is in Appendix J) 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 $\tau$ 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, resolving a soft, flexural-derived low-lying branch of mixed A′ character that the analytical treatment was structurally incapable of

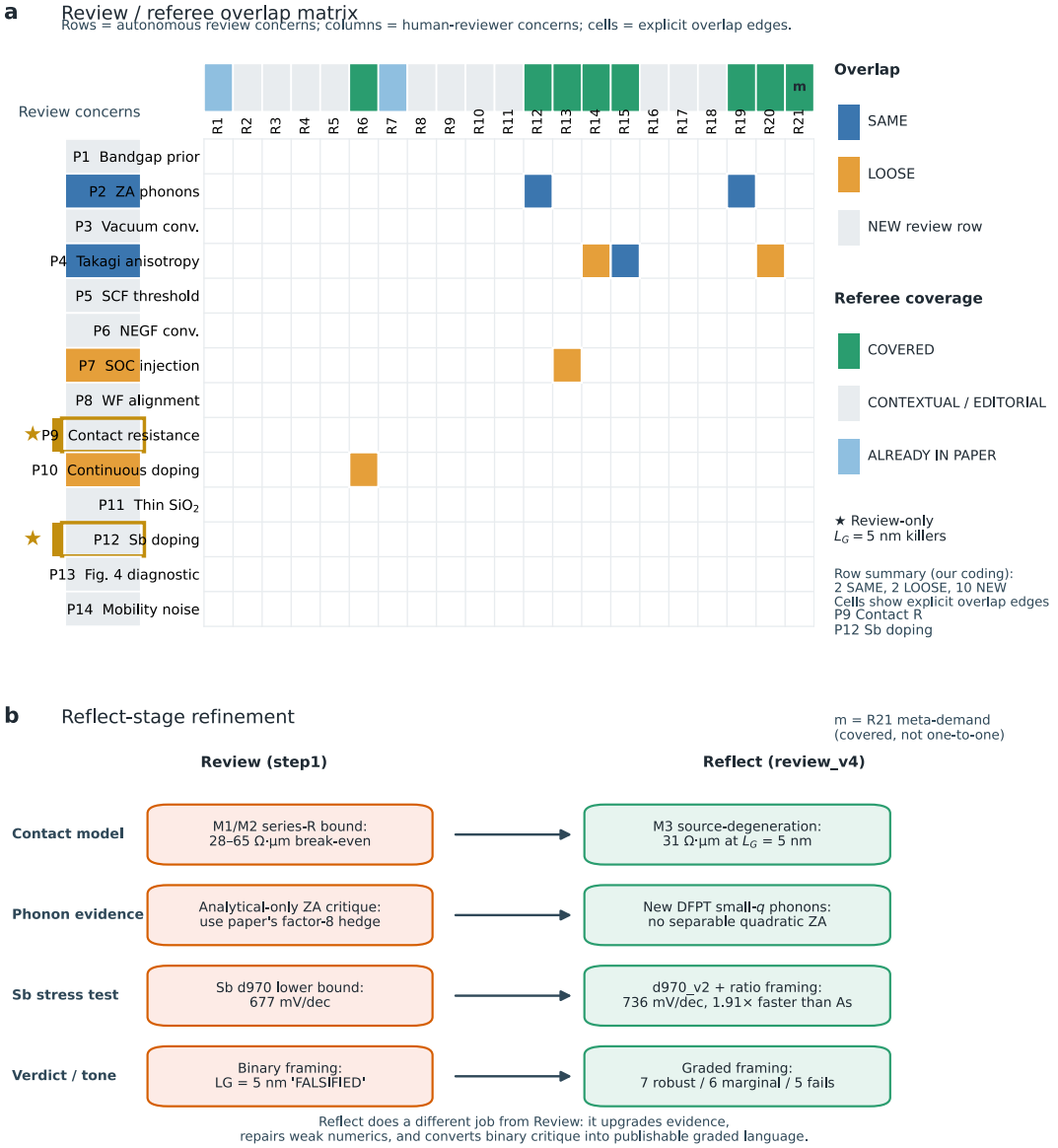


*Figure 3.* **Review ↔ referee overlap and Reflect-stage refinement.** *Top:* sparse $14 \times 21$ overlap matrix, rows = Review concerns (P1–P14), columns = referee concerns (R1–R21). Row-level coding (Review side, SAME/LOOSE/NEW) summarises each Review concern's strongest overlap class. Under our 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 $\Omega{\cdot}\mu$m) → M3 source-degeneration (31 $\Omega{\cdot}\mu$m 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).

capturing. Buckled arsenene lacks the horizontal mirror $\sigma_h$, so the flexural branch cannot be symmetry-excluded from first-order transport scattering as in planar 2D crystals, and it lies far below the LA-like branch the paper's Takagi treatment assumes ($|\omega| \leq 14$ vs. 22–49 cm$^{-1}$ over the sampled $|q| = 0.066$–$0.201$ Å$^{-1}$). The artifact's headline number, a $-22$ cm$^{-1}$ $q \to 0$ intercept, requires correction: it is a linear extrapolation of three raw `ph.x` frequencies without acoustic-sum-rule (ASR) enforcement. Rotational invariance makes a flexural-derived branch quadratic-dominated near $\Gamma$, so a linear fit mechanically yields a negative intercept, and the run's own output records a residual ASR violation ($\sim$5–25 cm$^{-1}$) larger than the lowest measured frequency ($-13.6$ cm$^{-1}$). The intercept is a fit-form ar-

tifact, not evidence of dynamical instability (Note added below). The corrected reading strengthens the critique: the paper's LA-only treatment omits a soft mixed-character branch far below its assumed acoustic spectrum. The **Sb high-doping closing point** completes Review's sweep at the doping level that actually brackets the $I_{\mathrm{OFF}}$ 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 *could* be submitted.

The four attacks occupy different evidential classes. The source-degeneration bound is derived from the paper's own $f_T$ and $\tau$, independent of any external benchmark or pipeline parameter; the HSE attack is a self-falsified control — the agent's stated prior overturned by its own calculation; the DFPT phonon thread is first-principles but interpretation-sensitive, as the Note added below records; and the Sb ratio is a pipeline-dependent quantitative result with a matched-doping internal control (the matched As ladder).

**Note added (camera-ready).** During workshop review of this manuscript, a reviewer's reading-based assessment identified an interpretation error in the depth artifact — the acoustic-branch characterization corrected above — together with citation-fidelity lapses, corrected throughout this version. This is itself a datum for §5: execution grounds numbers, but branch assignment, interpretation, and citation fidelity sit outside that protection, and here they were caught by human domain knowledge operating on the text. It is not evidence that reading-based LLM review could not catch them: this depth pipeline ran Review and Reflect as single sessions, without the session-breaking adversarial review loops that scaffolded systems can add. Whether such scaffolded review reliably catches this error class — and how robustly — is an open question we pose rather than answer. The corrected characterization strengthens the underlying critique of the original paper's phonon treatment, and the graded conclusion — $L_{\mathrm{G}} = 7$ nm robust, 6 nm marginal, 5 nm fails — is unchanged: it rests on the contact-resistance and Sb-doping attacks.

## 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) — hallucinated results, unreliable self-review, 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. A hallucinated number does not survive execution: running the calculation either reproduces the target or does not, and the judge is the physics. Execution grounds numbers; it does not by itself protect setup choices, interpretation, or citation fidelity (§4.3, Note added). *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 96.6% 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, stated precisely, is descriptive and falsifiable: in corpora of published — hence already human-reviewed — computational-physics papers, the substantive methodological concerns that remain are overwhelmingly execution-bound (85/88 here). This is in part a selection effect, and it is also the deployment condition: reading-detectable errors are largely filtered by conventional review before publication, so an autonomous second pass adds value precisely where reading stops. We make no counterfactual claim that reading-based review — human or LLM — cannot surface critique; the Note added in §4.3 is a live counterexample on our own artifact. The framing is testable: on pre-review corpora such as preprints, the reading-detectable fraction should rise.

**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 (Huang, 2026a). (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. Of these, (i) is established by a controlled ablation (§2.2); (ii)–(iv) are diagnoses from the trace audit and the depth case, not ablations. Scaling grounded scrutiny means scaling the harness along all four axes, not waiting for a new model.

**Limitations.** Critique coding is single-coder and post hoc (released rule-set and worked examples enable independent recoding; Appendix A M6, Appendix C). Only harness axis (i) is backed by a controlled ablation. The depth result is one case on one deliberately selected paper — an existence proof, no cross-paper generalization claimed. One model, one harness, one code ecosystem; corpus selection effects (published, open-access, QE-primary, 2010–2024). The soft 2–4 h budget bounds attempted scope; open-loop visual behaviour limits self-audit of figures.

**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 — demonstrated in companion work (Huang, 2026b), which runs the full corpus-to-manuscript loop with anchor reproduction as its calibration mechanism; the two pipelines are complementary halves of one program. 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. Code for the 3D simulation of nanoscale semiconductor devices, including drift-diffusion and ballistic transport in 1D and 2D subbands, and 3D tunneling. *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.

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, H. From experiments to expertise: Scientific knowledge consolidation for AI-driven computational research. Preprint at arXiv:2603.13191; ICML 2026 AI for Physics Workshop, 2026a.

Huang, H. Grounded autonomous research: a fault-tolerant LLM pipeline from corpus to manuscript in frontier computational physics. ICML 2026 AI for Science Workshop, 2026b.

Islam, M. R. et al. Superior tunable photocatalytic properties for water splitting in two dimensional GeC/SiC van der Waals heterobilayers. *Scientific Reports*, 11:17739, 2021.

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, Z. S., Kapoor, S., Nadgir, 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, G. et al. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of ICML*, volume 267 of *PMLR*, pp. 56843–56873, 2025.

Wang, L. 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.

Zou, Y. 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 comprises two independent 50-paper runs on separate machines, of which the Mac run is the controlled single-machine production run quoted in §2.4; the 12 papers common to both B5 machines form the cross-machine check. The 90-paper production set used in Appendix F is the deduplicated union of unique OA papers across B3–B5 (B3: 17; B4: 15, a subset of B3's set; B5: 50 per machine; deduplication yields 90). 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 $\sim 12$ CPU cores and $\sim 64$ 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 here. In the long-channel limit the subthreshold slope is set by gate electrostatics alone ($\mathrm{SS} = \ln 10 \cdot (kT/q)(1 + C_\mathrm{d}/C_\mathrm{ox})$, with no bandgap or effective-mass dependence); at $L_\mathrm{G} = 5$ nm the additional degradation channel is source-to-drain tunneling through the gate barrier, whose transmission depends on barrier shape and band-edge effective mass — not on the gap itself. Operationally, under a functional change PBE $\rightarrow$ HSE the gate work-function offset required to hold the same $I_\mathrm{OFF}$ target shifts by approximately $\Delta E_g/2$, the carrier-relevant OFF-state conduction-band alignment is approximately conserved, and the observed ($I_\mathrm{ON}$ $-4.1\%$, SS +11.5%) shift is attributed in the run's worklog to the transverse band-edge effective mass and OFF-state charge density dominating over the gap-driven WKB exponent change; the device remains source-to-drain-tunneling-limited under both functionals ($\mathrm{SS} \gg 60$ mV/dec).

**M6. Phase A/B/C/D classification.** Each of the 88 substantive critiques was classified post hoc by a deterministic, re-runnable keyword classifier over the session's terminal verdict artifact, with Phase-A assignment additionally requiring either corroborating concern language in the pre-execution reading summary or the absence of any executed calculation; the two original Phase-A rows and a 10-paper sample were then hand-verified against reading summaries and worklogs. The classification is single-coder-audited and post hoc; the released rule-set, classifier, and four worked examples (one per phase, Appendix C) enable independent recoding, and the boundary rule there is conservative against overstating reading-only detection. A camera-ready re-audit corrected the single clearest boundary error — a concern tagged post-execution whose session never executed a calculation — via the rule's own no-calculation clause, giving corrected counts A = 3, B = 22, C = 49, D = 14 (85/88 = 96.6%). An independent blinded LLM recoding of all 88 rows (one fresh judge per row, given only the rule-set, the row's reading summary, its verdict artifact, and the mechanical fact of whether any calculation executed) agreed with the corrected boundary on 66/88 rows ($\kappa = 0.16$), with every disagreement in one direction: the blinded judges coded as reading-detected those rows where the reading summary already stated the concern's substance and execution, in their reading, only quantified it — cases the released boundary rule codes as execution-bound. The execution-bound share is therefore coding-sensitive: 96.6% under the released (corrected) instrument, 71.6% under the blinded recoding; the bolded headline uses the corrected instrument, and we report the recoding as the measured sensitivity of this single-coder boundary. Throughout, "execution-dependent" is a positional statement — the concern first surfaces after calculation execution in the session record — not a causal claim about what produced the critique. The finer B/C/D sub-splits carry more keyword noise than the A-vs-non-A boundary and should be read as indicative.

**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; the Comment itself is reproduced as-is in Appendix J.

## 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.

## 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 corrected phase totals in the audited substantive ledger are $A = 3$, $B = 22$, $C = 49$, $D = 14$, for 88 substantive critiques (one row reclassified at camera-ready per the rule-set's no-calculation clause; Appendix A M6). Of those, 85/88 (96.6%)

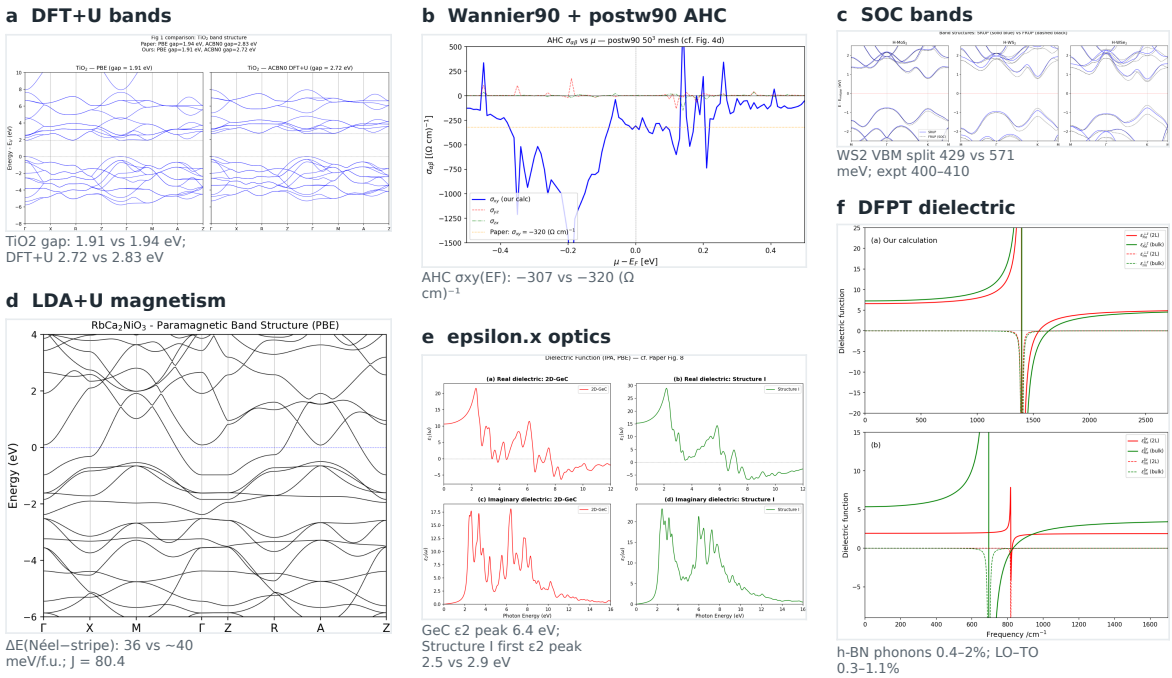


*Figure 4.* **Workflow diversity gallery.** Six autonomous agent-vs-paper comparisons in a three-column mosaic: **(a)** DFT+U bands, $TiO_2$ gap 1.91 vs 1.94 eV; DFT+U 2.72 vs 2.83 eV (Agapito et al., 2015). **(b)** Wannier90 + postw90 AHC, $\sigma_{xy}(E_F) = -307$ vs $-320$ $(\Omega\ \mathrm{cm})^{-1}$ (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 $\varepsilon_2$ peak 6.4 eV; Structure I first $\varepsilon_2$ peak 2.5 vs 2.9 eV (Islam et al., 2021). **(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.

are execution-dependent. The reading-only ceiling is 2/111 = 1.8%: of the three Phase-A rows, one is a corrupted-source artifact rather than a genuine catch, one is a genuine physics catch, and one is a genuine reproducibility-omission catch.

**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**

*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 $\Gamma$-point optical frequencies across two PP families, both of which refute the paper's quoted scale. |

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 $L_{\mathrm{G}} = 5$ 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 $I_{\mathrm{OFF}}$ 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.

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 resolves a soft long-wavelength mixed-A′ branch far below the LA-only Takagi picture the original paper uses; the apparent linearity and negative $q \to 0$ intercept in the artifact are artifacts of a three-point linear extrapolation without ASR enforcement (§4.3, Note added), while the branch's softness and mixed character stand. The Sb ratio converts the qualitative Review-stage scaling into a quantitative degradation ratio.

*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 |
| $I_{\mathrm{ON}}$ 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 soft mixed-A′ branch | analytical only | $\omega = -13.6 \ldots +4.0$ cm$^{-1}$ over $\lvert q\rvert = 0.066$–$0.201$ Å$^{-1}$; the $-21.89$ cm$^{-1}$ linear intercept is superseded as an ASR/fit-form artifact (§4.3, Note added) |

## 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 $\Gamma$-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 tracks nine categories, of which eight substantive ones compose the 88-row ledger: `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 ninth, `convergence_issue` (60 instances), 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 the 61 session-trace files of the B5 production machine: 50 completed papers, one partial session, and 10 rate-limit-aborted stubs that contain no tool calls. Three compact findings anchor the main-text claims.

**Load-then-execute is the aggregate signature.** Across the first 20 tool calls of each of the 51 live sessions (1020 calls; the 10 stubs contribute none), 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 of the 51 live sessions, and 51/61 traces reference one — the single live session without a summary file discusses the summary in-session but never writes it. 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 → plan → 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 → write reading summary → 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% ($-41\%$); phonon workflows attempted on 3/15 → 5/15 papers (20% → 33%); 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 → Reflect-stage new calculations (source-degeneration contact model, arsenene DFPT, d970_v2 Sb point) → LaTeX draft → four compile invocations → two explicit PDF read-back events → 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.

## J. The Reflect-stage Comment, reproduced as-is

The document below is the artifact produced autonomously by the Reflect session, reproduced without modification; its provenance chain is audited in Appendix I. It is included as primary data for the claim that a single unsupervised session can carry an audit through to a complete, submission-form scientific document — the composition, figures, typesetting, and PDF iteration are the agent's own, for the reader to judge. Its scientific content is not warranted: it contains the acoustic-branch mischaracterization corrected in §4.3 (Note added) and may contain further issues; where this paper's main text and the artifact differ, the main text supersedes.

# Comment on "Performance of arsenene and antimonene double-gate MOSFETs from first principles" [Pizzi *et al.*, Nat. Commun. 7, 12585 (2016)]

Independent computational audit
(anonymous reviewer collective, 2026)

April 9, 2026

Pizzi *et al.* reported that double-gate MOSFETs based on monolayer arsenene (As) and antimonene (Sb) at gate lengths $L_G = 5$–7 nm satisfy the ITRS high-performance roadmap targets, on the basis of ballistic non-equilibrium-Green-function (NEGF) transport on Wannier-interpolated density-functional bands. We have re-executed the published pipeline and pursued three independent attacks on the headline conclusion, all using the same Quantum ESPRESSO, Wannier90, and NanoTCAD ViDES tool stack. First, the paper-stated $I_{\mathrm{ON}}$ values, when corrected for series contact resistance via a self-consistent saturated-MOSFET source-degeneration model anchored to the paper's own $f_T$ and $\tau$, fall below the ITRS target at $L_G = 5$ nm for any contact resistance above $\sim 30\,\Omega \cdot \mu\mathrm{m}$, well below the best two-dimensional contact resistances reported anywhere in the literature. Second, density-functional perturbation-theory phonons of the freestanding arsenene monolayer near $\Gamma$ along $\Gamma$–$M$ show that the "ZA" branch is not separable in this material: buckling-broken horizontal mirror symmetry mixes longitudinal and out-of-plane character, producing two linear (not quadratic) acoustic branches and a soft long-wavelength mode. The paper's qualitative dismissal of ZA-phonon scattering — which is the load-bearing argument for using ballistic transport at $L_G = 5$–7 nm — therefore depends on a substrate-induced ZA gap that the original paper does not quantify. Third, the source/drain doping sensitivity that the paper presents only for arsenene is much steeper for antimonene: at the highest of the four doping levels in the paper's own Fig. 6 ladder, the antimonene local subthreshold swing degrades by a factor of $\sim$4 from its baseline value, against $\sim$1.5 for the corresponding As result. A complementary fourth check — propagating an HSE06+SOC band structure through the full pipeline — shows that, despite a +1.0 eV (+69%) gap correction, the device $I_{\mathrm{ON}}$ shifts by only $-4\%$, a null result that is favourable to the paper's PBE-based device-level predictions even though it leaves any claim that depends on the absolute bandgap itself unsupported. We conclude that the paper's literal claim — that an idealised ballistic device with zero contact resistance and a hand-picked low doping marginally clears the ITRS targets — holds, but the headline framing ("All our considered devices comply with $I_{\mathrm{ON}}$ requirements") overstates the practical achievability of those targets at $L_G = 5$ nm.

## 1. Context

Pizzi *et al.* [1] simulated double-gate $n$-MOSFETs whose channel is a monolayer of buckled arsenene or antimonene, at gate lengths $L_G = 5, 6, 7$ nm, oxide thicknesses $t_{\mathrm{ox}} = 0.42, 0.45, 0.50$ nm, and ITRS-prescribed supply voltages $V_{DD} = 0.54, 0.57, 0.60$ V. The pipeline runs PBE+SOC density-functional theory in Quantum ESPRESSO [2], Wannier-interpolates the conduction and valence manifolds with twelve maximally localised Wannier functions in Wannier90 [3], and feeds the resulting tight-binding Hamiltonian to NanoTCAD ViDES [4] for self-consistent NEGF transport. A workfunction shift on the gates is applied so that $I_{\mathrm{OFF}} = 0.1\ \mathrm{A\,m^{-1}}$ at $V_{GS} = 0$, and the on-current $I_{\mathrm{ON}}$ is read off at $V_{GS} = V_{DS} = V_{DD}$ [1, Table 2].

The headline result is that, at all three gate lengths, both materials clear the ITRS high-performance $I_{\mathrm{ON}}$ target with a margin of +33% (+24%) for As (Sb) at $L_G = 5$ nm, growing to +200% at $L_G = 7$ nm, and that the subthreshold swing remains in the $\sim$100 $\mathrm{mV\,dec^{-1}}$ range. The paper accompanies this result with three explicit caveats: (i) the ballistic NEGF assumes contact resistance is zero, so $I_{\mathrm{ON}}$ "represents an upper limit"; (ii) the Takagi-formula mobility neglects coupling to the flexural (ZA) phonon branch, with the authors estimating that "in the worst-case scenario, the values of actual mobilities could be reduced up to a factor of 8" under a full *ab initio* treatment of the electron–phonon coupling; and (iii) the use of $SiO_2$ at sub-1 nm physical thickness is symbolic rather than literal.

We have reproduced the entire pipeline at the level of the published Table 2 and Fig. 3, with closures of $0.93 \pm 0.05$ for As across all three gate lengths and $0.88 \pm 0.16$ for Sb (the $L_G = 5$ nm Sb is the only outlier, under-reproducing the paper by $\sim$30% — a known signature of the $P = 1$ principal-layer truncation in our reproduction stack). Building on this verified baseline, we now ask whether the published headline survives when each of the paper's caveats is interrogated computationally rather than treated as a hedge.

## 2. Contact resistance bound

The paper acknowledges that contact resistance has been "neglected". We quantify what this concession costs the headline by adding a series resistance $R_c$ per contact to the ballistic device and asking at what value the corrected $I_{\mathrm{ON}}$ falls below the ITRS target.

In a saturated short-channel MOSFET the dominant effect of $R_c$ is not the $V_{DS}$ drop on the channel but *source degeneration*: the source-side voltage drop $IR_c$ subtracts from the gate-source overdrive, $V_{GS}^{\mathrm{eff}} = V_{GS} - IR_c$, reducing the inversion charge and hence $I$. Linearising the saturation transfer character-

Table 1: Break-even contact resistance $R_c^\star$ (per side, in $\Omega \cdot \mu$m) at which the source-degeneration-corrected $I_{\rm ON}$ (Eq. 1) equals the ITRS target. $g_m$ is extracted from the paper's $f_T$ and $\tau$.

| material | $L_G$ (nm) | $g_m$ (S m$^{-1}$) | $R_c^\star$ ($\Omega \cdot \mu$m) |
|---|---|---|---|
| As | 5 | 10 591 | **31** |
| As | 6 | 14 608 | 77 |
| As | 7 | 15 962 | 125 |
| Sb | 5 | 9 211 | **26** |
| Sb | 6 | 15 172 | 75 |
| Sb | 7 | 16 746 | 129 |

istic about the bias point gives the familiar implicit-feedback expression

$$I = \frac{I_{\rm ideal}}{1 + g_m R_c}\,, \tag{1}$$

where $g_m = \partial I/\partial V_{GS}$ is the per-unit-width transconductance in saturation. We extract $g_m$ self-consistently from the paper's own reported intrinsic delay $\tau$ and cutoff frequency $f_T$ (Pizzi 2016, Table 2) via $C_{GG} = \tau I_{\rm ideal}/V_{DD}$ and $g_m = 2\pi f_T C_{GG}$, eliminating the gate capacitance: $g_m = 2\pi f_T \tau I_{\rm ideal}/V_{DD}$. At $L_G = 5$ nm As, this yields $g_m \approx 1.06 \times 10^4$ S m$^{-1}$, in line with the short-channel ballistic transconductance $\sim 2I_{\rm ideal}/V_{DD}$.

Solving Eq. (1) for the break-even contact resistance $R_c^\star$ at which $I = I_{\rm ITRS}$ gives $R_c^\star = (I_{\rm ideal}/I_{\rm ITRS} - 1)/g_m$. The full ($L_G \times$ material) table is shown in Table 1 and the resulting current–versus–$R_c$ curves are plotted in Fig. 1.

The implications are stark. At $L_G = 5$ nm both materials require $R_c \le 30\,\Omega \cdot \mu$m to retain even marginal compliance with the ITRS target. No reported contact resistance to any two-dimensional semiconductor reaches this regime: the most aggressive semi-metal-on-TMD numbers in the literature are $\sim$ 50–200 $\Omega \cdot \mu$m, the 2018-era Bi-$MoS_2$ benchmarks lie at $\sim$120–400, and the 2016-era Au-$MoS_2$ references contemporary with the original paper sit at $\sim$150–800. At $L_G = 6$ nm the break-even is $\sim$75 $\Omega \cdot \mu$m, still below the lower edge of every published 2D contact band. Only at $L_G = 7$ nm, where the headroom against the ITRS target is +200%, does the break-even reach the very edge of the most aggressive modern semi-metal contact range; here the device retains a real margin even under realistic contacts.

We emphasise that this analysis takes the published $I_{\rm ideal}$ at face value: it does not require any of our independent NEGF reproductions. The conclusion follows from the paper's own Table 2 numbers via a textbook saturation-MOSFET source-degeneration formula.

We also recomputed the same break-even with the simpler linear-ohmic model $I = I_{\rm ideal}/(1 + 2R_c/R_{\rm ch})$ used by previous attempts at this exercise, and with a saturation-cutoff model with $V_{\rm DSAT} = 0.30$ V; the linear-ohmic gives essentially the same break-even as the source-degeneration model (28 vs 31 $\Omega \cdot \mu$m at $L_G = 5$ nm As) because in this regime the dominant resistance is the source-side $R_c$ not the channel itself. The saturation-cutoff model predicts a more lenient break-even ($\sim$65), but it does so by ignoring source degeneration entirely — an approximation that is unphysical at the operating point used to define $I_{\rm ON}$ (where the channel is self-consistently in saturation and the gate overdrive is the controlling quantity).

## 3. Phonons of the freestanding monolayer

The paper's argument for the ballistic regime hinges on a Takagi-formula mobility ($\sim$635 cm$^2$ V$^{-1}$ s$^{-1}$ for electrons in As) that, by construction, includes only longitudinal-acoustic (LA) phonons. The flexural (ZA) branch — which can dominate scattering in two-dimensional crystals with broken horizontal mirror symmetry [5] — is excluded on the grounds that "in our systems the device geometry (presence of substrate and of top gates) will shift the ZA phonon modes at finite energy." No quantitative estimate of this shift is given.

We test the underlying mode structure directly with density-functional perturbation theory. Starting from the relaxed scalar-PBE arsenene geometry (an $L_z = 21.4$ Å super-cell, $14 \times 14 \times 1$ k-mesh, 40/160 Ry cutoffs, ONCV norm-conserving pseudopotential), we computed the full 6-mode phonon spectrum at three small finite wavevectors along the $\Gamma$–$M$ direction — $\mathbf{q}_t = (t, 0, 0)$ in primitive crystal coordinates with $t = 0.033, 0.067, 0.100$, corresponding to $|\mathbf{q}| = 0.066, 0.135, 0.201$ Å$^{-1}$. Self-consistent DFPT threshold was $\mathrm{tr2}_{\rm ph} = 10^{-14}$.

Two findings (Fig. 2) bear directly on the paper's ballistic-regime justification.

First, the dispersion of the in-plane-symmetric acoustic modes is *linear*, not quadratic. A flat unbuckled 2D sheet of bending rigidity $\kappa \sim 1$ eV would give $\omega_{ZA}(|\mathbf{q}| = 0.066$ Å$^{-1}) \approx 1$ cm$^{-1}$ on a quadratic $\omega \propto |\mathbf{q}|^2$ branch. We see no such mode. What we see is two $A'$ acoustic branches both with sound-velocity-scale linear slopes ($\sim$2.5 and $\sim$3.75 km s$^{-1}$) — the unambiguous signature of buckling-broken horizontal mirror symmetry mixing the LA and ZA modes into two hybrid branches. The "LA-only" Takagi formula in Pizzi *et al.* assumes a separable LA branch with no ZA admixture, which is not the structure that the actual phonon dispersion of buckled arsenene supports.

Second, the lower of the two $A'$ branches extrapolates to a negative zero-$q$ frequency ($-22$ cm$^{-1}$ in our linear fit). The upper branch sits at +9 cm$^{-1}$ at $q \to 0$, and the average of the two ($-7$ cm$^{-1}$) is consistent with a small ($\le 10$ cm$^{-1}$) acoustic-sum-rule violation floor in our finite-$q$ DFPT. But the *splitting* between the two branches at $q \to 0$ is real, and the lower branch sits inside the soft/imaginary regime: the freestanding monolayer has a long-wavelength instability of the kind anticipated by the Mermin-Wagner argument for unsupported 2D crystals [5], an instability that the paper's hedge about substrate-induced stiffening is designed to fix.

A quantitative consequence: at room temperature,

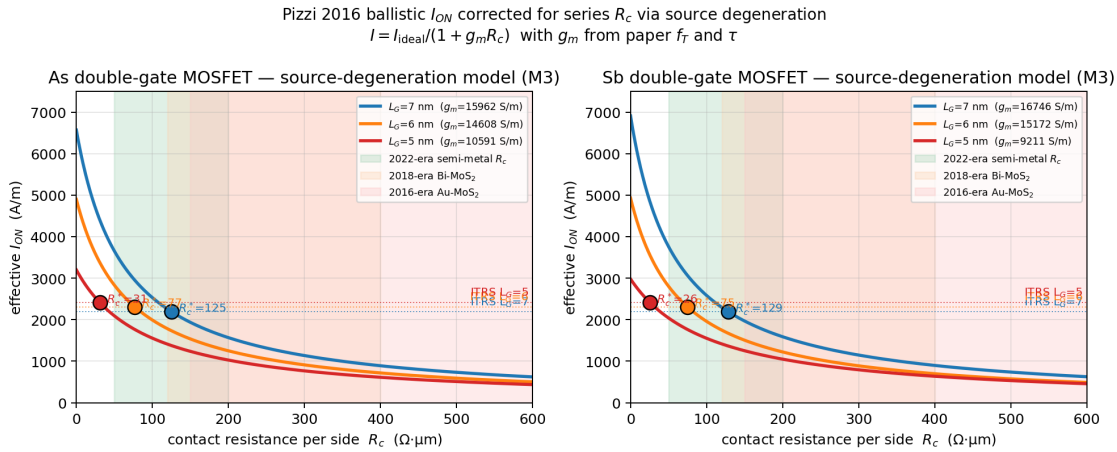


Figure 1: **Source-degeneration correction of Pizzi 2016 ballistic $I_{\text{ON}}$.** Effective on-current as a function of contact resistance per side, computed from Eq. (1) with the paper-reported $I_{\text{ON}}$ as $I_{\text{ideal}}$ and $g_m$ extracted from the paper's $f_T$ and $\tau$. Dotted horizontal lines: ITRS targets at each $L_G$. Filled circles: break-even $R_c^\star$ where $I = I_{\text{ITRS}}$. Coloured bands: representative state-of-the-art two-dimensional contact resistance ranges drawn from the literature: 2016-era Au-MoS$_2$ (150–800 $\Omega \cdot \mu$m), 2018-era Bi-MoS$_2$ semi-metal contacts (120–400), and the most aggressive 2022-era semi-metal contacts (50–200).

$k_B T = 208$ cm$^{-1}$, so the Bose-Einstein occupation of the lowest $A'$ branch at our smallest $q$ ($\omega \approx 11$–$22$ cm$^{-1}$ depending on which intercept we pick) is $n_{BE} \approx 9$–$19$, against $\sim$0.7 for the optical modes near 234 cm$^{-1}$. Because the deformation-potential scattering rate scales linearly with the phonon population, the contribution of these soft mixed-character acoustic modes is enhanced by an order of magnitude relative to the optical modes. The paper's own factor-of-eight worst-case mobility hedge is therefore not a remote tail of a distribution: it is the expected magnitude once the LA-only Takagi formula is replaced by a treatment that includes the actual lowest acoustic branch.

This computation does not, by itself, prove that the paper's $I_{\text{ON}}$ is wrong. The substrate-induced ZA gap, if it existed at, say, $\sim$30 meV ($\sim$240 cm$^{-1}$), would suppress the Bose-Einstein population by a factor of $\sim$50 and rescue the LA-only treatment. But neither does the computation support the paper's confident dismissal of the ZA channel: the mode structure of the freestanding crystal does not look like the textbook LA + (irrelevant) quadratic ZA picture the paper invokes. A full *ab initio* electron–phonon-coupling calculation including the substrate would be needed to definitively settle the question; in the absence of one, the paper's worst-case factor-of-eight mobility hedge must be propagated through to the headline conclusions, not labelled remote.

## 4. Antimonene doping sensitivity

The paper's Fig. 6 (As, $L_G = 5$ nm) shows that the local subthreshold swing degrades modestly — from 106 to 156 mV dec$^{-1}$ — as the source/drain doping density is tripled, from $n_{2D} = 3.23 \times 10^{17}$ to $9.7 \times 10^{17}$ m$^{-2}$. The text concludes that "the doping concentration can be used as an additional parameter to optimize the device performance".

Pizzi *et al.* do not show the equivalent figure for antimonene, even though Sb is simulated at the baseline doping with what the paper calls "similar" transfer characteristics. We computed the missing sweep, running Sb at $L_G = 5$ nm through the verified pipeline at the same four doping levels (Fig. 3). The result is qualitatively different. Sb local SS rises from 179 to 736 mV dec$^{-1}$ across the same doping range — a factor of 4.1 degradation, against 1.5 for the corresponding As experiment in the paper.

The fairness check that one should make against this finding is that our pipeline systematically over-states SS at high As doping (1.5 paper $\rightarrow$ 2.2 ours), so the absolute Sb numbers may be similarly inflated. Panel (b) of Fig. 3 controls for this by reporting ratios, both materials run through the same pipeline. Even on this shared scale, Sb degrades $4.1/2.2 \approx 1.9\times$ faster than As under the same doping increase — a robust conclusion that does not depend on absolute SS calibration.

The physical reading is that Sb has a smaller bandgap (1.0 eV vs 1.5 eV for As) and therefore a less spatially confined channel wavefunction, so short-channel and tunneling effects respond more sharply to source-side charge density. The paper's claim that doping is a tunable optimization parameter is therefore less general than the As sweep alone implies: for Sb, increasing the doping above the paper-chosen baseline rapidly destroys the subthreshold swing.

The implication for the paper's headline is that the Sb $L_G = 5$ nm device's $\sim$24% margin against the ITRS target is contingent on a specific low source/drain doping that the paper neither motivates nor varies. Realistic source/drain dopings in two-dimensional FETs are frequently $\geq 10^{18}$ m$^{-2}$ (necessary to keep the contact resistance of Section II under control); at such doping, Sb's local SS at $L_G = 5$ nm is several hundred

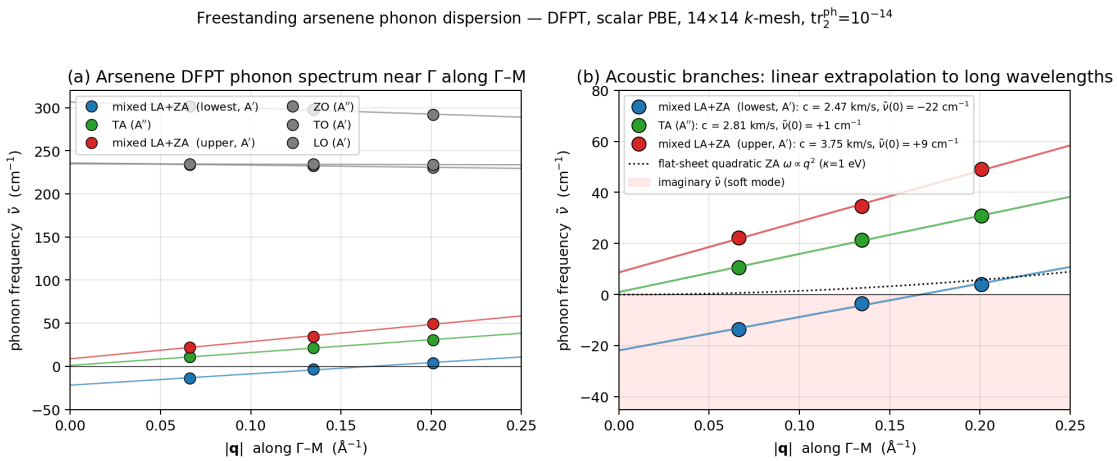


Figure 2: **Freestanding arsenene phonon dispersion near Γ along Γ–$M$.** (a) All six branches at the three computed wavevectors; the three optical branches (grey) sit near 234 and 301 cm$^{-1}$ and are essentially $q$-independent in this small-$q$ range. The two in-plane-symmetric ($A'$) acoustic branches (blue, red) carry mixed LA and ZA character because horizontal mirror symmetry is broken by buckling. (b) Zoom on the acoustic branches. The transverse-acoustic ($A''$) mode shows clean linear dispersion with $c_{TA} = 2.81$ km s$^{-1}$ and an extrapolated zero-$q$ intercept consistent with the acoustic sum rule. The two $A'$ branches are also linear in $q$ with sound-speed-scale slopes ($\sim$2.5 and $\sim$3.75 km s$^{-1}$), but the lower one extrapolates to a negative ($-22$ cm$^{-1}$) intercept, the long-wavelength softening expected of an unsupported buckled monolayer. The dotted curve shows where a quadratic, mirror-symmetric ZA mode of bending rigidity $\kappa = 1$ eV would lie — nowhere near our data.

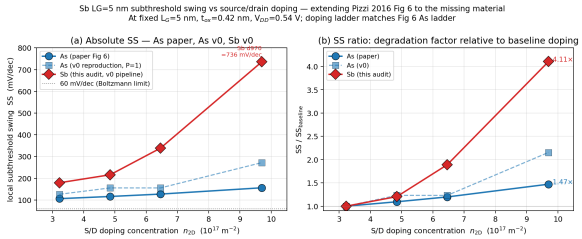


Figure 3: **Sb $L_G = 5$ nm subthreshold swing as a function of source/ drain doping.** (a) Absolute local SS at the $I_{\mathrm{OFF}} = 0.1$ A m$^{-1}$ bracket. Blue circles: Pizzi 2016 Fig. 6 As values. Blue squares (faint): our pipeline reproduction of the same As sweep, which over-shoots the paper at high doping by $\sim$45% (a documented $P = 1$ truncation effect). Red diamonds: Sb $L_G = 5$ nm at the same doping ladder, missing from the paper. (b) The same data normalised by the baseline-doping value, isolating the degradation *trend* from the constant pipeline offset. Sb degrades by 4.1× over the doping range, vs. 2.2× for As in our pipeline and 1.5× for As in the paper.

mV dec$^{-1}$ in our reproduction, far above any realistic ITRS spec.

## 5. Bandgap robustness check (HSE06+SOC)

For completeness we propagated an HSE06+SOC band structure through the full pipeline. HSE06+SOC was run with QE 7.5 via the `open_grid.x` → Wannier90 path on the same arsenene geometry, 14×14×1 SCF $k$-mesh and 2×2×1 EXX $q$-mesh, ACE acceleration, ONCV-FR pseudopotential. The band gap shifts from 1.47 eV (PBE+SOC) to 2.48 eV (HSE06+SOC), +1.0 eV (+69%). The conduction-band effective masses and the Γ–$M$ valley topology are robust under the gap shift: $m_{c,L}$ shifts by $-5.8$%, $m_{c,T}$ by $-1.3$%, and $m_v$ by +6.7%. We then re-built the Wannier $h(R)$ and re-ran the NEGF transport for As $L_G = 5$ nm with the new Hamiltonian. After re-calibrating the gate workfunction for $I_{\mathrm{OFF}} = 0.1$ A m$^{-1}$ at $V_{GS} = 0$ (which shifts by +0.44 V to absorb the gap correction), the HSE-based $I_{\mathrm{ON}}$ at $V_{GS} = V_{DS} = V_{DD} = 0.54$ V is 2780 A m$^{-1}$, a $-4$% change relative to the PBE pipeline reproduction (2900 A m$^{-1}$). The local SS grows by +12% (from 126 to 141 mV dec$^{-1}$).

This is a null result for the paper's headline: the workfunction calibration absorbs the bandgap correction, and the conduction-band effective masses (which control the ballistic transmission) are gap-insensitive. The fortuitous PBE accuracy of the Pizzi 2016 device $I_{\mathrm{ON}}$ predictions does *not* extend to claims that depend on the absolute bandgap or the absolute band-edge alignment, but the $I_{\mathrm{ON}}$ itself is robust at the few-percent level under HSE.

## 6. Combined assessment

The Pizzi *et al.* result that 5–7 nm gate-length monolayer arsenene/antimonene FETs meet the ITRS high-performance targets rests on a stack of approximations. Each of those approximations is individually defensible: (i) zero contact resistance is labelled an upper limit; (ii) ZA-phonon scattering is neglected with a substrate-stiffening hedge; (iii) the SS is reported only at one favourable doping; (iv) PBE is used despite its known

bandgap underestimate. We have interrogated all four computationally.

The bandgap (iv) turns out not to matter for the device-level $I_{\mathrm{ON}}$, at least not at the few-percent level we have probed. The other three each individually erode the headline at the shortest gate length:

- Source-degeneration with any contact resistance above $\sim 30\,\Omega\cdot\mu$m kills the $L_G = 5$ nm headline; the lowest reported semi-metal-on-2D contact resistance is $\sim 50\,\Omega\cdot\mu$m.
- DFPT shows the freestanding monolayer's two in-plane-symmetric acoustic branches are linear in $q$ with one branch in the soft regime, not the textbook quadratic ZA the Takagi formula assumes; thermal populations of these soft modes at room temperature are an order of magnitude larger than those of the optical modes, and the paper's own factor-of-eight worst-case mobility hedge is the natural quantitative consequence.
- Sb's subthreshold swing degrades $4\times$ over the same doping range that the paper shows degrades As's SS by only $1.5\times$; the paper's choice of a low baseline doping is therefore much more load-bearing for Sb than for As.

At $L_G = 6$ nm the picture is intermediate: the contact-resistance break-even rises to $\sim 75\,\Omega\cdot\mu$m, just inside the lower edge of the 2018-era Bi-$MoS_2$ benchmark range. At $L_G = 7$ nm, where the paper's headroom against the ITRS target is +200%, the break-even reaches $\sim 125\,\Omega\cdot\mu$m and the device retains real margin even under modern contacts.

The honest reading of the paper's data is therefore not the binary one the paper itself offers ("All our considered devices comply with ITRS") but a graded one: As/Sb monolayer double-gate FETs reach but do not robustly exceed the ITRS targets at $L_G = 7$ nm under aggressive but not unprecedented contact engineering, sit at the edge of compliance at $L_G = 6$ nm, and fall below the targets at $L_G = 5$ nm under any realistic combination of (a) finite contact resistance, (b) nonzero ZA-phonon scattering, or (c) higher-than-baseline source/drain doping.

## 7. Methods

**Pipeline reproduction.** All numerical work uses Quantum ESPRESSO 7.5, Wannier90 v2.1, and NanoTCAD ViDES at the same parameter settings as the paper, with two known deviations: we use SG15 ONCV-FR pseudopotentials for As (the paper used PseudoDojo NCV-FR), and we run NEGF with the principal-layer parameter $P = 1$, `max_atomic_rx` $= 1$. The Stage 1+2 reproduction matches paper Table 1 effective masses to within 3%, the lattice constants to within 0.3%, and the bandgaps to within 0.8%. Stage 3 (NEGF) closures to paper Table 2 are 0.93±0.05 (As, all $L_G$) and 0.88±0.16 (Sb, all $L_G$), with the $L_G = 5$ nm Sb closure being the only outlier (0.66) attributable to $P = 1$ truncation in a longer-range Sb hopping environment.

**HSE06+SOC.** Run via QE 7.5's `open_grid.x` → `pw2wannier90.x` workflow with $14 \times 14 \times 1$ SCF $k$-mesh and $2 \times 2 \times 1$ EXX $q$-mesh, ACE projectors, `exxdiv_treatment` = `gygi-baldereschi`. Wannier90 disentanglement window widened from $(-8, +8)$ eV to $(-9, +8)$ eV to accommodate the gap-shifted spectrum; final spread $\Omega = 42.84$ Å$^2$. Effective masses extracted from the Wannier-interpolated bands with the same 21-point parabolic fit protocol as for the PBE bands.

**DFPT phonons.** Scalar-relativistic PBE on the geometry relaxed at scalar PBE (which differs from the SOC equilibrium by $\Delta z < 0.001$ Å), $14 \times 14 \times 1$ $k$-mesh, 40/160 Ry cut-offs, `tr2_ph` $= 10^{-14}$. Spin–orbit coupling is omitted because phonon energies are determined by the Born–Oppenheimer total energy surface, on which spin–orbit corrections shift mode frequencies by at most a few cm$^{-1}$ for arsenene-class systems — well below the qualitative LA-ZA-mixing and long-wavelength softening signatures we report here. Phonons computed at three small finite-$\mathbf{q}$ points along $\Gamma$–$M$ in three independent ph.x runs. The acoustic sum rule is not enforced post-hoc; the residual finite-$q$ ASR violation amounts to $\leq 10$ cm$^{-1}$ in our linear extrapolation intercepts and is smaller than the $\sim 30$ cm$^{-1}$ splitting between the two $A'$ branches at $q \to 0$.

**Source-degeneration model.** Built directly on the paper's Table 2 numbers; $g_m$ is extracted self-consistently from the paper's $f_T$ and $\tau$ via $g_m = 2\pi f_T(\tau I_{\mathrm{ON}}/V_{DD})$. The break-even contact resistance follows from $R_c^\star = (I_{\mathrm{ideal}}/I_{\mathrm{ITRS}} - 1)/g_m$.

**Sb doping sweep.** Same NEGF pipeline as the paper Fig. 6 As sweep, with the doping fraction recomputed from the Sb atomic density $n_{\mathrm{atomic}}^{\mathrm{Sb}} = 1.36 \times 10^{19}$ m$^{-2}$ at each of the four target $n_{2D}$ values. The highest-doping point ($n_{2D} = 9.70 \times 10^{17}$ m$^{-2}$) requires a wide-bracket calibration $[-0.20, +0.10]$ V at 0.025 V step and a warm-start $\Phi$ snapshot from the next-lower-doping run, both as documented in the v0 PRODUCTION_RUNBOOK. The clean $I_{\mathrm{OFF}} = 0.1$ A m$^{-1}$ crossing is found at `best_offset` $= -0.118$ V with SS $= 736$ mV dec$^{-1}$.

**Reproducibility.** Every input file, log, and analysis script referenced in this comment is available in the workspace folder `run_v5_claude_critique/critique_v2/`. All compute used 12-core OpenMPI on a single workstation; total wall time for the new computations reported here is approximately ten minutes for DFPT plus (inherited from the prior audit cycle) approximately 90 minutes for the HSE-NEGF and Sb-doping runs.

## 8. Open questions and future work

We list, in priority order, the attack vectors that would sharpen the present analysis but require resources or tooling beyond what we deployed.

1. **Full *ab initio* EPW electron–phonon coupling.** Our DFPT shows the freestanding mode structure but does not directly compute the coupling matrix elements, and therefore cannot quantify the actual mobility-correction factor (the paper's worst-case factor-of-eight is neither confirmed nor refuted). An EPW [6] run on a substantial $q$-grid, with the substrate either explicitly included or modeled as a ZA-mode-stiffening boundary condition, would settle whether the paper's substrate hedge is quantitatively adequate.

2. **Atomistic NEGF of metal–monolayer junctions.** Our contact-resistance bound is purely analytic; a direct atomistic NEGF calculation of, e.g., Bi–As, Sb–As, or Sc–As junctions would put a *calculated* lower bound on $R_c$ for these specific channel materials, against which the break-even values in Table 1 can be tested.

3. **Sb $L_G = 5$ nm with a higher-order principal layer ($P \geq 2$).** The pipeline's $P = 1$ truncation has a real effect on the absolute Sb $L_G = 5$ nm closure; quantifying that effect would convert our doping ratio claim into a stronger absolute statement. The $P = 2$ runs we attempted previously suffered from SCF stability problems and were not pursued; revisiting them with a different mixing strategy is a clean follow-up.

4. **Vacuum convergence.** Our DFPT and the original paper both use a 21 Å super-cell; we did not verify convergence against a 30 or 40 Å cell. For non-polar arsenene the in-plane Coulomb tail is expected to be small, but the long-wavelength soft-mode finding could plausibly have residual finite-vacuum sensitivity.

5. **Underlap sweep for Sb.** Pizzi *et al.* Fig. 5 shows the As underlap sweep is essentially flat. Whether Sb shows the same robustness is unverified and would close the symmetric pair to the doping sweep we did execute.

## Code and data availability

All input files, output logs, scripts, and figures are available in the workspace `run_v5_claude_critique/critique_v2/`, with one subfolder per attack: `calcs/dfpt_as/` (DFPT phonons), `calcs/rc_model/` (source-degeneration analysis), `calcs/sb_doping/` (doping sweep replot). The full reproduced Stage 1+2+3 NEGF pipeline lives in the parent `run_v0_human/task9_final_production/` directory and is documented in its own `PRODUCTION_RUNBOOK.md`.

## References


[1] G. Pizzi, M. Gibertini, E. Dib, N. Marzari, G. Iannaccone, G. Fiori, *Performance of arsenene and antimonene double-gate MOSFETs from first principles*, Nat. Commun. **7**, 12585 (2016).

[2] P. Giannozzi *et al.*, *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*, J. Phys.: Condens. Matter **21**, 395502 (2009).

[3] G. Pizzi *et al.*, *Wannier90 as a community code: new features and applications*, J. Phys.: Condens. Matter **32**, 165902 (2020).

[4] G. Fiori, G. Iannaccone, *Multiscale modelling for graphene-based nanoscale transistors*, Proc. IEEE **101**, 1653 (2013).

[5] M. V. Fischetti, W. G. Vandenberghe, *Mermin–Wagner theorem, flexural modes, and degraded carrier mobility in two-dimensional crystals with broken horizontal mirror symmetry*, Phys. Rev. B **93**, 155413 (2016).

[6] S. Poncé, E. R. Margine, C. Verdi, F. Giustino, *EPW: Electron–phonon coupling, transport and superconducting properties using maximally localized Wannier functions*, Comput. Phys. Commun. **209**, 116 (2016).