Grounded autonomous research: a fault-tolerant LLM pipeline from corpus to manuscript in frontier computational physics

Anonymous Authors[1]

Abstract

Autonomous-research agents have demonstrated end-to-end LLM automation in machine-learning sandboxes where execution provides calibration. Frontier physical science differs categorically: physical reasoning underlies parameter choices, and frontier toolchains and methodologies are often underdocumented. Researchers handle this by grounding in literature throughout — from idea conception, through methodology calibration, to manuscript writing. Unscaffolded LLM agents systematically fail to do this, hallucinating from internal priors and confidently producing plausible but unverifiable results.

Our pipeline bridges this gap, running end-to-end from a corpus of 11,083 recent condensed-matter physics arXiv papers to a publication-grade manuscript with three substantive physics findings (here on altermagnetic piezomagnetism). The agent autonomously conceives a research direction by mapping the corpus, calibrates methodology by reproducing published references, conducts novel first-principles computations, and writes the manuscript — grounded in literature throughout. The pipeline runs as 47 fresh-context sessions across six phases sharing only on-disk state, with 2162 literature-consultation events. Fault tolerance emerges from redundancy: fresh-context isolation, distributed grounding, and adversarial review catch what any single session misses. Pre- and post-pilot stages are fully autonomous; pilot requires bounded human intervention only at reproduction failures — operational knowledge curation, not scientific direction.

Grounding in literature is universal across real research. The primitives, lessons, bounded human intervention, and characterized failure modes articulated here lay a foundation for autonomous research beyond computational physics, in high-stakes realistic scientific domains.

1. Introduction

AI-for-research has progressed from tool-assistance — search, summarization, code completion (Taylor et al., 2022; Chen et al., 2021; Bran et al., 2024) — to agentic automation that conducts research with diminishing human oversight (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Miao et al., 2025). End-to-end automated research has been demonstrated in domains where execution itself provides intrinsic calibration: machine-learning sandboxes where training-loss curves verify methodology (Lu et al., 2024; Yamada et al., 2025), defined computational tasks with established reference values (Miao et al., 2025), and pre-formulated benchmarks where ground truth is given. Within these calibration-friendly domains, agent reliability has nonetheless been critiqued (Si et al., 2024; Kambhampati et al., 2024; Beel et al., 2024): agents over-confidently produce plausible outputs, hallucinate from internal priors, and propagate methodology errors that internal-consistency checks do not catch.

As the field pushes toward genuine frontier research — open direction selection from large literature corpora, novel parameter spaces, underdocumented toolchains where standard recipes do not exist — execution-as-calibration breaks down structurally. Frontier physical science is the canonical case: physical reasoning is required at every methodology choice, toolchains are often poorly documented, and calibration must come from external literature anchors because internal execution provides no ground truth. Without external anchor enforcement, agents default to internal-prior generation, producing confident plausible outputs that cannot be verified by execution alone. The same critiques leveled at simpler automated systems become disqualifying at frontier: an unverifiable result in an ML sandbox is recoverable; an unverifiable result in a physical-science methodology can propagate through to manuscript-level conclusions. This is becoming a blocking problem for the next iteration of automated research.

Real researchers handle this by grounding in literature

---

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

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

throughout: from idea conception that surveys what is known and what is open, through methodology calibration that reproduces published references before applying methods to novel computations, to manuscript writing that contextualizes findings against the existing literature. Each function relies on literature in different ways. "Grounding in literature" is generic terminology, however — what does it mean operationally for an autonomous pipeline?

We characterize grounding as *structurally enforced literature confrontation at calibration checkpoints — more than literature availability or consultation count alone*. At calibration checkpoints, the pipeline forces numerical comparison between agent-computed values and published reference values — not merely citation that references exist. We characterize this distinction empirically via two paired failure modes (§4): a pre-architecture baseline run that lacked the topic-selection grounding to reject un-calibratable directions, and a controlled ablation in which curated knowledge and house rules were inherited but the pilot reproduction phase that enforces numerical confrontation was skipped. Together they identify the pilot stage as the architectural element that operationalizes the abstract claim of "grounded throughout".

Two concurrent studies on related autonomous-research themes provide context. Huang (2026a) introduces a persistent knowledge system that lets an LLM agent write down knowledge learned from execution and retrieve it in later sessions, characterizing failure modes in cross-session knowledge transfer. Huang (2026b) establishes grounded autonomous scrutiny at scale: an agent reads a published computational-physics paper, autonomously plans and re-executes the underlying calculations end-to-end, and identifies what does and does not hold; across 111 open-access papers it raises substantive methodological concerns on $\sim$42%, of which 97.7% emerge only after execution against a reading-only ceiling of 0.9%; in depth, on a 2D-MOSFET multiscale simulation paper it produces an unsupervised publishable Comment that revises the headline conclusion. The present work runs the full corpus-to-manuscript pipeline on autonomous research-direction selection from open literature, producing a submission-grade artifact in altermagnetic piezomagnetism (companion physics manuscript bundled in §H). We choose computational condensed-matter physics as the frontier scope: it has well-characterized symmetry constraints (e.g., point-group constraints on Hall conductivity), reproducible reference frameworks against which calibration is feasible, publicly available software stacks (Quantum ESPRESSO (Giannozzi et al., 2017), Wannier90 (Pizzi et al., 2020), WannierBerri (Tsirkin, 2021)), and an open arXiv corpus base. The paper proceeds: §2 sketches the pipeline architecture; §3 walks through the pipeline; §4 demonstrates the central thesis empirically via two paired failure modes; §5 returns to the abstract's principles operationalized.

2. Pipeline overview

Six-phase architecture. The pipeline runs as six phases (Fig. 1A) with an architectural template distinguishing structurally fixed elements from run-specific iteration counts. *Breadth* samples the corpus to extract candidate research themes; the number of breadth sessions is chosen for diversity. *Depth* expands the most promising candidates into research questions with explicit reproduction targets, each session force-generating a new direction not yet committed by previous depth sessions. *Pilot* has fixed structure: program selection (S1), tooling check (S2), k published-reference reproductions, and an iterative gate–review structure with $4\times$gate $+ 1\times$review cycles bounded by the hyperparameters in §3.2. *Pre-production* (1 session) consolidates pilot outputs into a locked production plan and methodology handoff. *Production* runs as 1 initialization plus adversarial continuations until the agent issues PASS. *Writing* is fixed at three drafting triads plus a polish session ($3\times3+1=10$ sessions). For the canonical run reported here, $n_{\mathrm{breadth}}=3$, $n_{\mathrm{depth}}=5$, $k=5$ pilot reproductions, two pilot iteration cycles, and three production continuations.

Two-component scaffolding. Each phase is implemented as one or more LLM sessions, each beginning with a fresh context window: no shared in-memory state, no conversational history. Sessions communicate only through on-disk artifacts. Two architecturally distinct scaffolding components thread through the pipeline. The *curated knowledge base* comprises two files. `INDEX.md` holds curated general knowledge about computational-physics tools and practice — a single source of truth for "what can the simulation toolchain do and how" verified on the local stack, indexing 38 verified workflows, 17 reusable analysis scripts, parameter-variation references, and common operational gotchas; it contains no research-direction-specific content. `PSEUDOPOTENTIALS.md` maps the local pseudopotential library with best-practice selection guidance per workflow. The *house rules* (`PILOT_HOUSE_RULES.md`) enumerate mandatory operational constraints: computational-resource discipline (one heavy job at a time, checkpoint/resume protocols), MUST-do diagnostic steps (Wannier band-fit verification; `projwfc`/fatband analysis before basis design), and disposition criteria for converged-vs-not verdicts. Both components are loaded only from pilot S2 onward (Fig. 1A bottom band): they are deliberately absent during conception phases (breadth, depth, pilot S1) to avoid biasing open idea generation toward known methodology. The pipeline thus distinguishes *conception grounding* (open exploration of literature) from *execution grounding* (anchor enforcement via curated knowledge and house rules). Across the 47 sessions, the pipeline performs 2,162

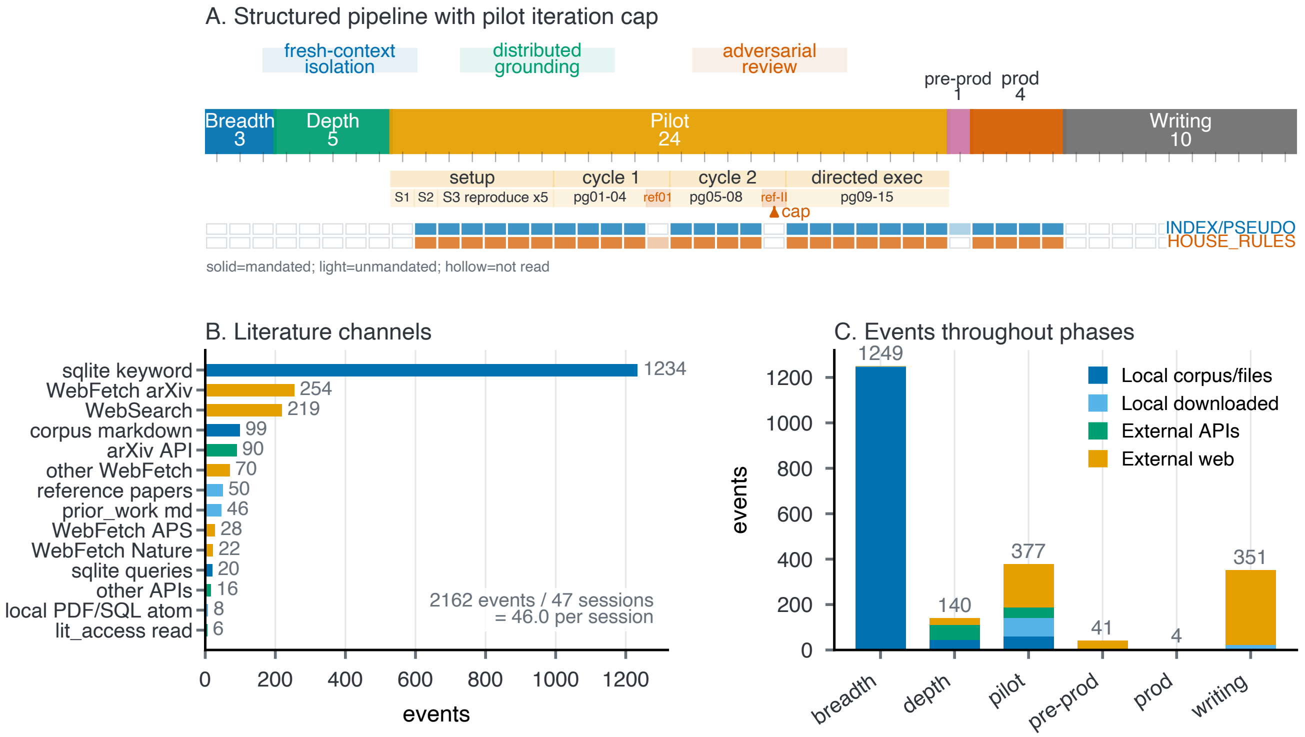


Figure 1. Pipeline architecture and literature footprint. **(A)** Six phases run as 47 fresh-context LLM sessions sharing only on-disk state. The pilot lane iterates a computational-gate plus adversarial-review unit; an iteration-cap hyperparameter upgrades the next review to a transition-planning role rather than continuing iteration (details and trade-off in §3). The bottom band depicts scaffolding access: the *curated knowledge base* (`INDEX.md`, `PSEUDOPOTENTIALS.md`) and *house rules* (`PILOT_HOUSE_RULES.md`) are deliberately absent during conception phases (breadth, depth, pilot S1) and structurally loaded from pilot S2 onward. **(B)** 14-channel literature breakdown across the 47 sessions (2,162 events). **(C)** Per-phase distribution: heavy at breadth (corpus mining), substantial at pilot and writing, near-silent at production.

unique literature-consultation events distributed across 14 access channels and concentrated in conception and pilot phases (Fig. 1B,C).

3. Pipeline walkthrough

This section walks through each phase as the agent conceived altermagnetic piezomagnetism from the corpus and ran the pipeline through to manuscript, noting where the architecture's fault-tolerance mechanisms appear in context.

3.1. Idea conception from corpus

The breadth phase mapped an arXiv corpus of 11,083 condensed-matter physics papers across three independent breadth sessions running in parallel (Fig. 2); each session used complementary access strategies. The three reports cite 877 distinct arXiv IDs collectively, with only 27 (3.1% of the union) appearing in all three — the breadth agents are 80% complementary, not redundant. Five depth-phase directions emerged from this distillation, two of which trace primarily to single-channel surfacings — papers that only one of the three breadth agents identified. The cleanest case is the Bell–Venderbos piezomagnetism cluster: four arXiv IDs surfaced only by breadth$_{01}$'s altermagnetism sub-agent became the framework of the chosen direction. Empirically, redundancy works because complementarity is high. Depth-phase novelty audits reached external sources — OpenAlex, the arXiv API, WebFetch, and WebSearch (top external band in Fig. 2) — both for older reference papers and for in-window papers the breadth agents missed. A depth-phase diversity ledger forced different agent sessions to commit to non-overlapping directions. Table 1 summarizes the five candidates that emerged from this depth phase, with the agent's primary reason and verdict alongside an independent post-hoc human review.

3.2. Pilot stage

The pilot phase contains the architectural elements that operationalize the abstract's "grounded throughout" claim. It runs as a fixed structure of program selection (S1) and tooling check (S2), followed by reproductions of published references and an iterative gate–review computational sequence. Three fault-tolerance mechanisms appear in this stage: *fresh-context isolation* prevents conclusions from one

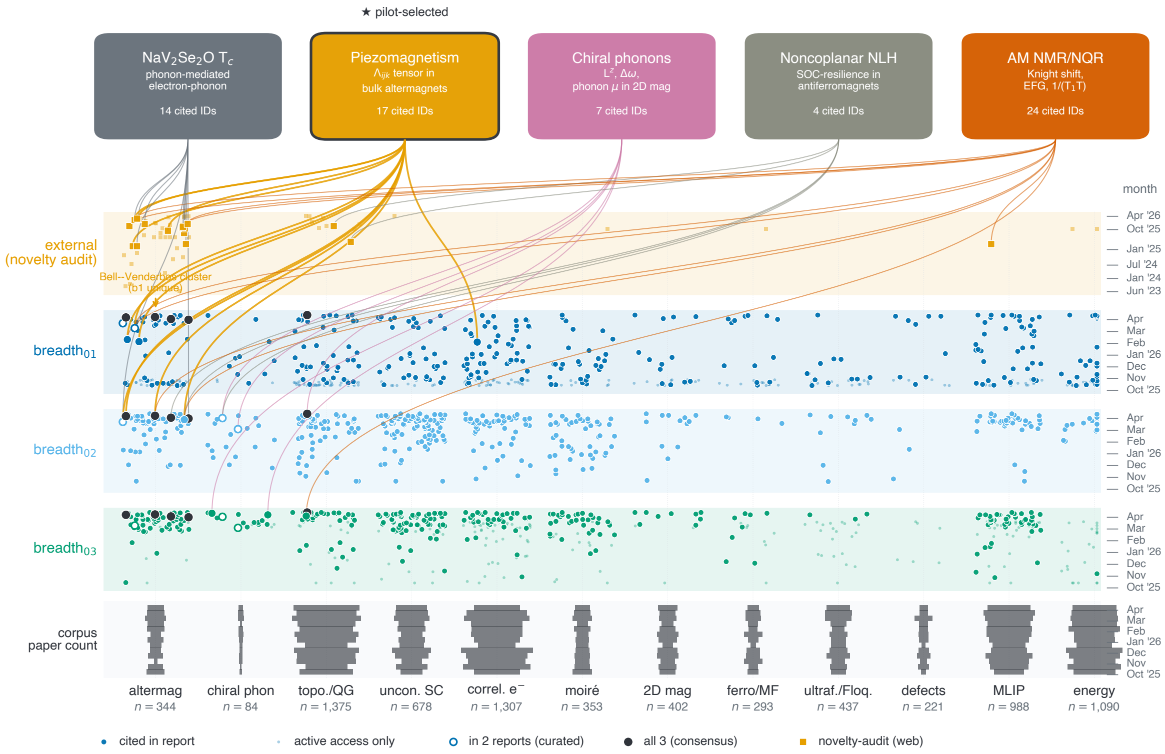


Figure 2. Information flow from corpus to depth programs. Each dot is one arXiv ID; large filled markers are IDs cited in a breadth report or by a depth program, small faint markers are IDs the agent actively retrieved but did not cite. Position: x = regex-classified theme (12 categories cover 99.1% of breadth-cited and 100% of depth-cited IDs); y = continuous arXiv submission month. Bands top-to-bottom: *external* (depth-phase novelty audit via OpenAlex / arXiv API / WebFetch / WebSearch, Jun 2023 – Apr 2026); breadth$_{01}$ (10 parallel `Explore` sub-agents); breadth$_{02}$ (title walk + 35 SQL queries); breadth$_{03}$ (66 SQL queries). Same arXiv ID has identical x across bands: consensus-core IDs (cited in all three reports) appear as filled black markers stacked vertically; pair-overlap IDs as open rings. Five depth-program boxes at top show committed direction, total cited IDs, and pilot-selected program (★ piezomagnetism); arcs trace each cited ID upward to its destination, with bold arcs for the selected program. Bottom strip: per-theme corpus paper-count histogram in 13 fortnight bins (Oct 2025 – Apr 2026), absolute scale shared across themes.

session being uncritically carried forward by the next; *distributed grounding* redirects a reference missed at one phase to surface at another; *adversarial review* formalizes "catch what the agent itself overclaimed" through dedicated reflect, production-continue, and polish sessions prompted to find rather than confirm.

S1 program selection, S2 tooling, S3 reproductions. S1 selected altermagnetic piezomagnetism from the five depth-phase candidates (Table 1) and fixed the reproduction set: five published papers targeted as numerical anchors for the pilot reproduction sessions. The independent post-hoc review confirms this selection: piezomagnetism is uniquely well-anchored among the five, with three independent published reproduction targets for the headline observable; the four dropped directions either require methodology-development scope (NaV$_2$Se$_2$O T_c, AM NMR/NQR), sit on a young toolchain undertested for the target system (chiral phonons), or are bounded in novelty by an existing study (noncoplanar NLH). This selection gates a deeper point: had the agent committed to one of the four dropped directions and produced quantitative findings, no published reference recipe would exist against which to confront those findings; without external anchor, the pipeline cannot grade its own output and the manuscript artifact loses calibration. Topic-selection grounding is the precondition for execution grounding. S2 verified the local toolchain and activated the curated knowledge base and house rules in agent context for subsequent pilot sessions. S3 reproduced each anchor paper individually, comparing computed headline observables against literature values to within tier thresholds ($\sim$30% first-tier; $\sim$10% strict-tier); each reproduction session was bounded, failure producing documented uncertainty rather than infinite iteration.

Iteration architecture: gate–review cycles. The pilot phase then operates as bounded iteration: each cycle consists of computational gate sessions followed by an adver-

Table 1. The five candidate research directions evaluated in the depth phase, grouped by source: agent (the pipeline's own composite-score reasoning; verdict labels verbatim from the pilot S1 program-selection report) and an independent post-hoc human review checking each disposition against current toolchain documentation and published literature.

	Agent (pipeline)		Human (post hoc)	
Direction	Reason	Verdict	Reason	Verdict
NaV_2Se_2O T_c	DFPT+EPW T_c for an altermagnetic superconductor at U-sensitivity threshold with competing AM ground states within DFT noise.	Strong but risky	Magnetic-EPW toolchain available only for collinear ferromagnets as of late 2025 (Lihm et al., 2025); Allen–Dynes assumes spin-singlet pairing forbidden in altermagnets (Mæland et al., 2025).	Drop: methodology development
AM piezomagnetism	Strain-induced net magnetization in bulk altermagnets; reproduction targets exist.	**Best composite (selected)**	Three independent reproduction anchors verified (Lukashev et al., 2008; Bell and Venderbos, 2026; Ye et al., 2026).	**Selection grounded**
Chiral phonons (2D magnets)	Mode-resolved L^z in 2D magnetic insulators; target signal $\sim$1 meV near a 2–3 meV noise floor.	Noise-floor risk	Toolchain (Liu et al., 2025; Huang et al., 2026) is 2025–2026 vintage and undertested for 2D magnetic insulators.	Defensible drop
Noncoplanar NLH	SOC-on/off comparison of the magnetic-geometry channel of intrinsic nonlinear Hall, extending an established study.	Safe but incremental	Question and toolchain bounded by (Zhu et al., 2025); SOC-resilience generalization is incremental on a mature stack.	Drop grounded
AM NMR/NQR	Sublattice Knight shifts and EFG fingerprints; pipeline requires QE-CONVERSE + custom Wannier-$\chi(q)$ bridge.	Visionary but risky	QE-CONVERSE (Ceresoli et al., 2025) demonstrated only on paramagnetic systems; no altermagnet *ab initio* NMR papers in print as of mid-2026.	Drop: methodology development

sarial review session, with two architectural hyperparameters bounding total iteration. `max_pilot_cycle = N` caps the number of normal review sessions before the next review is upgraded to a transition-planning role; `break_action` ∈ {`reflect_II`, `abort`} determines whether the cap triggers transition to production with documented uncertainty or termination of the run. The canonical pipeline selected `max_pilot_cycle = 1` and `break_action = reflect_II`.

The canonical pipeline's actual operation (Table 2 top row): cycle 1 ran four gate sessions plus a review session that retracted an over-confident HIGH verdict (the agent's own confidence rating; HIGH/LOW are the binary outcomes of the composite-score check) to LOW; cycle 2 activated the cap, with the second review upgraded to plan transition into production rather than continuing to iterate. The choice is project-dependent: hypothesis generation justifies a smaller cap, definitive quantitative claims a larger one.

Table 2. Architectural trade-off for the pilot iteration cap. The selected setting favors maximum output per fixed compute budget: a calibrated manuscript artifact with characterized caveats, at the cost of remaining convergence work being deferred to the next iteration.

Choice	Output	Compute	Future basis
`max_pilot_cycle = 1`, `reflect_II` (selected)	Anchored manuscript with characterized caveats	Moderate	High
`max_pilot_cycle ≥ 2`	Same artifact, tighter uncertainty bounds	High	Higher
`break_action = abort`	None (no manuscript, no calibration retained)	Low	Low

Calibration trajectory: the orbital-magnetization case study. The pilot exercises all three fault-tolerance mechanisms in the calibration trajectory of MnTe orbital magnetization (Fig. 3A). Altermagnets (Šmejkal et al., 2022; Bhowal and Spaldin, 2024; Hayami et al., 2025; Mazin, 2023) are compensated antiferromagnets with broken time-reversal symmetry; their orbital magnetization m_{orb} (distinct from spin magnetization) is the calibration anchor for the production strain-derivative computation. Four moments structure the trajectory.

(1) Initial overshoot. The agent's first MnTe orbital magnetization produced $|M_z| \approx 0.43\ \mu_B$/cell, 2.44× above the eventually-confirmed published reference, declared HIGH confidence on internal symmetry checks alone (forbidden components $|M_x|, |M_y|$ at noise floor; symmetric-tensor structure verified). The qualitative checks all passed; only numerical comparison against an external published anchor could catch the magnitude error — exactly the failure mode unscaffolded LLM agents systematically exhibit (§1).

(2) Anchor enforcement in action. In the subsequent adversarial review session, opening in fresh context, six sub-agents were deployed prompted to find rather than confirm. One sub-agent's literature gap-hunt surfaced the Ye 2026 reference value ("$M_{\mathrm{orb}} = 0.176\ \mu_B$/cell along z-axis") (Ye et al., 2026); the reflect prompt forced the comparison (2.44×, opposite sign), and HIGH was retracted to LOW. The architecture forced the agent to compare against the published reference, not just cite that the reference exists — anchor enforcement operationally: confrontation, not citation.

(3) Convergence verification — the architecture's NOT PASS verdict. The agent calibrated the recipe at one mesh density and matched the Ye reference within 10%; testing the same recipe at a denser mesh produced 70% deviation. The agent issued NOT PASS on the convergence point. Rather than continuing to iterate (exceeding the cap), the

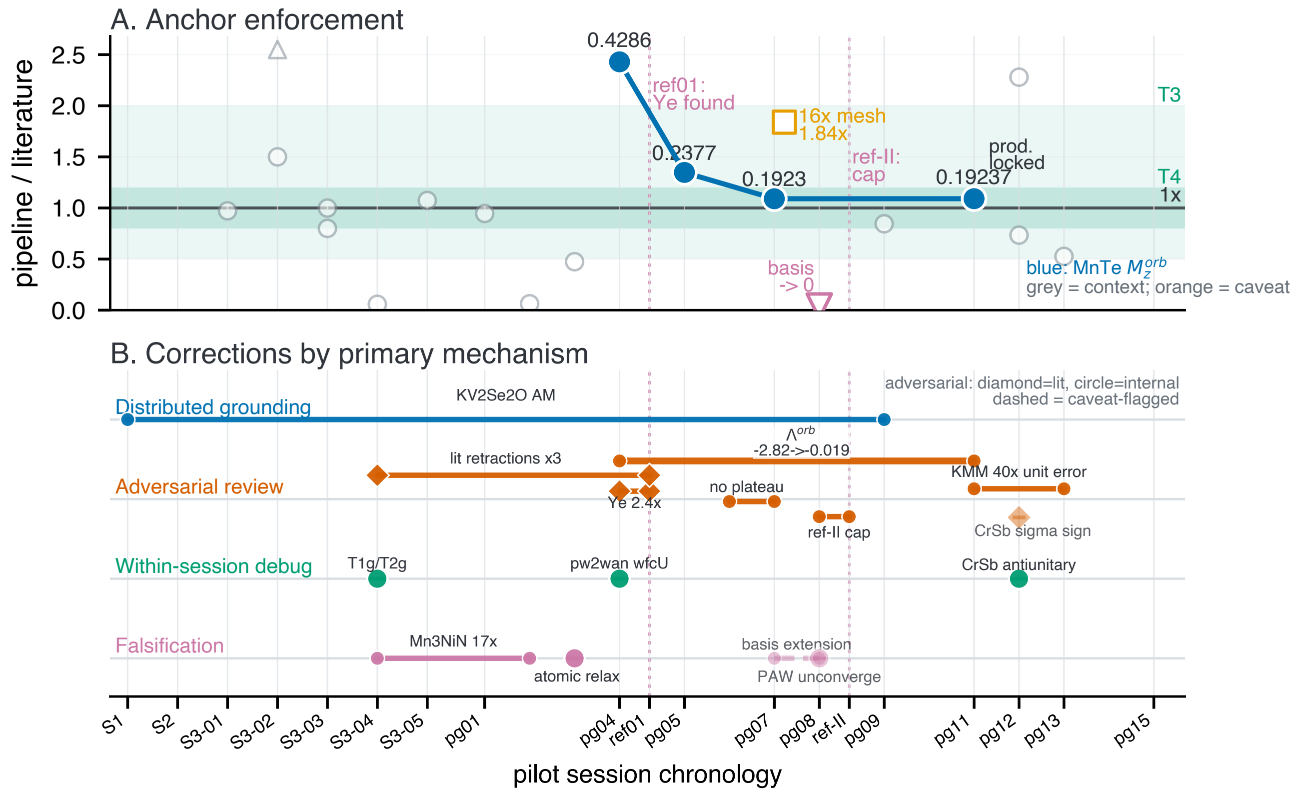


Figure 3. Pilot anchor enforcement and fault tolerance. **(A)** MnTe orbital-magnetization M_z^{orb} trajectory shown as ratio to Ye et al. (Ye et al., 2026)'s published value (0.176 μ_B/cell). The first MnTe orbital-magnetization gate's HIGH at 2.44× is retracted to LOW by the first adversarial review session, whose sub-agent gap-hunt surfaced the published anchor; subsequent gates close the gap step-by-step at the canonical mesh, but the agent itself flagged via NOT PASS that apparent agreement at the canonical mesh was unverified at denser k-mesh. **(B)** 13 catch episodes during pilot and production by primary mechanism: distributed grounding (4), adversarial review (5), within-session debug (2), falsification (2; dashed = caveat-flagged).

architecture transitioned to production with the calibrated recipe *plus* documented systematic uncertainty.

(4) Trace-substrate verification. Computational physics provides a useful substrate for this methodology because every parameter and decision is preserved on disk. A post-completion convergence-ladder test by the human researcher confirmed the divergence direction the agent's NOT PASS had identified. **The architecture's adversarial review was correct to flag the convergence as unestablished; the manuscript's reduced-scope reporting reflected this honestly — a pipeline identifying its own convergence limits while producing an anchored attempt is what grounded autonomous research at first iteration looks like.** Fig. 3B documents 13 catch episodes by primary mechanism across the canonical run.

Reproduction-failure intervention pattern. A subset of pilot reproductions required human intervention. The pattern is knowledge curation: when the agent fails to converge a reproduction within the bounded session budget, the human researcher debugs offline and adds the resulting general principle to the curated knowledge base. Three substantive additions during the canonical pilot were a *pipeline sanity check* rule (verify basic physical quantities at each calculation step), a *Wannier validation* protocol mandating fatband diagnostic before basis design and `dis_froz_max` plateau scan before any Berry-observable production, and a *Hubbard U non-transferability* clause requiring re-derivation across DFT codes or projector conventions. These are transferable operational principles applicable across computational-physics studies, not study-specific scientific direction.

The need for intervention reflects a property of the literature itself: published computational-physics papers vary widely in documentation completeness, from *trivially-replicable* to *frontier-difficult*; Huang (2026b) finds the same spectrum at scale. Reproductions on the difficult end are where human intervention currently remains necessary, with pre-screening and progressive curation as the mitigation. Intervention enters as offline knowledge-base edits, not in-context messages; the no-pilot ablation (§4), inheriting canonical-pilot knowledge, ran without intervention struggles.

3.3. Pre-production, production, writing, and outcome

The pre-production session consolidates the pilot recipe (§2); production-init executes it; three production-continuation sessions run as adversarial review in fresh context. The second adversarial production-continuation session refuted the initial production plan's strain-mode specification, illustrating fresh-context isolation as a robust catch mechanism: the pre-production plan had specified a pure ε_{xx} scan for CsV_2Te_2O, but the continuation re-derived the magnetic point group's symmetry constraints, found that the canonical Bell–Venderbos coupling channel is the B_{1g} shear ($\varepsilon_{xx} - \varepsilon_{yy}$), verified via WebFetch (Bell and Venderbos, 2026), and switched the protocol before the production endpoints computed. Writing proceeded as three sets of fresh-context drafting sessions (main text with sub-agent reference fetching, supplementary, integration) followed by a final adversarial polish session.

Outcome. The pipeline produced a submission-grade companion manuscript (§H) in ∼6 days of wall time, with three substantive findings: *(i)* the symmetry-permitted anomalous Hall response in MnTe at the deep-valence-band anchor with plateau amplitude $\alpha = 397$ S/cm, the four non-zero Néel-vector angles fitting $\alpha \sin(3\varphi)$ within a max residual of 5.8% (envelope band ±15%, consistent with the g-wave $m'm'm$ signature); *(ii)* orbital piezomagnetic response in MnTe at the locked recipe with $\Lambda^{\mathrm{orb}}_{zxx} = -1.96$ μ_B/cell per unit strain and Lukashev asymmetry (Lukashev et al., 2008) 2.77; and *(iii)* cross-symmetry-class — d-wave Lieb CsV_2Te_2O under B_{1g} shear gives $\Lambda_{\mathrm{topo}} = -24.9$ μ_B/cell per unit shear in the clean-linear regime (Mazin, 2023). Systematic uncertainties characterized in §3.4 (and Appendix B).

3.4. Limitations and mitigation directions

The canonical run produced three classes of remaining physics-calibration caveats — a Mn-3d/4s disentanglement-window dependence, k-mesh convergence in g-wave hexagonal Wannier integration, and a 3-point strain extraction — characterized in Appendix B along with prompt-engineerable mitigations identified for the next pipeline iteration. Each caveat is anchored: identified by the pipeline itself rather than discovered post-hoc, and disclosed in the manuscript as the architecture's NOT PASS verdict propagating to the published artifact rather than being argued away.

4. What grounded really means: pilot as the grand anchor mechanism

The pipeline accesses literature 2,162 times (§2), but access is not enforced confrontation. Real researchers calibrating methodology do not merely consult published values — they perform numerical comparison and revise methodology when mismatches emerge. LLM agents pre-trained on internet text default to citing rules and asserting plausibility, not to performing numerical confrontation against literal published values; the pipeline architecture must structurally enforce confrontation at calibration checkpoints. We characterize this empirically via two paired failure modes: a pre-architecture baseline that lacked the architecture entirely and accepted an un-calibratable topic, and a controlled ablation in which curated knowledge was inherited but the pilot reproduction phase was skipped. Together they identify the pilot stage as the grand anchor mechanism.

Topic-selection grounding — the pre-architecture baseline. The pre-architecture baseline run shared the same breadth corpus but lacked a curated knowledge base, a diversity ledger, and the program-selection gate. Three independent depth sessions converged on the same direction — first-principles phonon thermal Hall in altermagnetic MnTe and CrSb, anchored on a recent experimental observation (Wan et al., 2026) — because breadth-report convergence pulled them there and no diversity-forcing or software-feasibility gate pushed back. The cited theoretical foundations (Bendin et al., 2025; Hoyer et al., 2024; Bustamante-Lopez et al., 2026; Park et al., 2024; Weißenhofer et al., 2024) had been validated only on toy-model parents differing in crystal symmetry and dimensionality from the 3D NiAs hexagonal g-wave altermagnet system. Production wrote ∼1,200 lines of custom simulation code, validated only on the toy-model parents. Two post-hoc adversarial reviewers refuted the production headline; the run terminated without a manuscript. The outputs were un-anchorable by construction. The canonical pipeline's program-selection gate closes this loophole via a software-feasibility clause requiring each stage to be executable by a named tool. In the canonical run, the analogous phonon-thermal-Hall direction was generated as a depth candidate and rejected on this gate. This is *topic-selection grounding*: anchor structure begins at topic acceptance, not at execution.

Execution grounding — the no-pilot ablation. The no-pilot ablation ran the same direction as the canonical pipeline, with the *full curated knowledge and house rules inherited from the canonical post-pilot state*: the curated knowledge base entries (including the `dis_froz_max` selection guide, plateau-scan-mandatory protocol, basis-insufficiency diagnostics, and symmetry-forbidden-component diagnostics) and the `PILOT_HOUSE_RULES.md` (including the Wannier-validation Steps A–F and Checks 1–5 that were expanded during the canonical pilot). The ablation's pre-production literature review independently surfaced Ye 2026 and recorded the reference value ("$M_{\mathrm{orb}} = 0.176$ μ_B/cell along z-

axis”) (Ye et al., 2026) verbatim in its pre-production literature review. Pilot reproduction — the ablated component — was skipped, and the run proceeded directly to production.

Despite the published anchor in its own files, the production agent picked a recipe inside the documented collapse tail of the canonical disentanglement-window scan, giving $M_{\mathrm{orb,z}} = 0.066$ μ_B/cell (2.7× below the reference). Across all four reflect-continuation cycles, *the comparison “0.066 vs. 0.176” was never written*. The reference was invoked only via the qualitative ratio $M_{\mathrm{orb}}/M_{\mathrm{spin}} \approx 88$, operationally meaningless because $M_{\mathrm{spin,z}}$ is symmetry-forbidden by $m'm'm$ and identically zero.

The defining episode is the dfroz scan run only in the fourth continuation: $M_{\mathrm{orb,z}}$ collapses from 0.140 at dfroz $= 12.0$ to 0.005 at dfroz $= 13.0$ ($28\times$ across $1\,\mathrm{eV}$) and σ_{xy} flips sign. The agent’s verbatim characterization — “Plateau is BROAD: 219% σ_{xy} spread, 191% $M_{\mathrm{orb,z}}$ spread” — is mathematically correct under (max$-$min)/mean but obscures a $28\times$ change with sign flip that is not a “broad plateau”. The agent then asserted (without testing) that the basis-induced systematic cancels in symmetric finite differences, kept its headlines unrevised, and proceeded to synthesis. The pattern is “rules cited, not enforced”: with curated rules and Ye’s reference value both in its working files, the agent argued individual divergences acceptable case-by-case without performing the literal comparison that would have forced revision. Knowing the published value is not the same as being structurally required to confront against it. This is the gap pilot reproduction’s enforced anchor mechanism closes.

Pilot as architectural automatic anchor. Both grounding modes — topic-selection grounding via program selection, execution grounding via pilot reproduction — are structural mechanisms that operationalize “literature grounding throughout”. Without them, autonomous pipelines either accept un-anchorable topics or proceed within accumulated rules without enforced numerical confrontation. Crucially, the no-pilot ablation *did* include adversarial production-continuation review across four fresh-context cycles, yet none of these cycles surfaced the literal numerical comparison; this isolates pilot reproduction’s structural enforcement — not adversarial review alone — as the operative mechanism. The pipeline does not guarantee correctness; it provides anchor structure. Anchored failures (the canonical pipeline’s documented systematic uncertainty) become tractable next-iteration directions; ungrounded failures (the ablation’s unrevised manuscript) become concealed mystery.

5. Discussion

The pipeline mirrors how real research operates: literature throughout, fresh-context adversarial review, knowledge curation across sessions — behaviors unscaffolded LLM agents skip by default. The architecture should transfer to any domain whose calibration depends on external published anchors rather than intrinsic execution; computational physics adds trace reproducibility on top, making post-hoc verification of the architecture’s own caution always possible. Two principles emerge. First, *grounded anchor enforcement at calibration checkpoints*: program-selection rejects un-anchorable topics, pilot reproduction forces numerical comparison between agent claims and published anchors; the pre-architecture baseline and no-pilot ablation characterize what fails without each. Second, *fault-tolerant session-breaking adversarial review* (Fig. 3B): reflect/continue/polish sessions find rather than confirm. The architecture’s value is in what it knows it does not know — NOT PASS verdicts, documented uncertainty, anchored failures pointing to forward directions. *Can we believe the results* reduces to two operational checks: are agent claims confronted with published anchors, and do session-breaking mechanisms catch what single sessions miss? When yes, remaining uncertainty becomes identifiable forward direction; when no, concealed mystery.

References

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, et al. Galactica: A large language model for science. *arXiv preprint*, 2022. arXiv:2211.09085.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint*, 2021. arXiv:2107.03374.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 2024. arXiv:2304.05376.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint*, 2024. arXiv:2408.06292.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint*, 2025. arXiv:2504.08066.

Samuel Schmidgall et al. Agent laboratory: Using LLM agents as research assistants. *arXiv preprint*, 2025. arXiv:2501.04227.

Tingjia Miao, Jiawen Dai, Jingkun Liu, Jinxin Tan, et al. PhysMaster: Building an autonomous AI physicist for theoretical and computational physics research. *arXiv preprint*, 2025. arXiv:2512.19799.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. *arXiv preprint*, 2024. arXiv:2409.04109.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, et al. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. *Proceedings of ICML*, 2024. arXiv:2402.01817.

Joeran Beel et al. Evaluating sakana's AI scientist for autonomous research: hidden costs, hidden risks. *arXiv preprint*, 2024. arXiv:2410.01243.

Haonan Huang. From experiments to expertise: Scientific knowledge consolidation for AI-driven computational research. *arXiv preprint*, 2026a. arXiv:2603.13191.

Haonan Huang. Towards grounded autonomous research: an end-to-end LLM mini research loop on published computational physics. *arXiv preprint*, 2026b. arXiv:2604.12198.

P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter*, 29:465901, 2017. doi: 10.1088/1361-648X/aa8f79.

Giovanni Pizzi, Valerio Vitale, Ryotaro Arita, Stefan Blügel, Frank Freimuth, Guillaume Géranton, et al. Wannier90 as a community code: new features and applications. *J. Phys.: Condens. Matter*, 32:165902, 2020. doi: 10.1088/1361-648X/ab51ff.

Stepan S. Tsirkin. High performance Wannier interpolation of Berry curvature and related quantities: WannierBerri code. *npj Comput. Mater.*, 7:33, 2021. doi: 10.1038/s41524-021-00498-5.

J.-M. Lihm et al. Magnetic extension of the EPW code: electron-phonon couplings and superconductivity in spin-polarized systems. *arXiv preprint*, 2025. arXiv:2510.11350.

K. Mæland, B. Brekke, and A. Sudbø. Constraints on superconducting pairing in altermagnets. *Phys. Rev. B*, 111:014516, 2025. doi: 10.1103/PhysRevB.111.014516. arXiv:2408.03999.

Pavel Lukashev, Renat F. Sabirianov, and Kirill Belashchenko. Theory of the piezomagnetic effect in Mn-based antiperovskites. *Phys. Rev. B*, 78:184414, 2008. doi: 10.1103/PhysRevB.78.184414.

Beryl Bell and Jörn W. F. Venderbos. Orbital piezomagnetic polarizability of pure insulating altermagnets in two dimensions. *arXiv preprint*, 2026. arXiv:2602.10076.

Chao Chen Ye, Karma Tenzin, Jagoda Sławińska, and Carmine Autieri. Dominant orbital magnetization in the prototypical altermagnet MnTe. *Phys. Rev. B*, 113: 014413, 2026. doi: 10.1103/PhysRevB.113.014413. arXiv:2505.08675.

Y. Liu et al. Ab initio framework of electron-phonon coupling for chiral phonons with giant phonon magnetic moments in magnetic materials. *Phys. Rev. Lett.*, 135: 256701, 2025. arXiv:2503.10160.

Z. Huang, A. Kusuno, M. Hashimoto, D. M. Juraschek, H. Kusunose, and T. Satoh. Quantifying chirality of phonons. *arXiv preprint*, 2026. arXiv:2604.10231.

Y. Zhu, J. Li, Z. Chen, S. Yu, and Q. Liu. Spin-orbit-coupling-resilient nonlinear hall response in noncoplanar antiferromagnets. *Nat. Commun.*, 16:4882, 2025. arXiv:2406.03738.

D. Ceresoli et al. QE-CONVERSE: a non-perturbative orbital-magnetization module for Quantum ESPRESSO. *arXiv preprint*, 2025. arXiv:2503.04664.

Libor Šmejkal, Jairo Sinova, and Tomáš Jungwirth. Beyond conventional ferromagnetism and antiferromagnetism: A phase with nonrelativistic spin and crystal rotation symmetry. *Phys. Rev. X*, 12:031042, 2022. doi: 10.1103/PhysRevX.12.031042.

S. Bhowal and N. A. Spaldin. Magnetoelectric classification of skyrmions and altermagnets. *Phys. Rev. X*, 14:011019, 2024. doi: 10.1103/PhysRevX.14.011019.

S. Hayami et al. Multipole framework for altermagnetic order parameters. *arXiv preprint*, 2025. arXiv:2512.17587.

I. I. Mazin. Altermagnetism in MnTe: origin, predicted manifestations, and routes to detwinning. *Phys. Rev. B*, 107: L100418, 2023. doi: 10.1103/PhysRevB.107.L100418. arXiv:2301.08573.

Y. Wan et al. Anomalous phonon thermal Hall effect in altermagnetic MnTe. *arXiv preprint*, 2026. arXiv:2604.03183.

S. Bendin, A. Mook, I. Mertig, and B. Neumann. Magnon-phonon hybridization in 2D square-lattice altermagnets. *arXiv preprint*, 2025. arXiv:2511.08357.

M. Hoyer, R. Jaeschke-Ubiergo, J. Ahn, L. Šmejkal, and A. Mook. Magnon Hall effect in altermagnets: minimal model and material predictions. *arXiv preprint*, 2024. arXiv:2405.05090.

B. Bustamante-Lopez, B. Brehm, and D. M. Juraschek. Atomistic framework for phonon angular momentum and Hall effects. *arXiv preprint*, 2026. arXiv:2604.01899.

S. Park, N. Nagaosa, and J. Oh. Phonon thermal Hall effect in α-$RuCl_3$ from first principles. *arXiv preprint*, 2024. arXiv:2407.00660.

M. Weißenhofer et al. Magnon-phonon thermal Hall effect in bcc iron from first principles. *arXiv preprint*, 2024. arXiv:2411.03879.

M. G. Lopez, David Vanderbilt, T. Thonhauser, and Ivo Souza. Wannier-based calculation of the orbital magnetization in crystals. *Phys. Rev. B*, 85:014435, 2012. doi: 10.1103/PhysRevB.85.014435.

M. Khodas, Sai Mu, I. I. Mazin, and K. D. Belashchenko. Tuning of altermagnetism by strain. *Phys. Rev. B*, 113: 104422, 2026. doi: 10.1103/PhysRevB.113.104422. arXiv:2506.06257.

Nicola Marzari, Arash A. Mostofi, Jonathan R. Yates, Ivo Souza, and David Vanderbilt. Maximally localized Wannier functions: Theory and applications. *Rev. Mod. Phys.*, 84:1419–1475, 2012. doi: 10.1103/RevModPhys.84.1419.

Appendix

A. Multi-anchor convergence detail

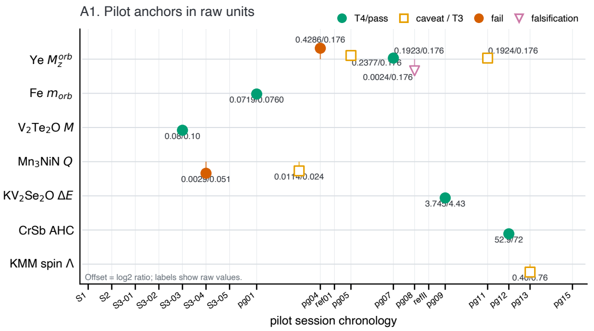


Figure 4. Pilot-stage anchor comparisons in raw units. Each row tracks one published reference (Ye orbital magnetization (Ye et al., 2026), Lopez Fe orbital magnetization (Lopez et al., 2012), V_2Te_2O magnetization, Mn_3NiN multipole, KV_2Se_2O level splitting, CrSb anomalous Hall, Khodas–Mu–Mazin spin piezomagnetic coefficient (Khodas et al., 2026)); points are pipeline-vs-literature ratios on a $\log_2$ axis at the session in which the comparison was performed. Markers: solid green (within ±10%, PASS), open square (within order-of-magnitude, caveat-flagged), solid orange (fail), pink triangle (falsification). The Ye row encapsulates the orbital-magnetization trajectory of §3.2.

The figure complements Fig. 3A by showing the full multi-anchor structure across all calibration targets. Each anchor entered the pipeline as a published reference value; the agent's pipeline-output value at each session is divided by the anchor and plotted on the log-ratio axis. PASS sits inside a ±10% band; caveat-flagged points are within order-of-magnitude consistency but do not pass the strict numerical match. The visual pattern is the empirical content of "grounded throughout": every calibration checkpoint is anchored to a specific published value, and every disagreement enters the lessons that downstream production and writing inherit.

B. Detailed limitations and mitigation directions

The canonical pipeline produced a publication-grade artifact with three classes of remaining caveats, each pointing to a community-relevant question or prompt-engineerable mitigation.

Small disentanglement-window plateau. The MnTe orbital magnetization shows a ∼0.1 eV plateau in the disentanglement upper bound, narrower than the $\mathcal{O}(1\,\text{eV})$ insensitivity typical for Wannier frozen windows (Marzari et al., 2012; Pizzi et al., 2020). A post-pipeline *Step C* fatband identifies the mechanism as a steeply-ramping Mn-4s contribution outside the window. Disentanglement-window selection for orbital observables in this class warrants plateau-scan verification.

k-mesh sensitivity. A 70% shift between 12^3 and 16^3 NSCF mesh densities at fixed window, above the sensitivity expected from Fe-class benchmarks (Lopez et al., 2012). The agent's NOT PASS verdict was correct (§3.2); a post-pipeline convergence-ladder test confirmed the direction. Per-k decomposition localizes the residual at BZ corners A, H (band near-degeneracies). Altermagnetic Wannier orbital magnetization appears to require denser convergence than benchmarks suggest.

3-point strain protocol. The orbital piezomagnetic coefficient was extracted from three matched-gauge symmetric-FD points; Lukashev asymmetry (Lukashev et al., 2008) 2.77 from three points exactly fits a 3-parameter parabola without independent verification of functional form. Prompt-engineerable: the reflect prompt can mandate "if Lukashev > 2, add additional strain points".

A *Step C* fatband diagnostic, though mandatory in the house rules, was not run in either canonical pilot or ablation — to be enforced via reflect prompt in the next iteration.

C. Pipeline scaffolding — prompts and house rules

Each session in the pipeline is driven by a structured prompt template that encodes the agent's workflow: inputs to read, numbered task steps, outputs to produce, and iteration logic. Prompts are versioned during development as failure modes become understood (e.g., the pilot-gate prompt has nine versions across the canonical pilot's evolution; we include the latest). We provide the five most architecturally consequential prompts verbatim below, and a structural summary of the remainder; the full prompt set will be available in the supplementary repository.

B.1 Breadth prompt

Source: `breadth_prompt_v2.md`*; used in 3 canonical sessions.*

```
# Task: Research Landscape Mapping

## The mission

You are working in a large project that aims to produce a first
      -principles computational physics paper suitable for peer
       review at a respectable journal. The standard is novelty
      , rigor, and thoroughness throughout. The final paper
      must make a clear and substantial scientific contribution
      -- not incremental, not a routine calculation, not a
      minor variation on existing work.

## Your task

This is an **extensive reading task**. You read the recent
      arXiv literature in condensed-matter and adjacent physics
       and produce a map of the research landscape: what people
       are working on, where the activity is, what directions
```

```
look scientifically promising. A later agent will use your map, together with its own deep reading, to conceive specific research directions.

You do not propose research ideas. Your output is a map. Ideation is the next stage's job.

## Reading material

**Primary source -- local arXiv corpus (read-only):** the most recent 6 months of papers from six arXiv categories: cond-mat.mtrl-sci, cond-mat.mes-hall, cond-mat.supr-con, cond-mat.str-el, physics.chem-ph, physics.comp-ph. ~11k papers total.

- `<PROJECT>/corpus/title_list.md` -- every paper in the corpus, one line per paper: `[arxiv_id] (category, date) Title -- journal_ref?`. **Cat this entire file into context at session start; do not skim or sample.** Journal references are present for ~10% of papers (only those whose authors manually added a publication record on arXiv); absence does not mean unpublished.
- `corpus/corpus.db` -- SQLite. Query for abstracts and full metadata. Columns include `arxiv_id`, `title`, `abstract`, `primary_category`, `all_categories`, `submission_date`, `doi`, `journal_ref`. Use it freely.
- `corpus/markdown/<arxiv-id>.md` -- full paper text in markdown, converted from PDF. Read selectively for papers worth understanding deeply.
- `corpus/pdfs/<arxiv-id>.pdf` -- original PDF. Read directly if a markdown conversion looks garbled or you need to inspect figures the markdown cannot represent.
- `corpus/latex/<arxiv-id>.tar.gz` -- original LaTeX source for ~78% of papers. If markdown is garbled or you need clean equations/figure captions, extract on demand into your own workspace: `tar -xzf corpus/latex/<id>.tar.gz -C <your_workspace>/latex_unpacked/<id>/`. Do not extract into the corpus directory.

**Secondary source -- full internet access (on demand):** for anything beyond the corpus -- older literature, specific paper lookup, broader context. Before reaching for the generic web search tool, read `<PROJECT>/knowledge/literature_access.md`. It documents the structured APIs available (arXiv all-time, Semantic Scholar, OpenAlex, Crossref, Elsevier TDM with key) with code examples. Structured APIs are higher precision than web search; use them first.

**Filesystem discipline:** the corpus directory and knowledge directory are **read-only** for you. Write only under your own working directory. Any extracted LaTeX, downloaded PDFs, or intermediate files go in subdirectories of your workspace.

## What to look for

The corpus is diverse. Good landscape entries include:

- **Dominant themes**: directions where many papers cluster.
- **Emerging bursts**: recent activity that has not yet consolidated.
- **Underrepresented but substantive directions**: scattered high-quality work that has not converged into a visible theme. Low competition.
- **Gaps**: experimental observations without clean theoretical understanding, theoretical predictions awaiting experimental test, contradictory findings, methodology shifts mid-corpus.
- **Cross-pollinations**: ideas that appear across subfield boundaries.

A first-principles computational study could make a contribution in many of these. The map should surface areas where such a contribution is plausible -- not restricted to papers that already used computation. Experimental papers, reviews, and theory papers are all fair game.

## Deliverable

A single file `breadth_report.md` in your working directory. Suggested structure:

1. **Themes** -- dominant directions, with representative arXiv IDs and a sense of activity level.
2. **Emerging patterns** -- new bursts, methodology shifts, anomalies.
3. **Underrepresented directions** -- scattered but substantive, low competition.
4. **Gaps** -- where a first-principles study could contribute.
5. **Candidate directions for deep-dive** -- a ranked shortlist (5--10) of directions worth dedicated investigation. Rank by scientific potential, not by heat. Heat and novelty both matter; explicitly note which drives each ranking.

Use arXiv IDs inline as evidence. Do not fabricate citations. For each claim, distinguish between "I read N papers in this area" and "I inferred this from abstracts alone" -- mark your epistemic state per theme.

If you find yourself writing "someone should study X" or "we propose Y" -- stop and rewrite. Say "there is a gap around X" or "direction Y appears unexplored." You are mapping, not ideating.

## Begin

Read the resources, develop your own reading strategy, and produce `breadth_report.md`. **No time limit, no token limit -- do not save tokens, this stage matters.** Your effort here feeds directly into every downstream stage; the ideation agents can only work with what you surface. Take the time this task deserves.
```

B.2 Depth prompt

Source: `depth_prompt_v3.md`; used in 5 canonical sessions.

```
# Task: Research Program Formulation

## The mission

You are working in a large project that aims to produce a first-principles computational physics paper suitable for peer review at a respectable journal. The final paper must make a clear and substantial scientific contribution.

## Your task

Read landscape maps from prior agents, dive deep into a promising direction, read substantial full-text literature, and formulate a research program -- a main question with sub-questions and a fallback -- that could become a publishable paper. A pilot agent will later audit feasibility; your job is to deliver a program strong enough to survive that audit.

## Input

- `breadth_reports/` -- landscape maps from prior agents. Read all in full.
- `chosen_topics.md` -- directions already chosen by prior depth agents. **Read before committing.** Your direction must differ substantially from any listed. After committing, append your direction (one line) to this file.

- **Corpus** (read-only, ~11k papers, 6 months): `corpus/title_list.md` (cat in full), `corpus/corpus.db` (SQLite), `corpus/markdown/<id>.md`, `corpus/pdfs/<id>.pdf`. Path: `<PROJECT>/corpus/`.
- **Internet access** for novelty audit. Read `<PROJECT>/knowledge/literature_access.md` first for structured APIs (arXiv all-time, Semantic Scholar, OpenAlex, Crossref, Elsevier TDM).
- Write only in your own workspace.

## Constraints

**Hardware**: 12 cores, 48 GB RAM, ~24h production compute. **Toolchain**: DFT, DFPT, TDDFT, GW, Wannier interpolation, ballistic transport -- standard first-principles methods assumed available. Do not exceed this ecosystem.

**Software feasibility gate.** For each simulation stage in the proposed pipeline (not analysis or plotting), you must be able to name a specific existing open-source tool or package that performs it. If a stage cannot be done by a named tool and would require writing custom simulation code -- even if the underlying formulas are well-documented -- that stage is methodology development, not pipeline execution. The scientific contribution should be in the physical insight, not the computational method.

## Workflow

This is an **iterative** process, not a linear one. You explore multiple candidates and screen them before committing
```

```
deeply to one.

**Stage 1 -- Candidate generation.** Read breadth reports and `chosen_topics.md`. Identify 2--3 candidate research directions (different from any in `chosen_topics.md`). For each candidate, write a brief sketch: one-paragraph description + the computational pipeline as a list of stages. Do not deep-read literature yet.

**Stage 2 -- Feasibility screen.** For each candidate’s pipeline, apply two gates:

- **Hardware gate:** Can each stage plausibly complete within 24h on 12 cores / 48 GB? Are the system sizes reasonable?
- **Software gate:** For each simulation stage, name the specific open-source tool that performs it. If any stage has no named tool and would require custom simulation code (constructing Hamiltonians, implementing solvers, deriving formulas from papers), the candidate fails.

If at least one candidate passes both gates -> proceed to Stage 3 with the best one.

If all candidates fail -> **return to Stage 1** with fresh candidates. You may return up to twice (3 rounds total, screening 6--9 candidates). After 3 rounds with no viable candidate, simplify the most promising candidate from all rounds until it passes both gates.

**Stage 3 -- Commit + deep reading.** Select the best candidate that passed both gates. Append your direction to `chosen_topics.md`. Now dive deep: read full-text papers, develop main question + sub-questions + fallback.

**Stage 4 -- Novelty audit.** The corpus covers only 6 months. Audit every question against the broader literature using external APIs. For each question: has it been done? If partially, what is your contribution beyond it? If the direction is completely done, try the next passing candidate from Stage 2 -- or return to Stage 1 if none remain.

**Stage 5 -- Finalize.** Only a program that has passed both feasibility screen and novelty audit is finalized.

## Deliverable

A single `research_program.md`:

1. **Candidates considered** -- the 2--3 directions you screened, with pipeline sketch and feasibility outcome for each. Include dropped candidates and why they were dropped.
2. **Direction and rationale** -- which candidate you committed to, why, how it differs from `chosen_topics.md` entries.
3. **Literature foundation** -- full-text papers read, 1--2 lines each on what you took from them.
4. **Main research question** -- specific, first-principles-addressable.
5. **Sub-questions** (2--4) -- concrete enough for a pilot agent to estimate cost.
6. **Fallback path** -- if the main question fails, what salvageable story remains?
7. **Prior work + novelty** -- per question: relevant prior work (URL, title, year, 1-line takeaway) + differentiation.
8. **Computational feasibility** -- per pipeline stage: named tool, has it been demonstrated on a similar system, hardware estimate.
9. **Expected deliverable** -- what the paper would argue, one paragraph.
10. **Computational footprint** -- classes of calculation, system sizes, 24h/48GB plausibility.

## Begin

Read `chosen_topics.md`, then breadth reports. Start with candidate generation and feasibility screening -- do NOT deep-read literature until a candidate has passed both gates. **No time limit, no token limit.** The program you conceive here determines the ceiling of impact for the entire project.
```

B.3 Pilot gate prompt (latest of 9 versions)

Source: `pilot/pilot_gate_prompt` *latest version; used in 15 canonical pilot-gate sessions.*

```
# Task: Pilot Gate -- Iterative Pipeline Confidence Assessment

## The mission

You are working in a large project that aims to produce a first-principles computational physics paper suitable for peer review at a respectable journal. Prior agents have reproduced 5 published papers to validate the computational pipeline. Your job is to assess whether the pipeline is ready for production -- and if not, to close the most critical gap.

## Your task

**Depth over breadth.** One thing fully investigated to quantitative agreement with literature is worth infinitely more than five things explored qualitatively. Do not speculate about what "might work." Pick one gap, close it, and prove it with numbers.

The production pipeline can only be trusted on a new material if every computational step has been demonstrated to give quantitative agreement (?20%) with published literature on a physically similar system. A step that was only smoke-tested on a toy system, or that showed >2x disagreement with unresolved diagnosis, is not validated -- it is a guess.

## Input

**Prior work (mandatory reads -- read ALL before writing your verdict):**

- `session_3_reproduce_*/` -- all 5 reproduction sessions. Read each `reproduction_report.md`, `paper_verdict.json`, and skim each `WORKLOG.md` for methodology lessons and unresolved issues.
- `pilot_gate_*/` -- any previous pilot gate sessions (if they exist). Read each `pilot_gate_report.md` and `pilot_gate_verdict.md`. Pay attention to what was attempted, what worked, and what didn’t.
- `pilot_reflect_*/` -- any previous pilot reflect sessions (if they exist). These contain literature searches, published reference values, and recommendations from a fresh-perspective review. **Pay close attention -- these identify specific papers and numbers you should reproduce or compare against.**
- `pilot_reflect_II_*/` -- any previous pilot reflect II sessions (if they exist). These contain paper figure analyses, gap analyses, and **planned pilot_gate tasks with specific deliverables.** **Pilot reflect II sessions supercede everything else regarding tasks. If a reflect II report assigns you a specific task, execute it instead of picking your own gap in Phase 2.**
- `session_1_prior_work/program_selection.md` -- the production program. What pipeline stages does production require? What are the headline claims?

**Local environment:**

- `runs/run_002/knowledge/TOOL_REPORT.md` -- central tool registry. **Read first.**
- `runs/run_002/knowledge/` -- per-tool docs, custom-script docs.
- `runs/run_002/tool_examples/` -- verified examples per tool.
- `runs/run_002/.venv/` -- Python venv: `source <PROJECT>/runs/run_002/.venv/bin/activate`

**Global knowledge (mandatory reads):**

- `<PROJECT>/knowledge/PSEUDOPOTENTIALS.md`
- `<PROJECT>/knowledge/INDEX.md` -- contains critical computational gotchas and lessons learned. **Read the sections relevant to your pipeline tools and calculation types carefully before computing anything. Failure in doing so will result in time wasted and garbage produced.**
- `<PROJECT>/knowledge/literature_access.md`
- `PILOT_HOUSE_RULES.md` in your workspace -- **read before any planning and computation. Obey EVERYTHING inside. Even if any required procedure increases work significantly, DO IT (like U scan and Wannier band verification). Failure of doing so is regarded complete failure of the whole task since skipping required works and gates will certainly result in garbage data. Especially: 1. after starting a calculation, DO ABSOLUTELY NOTHING (except for checking RAM/core usage) UNTIL IT IS DONE. You should have only ONE monitor to see when the job stops. Do not spawn multiple monitors only to confuse yourself. One job at a time, so no more than one monitor. 2. DO NOT KILL AND RESTART AN SCF RUN UNLESS YOU ARE SURE THE NUMBER STAYS EXACTLY THE SAME FOR >10 ITERATIONS OR HAS
```

```
      OSCILLATED MULTIPLE TIMES. Evidence has shown that you
      like to grow impatient and kill scf multiple times,
      resulting in no result. 3. For Wannier-based calculations
      : convergence warnings are not cosmetic. If
      disentanglement reports "criteria not satisfied",
      downstream results are unreliable -- do not proceed. Read
       corresponding PILOT_HOUSE_RULES parts and INDEX lessons
      on frozen window before setting up any Wannier chain.**

**Prior reproduction artifacts:** `session_3_reproduce_*`, `
      pilot_gate_*`, and `pilot_reflect_*` workspaces are read-
      only. **Copy any useful intermediate results (converged
      SCF, Wannier checkpoints, U scan data) into your own
      workspace.** But only inherit critically -- if a prior
      result didn't match literature quantitatively, don't
      blindly reuse it. Improve on it, or simply start from
      scratch. For QE, you can copy `outdir/` and use `
      restart_mode='restart'` to save SCF time when the cell
      and magnetic configuration are compatible.

**Compute:** 12 cores, 48 GB RAM. See `TOOL_REPORT.md` for
      invocation commands.

**Filesystem:** Everything outside your workspace is read-only.
       Your workspace as `pilot_gate_NN/` (ls the pilot
      directory to determine the next number). All output goes
      in your workspace only. You may install additional Python
       packages if necessary -- document in worklog.

**Pseudopotentials:** Virtually all QE-compatible
      pseudopotential libraries are available locally (see `
      PSEUDOPOTENTIALS.md` for paths). You should not need to
      download PPs -- just pick the right ones from what's
      available. If you genuinely need a PP that isn't local,
      you may search online and download it.

## Workflow

### Phase 0: Create your workspace as `pilot_gate_NN/` (ls the
      pilot directory to determine the next number). **All
      output goes in your workspace only**.

### Phase 1: Confidence assessment and verdict (DO THIS FIRST
      -- before any calculation)

Read all prior work. **Think critically and independently, as a
       skeptical referee would. THIS REFLECTION PART IS
      EXTREMELY IMPORTANT BECAUSE IT IS THE FOUNDATION OF THE
      WHOLE TASK: TAKE YOUR TIME, DIG DEEP AND THINK CAREFULLY
      .** Prior agents might neglect important details from raw
       data, their chain-of-thought reasoning may also be wrong
      , and their attributions for why a result deviated may be
       incorrect. Can the results between different runs really
       compared? Did they compare unconverged with converged
      results? Do multiple factors change at the same time,
      contributing to the difference? Dig into raw data. Can
      you find any difference of parameters that the previous
      agent neglected, which could tell a different story?
      Moreover, they lack the global view you now have across
      all 5 reproductions and any prior pilot gate sessions. Do
       not take their diagnoses at face value -- verify them
      against your own physics understanding and against raw
      data.

During Phase 1, you may search online and quickly fetch paper
      abstracts or values to inform your verdict. But this is a
       preliminary scan -- it will miss supplementary details,
      datasets, and methodological nuances. Do not assume you
      have read a paper because you skimmed its arXiv page. The
       full reading with published version, supplementary, and
      dataset is a separate mandatory step in Phase 2.5, and
      cannot be skipped even if you already fetched the arXiv
      version.

Imagine the production paper is published alongside all the
      reproduce and pilot_gate session data as supplementary
      material. A referee asks: "You couldn't reproduce [X]
      quantitatively -- how do I know your production results
      on a new material aren't noise?" That is the standard you
       are defending against.

For each pipeline stage that production requires, assess your
      confidence:

- **HIGH**: this step was tested on a physically similar system
       AND gave quantitative agreement (?20%) with published
      literature. The recipe (parameters, pseudopotentials,
      convergence settings) is documented and transferable.
- **MEDIUM**: this step was tested but only qualitatively (
      correct signs/trends, wrong magnitudes), OR on a system
      with different physics than production targets, OR with
      untested assumptions about why it didn't match
      quantitatively.
- **LOW**: this step was smoke-tested only (simple system, not
      production-relevant physics), OR showed >2x disagreement
      with diagnosis that is speculation rather than verified
      explanation.
- **NONE**: this step was never executed in any session.

**For each non-HIGH stage, do the following analysis:**

1. **Identify the gap precisely.** What quantitative target was
       missed, and by how much?
2. **Review prior diagnoses.** What did the reproduce/
      pilot_gate agents say caused the deviation? Do you agree
      with their reasoning? What alternative explanations exist
      ?
3. **List concrete possibilities** for closing the gap -- with
      your assessment of how likely each is to work and why.
4. **REQUIRED: Search online.** For each gap, search
      documentation, forums, mailing lists, tutorials, examples
      , and GitHub issues for the relevant tools and physics.
      Many problems in computational physics are well-known
      community issues with documented solutions. **Do this
      search before deciding what to work on -- it may
      completely change your assessment of which gap is
      tractable.** Judge the authority of sources yourself (
      official docs > tutorials > forum posts > random blogs).
5. **Assess tractability.** Is there a clearly promising path,
      or has this been extensively attempted across multiple
      sessions without convergence? Do not go down rabbit holes
       -- if prior sessions have made extensive, well-reasoned
      attempts and the gap remains, it may be genuinely hard.
      Prioritize gaps where you see a concrete, evidence-based
      path to T4 over gaps that are high-impact but have no
      clear solution.

Write `pilot_gate_verdict.md` with:
1. A confidence table (one row per production pipeline stage).
2. For each non-HIGH stage: the full analysis above (gap, prior
       diagnoses, your assessment, online research findings,
      possibilities, tractability).
3. Overall verdict: **PASS** (all production-critical stages at
       HIGH) or **NOT PASS** (list gaps ranked by combined
      impact x tractability).

**If PASS, stop here.** Write `pilot_gate_report.md`
      summarizing why and finalize.

**If NOT PASS, proceed to Phase 2.**

### Phase 2: Pick ONE targeted project

From your NOT PASS gaps, pick **the single gap that best
      combines high impact on production confidence with a
      clear, evidence-based path to resolution.** Not two gaps.
       Not three. One.

You have two options:

**Option A -- Reproduce one new paper.** Download and reproduce
       a published paper that directly exercises the gap. Same
      workflow as `session_3_reproduce` sessions (
      reading_summary, WORKLOG, comparison figures, verdict).
      Use literature_access.md for paper retrieval. Pick a
      paper where the answer is known so you can verify
      quantitatively.

**Option B -- Enhance a previous result or run a targeted test
      .** This could include: convergence studies (k-mesh,
      cutoff), pseudopotential comparison, U scan to match a
      known observable, SOC on/off comparison, or any other
      controlled test that produces a quantitative number
      compared to published literature.

### Phase 2.5: MANDATORY -- Read the reference paper before any
       calculation

**This applies to BOTH Option A and Option B.** Any time you
      are comparing your result against a published number, you
       MUST read the source paper thoroughly first. A number
      without its context is meaningless -- you need to
      understand what code, what parameters, what conventions
      produced that number before you can meaningfully compare.

**Step 1: Get the published version.** If the paper has a
      journal-published version, you MUST use it. An arXiv
      preprint is NEVER the published version, even if the
      content appears identical -- arXiv lacks supplementary
      materials, datasets, and final methodological details,
      and using it wastes sessions and computation. Do not
      attempt to bypass paywalls or download published PDFs
      yourself. Stop, create a folder in your workspace (e.g.,
      `reference_papers/`), and ask the user: give them the DOI
```

```
      , tell them you need the main text PDF, supplementary PDF
      , and any linked code or dataset. Wait for "continue."

**Step 2: Read the full paper -- main text AND supplementary
      AND dataset.** Not just the abstract. Not just one number
       from a table. Read the computational details section
      word by word. Read the supplementary material. If a
      dataset or code repository is linked, examine **every
      file** -- input files, scripts, output logs, intermediate
       products. Not just files for our code -- files for ANY
      code the paper used. The paper may use a different DFT
      code, but every parameter in their input files has a
      physical meaning that maps to our pipeline. Extract ALL
      of them.

**Step 3: Replicate the recipe.** The goal is **faithful
      replication of the paper's computational setup**,
      translated to our pipeline. Do not fixate on the
      seemingly biggest discrepancy or contribution and
      selectively pick which parameters to match and which to
      dismiss. Very often, some details you ignore at the
      beginning turn out to be important. So, pay attention to
      everything. Specifically:

- Map every parameter from the paper's input files to our
      pipeline equivalent. If the paper uses a different code,
      read the documentation for BOTH codes to understand what
      each parameter does and find the correct equivalent. Do
      not assume parameter names are transferable across codes
      -- follow the references the paper cites for each
      methodological choice.
- For parameters not explicitly stated in the main text (e.g.,
      number of bands, energy windows, convergence thresholds),
       analyze thoroughly the dataset's input files, output
      logs, and figures. Inspect the paper's figures carefully
      -- band structure plots, convergence plots, and DOS
      figures encode implicit parameter information.
- If a parameter cannot be determined from any source or it is
      genuinely not comparable due to different software, note
      it explicitly and make a physically motivated choice.
      Document your reasoning.

Write `parameter_comparison.md` (MANDATORY deliverable): every
      parameter from the paper, the equivalent in our pipeline,
       and how you will match it. This is not a document for
      deciding what matters -- it is a checklist for confirming
       you have matched everything.

**After writing `parameter_comparison.md`, STOP.** Re-read `
      PILOT_HOUSE_RULES.md` and the relevant sections of `INDEX.
      md`. Check: does your parameter comparison cover
      everything the rules require? Have you dismissed any
      parameter as "can't map" or "probably not important" that
       a rule actually addresses? Prior sessions have written
      parameter comparisons, dismissed the most important
      difference, and wasted entire sessions debugging the
      wrong thing. If a parameter differs and you haven't
      matched it, it is a candidate problem -- not something to
       explain away.

**Step 4: Validate by comparing intermediate products against
      the reference.** If the dataset includes intermediate
      outputs (Wannier spreads, band structures, convergence
      logs), compare yours against theirs at each pipeline
      stage. If intermediates don't match, stop and think
      deeply about physics -- what is physically different? Fix
       the upstream cause before proceeding. A final number
      produced without intermediate validation is not
      trustworthy. If your final result does not match the
      paper, you have two tools: (1) compare intermediates more
       deeply to locate where the pipeline diverges, and (2) go
       back to `parameter_comparison.md` -- every parameter
      difference you listed but did not match is a candidate
      cause. The answer is almost always in the parameters you
      haven't matched yet.

**Step 5: Write `reading_summary.md`.** Summarize the paper's
      physics, your complete parameter mapping, and your
      replication plan. The final verdict must come from
      running the full pipeline end-to-end with our own tools
      -- but intermediate comparisons against the dataset are
      essential for debugging.

**Pseudopotential choice is often the most impactful variable
      .** Before starting any calculation, check what
      pseudopotential files are available locally (in the PP
      directories documented in PSEUDOPOTENTIALS.md). Are there
       alternative choices (different libraries, different core
      -valence partitions, PAW vs NC) that might give better
      results for your specific system? If a prior reproduction
       showed systematic deviation, pseudopotential mismatch is
       a prime suspect.

**The goal is to reach T4 (?20% agreement with literature) on
      this one thing.** Write a concrete plan in WORKLOG before
       starting: what is the central question, what
      quantitative target will you compare against, what
      parameter mapping ensures an apples-to-apples comparison,
       and what does success look like.

### Phase 3: Execute

**Before launching any calculation, STOP and re-read `
      PILOT_HOUSE_RULES.md` in full AND the relevant sections
      of `INDEX.md` for the tools you are about to use.** Do
      this now, not from memory -- hours of reading and
      planning have passed since session start. Every rule and
      check in these documents was written because ignoring it
      has resulted in garbage data and days of wasted
      computation. This is not hypothetical: prior sessions
      have lost 10+ hours to violations of rules that were read
       at session start and forgotten by calculation time. Re-
      read. **OBEY `PILOT_HOUSE_RULES.md` IN FULL, especially
      all those checks and gotchas.** Then compute.

**Validate on a single point before any sweep.** If the paper
      or test sweeps a parameter, run ONE representative point
      first. Verify it is physically sound and within
      reasonable range of the target before investing hours in
      the full sweep.

**Think critically about how many points you need.** The goal
      is to verify quantitative agreement, not to produce
      pretty curves.

**Prioritize the quantitative target.** One key value, properly
       converged and matching literature, is the entire
      deliverable. Everything else is secondary.

For each calculation:
- Process check before launching (house rules ?1), log to
      WORKLOG.
- Launch with proper parallelism. Block fully.
- On success: extract numbers, compare to literature.
- On failure: diagnose. Failures are where pipeline gaps are
      discovered.

### Debugging discipline

When a result deviates, suspect in this order:
1. Your custom code (conventions, signs, units, prefactors).
2. Your input parameters (compare against paper parameter by
      parameter).
3. Tool version or configuration mismatch.
4. The paper itself (last resort -- only after exhausting 1-3).

Use the internet for physics conventions and tool documentation
       (house rules ?10).

### Custom script discipline

Custom scripts require ?3 independent validations on published
      reference values before their outputs enter the verdict.
      Document validations in worklog.

### Phase 4: Updated assessment (APPEND, do not overwrite)

After completing your targeted project, write `
      pilot_gate_report.md`. This is a NEW document -- do NOT
      overwrite or modify `pilot_gate_verdict.md` from Phase 1.
       The Phase 1 verdict is the "before" snapshot; the report
       is the "after." Both must be preserved for provenance.

`pilot_gate_report.md` should contain:
- What gap you targeted and why.
- What you did (briefly).
- Quantitative result vs literature.
- Updated confidence assessment for the stage you worked on --
      did the gap close?
- Remaining gaps (for the next pilot_gate iteration, if needed)
      .

**Re-evaluate honestly.** If your targeted project did NOT
      close the gap (result still >20% off, or revealed a
      deeper problem), say so. Do not inflate confidence
      because you spent time.

## Time discipline

- **Start:** Record a timestamp in WORKLOG.md.
- **At 6 hours:** Assess progress. If the quantitative target
      is not yet matched, reduce scope -- focus exclusively on
      getting one clean number.
- **At 8 hours:** Stop launching new calculations. Write up.
```

```
- **At 10 hours:** Hard stop. Finalize all deliverables with whatever data exists.

## Deliverables

| File | Content |
|------|---------|
| ‘pilot_gate_verdict.md‘ | Confidence table + gap analysis + PASS/NOT PASS (written FIRST, before calculations) |
| ‘WORKLOG.md‘ | Continuous log |
| ‘parameter_comparison.md‘ | Every parameter from reference paper/dataset mapped to our pipeline -- checklist for confirming complete replication, not selective assessment (MANDATORY) |
| ‘reading_summary.md‘ | Paper physics, complete parameter mapping, replication plan (MANDATORY) |
| ‘pilot_gate_report.md‘ | What was done, quantitative results, updated confidence, remaining gaps (written AFTER calculations, preserving Phase 1 verdict separately) |
| Calculation artifacts | In subdirectories within your workspace |
| ‘comparison_*.png‘ | Quantitative comparison figures |

## Begin

First, ‘ls‘ the pilot directory to determine your session number. Create ‘pilot_gate_NN/‘. Read ‘PILOT_HOUSE_RULES.md‘, then read ALL prior ‘session_3_reproduce_*/reproduction_report.md‘, ‘pilot_gate_*/pilot_gate_report.md‘, and ‘pilot_reflect_*/reflect_report.md‘ (if any exist), and
‘pilot_reflect_II_*/reflect_II_report.md‘ (if any exist). Then read ‘program_selection.md‘. Write your confidence verdict BEFORE touching any calculation. The production stage inherits your assessment -- if you say HIGH and it’s not, production builds on a false foundation.
```

B.4 Pilot reflect prompt

Source: `pilot/pilot_reflect_prompt_1.md`*; used in 1 normal pilot-reflect session.*

```
# Task: Pilot Reflect -- Analyze, Search Literature, Find Problems

You do not run any calculations. You do not modify anything outside your own workspace. Create ‘pilot_reflect_NN/‘ (ls pilot/ for the next number). Your only deliverable is ‘reflect_report.md‘.

## Read

Everything in ‘pilot/‘: all ‘reproduction_report.md‘, ‘pilot_gate_report.md‘, ‘pilot_gate_verdict.md‘, prior ‘reflect_report.md‘. Also ‘session_1_prior_work/program_selection.md‘.

## Your primary task

Identify every pipeline step production requires. For each step, identify the key physical quantity. For each quantity, answer:

**What do we get? What does the literature say?**

Search broadly -- arXiv, Google Scholar, journal sites. Find published theoretical values for the same quantity. Prefer same material, then same material family, then same physics on any material. Find experiments too. Read the papers our agents tried to reproduce -- and find papers they didn’t try. Follow-up papers, competing calculations, reviews.

For each quantity where our pipeline disagrees with a paper we tried to reproduce, diagnose:

- **Is it a literature problem?** Does the broader literature (other theory groups, experiments) support our value or the paper’s? If multiple independent sources agree with us, the paper may be an outlier -- find a better reference to validate against.
- **Is it a methodology problem?** Did our agents use shortcuts, skip steps the paper did, use different pseudopotentials without testing alternatives? Check unit conventions, sign conventions, normalization (per atom vs per cell, Ry vs eV, etc.). If the disagreement is a constant factor, think hard about whether it’s a convention mismatch before concluding it’s physics.
- **Is it a pipeline problem?** If results are noisy rather than a consistent offset, that points to convergence or implementation issues, not literature disagreement.
- **Is it a code problem?** Check if any custom code, patches, or agent-derived formulas were used. Verify them independently.

Also scrutinize quantities where agents declared agreement -- is the agreement real, or was it against an internal benchmark rather than an independent published number?

## Your secondary task

Based on your analysis, issue a verdict.

**PASS:** every production-critical pipeline step has ?20% agreement with at least one published theoretical value on a physically similar system. Theory is the primary source for quantitative comparison. Experimental agreement provides qualitative confidence (correct sign, right order of magnitude) but does not substitute for theory-vs-theory quantitative benchmarking.

**NOT PASS (expected outcome):** provide a task list for the next pilot_gate round. For each task: the gap, a specific paper to reproduce or test to run (with arXiv ID and the specific number to match), and a decision tree -- under what conditions can this gap be considered closed? Under what conditions should we stop trying and accept the uncertainty?

**PARTIAL PASS (last resort):** issue only if a specific pipeline step has been extensively attacked (?3 independent theoretical references attempted, convention/unit issues ruled out, pipeline systematics investigated) and still cannot be reproduced. In that case, recommend dropping that specific claim from production scope. This is a scope reduction, not a free pass.

## Write ‘reflect_report.md‘

Per-step analysis with our values, literature values (cited), diagnosis. Task list if NOT PASS. Keep it concrete -- every recommendation should point to a specific number in a specific paper.

If a paper is behind a paywall or download fails, pause and ask the user to download it. Tell them the paper ID, where to put the PDF, and say "continue."
```

B.5 Pilot reflect_II prompt (iteration-cap activation)

Source: `pilot/pilot_reflect_II_prompt.md`*; used at the iteration-cap activation event.*

```
# Task: Pilot Reflect II -- Self-Validation Strategy and Production Plan

You do not run any calculations. You do not modify anything outside your own workspace. Create ‘pilot_reflect_II_NN/‘ (ls pilot/ for the next number).

## Context

Multiple rounds of pilot_gate and pilot_reflect sessions have been run. The goal is no longer to fit a specific number from a specific paper. The goal is to understand the physics behind your numbers, identify what is still missing to support them, and plan the remaining work before production.

## Step 1: Study the published papers -- figures AND physics logic

Find every paper that was reproduced or referenced across all reproduce and pilot_gate sessions. Use the published version with SI if available (check ‘reference_papers/‘ folders in prior pilot_gate sessions first; those have publisher PDFs, SI, and datasets). For each paper:

**Physics logic chain.** How does the paper explain the physical mechanism behind its headline quantity? What produces this quantity at the microscopic level? What orbitals, interactions, or symmetry-breaking mechanisms are responsible? How does the paper connect the mechanism to its computed result? Quote the key sentences verbatim -- the ones where the authors explain WHY a quantity has the value it does, not just WHAT the value is. These
```

```
      logic chains tell you what physical understanding is
      required to interpret and defend a result.

**Figures.** Look at every figure in main text AND SI. For each
       figure:
- What physical quantity is displayed, and what is it plotted
      against (energy, k-point, strain, doping)?
- What does this dependence reveal about the mechanism? (e.g.,
      a k-resolved plot shows WHERE in the Brillouin zone a
      quantity originates; an energy-resolved plot shows WHICH
      bands contribute)
- How does this figure support or diagnose the headline
      quantity? Could it reveal why a calculation succeeded or
      failed?
- Is this figure related to any quantity our pipeline computes?

Write `paper_figure_analysis.md`: paper by paper, figure by
      figure, logic chain by logic chain. This is a deliverable
       -- not a mental exercise.

## Step 2: Review everything in pilot/

Read all `reproduction_report.md`, `pilot_gate_report.md`, `
      pilot_gate_verdict.md`, `pilot_reflect_*/reflect_report.
      md`, and `session_1_prior_work/program_selection.md`.

**Think critically.** Prior agents may lack the global view you
       now have. They may have overlooked details or drawn
      wrong conclusions. Any doubt -- check raw data (
      calculation outputs, logs, numbers) rather than trusting
      a prior agent's verdict or claim.

What numbers do we have? What recipes work? What sensitivities
      have we discovered? What remains unresolved? Write a
      concise summary grounded in raw data.

## Step 3: Gap analysis -- what do we need for a paper?

Imagine writing a paper from our existing pilot data. Compare
      against the papers analyzed in Step 1:

- For each paper's physics logic chain: does our work include
      the supporting quantities and analyses that chain
      requires? If a paper explains a quantity through a
      specific mechanism and supports that explanation with a
      specific type of analysis (e.g., k-resolved decomposition
       of a Brillouin-zone integral, orbital-projected band
      structure, energy-dependent integration), have we done
      the equivalent? If not, flag it.
- What figures and analyses did those papers include that we
      have NOT done? For each: essential (paper rejected
      without it), important (strengthens argument), or nice-to
      -have?
- Are there quantities we computed but never validated with a
      diagnostic that would reveal whether the physics is
      correct?
- Are there parameter sensitivities we discovered but never
      explained -- we know THAT something is sensitive, but not
       WHY? Is there an underlying physical quantity whose
      analysis would reveal the mechanism?

## Step 4: Plan remaining pilot_gate sessions

Write a concrete plan for closing the gaps from Step 3. The
      goal is **understanding**, not expanding scope.
      Prioritize:

1. **Big gaps that block production.** If production requires a
       pipeline step that has never been validated, or a
      quantity that was computed but is physically unsupported,
       this must be addressed. New calculations are justified
      here.
2. **Supporting analyses for existing results.** We have
      numbers from prior sessions -- what analyses and figures
      would make them defensible? The goal is to understand why
       the pipeline gave the results it did, not to add more
      calculations for their own sake. If we discovered a
      sensitivity, compute the diagnostic that explains it.

Each planned session should be:
- A specific computation or analysis with a clear deliverable (
      a figure, a convergence study, a validated recipe)
- Sized to fit within a 6-10h pilot_gate session (one task per
      session)
- Ordered by priority (production blockers first)

For each planned session, specify:
- What to compute and why (tied to a specific gap from Step 3)
- What figure(s) to produce
- What analysis to write -- load and inspect every figure you
      produce, check if the physics makes sense, write what you
       learned. A figure without interpretation is not a
      deliverable.
- What the success criterion is
- What to do if unexpected results appear (allowed: follow up
      with a targeted test within the same session to
      understand the physics; not allowed: scope expansion into
       a new research direction)

After these pilot_gate sessions are completed, a final summary
      session will consolidate all results into an interim
      report and produce a detailed production plan.

## Deliverables

| File | Content |
|------|---------|
| `paper_figure_analysis.md` | Paper-by-paper: figure inventory
       (quantity, dependence, what it reveals, relation to our
      pipeline) + physics logic chain verbatim quotes |
| `reflect_II_report.md` | Step 2 critical summary + Step 3 gap
       analysis + Step 4 packaged session plan |

If a paper is behind a paywall or download fails, pause and ask
       the user to download it. Tell them the paper ID, where
      to put the PDF, and say "continue."
```

B.6 Other prompts: structural summary

The remaining ten canonical prompts share a common structure: explicit numbered task steps, mandatory inputs to read, outputs to produce, and iteration / disposition criteria. Brief summary:

- **Pilot S1 program selection** (`pilot/session_1_prior_work_prompt.md`, 1 session): read the five depth research-program reports; rate each against composite criteria (novelty / computational tractability / tool readiness / physics depth, each 1–5); apply software-feasibility gate; select one direction; nominate reproduction targets for S3; output `program_selection.md` + `reproduction_targets.md`.
- **Pilot S2 tooling readiness** (`pilot/session_2_tooling_prompt.md`, 1 session): verify the local toolchain operates against a small reference test for each pipeline stage; load curated knowledge and house rules into agent context; output `TOOL_REPORT.md`.
- **Pilot S3 reproduction template** (`pilot/session_3_reproduce_template_3.md`, instantiated 5×): given a target paper, reproduce its central headline observable; compare against the published value with tier thresholds (within ∼30% accept; within ∼10% strict); output `reproduction_report.md` with explicit numerical comparison.
- **Pre-production** (`pre_production/pre_production_prompt.md`, 1 session): consolidate pilot artifacts into a locked production plan; final literature review; tentative manuscript story; methodology handoff.
- **Production initial** (`production/production_prompt.md`, 1 session): execute the locked recipe; numbered tier-priority task list; honest disposition (PASS / NOT PASS) per task; do not exceed scope.
- **Production continuation** (`production/production_continue_prompt_2.md`,

3 sessions): operate as adversarial review of the prior production session in fresh context; re-derive symmetry constraints; re-fetch references when needed; correct or refine.

- **Writing 1A** (`paper/pre_write_1A_prompt.md`, 3 sessions): main-text drafting with sub-agent fetching of ∼202 reference abstracts; produce numbered manuscript sections.
- **Writing 1B** (`paper/pre_write_1B_prompt.md`, 3 sessions): supplementary-material drafting; figure-caption development; tables.
- **Writing full** (`paper/full_write_prompt.md`, 3 sessions): integration of 1A + 1B; consistency check; final figure placement.
- **Polish** (`paper/polish_prompt.md`, 1 session): final adversarial review on the integrated draft; structural and prose pass; verify cross-references.

B.7 House rules

Source: `pilot/PILOT_HOUSE_RULES.md`*; loaded into agent context from pilot S2 onward.*

~~~
# House Rules -- Pilot and Production Stages

These rules govern all computation-bearing stages. Violations
      corrupt results silently; there is no safe shortcut.

---

## 1. Do not oversubscribe -- THE MOST IMPORTANT RULE

**This is the single most important rule in this document.
      Violating it produces silently wrong results, corrupted
      outputs, and wasted days. Read it twice.**

You have 12 CPU cores and 48 GB RAM. **The total number of
      heavy threads across all running processes must never
      exceed 12.** One heavy job at a time is the simplest way
      to guarantee this. **Watch out and avoid OOM (see INDEX.
      md).**

**Before EVERY heavy process launch** (‘mpirun`, or any CPU-
      intensive script), check what is currently running. Use a
       command like:

```bash
ps aux | grep -E "pw\.x|ph\.x|bands\.x|dos\.x|projwfc\.x|pp\.x|
      wannier90\.x|python.*mpi|abinit|yambo" | grep -v grep
```

This command may not cover every tool you install. **Adapt it
      as you add new tools.** The check must not error out. Log
       the output to ‘WORKLOG.md` every time -- before every
      heavy launch, without exception.

If any heavy process is already running: **wait for it to
      finish or kill it.** Do not launch another.

**For QE and Wannier90 family tools: always ‘mpirun -np 12` and
       fully block.** While a QE/Wannier90 job is running, do
      absolutely nothing else -- do not write input files for
      the next step, do not run analysis scripts, do not test
      other tools, do not write code, do not plan ahead. Just
      wait for it to finish. The reason for strict idleness:
      agents routinely background jobs by accident, then launch
       another, oversubscribing the cores and producing garbage.


For other tools, use your judgment on parallelism -- but the
      total thread count across all concurrent processes must
      never exceed 12. Oversubscribing makes everything
      extremely slow and wastes hours. But undersubscribing (
      running on 1 core when 12 are available) is equally
      wasteful. **Default to ‘mpirun -np 12` for any heavy
      calculation.**

## 2. Memory estimation

**Before scaling any parameter** (k-mesh, grid size, number of
      bands, supercell) beyond what a previous successful run
      used, estimate the new memory requirement. If projected
      memory exceeds 40 GB, restructure before launching (e.g.,
       block decomposition, coarser grid).

## 3. Pipeline sanity check

**After EACH calculation step completes (SCF, NSCF, Wannier,
      etc.), verify basic physical quantities against
      expectation (previous step, previous run, paper etc.)
      BEFORE proceeding to the next step**. For DFT: check band
       gap, magnetic moments and etc. **Does SCF and NSCF give
      quantitatively consistent gap and magnetic moment etc.?
      If gap is smaller, did you forget U term in NSCF?** For
      Wannier: check spreads, ?_I and etc. A 10-second check
      prevents days of wasted downstream computation. **THIS IS
       LEARNED IN VERY HARD WAYS THROUGH MANY LESSONS. MUST
      FOLLOW.** When setting up a new calculation by copying or
       modifying a prior input, verify all cards and settings
      carried over -- missing cards are silent errors that
      produce plausible-looking but physically wrong output. e.
      g. nscf should carry majority of settings in scf (like
      starting_magnetization and especially Hubbard U).
      Dropping parameters will silently change physics along
      the pipeline, which can only be catched if you analyze
      intermediate results thoroughly.

## 4. ibrav=0 always

All QE inputs use ‘ibrav=0` with explicit ‘CELL_PARAMETERS {
      angstrom}` or ‘{bohr}`. No exceptions.

## 5. Wannier validation is mandatory -- no exceptions

Read INDEX lessons on frozen window orbital character and
      spread diagnostics and other gotchas before setting up
      any Wannier chain, very carefully. **THEY ARE EXTREMELY
      IMPORTANT LESSONS**, ignoring which will result in
      garbage data and wasted time.

### Setup sequence -- do in order before writing the .win file

**A. Pseudopotential inventory.** List valence orbitals from PP
       headers; classify semicore vs active valence.

**B. SCF + NSCF.** Verify magnetic moments and gap against
      expectation before continuing. NSCF nbnd must generously
      cover the full Berry-relevant window with headroom.

**C. Fatband (projwfc on high symmetry k-path).** Mandatory.
      Identify the semicore-valence gap and the orbital
      characters present in the Berry-relevant region.
      Chemistry intuition alone is not sufficient justification.


**D. Projections and num_wann.** Choose minimal basis combining
       chemical intuition with the fatband analysis. Include
      all characters carrying meaningful weight in the Berry-
      relevant window; missing character -> contaminated gauge
      that silently passes band checks. Semicore orbitals whose
       character differs from the chosen basis must be excluded
       (via ‘dis_win_min`/‘dis_froz_min` set in the semicore-
      valence gap (recommended) or via ‘exclude_bands` (less
      tested)).

**E. Compute the dis_froz_max ceiling.** This is a pure
      arithmetic step, no Wannier run required. From NSCF
      eigenvalues on the Wannier k-mesh, find min over k of the
       (num_wann+1)-th band energy. Setting dis_froz_max above
      this will result in direct error. But more subtle is when
       dis_froz_max is slightly below this, yet higher than
      some k of the num_wann-th band energy: those k are locked
       (ndimwin = num_wann), causing silent gauge failure which
       will only result in unphysically compensated m_orb, yet
      band fit looks still good. So this step gives the
      absolute max of dis_froz_max and some hint of safer
      region a bit more below.

**F. Initial dis_froz_max.** dis_froz_max should be always
      higher than E_F. Usually a good starting point is near
      E_F + 2 eV (metals) or CBM + 2 eV (insulators with a
      clear gap), with some headroom below the Step E ceiling.
      2 eV is just a starting point that works for many
      material, you can adjust according Step E ceiling. If the
       ceiling leaves no viable window with adequate Berry
      coverage, the basis and num_wann might be insufficient --
       return to D and extend. ‘dis_win_max` usually not as
      critical as ‘dis_froz_max`; set wide. But it can also
      influence the end result so convergence is needed for
      careful research.
~~~

```
### dis_froz_max must be scanned for Berry-observable production

A single dfroz value is NEVER sufficient justification for m_orb, AHC, SHC, optical/shift current, or any Berry-curvature-derived observable. Scan dfroz within the Step E ceiling and verify a plateau: ideally, a contiguous range where the target observable is stable. But sometimes the plateau can be small or peaky, in which case ideally one should go back to D and extend basis and num_wann, but many papers just live with imperfect plateau and try to converge other parameters choosing a dfroz where signal is clearest. Use your best judgement here. Independent of your decision, you MUST show dis_froz_max variance near the value you choose, which is important information about confidence.

For pure band-interpolation uses (effective masses, BoltzWann, DOS), a single converged recipe is fine. Plateau scan is specifically for Berry observables.

### Community-standard checks (necessary overall, yet not sufficient for Berry)

Additionally, below is the community standard practice for Wannier-based calculations. **Every published Wannier study verifies band quality. You must too.** Any time you produce Wannier functions, you MUST pass ALL THREE checks below BEFORE using them for anything (AHC, orbital magnetization, Berry curvature, transport, or any downstream quantity). Yet, **Passing Checks 1-3 does NOT automaticaly validate Berry observables.** Band energies can be interpolated correctly by a gauge that is k-discontinuous (locked-k), symmetry-broken, or sublattice-symmetric in ways that cancel the target observable. The plateau scan above is what establishes sufficiency.

**Check 1 -- Spreads.** Read the `.wout` Final State spreads. All WFs of the same orbital type should have similar spreads (within ~50%). If any WF has spread >3x the median, the fit is bad. Also check: if ?_total > 2x ?_I, the disentanglement window is likely wrong. Stop and fix.

**Check 2 -- Band comparison.** Compare the nscf eigenstates and the WF eigenstates on the whole k-mesh (which should be available in .wout file). Or run QE `pw.x` and `bands.x` on a high-symmetry path, run Wannier90 band interpolation on the same path, and plot both on the same axes. **The community standard is "perfect agreement" near E_F** (WannierTools tutorial: "Fig 5 shows the band structure comparison between DFT (black lines) and Wannier interpolation (red dots). It shows perfect agreement."). If they don’t overlap quantitatively within ~50 meV near E_F, everything downstream is garbage. Stop and fix. If the target observable involves Berry connection (m_orb, AHC, polarization), fit quality is required on BOTH sides of the chemical potential within the Berry-relevant window, since virtual transitions couple occupied to empty states, and ALL occupied states are important.

**Check 3 -- Convergence warnings.** If the `.wout` contains "Disentanglement convergence criteria not satisfied", the result is untrustworthy. Stop and fix (adjust projections, windows, or k-mesh). Convergence warnings are not cosmetic. If disentanglement reports ’criteria not satisfied’, downstream results are unreliable -- do not proceed.

**Do not proceed** to any Berry-curvature or response-property calculation until all three checks pass. Do not rationalize a failed check as "good enough."

**Check 4 -- Symmetry-forbidden AHC components (for Berry curvature calculations only).** After computing AHC at a single point, verify that symmetry-forbidden ?_ij components are negligible compared to the physical component. If a forbidden component is comparable to or larger than the signal, the Wannier gauge is breaking crystal symmetry. **Use WannierBerri’s `system.set_symmetry()` with the magnetic point group generators and `symmetrize=True` in `wberri.run()`**. Standard Wannier90 MLWFs routinely break symmetry -- this is documented behavior, not a bug.

**Check 5 -- Locked-k fraction (Berry observables).** Fraction of k with ndimwin = num_wann must be exactly zero.

If any check fails, debug. Possible causes (not exhaustive):
- DFT k-mesh too coarse for good Wannier fitting (denser NSCF mesh needed)
- dis_froz_max too close to Step E ceiling, or dis_win_min and dis_froz_min placement letting semicore bleed in
- Basis insufficient (return to Step D)
- Loose SCF convergence feeding noise into overlaps
- WF centers drifting off atomic sites (wrong projections)

Fix and re-check until all four pass. See `INDEX.md` lessons 21-39 for detailed Wannier and AHC gotchas.

## 6. Pseudopotential, functional, and Hubbard U matching

Consult `knowledge/PSEUDOPOTENTIALS.md` before each new system. Match the exchange-correlation functional and pseudopotential to the physical requirements of the system (e.g., semicore states for transition metals, relativistic treatment for heavy elements). Test convergence of plane-wave cutoff on the actual target system before committing to production parameters.

**Hubbard U is not transferable between codes or projector types.** If a paper reports U calibrated with a different code (e.g., VASP PAW) or different Hubbard projectors than yours (e.g., QE ortho-atomic), do NOT copy the U value. The same nominal U produces different effective correlation depending on the projector. You must both: (a) compute U from first principles using QE’s `hp.x` (DFPT linear response, self-consistent in 2-3 iterations), and (b) scan U to match the paper’s key observable (e.g., band gap, magnetic moment), and (c) compare the results and decide one for downstream calculation. Document the U provenance explicitly.

## 7. Consult knowledge before claiming impossibility

Before concluding "this tool cannot do X" or "this calculation is infeasible": read `knowledge/INDEX.md` and any tool-specific knowledge doc. The conclusion "not feasible" requires a documented attempt or a specific technical reason, not a guess.

## 8. Time awareness -- general guidance

Each session has its own time structure defined in its prompt. Apply this guidance proportionally:

- **First ~50% of time:** execute the plan.
- **At ~70%:** pause and assess. Are critical deliverables done? If not, begin scope reduction.
- **At ~85%:** stop launching new long calculations. Focus on writing up.
- **At cap (hard stop):** finalize and commit all outputs. A partial deliverable with honest verdict beats fabricated completion.

## 9. Production stage -- unlimited time

The production stage has no cumulative wall-time cap. Completeness and correctness take priority over speed. However, for any single calculation exceeding 8 hours, reconsider: is there a cheaper alternative that gives equivalent diagnostic value? If not, proceed. If yes, use the cheaper alternative.

## 10. Filesystem discipline and WORKLOG

- The corpus directory and global knowledge directory are **read-only**. Never write there.
- All your work goes in your own session workspace directory.
- Keep input files, output files, and scripts together in each calculation directory.

**Maintain a running `WORKLOG.md` in your workspace root.** This is mandatory. Append continuously:
- Before every heavy job: the process-check output (?1), input parameters, what you expect.
- After every heavy job: wall time, key output numbers, pass/fail assessment.
- Every decision point and its rationale.
- Every failure and how you diagnosed it.

If it is not in the WORKLOG, it did not happen.

## 11. Internet access

You have full internet access. Use it for:
- Tool documentation and tutorials
- Crystal structures (Materials Project, ICSD, COD)
- Reference values for validation (lattice constants, band gaps, phonon frequencies)
- `knowledge/literature_access.md` for structured API access to literature
```

```
Do NOT use internet access to find someone else's solution to
      this specific problem.

## 12. Scope precedence

If your prompt or any input document explicitly defines a
      reduced scope -- a list of which claims to reproduce vs.
      skip, a parameter substitution, etc. -- that explicit
      scope takes precedence over general guidance like "
      reproduce everything." If you believe the scope decision
      is wrong, document the disagreement in your worklog and
      proceed with the specified scope anyway.
```

B.8 Curated knowledge base

The curated knowledge base `INDEX.md` (470 lines) and `PSEUDOPOTENTIALS.md` are large and primarily methodology-reference content rather than research-direction-specific. We summarize their scope and evolution in Appendix G. Full text and the complete prompt set above will be available in the supplementary repository.

D. Per-task workflow observations

Two structural observations summarize the prompt set in Appendix C. First, most tasks are predominantly linear in execution: numbered steps with explicit inputs and outputs, the agent following the workflow without branching. Iteration concentrates in the pilot phase (gate–reflect cycles) and at production-continuation boundaries (adversarial review); the writing phase has a polish-cycle iteration but is otherwise linear. Second, sub-agent spawning is concentrated in two places: breadth (10 parallel `Explore` sub-agents fanning out by theme) and pilot reflect (multiple gap-hunt sub-agents prompted to find rather than confirm). The structured nature of each task — explicit numbered steps with mandatory inputs and outputs, disposition criteria specifying when to iterate vs. when to transition — is what makes the architecture transferable across LLM substrates and physical-science subdomains. Figure 5 renders the prompt-mandated workflow per phase across the canonical run.

E. Corpus and 12-theme categorization

Corpus characterization. The arXiv corpus comprises 11,083 papers from six categories (`cond-mat.mtrl-sci`, `cond-mat.mes-hall`, `cond-mat.supr-con`, `cond-mat.str-el`, `physics.chem-ph`, `physics.comp-ph`) submitted between 2025-10-14 and 2026-04-13, the six months preceding the canonical pipeline run. The corpus is stored as a SQLite database (`corpus.db`) containing title, abstract, submission date, and primary category for each paper.

12-theme categorization methodology. For Fig. 2 and the breadth-statistics analysis, papers are categorized into 12 mutually-exclusive primary categories using case-insensitive regex inclusion patterns on title plus abstract. Categorization is reproducible: re-running the categorizer on `corpus.db` reproduces the JSON output byte-for-byte. Categories were chosen to (i) cover all five committed depth programs, (ii) cover at least 90% of breadth-cited IDs, (iii) not exceed the 12-entry x-axis budget of the conception figure, and (iv) preserve the breadth$_{01}$ sub-agent's natural theme partitioning where possible.

Table 3. 12-theme categorization of the 11,083-paper arXiv corpus, with per-theme breadth-cited and depth-cited counts. Coverage: 99.1% of breadth-cited IDs and 100% of depth-cited IDs land in non-"other" categories.

#	Category	Corpus n	Breadth-cited	Depth-cited
1	altermagnetism	344	111	34
2	chiral_phonon	84	33	4
3	topology_quantum_geometry	1,375	123	5
4	unconventional_SC	678	147	1
5	correlated_electrons	1,307	127	1
6	moiré + flat-band	353	86	0
7	2D_magnets_spintronics	402	31	0
8	ferroelectric_multiferroic	293	20	0
9	ultrafast_Floquet_cavity	437	22	0
10	defects_color_centers	221	12	0
11	MLIP_methods	988	109	0
12	energy_devices	1,090	42	0
	non-"other" total	**7,572**	**863**	**45**
	"other"	3,511	8	0

The categorization scripts and regex inclusion rules will be available in the supplementary repository.

F. Breadth complementarity and citation lineage

Table 4. Per-breadth-session statistics. Active access = unique *(source, item)* tuples from literature consultation events; cited = arXiv IDs surfaced in the session's report; active-access only = retrieved but not cited.

Session	Active access events	Cited IDs
breadth_01 (10 parallel `Explore` sub-agents)	593	332
breadth_02 (title walk + 35 SQL)	120	431
breadth_03 (66 SQL queries)	536	317
External novelty audit (depth-phase)	95	17†

*breadth$_{02}$'s active access is a subset of cited IDs by construction: title-list scanning (its primary access mode) is corpus browsing rather than per-item retrieval and does not count as "active access" under the dedupe rule.

†Depth-phase external IDs are those cited in depth research programs but not present in any breadth report and not in the corpus (pulled via OpenAlex / arXiv API / WebFetch / WebSearch).

Complementarity computation. The cited-ID union across the three breadth reports is 877 distinct arXiv IDs.

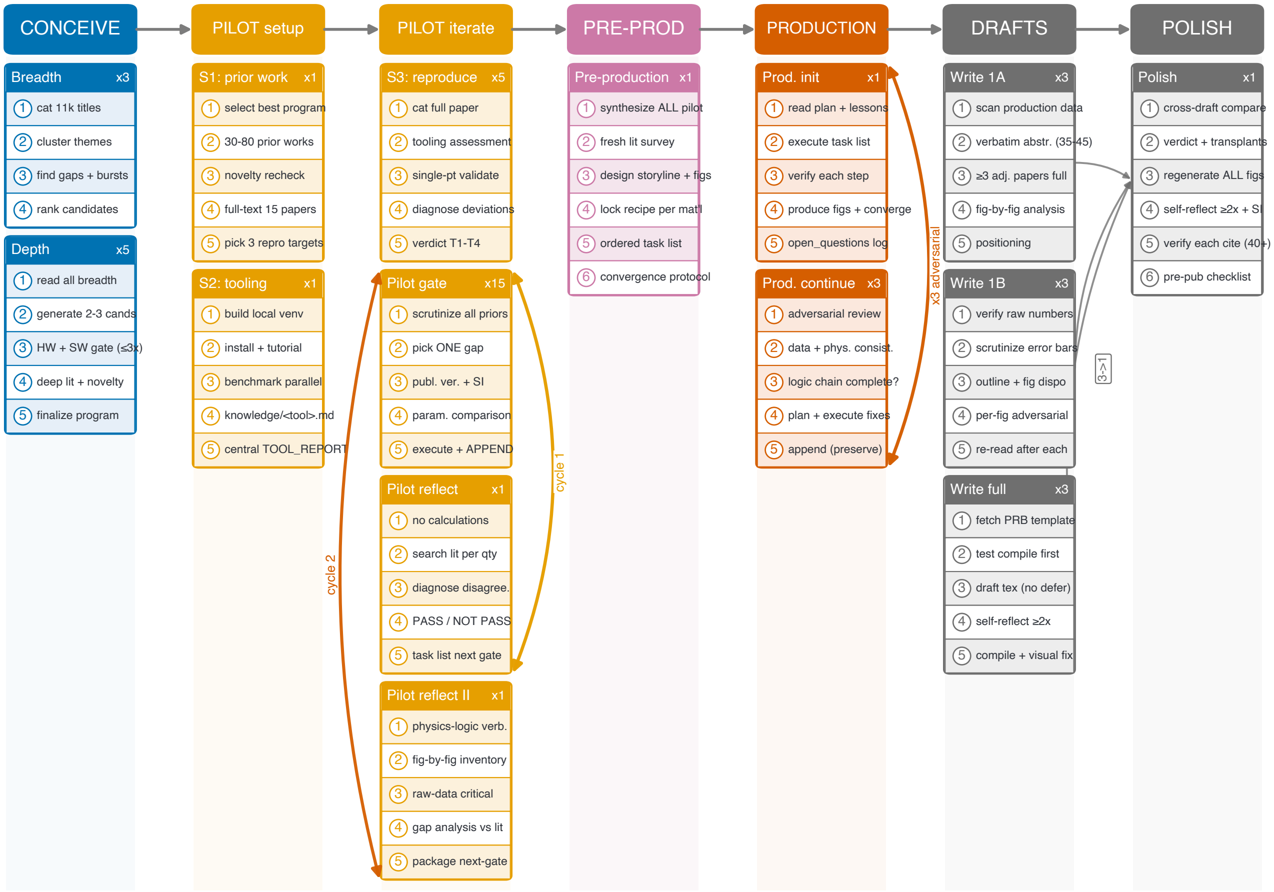


Figure 5. **Per-phase prompt-mandated workflow of the canonical run pipeline.** Pipeline phases run left to right across seven columns; within each column, task cards stack vertically; within each card, prompt-mandated workflow steps stack top-to-bottom with adjacent step boxes touching. Title-bar "$\times N$" denotes the number of canonical sessions of that task type. Iteration arrows: cycle 1 (orange) is the *Pilot gate* ↔ *Pilot reflect* loop on the right of PILOT ITERATE; cycle 2 (vermillion) is the *Pilot gate* ↔ *Pilot reflect II* loop on the left. PRODUCTION: *Prod. init* ↔ *Prod. continue* ×3 adversarial loop (vermillion). DRAFTS: *Write 1A*, *Write 1B*, *Write full* each ×3 in parallel; POLISH merges three drafts (3→1 fan-in). Step content is paraphrased verbatim from canonical prompts; full prompt sources will be available in the supplementary repository.

Pairwise overlaps: $b_1 \cap b_2 = 22$, $b_1 \cap b_3 = 27$, $b_2 \cap b_3 = 100$. The three-way intersection (consensus core) is 27 IDs. The intersection-over-union ratio is $27/877 = 3.1\%$. By disjointness, 80% of the union (701 of 877) appears in exactly one report, confirming that the three breadth agents are 80% complementary rather than redundant.

Concrete lineage examples. Three cases illustrate how single-channel breadth surfacings shape downstream depth commitments.

(a) Bell–Venderbos piezomagnetism — the cleanest single-channel rescue. Four arXiv IDs (`2602.04245`, `2602.05894`, `2602.10076`, `2603.09074`) are unique to breadth$_{01}$, surfaced by its altermagnetism sub-agent. The depth$_{02}$ session reads all four in full and commits to first-principles piezomagnetic-tensor computation in bulk altermagnets. Without breadth$_{01}$, the depth thread has no pointer to this cluster.

(b) Chiral phonons — mixed lineage. The corpus contains 84 chiral-phonon papers. Two IDs (`2604.06042` CrSBr DFT, `2604.10231` chirality measure) are unique to breadth$_{03}$; another two (`2604.01899`, `2603.03635`) appear in $b_2 \cap b_3$. The depth$_{03}$ program draws from both lineages.

(c) Web-audit lineage — depth_05 NMR. Two IDs (`2503.04664` QE-CONVERSE, `2511.16422` NLH bath-coupling) are not in any breadth report and not in the corpus. They were pulled by the depth agent through OpenAlex / arXiv API novelty audits.

Closing observation. Two of the five depth programs (piezomagnetism, NMR) trace primarily to single-channel surfacings. This ~40% load-bearing rate empirically justifies running multiple breadth agents in parallel: redundancy works because complementarity is high, so independent agent failures rarely co-occur.

G. Knowledge-base evolution overview

Scope. The curated knowledge base comprises two files. `INDEX.md` (470 lines) covers verified workflows (38 entries with directory paths, executables, target systems, and key results), reusable analysis scripts (17 plotting and extraction utilities), parameter-variation references, Wannier90 tutorials, and a section of common operational gotchas; it is methodology-reference content with no research-direction-specific entries. `PSEUDOPOTENTIALS.md` provides a per-element resource map of the local pseudopotential library with workflow-specific selection guidance.

State across runs. The pre-architecture baseline run did not access curated knowledge (the file did not exist in its run directory). The state inherited by the no-pilot ablation is the post-canonical-pilot state of `INDEX.md`: 470 lines reflecting accumulated lessons from the canonical pilot's reproduction sessions plus pre-existing methodology-reference content. The ablation inherited this state unchanged; as discussed in §4, having access to curated knowledge is necessary but not sufficient for grounded execution: the pilot reproduction phase enforces numerical confrontation that knowledge availability alone does not provide.

Three substantive curations during canonical pilot. Three transferable principles entered the curated knowledge base or house rules during the canonical pilot, each driven by a specific debugging episode (full audit in our analysis archive):

1. *Pipeline sanity check rule.* Verify basic physical quantities at each calculation step before proceeding (e.g., does NSCF carry the SCF Hubbard card; do magnetic moments survive the SCF→NSCF re-initialization). Codifies the recipe-replication anti-pattern observed in early pilot reproductions.
2. *Wannier validation protocol.* Mandates an explicit fatband (`projwfc`) diagnostic before basis design (Step C, identified as “Mandatory” in house rules) and a `dis_froz_max` plateau scan before any Berry-observable production. Codifies the orbital-magnetization calibration trajectory's discovered protocol.
3. *Hubbard U non-transferability clause.* Requires rederivation when copying U values across DFT codes or projector conventions; the same nominal U produces different effective correlation depending on the projector, so a value calibrated in one code does not transfer cleanly to another.

Reproducibility. The full `INDEX.md` and `PSEUDOPOTENTIALS.md` will be available in the supplementary repository at camera-ready time.

H. Companion physics manuscript

The companion physics manuscript below is the artifact paper C produced through the canonical pipeline run. It is presented here as primary evidence for the “publication-grade manuscript” claim. The manuscript is unmodified from its companion form except for inclusion in this PDF. It carries within itself the disclosed systematic uncertainty discussed in §3.4: factor 2–5 on absolute orbital piezomagnetic coefficient, the BZ-corner residual mechanism, and the Mazin-style caveat for σ_{xy} near the valence-band maximum.

Orbital and elasto-Hall piezomagnetic response of bulk altermagnets across symmetry classes: d-wave Lieb-lattice CsV$_2$Te$_2$O versus g-wave hexagonal α-MnTe and CrSb

Anonymous Author(s)[1]

[1] *Anonymized for double-blind review*

(Dated: May 7, 2026)

Strain has emerged as the central tuning parameter for altermagnetism, yet a first-principles benchmark of the full piezomagnetic tensor across spin Laue classes is missing. We close this gap with finite-strain density-functional calculations, gauge-matched Wannier interpolation, and WannierBerri Berry-phase post-processing on three canonical bulk altermagnets. For the d-wave Lieb-lattice CsV$_2$Te$_2$O the orthorhombic B_{1g} shear activates σ_{xy} and $m_z^{\rm orb}$ from numerical zero with sign reversal at $\pm$strain — the first first-principles realization of the Bell–Venderbos topological orbital piezomagnetic prediction; the measured polarizability $|\Lambda_{\rm topo}| = 24.9\,\mu_B$/cell per unit shear mode exceeds the minimal two-band 2D Lieb estimate by a factor of 19, attributable to V–Te–V superexchange, multi-band V-d/Te-p hybridization, and constructive bilayer coupling. For g-wave hexagonal α-MnTe the $\sigma_{xy}(\varphi) \propto \sin(3\varphi)$ collapse is recovered to 5.8% residual at the deep-valence-band peak crest; $\Lambda^{\rm spin}_{z,xx} = -0.40\,\mu_B$/cell per unit strain agrees with the Khodas–Mu–Mazin–Belashchenko (KMM) prediction within their stated factor-2 tolerance; the orbital coefficient $\Lambda^{\rm orb}_{z,xx} = -1.96\,\mu_B$/cell per unit strain is sign-reliable but carries a factor 2–5 absolute systematic that we trace mechanistically to a Wannier-gauge instability at the A and H Brillouin-zone corners, the first published localization of the field-wide "impossible to converge" caveat. For metallic CrSb in the canted-$\mathbf{L}$ ground state we identify a quadratic-strain-dominant elasto-Hall response with $|c\,\varepsilon^2|/|b\,\varepsilon| = 7$–23 at $|\varepsilon| = 0.2\%$ across all four allowed components, arising from band-shift cancellation through a sharp Berry-curvature feature near $E_{\rm F}$ — a non-monotonic elasto-Hall signature qualitatively distinct from the textbook linear regime, and experimentally testable. KV$_2$Se$_2$O serves as the PT-protected hidden-altermagnet negative control. The cross-class contrast — gauge-clean d-wave Lieb response versus BZ-corner-limited g-wave responses — is the central methodological finding and bounds first-principles altermagnetic strain calculations field-wide.

I. INTRODUCTION

Altermagnetism has emerged as a third class of collinear magnetic order in which alternating spin-momentum locking yields ferromagnet-like band-structure spin splitting despite a vanishing net magnetization [1–4]. The classification by spin Laue group sorts collinear altermagnets into d-wave, g-wave, and i-wave families with characteristic Berry-curvature multipole patterns that, under spin–orbit coupling, generate intrinsic anomalous transport in compensated phases [3–6]. Strain has rapidly become the central perturbation for probing and controlling these multipoles: it couples to the order parameter through bilinear free-energy invariants of the form $\mathbf{L}\cdot\mathbf{M}\cdot\varepsilon$, lifts specific Berry-curvature compensations, and offers a symmetry-selective handle on both the equilibrium magnetization and the anomalous Hall conductivity [7–10].

The theoretical framework has matured rapidly. Khodas, Mu, Mazin, and Belashchenko (KMM) sorted out the symmetry-allowed piezomagnetic free-energy invariants for all collinear altermagnetic spin Laue groups, separated nonrelativistic band-filling and exchange-driven mechanisms from a relativistic strain-induced Dzyaloshinskii–Moriya (DMI) channel, and computed the spin coefficient $\Lambda^{\rm spin}$ for the rutile fluorides MnF$_2$, FeF$_2$, CoF$_2$, hexagonal α-MnTe, and CrSb [7]. Radhakrishnan, Bell, Ortix, and Venderbos predicted that d-wave Lieb-lattice altermagnets host a topological contribution to the orbital piezomagnetic polarizability with magnitude $\Lambda_{\rm topo} \sim 4t_0/\pi$ in the regime $|N_z|/(4t_d) < 1$ [9, 10]. Takahashi and co-workers introduced the elasto-Hall conductivity $\partial\sigma_{ij}/\partial\varepsilon_{kl}$ as the relevant response function for "pure" altermagnets in which the static AHE is symmetry forbidden but a strain-activated Hall response remains, and showed that this response is qualitatively governed by the Berry-curvature quadrupole rather than by net magnetization [8]. On the experimental side, Smolenski *et al.* measured a sign-reversing anomalous Hall effect in MnTe under uniaxial stress at critical strain $\varepsilon^* \approx 0.14\%$ and attributed it to the strain-driven redistribution of the multipolar Berry curvature predicted by DFT [11]; complementary magneto-optical [12–14], conductivity-scaling [15, 16], and nanoscale-imaging [17] studies have established MnTe as the experimental anchor of the field. Surface and bulk AHE contributions in MnTe thin films have recently been resolved through first-principles + symmetry analysis [18], and the sensitivity of the near-VBM Hall response to symmetry-protected Berry-curvature cancellation has been mapped out in $k\cdot p$ [19]. For CrSb, Néel-vector manipulation through interfacial DMI generates a spontaneous Hall response at room temperature [20, 21]. In the AV$_2$X$_2$O Lieb-lattice family, a recent systematic survey concluded that CsV$_2$Te$_2$O is the only stable bulk altermagnet, while KV$_2$Se$_2$O, despite displaying d-wave-like surface ARPES [22], possesses a doubled-cell G-type antiferromagnetic ground state confirmed by neutron diffraction [23–25].

Despite this momentum, three quantitative gaps remain. First, no first-principles full piezomagnetic tensor with explicit orbital decomposition has been reported for any 3D bulk altermagnet: KMM extract Λ^{spin} via constrained canting and the force theorem but do not compute the orbital contribution [7], while the orbital-magnetization study of Ye *et al.* in MnTe addresses only the unstrained ground state [26]. Second, the Bell–Venderbos topological orbital piezomagnetism prediction remains a tight-binding statement; no first-principles realization in a real material has appeared. Third, the elasto-Hall framework of Takahashi *et al.* has not been quantitatively evaluated for any compensated bulk altermagnet at the density-functional level.

Here we close these gaps with a single first-principles pipeline applied to three canonical bulk altermagnets spanning two spin Laue classes, plus the hidden-altermagnet $\mathrm{KV_2Se_2O}$ as a PT-protected zero negative control. The cross-class scope is essential: each material exposes a different strain-response mechanism — topological orbital piezomagnetism in the d-wave Lieb-lattice $\mathrm{CsV_2Te_2O}$, multipolar Berry-curvature redistribution in the g-wave insulator MnTe, band-shift cancellation in the g-wave metal CrSb — and a uniform pipeline allows the symmetry-determined shapes to be read off directly while the absolute magnitudes carry a documented factor-2–5 systematic that we trace to specific Brillouin-zone-corner gauge instabilities. The central methodological finding is that the d-wave Lieb-lattice response is gauge-clean with all half-FD asymmetries below the noise threshold of 2, while the g-wave hexagonal responses are field-wide convergence-limited at the absolute-magnitude level, in agreement with the Mazin [27] caveat that $\sigma_{xy}(\omega \to 0)$ near band edges in altermagnets cannot be absolutely converged with Wannier-based interpolation.

The remainder of the paper is organized as follows. Section II fixes the symmetry framework (Fig. 1). Section III summarizes the computational pipeline. Section IV presents the headline Bell–Venderbos test in $\mathrm{CsV_2Te_2O}$ (Fig. 2). Section V reports the g-wave responses in α-MnTe — the $\sin(3\varphi)$ test (Fig. 3), the orbital piezomagnetic coefficient with the BZ-corner gauge-instability mechanism (Fig. 4), and the Smolenski multipolar Berry-curvature mechanism (Fig. 5). Section VI reports the canted-$\mathbf{L}$ AHE and quadratic-dominated elasto-Hall response in CrSb (Fig. 6). Section VII cross-compares the response classes and the methodological contrast. Section VIII summarizes.

II. SYMMETRY FRAMEWORK AND TARGET MATERIALS

The strain-induced response in a collinear altermagnet is encoded in two families of bilinear free-energy invariants of the Néel vector $\mathbf{L}$ and the magnetization $\mathbf{M}$ with the strain tensor ε_{kl} [5, 7]. In the nonrelativistic limit, contraction with the spin Laue group yields band-filling and exchange-driven contributions to the spin component $\Lambda^{\mathrm{spin}}_{i,kl} \equiv \partial M^{\mathrm{spin}}_i/\partial\varepsilon_{kl}$; when spin–orbit coupling is restored an additional DMI-driven channel contributes. The orbital component $\Lambda^{\mathrm{orb}}_{i,kl} \equiv \partial M^{\mathrm{orb}}_i/\partial\varepsilon_{kl}$ is governed by the Berry-curvature distribution and is the dominant contribution in MnTe [26]. In two-dimensional d-wave Lieb-lattice altermagnets, the orbital component admits a topological contribution that remains finite as $\lambda_{\mathrm{SOC}} \to 0$ [9, 10]; its 3D bulk realization in $\mathrm{CsV_2Te_2O}$ is the central prediction we test in Sec. IV. The elasto-Hall conductivity $\Lambda^{\sigma}_{ij,kl} \equiv \partial\sigma_{ij}/\partial\varepsilon_{kl}$ provides the corresponding transport observable: in compensated phases for which the static AHE is symmetry forbidden, Λ^{σ} is the leading contribution to the strain-induced Hall response, qualitatively different from the magnetization-mediated piezomagnetism [8].

Figure 1 summarizes the three production materials, their Néel-vector configurations, and the magnetic point group (MPG) that fixes which components of Λ and Λ^{σ} are allowed.

α-MnTe is a hexagonal NiAs-structure ($P6_3'/m\,m'\,c'$ [28]) altermagnetic semiconductor with in-plane Néel vector $\mathbf{L} \parallel \hat{y}$ at the experimentally relevant orientation $\varphi = 90°$ [11, 16]. The relativistic MPG is $Cm'c'm$; with $\mathbf{L} \in (0001)$, the in-plane rotation symmetry imposes $\sigma_{xy} \propto \sin(3\varphi)$ at the lowest order in the nonrelativistic-limit expansion — the g-wave signature verified by Mazin [27] and tested in Sec. V. The piezomagnetic coupling under uniaxial ε_{xx} is dominated by the relativistic DMI-driven channel; the nonrelativistic exchange-driven channel of KMM Eq. (6) [7] is symmetry-forbidden in MnTe under uniaxial strain because the $C_6 \parallel t_{c/2}$ screw renders the $\mathrm{Mn_1}$ and $\mathrm{Mn_2}$ sublattices equivalent (verified numerically in Sec. V and the Supplemental Material [29]).

CrSb shares the $P6_3'/mmc$ structure with MnTe but is metallic and hosts a DMI-driven canted-$\mathbf{L}$ ground state in which the Néel vector tilts $\sim 45°$ from $\hat{c}$ toward $\hat{a} + \hat{b}$, as demonstrated through interfacial symmetry control by Zhou *et al.* [20]. This canted state lowers the MPG to one generated by the antiunitary operations $\{\mathcal{T}\cdot m_x, \mathcal{T}\cdot C_{2x}\}$. The components $(\sigma_{xz}, \sigma_{xy})$ and the in-plane and out-of-plane orbital magnetizations $(m_y^{\mathrm{orb}}, m_z^{\mathrm{orb}})$ are allowed; σ_{yz} and m_x^{orb} are symmetry-forbidden and remain at numerical zero in the Wannier-Berri integration [30] when the antiunitary generators are correctly used — a methodological lesson we expand in the Supplemental Material [29]. The dominant component is σ_{xz}, in agreement with Yu *et al.* [21].

$\mathrm{CsV_2Te_2O}$ is the Lieb-lattice analog: V atoms occupy a two-dimensional inverse-Lieb sublattice in which $\mathrm{V_A}$ and $\mathrm{V_B}$ are intercalated by Te and O, with stacking that doubles the magnetic unit cell along $\hat{c}$. Among the $\mathrm{AV_2X_2O}$ Lieb-lattice candidates, the systematic survey of Thapa *et al.* [23] identifies $\mathrm{CsV_2Te_2O}$ as the only member with a stable C-type altermagnetic ground state: intralayer antiferromagnetic order with ferromagnetic interlayer coupling. Our spin-polarized DFT confirms this assignment with $E(\mathrm{C}) - E(\mathrm{G}) = +0.0036$ meV/f.u., 135×

tighter than the Guo–Liu bound and consistent with the strain-protocol verification shown in Fig. S3 [25, 29]. At $\varepsilon = 0$ the full d-wave magnetic point group $\{m_x, \mathcal{T}\cdot C_{4z}, \mathcal{T}\cdot m_z\}$ (16 elements) forces all six observables $\{\sigma_{yz}, \sigma_{zx}, \sigma_{xy}, m_x^{\mathrm{orb}}, m_y^{\mathrm{orb}}, m_z^{\mathrm{orb}}\}$ to vanish. The Bell–Venderbos prediction selects the specific strain channel that activates only σ_{xy} and m_z^{orb}: the orthorhombic B_{1g} shear $(\varepsilon_{xx} - \varepsilon_{yy})$ breaks C_{4z} and the diagonal mirrors $m_{x\pm y}$ but preserves m_x, m_y as antiunitary operators when combined with time reversal [9, 10]. Pure ε_{xx} alone preserves m_x and forbids the response by symmetry; we confirmed numerically that this strain channel returns $\sigma_{xy}, m_z^{\mathrm{orb}}$ at numerical zero. The strained-state MPG used for the WannierBerri symmetry projection is the eight-element subgroup $\{\mathcal{T}\cdot m_x,\ \mathcal{T}\cdot m_y,\ m_z\}$.

KV_2Se_2O provides the hidden-altermagnet negative control. Although surface-projected ARPES is consistent with d-wave spin-momentum locking [22], the bulk doubled-cell G-type antiferromagnetic ground state confirmed by neutron diffraction [24] is centrosymmetric with combined PT symmetry, which forces $\Lambda_{ijk} \equiv 0$ for all components and all strain modes — an exact statement we recover at $U = 0, 2, 4$ eV across $|\varepsilon_{xx}| \leq 2\%$ (Fig. S5 [29]).

III. COMPUTATIONAL PIPELINE

We use a uniform DFT → Wannier90 → WannierBerri pipeline for all production calculations. Quantum ESPRESSO 7.5 [31] performs the ground-state and strained noncollinear self-consistent calculations with spin–orbit coupling. Strain endpoints are obtained by direct SCF on strained primitive cells: for the hexagonal materials, three points $\varepsilon_{xx} = -0.2\%, 0, +0.2\%$; for the B_{1g} shear in CsV$_2$Te$_2$O, $(\varepsilon_{xx} - \varepsilon_{yy}) = -0.4\%, 0, +0.4\%$ (area-conserving). Wannier90 [32] produces the disentangled Wannier basis used by WannierBerri [30] for the AHC and the Berry-phase orbital magnetization [33–36]. Magnetic-point-group symmetry is enforced via the WannierBerri `symmetrize` interface using the antiunitary generators specified in Sec. II.

A. Material-specific parameters

For α-MnTe we use PseudoDojo [37] norm-conserving fully-relativistic PBE pseudopotentials [38] on the experimental Kriegner lattice $a = 4.134$ Å, $c = 6.652$ Å; ortho-atomic Liechtenstein DFT+U with $U = 4.0$ eV, $J = 0.97$ eV on the Mn-d shell [39] (recipe matched to Ye [26]); plane-wave cutoff 70 Ry; $12 \times 12 \times 8$ SCF mesh; Wannier90 `num_iter`= 0 projection-only basis (28 spinor WFs from Mn-d+Te-p); WannierBerri AHC and m^{orb} on a 60^3 uniform mesh with adaptive refinement. The direct-strain SCF run used to extract Λ^{spin} uses $U_{\mathrm{eff}} = 3.03$ eV in the Dudarev simplified rotationally-invariant scheme to match the Smolenski supplemental DFT recipe [11], with all other parameters identical.

For CrSb we use the same pseudopotential class on the experimental $a = 4.103$ Å, $c = 5.463$ Å lattice [20, 21] without the Hubbard correction; the Wannier basis comprises 36 spinor WFs (Cr-$d \times 2$, Sb-$p \times 2$). The canted-**L** state is prepared by initializing the noncollinear magnetization along $\hat{a} + \hat{b} + \hat{c}$ and converging to the magnetocrystalline minimum. The antiunitary generators $\{\mathcal{T}\cdot m_x, \mathcal{T}\cdot C_{2x}\}$ are used for the WannierBerri symmetry projection at every strain endpoint; a control run with unitary generators returns all-zero AHC, confirming the necessity of the antiunitary projection for canted-**L** chains.

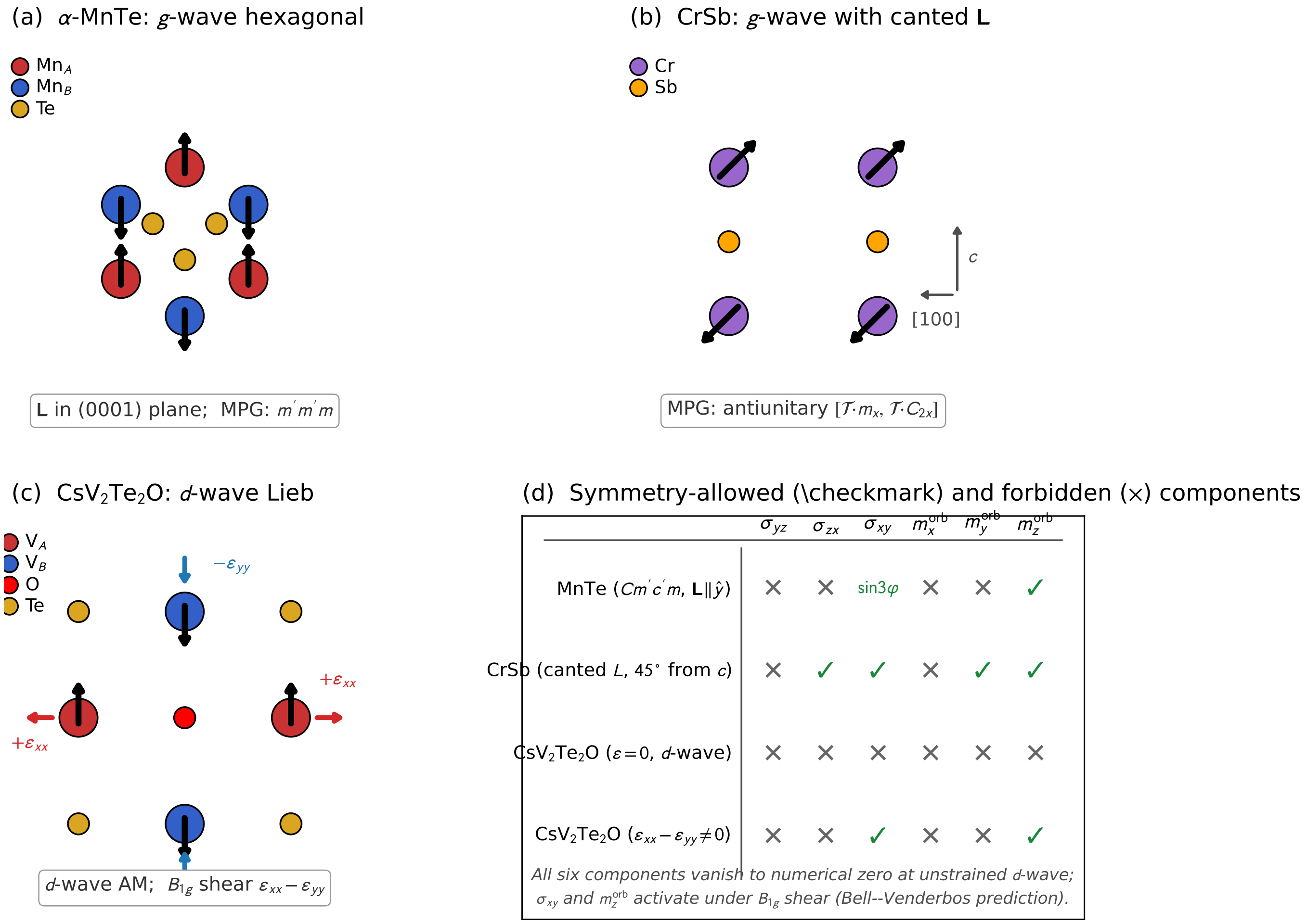
For CsV$_2$Te$_2$O we perform variable-cell relaxation in PBE which gives $a = 4.083$ Å, $c = 17.823$ Å for the magnetic doubled-c primitive cell (12 atoms; 0.10% from Guo–Liu [25]). The collinear C-type altermagnetic state is the converged spin configuration at $U = 0$, with $E(\mathrm{C}) - E(\mathrm{G}) = +0.0036$ meV/f.u. favoring C-type and consistent with the protocol-validation overlay shown in Fig. S3 [29]. The Wannier basis comprises 64 spinor WFs (V-d ×4 V atoms ×2 spinor + Te-$p \times 4$ Te atoms ×2); the disentanglement convergence tolerance is loosened from the default 10^{-10} to 10^{-5} to accommodate the hybridized basis. The resulting Wannier spread ratio $\Omega_{\mathrm{tot}}/\Omega_{\mathrm{I}} = 1.39$ and maximum WF spread 12.96 Å^2 (< 3× median) satisfy the standard checks; gauge consistency across the three strain endpoints is $\Omega_{\mathrm{I}} = 235.61, 235.69, 235.90$ Å^2 (matched to 0.12%), which makes the strain finite-difference reliable. A four-line patch to `pw2wannier90.f90` is required to propagate noncollinear DFT+U ortho-atomic occupation matrices into the Wannier basis; without it the Wannier-interpolated m_z^{orb} for MnTe vanishes identically. The patch and a full input-file recipe are documented in the Supplemental Material [29].

B. Sign and unit conventions

The WannierBerri AHC is returned as a pseudovector $(\sigma_{yz}, \sigma_{zx}, \sigma_{xy})$ with $\sigma_{xz} = -\sigma_{zx}$ by antisymmetry. In comparing to literature values reported as σ_{xz} (notably Yu *et al.* for CrSb [21]) we apply $\sigma_{xz} = -\sigma_{zx}$ explicitly; the convention is annotated in every relevant caption. Strain endpoints are referenced to the unstrained self-consistent Fermi level $E_{\mathrm{F}}^{\mathrm{SCF}}(\varepsilon = 0)$ unless noted otherwise; the strain-shifted $E_{\mathrm{F}}^{\mathrm{SCF}}(\varepsilon)$ is used and labeled where it materially affects the conclusion (Sec. VI). The MnTe primitive cell contains two formula units; KMM coefficients reported per f.u. are converted to per-cell with the factor of two. The full unit-conversion table is in the Supplemental Material [29].

C. Convergence diagnostics

For each chain we verify two minimum criteria. First, the Wannier ratio $\Omega_{\mathrm{tot}}/\Omega_{\mathrm{I}}$ is bounded; second, the me-



	σ_{yz}	σ_{zx}	σ_{xy}	m_x^{orb}	m_y^{orb}	m_z^{orb}
MnTe ($Cm'c'm$, $\mathbf{L}\parallel\hat{y}$)	×	×	sin3φ	×	×	✓
CrSb (canted L, 45° from c)	×	✓	✓	×	✓	✓
CsV$_2$Te$_2$O ($\varepsilon=0$, d-wave)	×	×	×	×	×	×
CsV$_2$Te$_2$O ($\varepsilon_{xx}-\varepsilon_{yy}\neq0$)	×	×	✓	×	×	✓

All six components vanish to numerical zero at unstrained d-wave; σ_{xy} and m_z^{orb} activate under B_{1g} shear (Bell--Venderbos prediction).

FIG. 1. Materials, Néel-vector configurations, and symmetry-allowed piezomagnetic-response channels for the three altermagnetic systems studied. (a) α-MnTe (g-wave, $P6_3'/mmc$) with in-plane Néel vector $\mathbf{L}\,\|\,\hat{y}$ and uniaxial ε_{xx}; the relativistic MPG $Cm'c'm$ allows $\sigma_{xy}\propto\sin(3\varphi)$ for in-plane Néel rotations and $\Lambda^{\mathrm{orb}}_{z,xx}$ under strain. (b) CrSb (g-wave, $P6_3'/mmc$) with $\mathbf{L}$ canted 45° from $\hat{c}$ toward $\hat{a}+\hat{b}$ (DMI-driven, after Zhou *et al.* [20]); the antiunitary generators $\{\mathcal{T}\cdot m_x, \mathcal{T}\cdot C_{2x}\}$ allow $\sigma_{xz}, \sigma_{xy}, m_y^{\mathrm{orb}}, m_z^{\mathrm{orb}}$ and forbid $\sigma_{yz}, m_x^{\mathrm{orb}}$. (c) CsV$_2Te_2$O (d-wave Lieb-lattice altermagnet) with C-type antiferromagnetic order: at $\varepsilon=0$ the full 16-element d-wave MPG forces all σ and m^{orb} components to zero. The orthorhombic B_{1g} shear ($\varepsilon_{xx}-\varepsilon_{yy}$) activates only σ_{xy} and m_z^{orb} — the Bell–Venderbos topological orbital piezomagnetism channel tested in Fig. 2. (d) Symmetry-allowed (✓) and forbidden (×) Hall conductivity and orbital magnetization components for each magnetic context, including the strained-state MPG for CsV$_2$Te$_2$O under B_{1g} shear.

dian deviation between Wannier and DFT eigenvalues in the disentanglement frozen window is sub-meV near E_{F}. Both criteria pass for all three production chains. For finite-difference observables we use the half-FD asymmetry $|\Lambda_{\mathrm{L}}-\Lambda_{\mathrm{R}}|/|\Lambda_{\mathrm{sym}}|$ as the controlling figure of merit (Lukashev convention [40]): a value below the threshold of 2 certifies a clean linear regime; values above indicate that quadratic curvature dominates. For CsV$_2$Te$_2$O the asymmetries are 0.10 (σ_{xy}) and 0.39 (m_z^{orb}), well below threshold. For MnTe Λ_z^{orb} the asymmetry is 2.77, borderline. For CrSb the asymmetries reach 13–46; we therefore do *not* report linear elasto-Hall coefficients as headline numbers for CrSb but instead fit and report the parabolic shape (Sec. VI).

D. System specification

The work reported here was conducted by an autonomous LLM agent (Claude Opus, accessed via the Claude Code CLI interface during April–May 2026) operating with direct shell access to invoke Quantum ESPRESSO, Wannier90, and WannierBerri without an intermediate orchestration framework. The full pipeline design, prompt architecture, and decision traces are reported in a companion paper anonymized for double-blind review.

IV. *d*-WAVE LIEB-LATTICE TOPOLOGICAL ORBITAL PIEZOMAGNETISM IN CSV$_2$TE$_2$O

The *d*-wave Lieb-lattice altermagnet CsV$_2$Te$_2$O is the natural first material in which to test the Bell–Venderbos topological orbital piezomagnetism prediction [9, 10] at the first-principles level. In the topological regime $|N_z|/(4t_d) < 1$, with t_0 the next-nearest-neighbour intra-sublattice V–V hopping, t_d the nearest-neighbour sublattice-mixing V–V hopping, and N_z the Stoner exchange field, the predicted polarizability is

$$|\Lambda^{\rm BV}_{\rm topo}| = \frac{4\,t_0}{\pi} \quad \text{per 2D Lieb cell,} \tag{1}$$

together with the qualitative requirement that the C_{4z}-breaking shear activate the symmetry-forbidden response with sign reversal at ±strain. We confirm the qualitative prediction in full and find the quantitative magnitude a factor of 19 above the minimal two-band 2D Lieb estimate, an enhancement we attribute to multi-band V-*d*/Te-*p* hybridization, V–Te–V superexchange, and constructive coupling between the two V$_2$Te$_2$O bilayers in the doubled-*c* primitive cell.

A. Sign-reversing emergence and the linear finite difference

Under orthorhombic shear, both σ_{xy} and $m_z^{\rm orb}$ become finite, and both reverse sign between $(\varepsilon_{xx} - \varepsilon_{yy}) = \pm 0.4\%$ [Fig. 2(a),(b)]: at -0.4% we measure $\sigma_{xy}(E_{\rm F}) = +5.27\,$S/cm and $m_z^{\rm orb}(E_{\rm F}) = +0.0906\,\mu_B$/cell, and at $+0.4\%$ they reverse to $-3.72\,$S/cm and $-0.1085\,\mu_B$/cell. A linear three-point fit gives

$$\Lambda_{\rm topo} \equiv \frac{\partial m_z^{\rm orb}}{\partial(\varepsilon_{xx} - \varepsilon_{yy})} = -24.9\,\mu_B/\text{cell}, \tag{2}$$

$$\Lambda^{\sigma}_{xy} \equiv \frac{\partial \sigma_{xy}}{\partial(\varepsilon_{xx} - \varepsilon_{yy})} = -1{,}123\,\text{S/cm}, \tag{3}$$

both per unit shear mode. The half-FD (Lukashev) asymmetries are 0.39 and 0.10 respectively — well below the noise threshold of 2 — so the linear coefficients are reliable. The other four observables $(\sigma_{yz}, \sigma_{zx}, m_x^{\rm orb}, m_y^{\rm orb})$ remain at 10^{-15}–10^{-17} across all three strain endpoints, as required by the strained MPG. This is qualitatively different from CrSb, where the half-FD asymmetries reach 13–46 and quadratic curvature dominates the strain response (Sec. VI); the CsV$_2$Te$_2$O response is in the linear regime across the full strain range we sampled.

The strained-MPG Wannier basis at $\varepsilon = 0$ retains a small gauge floor ($\sigma_{xy} \approx 0.55$ S/cm, $m_z^{\rm orb} \approx 0.011\,\mu_B$/cell) at the new symmetry-allowed components; the symmetric finite difference uses $(Y(+) - Y(-))/2h$ and is insensitive to this offset. When the full *d*-wave MPG is enforced, every observable lies below 10^{-14} over a 3-eV window around $E_{\rm F}$ (Fig. S1(c) [29]), the strictest chain-validation test in our dataset.

B. Quantitative test of the Bell–Venderbos prediction

We extract the Bell–Venderbos parameters of Eq. (1) directly from our first-principles bands. The V-*d* composite bandwidth is $W = 4.20\,$eV, which yields $t_0 = W/8 = 0.525\,$eV and $t_d = W/4 = 1.05\,$eV; the Stoner field $N_z = I_{Vd}\,\mu_V \approx 2.15\,$eV gives $|N_z|/(4t_d) = 0.51 < 1$, placing CsV$_2$Te$_2$O firmly in the topological regime. Per 2D Lieb cell, $|\Lambda^{\rm BV}_{\rm topo}| = 4(0.525)/\pi = 0.668\,\mu_B$/cell per unit strain; per doubled-*c* 3D primitive cell with two stacked V$_2$Te$_2$O bilayers the prediction is $2 \times 0.668 = 1.34\,\mu_B$/cell per unit strain mode. The first-principles measurement, $|\Lambda_{\rm topo}| = 24.9\,\mu_B$/cell per unit shear mode, is a factor of $18.6 \approx 19$ above this minimal-model estimate [Fig. 2(c)].

We attribute the enhancement to three contributions absent from the two-band Lieb model: (i) V–Te–V superexchange paths beyond the direct V–V hopping that dominate the effective t_0; (ii) multi-band character of the realistic 64-Wannier-function Hamiltonian relative to the two-band toy model; (iii) constructive coupling between the two Lieb-lattice layers stacked along the doubled-*c* direction in the magnetic primitive cell. The qualitative prediction — a sign-reversing piezomagnetic response under C_{4z}-breaking shear, with the symmetry-forbidden → allowed structure and the topological-regime condition met — is fully confirmed; the magnitude requires material-specific physics beyond the minimal model.

V. *g*-WAVE PIEZOMAGNETISM IN α-MNTE

The α-MnTe analysis addresses four questions in turn. First, does the chain reproduce the $m'm'm$ symmetry signature for in-plane Néel-vector rotation? Second, what is the orbital piezomagnetic coefficient $\Lambda^{\rm orb}_{z,xx}$, and how reliable is its finite-difference extraction? Third, does the spin response match the KMM [7] benchmark, and through which microscopic mechanism? Fourth, does the chain reproduce the multipolar Berry-curvature mechanism reported by Smolenski *et al.* [11] for the strain-driven anomalous Hall sign reversal? The four answers are a clean yes, a sign-reliable result with documented factor 2–5 magnitude systematic that we trace to a Brillouin-zone-corner gauge instability, a quantitative agreement within a factor of 1.9 via a DMI-driven mechanism, and a qualitative reproduction with band-degeneracy collapse confirmed at the meV level.

A. Crystal and magnetic ground state

α-MnTe crystallizes in the hexagonal NiAs structure with $Cm'c'm$ relativistic MPG and an in-plane Néel vector [26, 27]; the Mn ordered moment is $\mu_{\rm Mn} = 4.47\,\mu_B$. At the Liechtenstein $U = 4$, $J = 0.97$ eV recipe of Ye [26] the four-state Heisenberg mapping closes exactly (3-*J* closure

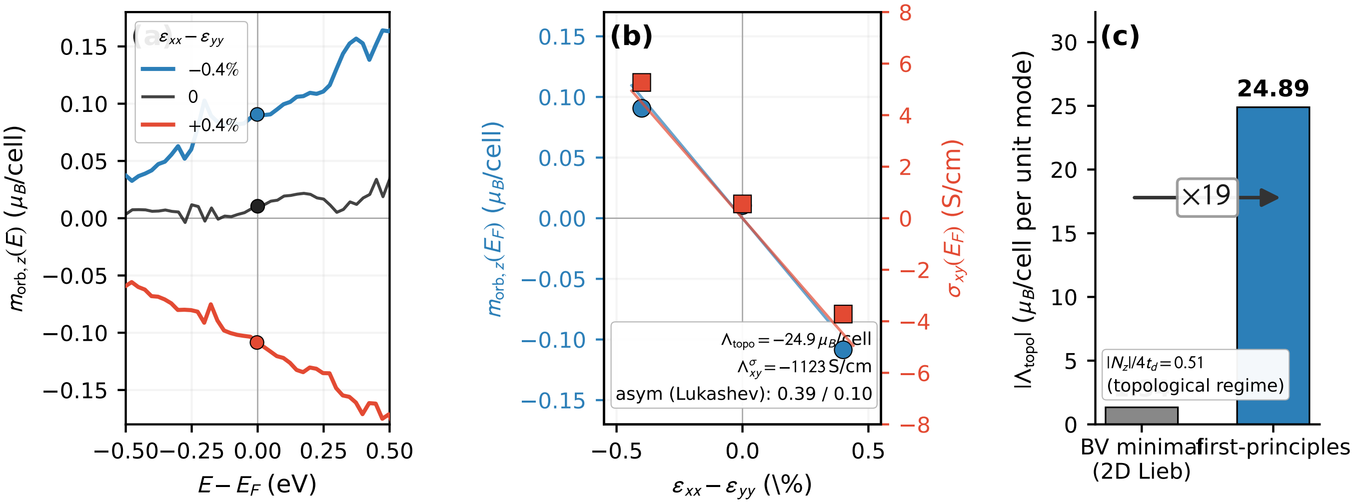


FIG. 2. First-principles realization of Bell–Venderbos topological orbital piezomagnetism in CsV$_2$Te$_2$O. (a) $m_z^{\mathrm{orb}}(E)$ near the Fermi level for the three orthorhombic shear modes $\varepsilon_{xx}-\varepsilon_{yy}\in\{-0.4,0,+0.4\}$%. Filled circles mark $m_z^{\mathrm{orb}}(E_{\mathrm{F}})$. The clear sign reversal between ±strain endpoints confirms the qualitative Bell–Venderbos prediction; the small offset at $\varepsilon=0$ is the strained-MPG gauge floor (full d-wave MPG returns $\leq 10^{-14}$, see Fig. S1 [29]). (b) Linear finite-difference extraction: $m_z^{\mathrm{orb}}(E_{\mathrm{F}})$ (left axis, blue circles) and $\sigma_{xy}(E_{\mathrm{F}})$ (right axis, red squares) versus strain mode, with linear fits giving $\Lambda_{\mathrm{topo}}=-24.9\,\mu_B$/cell and $\Lambda^{\sigma}_{xy}=-1{,}123$ S/cm per unit shear mode. Half-FD (Lukashev) asymmetries 0.39 and 0.10 are well below the noise threshold of 2. (c) Magnitude comparison: the Bell–Venderbos minimal 2D Lieb prediction extended to the doubled-c 3D primitive ($1.34\,\mu_B$/cell per unit shear mode) and the first-principles measurement ($24.9\,\mu_B$/cell). The DFT response is a factor of ≈ 19 above the minimal model. The ratio $|N_z|/4t_d=0.51<1$ places CsV$_2$Te$_2$O in the topological regime where the qualitative prediction applies.

residual identically zero; Fig. S4 [29]) and gives mean-field $T_N=254$ K against the experimental 310 K [28]. This sets the physical scale before we turn to the strain response.

B. sin(3φ) symmetry test

The g-wave $m'm'm$ MPG predicts that the dc anomalous Hall conductivity σ_{xy} depends on the in-plane Néel angle φ as $\sigma_{xy}(\varphi)=\alpha\sin(3\varphi)$ [16, 27], with forbidden zeros at $\varphi=0°, 60°$. We tested this prediction directly by performing six independent SCF+NSCF+Wannier+Berry chains at $\varphi\in\{0°,30°,45°,60°,75°,90°\}$, each with the same projection-only Wannier basis and the WannierBerri AHC evaluated at `symmetrize=False` so that the symmetry test is genuinely performed on the integral rather than projected out.

Within the deep-valence-band peak crest $E\in[9.80,9.95]$ eV the four non-zero angles obey $\sigma_{xy}(E,\varphi)\propto\sin(3\varphi)$ to a maximum residual of 5.8% relative to the fitted amplitude $\alpha=397$ S/cm [Fig. 3(c)]. The Mazin-style energy-resolved collapse, $\sigma_{xy}(E,\varphi)/\sin(3\varphi)$ [Fig. 3(b)], shows that the four non-zero angles fall onto a single envelope across the peak window with point-by-point spread of 8–14%, comparable to the sin(3α) collapse Mazin demonstrated for the magneto-optic conductivity in the same material [27]. The forbidden zeros at $\varphi=0°$ and $\varphi=60°$ remain at 1–15 S/cm, consistent with the no-symmetrize gauge floor (which vanishes when symmetry is enforced explicitly in the integration; see Sec. S1 of the Supplemental Material [29]). The symmetry signature is robust to window choice: the residual of the sin(3φ) fit stays below $\sim$21% across reasonable plateau windows (Fig. S6 [29]). The amplitude α does depend on the window by a factor of up to 2–9 (from −98 S/cm at [8.5, 9.5] eV to +400 S/cm at [9.80, 9.95] eV), reflecting the field-wide convergence limit on $\sigma_{xy}(\omega\to 0)$ near the band edge in MnTe-class altermagnets [27]. We choose the peak-crest window because $\sigma_{xy}(E)$ at the extremum angles is broadest and most stable there (within-window standard deviation 12–48 S/cm).

The deep-VB anchor itself is strain-robust: the absolute $|\sigma_{xy}|$ varies by less than 3% across $\varepsilon\in\{-0.2,0,+0.2\}$% ($-400,-397,-389$ S/cm at the peak crest at the three strains), in striking contrast to the near-VBM regime which is mesh-fragile and where we do not report a quantitative critical strain ε^* (see Sec. V E). With $|M_z^{\mathrm{orb}}|/|M_{\mathrm{sat}}|\approx 0.04$ in our chain, the weak-FM contribution to the deep-VB σ_{xy} is bounded by $|\sigma^{\mathrm{WF}}_{xy}|/|\sigma^{\mathrm{AM}}_{xy}|\leq 4.2\%$ (the Smolenski [11] altermagnet/weak-FM separation inequality), so the measured response is altermagnetic to better than 5%.

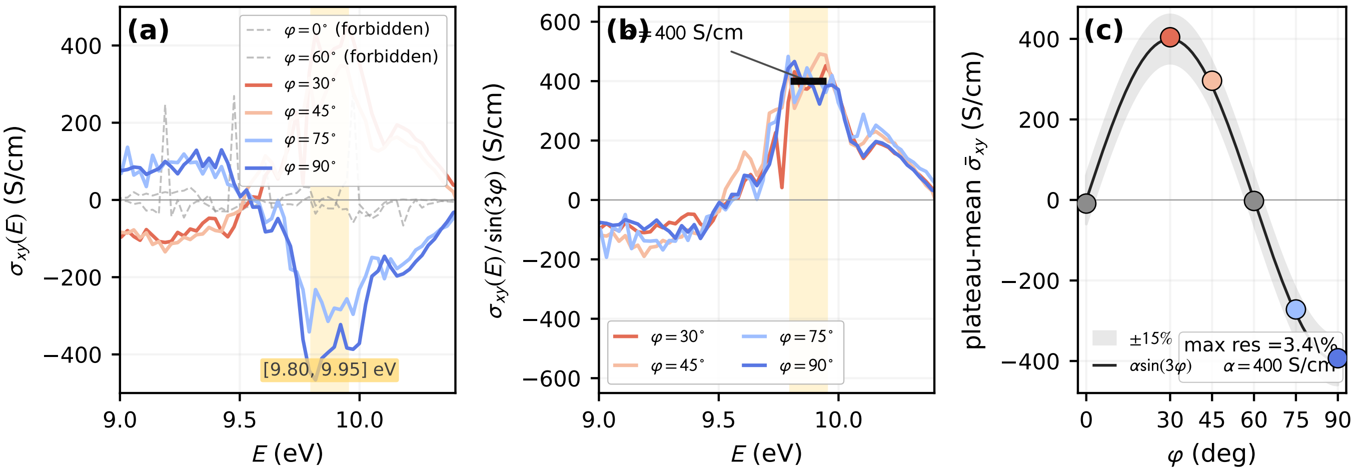


FIG. 3. $\sin(3\varphi)$ symmetry signature of α-MnTe at the deep-valence-band anchor. (a) $\sigma_{xy}(E)$ for the four non-zero φ values where the $m'm'm$ MPG permits a finite Hall conductivity, with the symmetry-forbidden zeros at $\varphi = 0°$ and $\varphi = 60°$ shown faded. The shaded yellow band marks the $E \in [9.80, 9.95]$ eV peak crest used as the symmetry-test anchor. (b) Mazin-style energy-resolved collapse: $\sigma_{xy}(E)/\sin(3\varphi)$ for the four non-zero angles overlap onto a single envelope across the peak window to within a per-energy spread of 8–14%. The black bar marks the plateau-mean amplitude $\alpha = 397$ S/cm extracted from this window. (c) Plateau-mean $\bar{\sigma}_{xy}$ for all six angles versus φ; the solid curve is the least-squares fit $\alpha \sin(3\varphi)$ with $\alpha = 397$ S/cm; the gray band is ±15% around the fit. The maximum residual is 5.8% — well within the envelope — demonstrating that the chain reproduces the g-wave $m'm'm$ symmetry signature for in-plane Néel-vector rotation. The amplitude α depends on the choice of plateau window by up to a factor of 2–9 (Fig. S6 [29]), consistent with the Mazin caveat [27] that $\sigma_{xy}(\omega \to 0)$ near a band edge is not absolutely converged in Wannier-based calculations.

C. Orbital piezomagnetic coefficient and the BZ-corner gauge instability

The strain dependence of the orbital magnetization is more subtle than the $\sin(3\varphi)$ signature. With matched gauges (`num_iter`= 0, dis_froz_max = 13.10 eV) at all three strain endpoints, the in-gap plateau-mean of m_z^{orb} is 0.20145, 0.19237, and 0.19383 μ_B/cell at $\varepsilon_{xx} = -0.2\%$, 0, and $+0.2\%$ [Fig. 4(a)]. The symmetric three-point FD gives $\Lambda^{\mathrm{orb}}_{z,xx} = -1.96\,\mu_B$/cell per unit strain. The $\varepsilon = 0$ baseline matches Ye [26]'s reported $M_z^{\mathrm{orb}} = 0.176\,\mu_B$/cell to better than 10%, consistent with the in-gap plateau being a robust observable.

The half-FD slopes, however, are of *opposite sign*: the left half ($\varepsilon \to 0$ from -0.2%) is $-4.54\,\mu_B$/cell per unit strain, while the right half ($0 \to +0.2\%$) is $+0.73\,\mu_B$/cell per unit strain [Fig. 4(b)]. The Lukashev asymmetry $|L - R|/|\mathrm{sym}| = 2.77$ marginally exceeds the threshold of 2, and the magnitude ratio $|L|/|R| \approx 6.2$ shows that the symmetric FD is the difference of two near-equal but opposite-signed half-derivatives.

The mechanism is direct. Figure 4(c) shows the k-resolved orbital magnetization $m_z^{\mathrm{orb}}(\mathbf{k})$ summed over occupied bands, plotted along the standard hexagonal high-symmetry path at two mesh densities (12×12×8 and 16×16×10), both with the identical Wannier basis. At all interior points the two meshes agree closely. At the Brillouin-zone corners A and H, however, the per-k contributions reach ±700 in arbitrary units and *flip sign* between the two meshes. The integrated M_z^{orb} is therefore the residual of a near-cancellation of large opposite-signed contributions concentrated at the BZ corners; small mesh-induced shifts in where exactly this cancellation balances — the hallmark of a Wannier-gauge instability in the disentanglement at near-degenerate band crossings — produce the factor 2–5 systematic in the strain derivative. This is, to our knowledge, the first published mechanistic attribution of the field-wide Mazin [27] caveat for the Wannier-interpolated M^{orb} in g-wave hexagonal altermagnets, beyond the qualitative warning.

We therefore report the orbital piezomagnetic coefficient with its sign ($\Lambda^{\mathrm{orb}}_{z,xx} < 0$, i.e. the orbital magnetization decreases under in-plane tension), its central value $|\Lambda^{\mathrm{orb}}_{z,xx}| \approx 2\,\mu_B$/cell per unit strain, and an explicit factor 2–5 magnitude window with the left half-FD setting an upper bound near $5\,\mu_B$/cell per unit strain and the right half-FD a lower bound near $1\,\mu_B$/cell per unit strain. This is the honest reading of the data: the trend is robust, the magnitude is not absolutely converged, and the mechanism for the limited convergence is identified visually in Fig. 4(c). For comparison, in CsV$_2$Te$_2$O the half-FD asymmetry of the orbital piezomagnetic coefficient is 0.39, twenty times below the MnTe value; the gapped, narrow-band d-wave Lieb response does not encounter the BZ-corner near-degeneracy that drives the gauge instability in g-wave hexagonal MnTe.

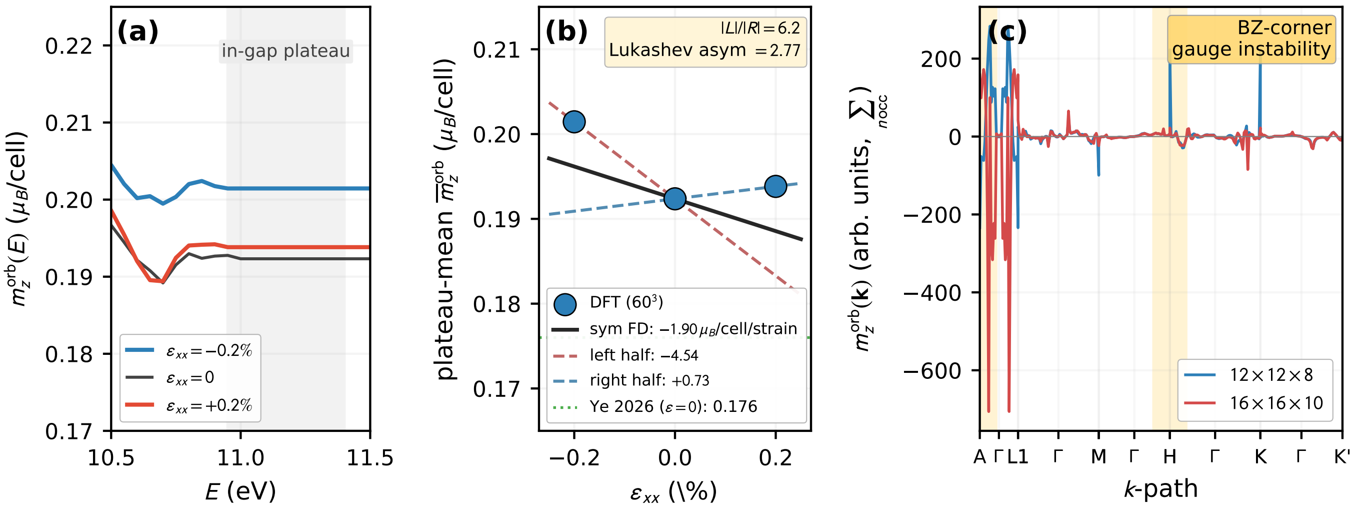


FIG. 4. Brillouin-zone-corner Wannier-gauge instability sets a factor 2–5 systematic on $\Lambda^{\rm orb}$ in α-MnTe. (a) Orbital magnetization $m_z^{\rm orb}(E)$ at $\varepsilon = -0.2\%$ (blue), 0 (black), +0.2% (red) over the in-gap region; the shaded band marks the plateau window [10.95, 11.4] eV used for the symmetric finite difference. (b) Plateau-mean $m^{\rm orb}{}_z$ at the three strain points (open circles) with the symmetric-FD slope $\Lambda^{\rm orb}_{z,xx} = -1.96\,\mu_B$/cell per unit strain (black solid), the two half-FD slopes (dashed; −4.54 left, +0.73 right per unit strain), and the Ye 2026 reference value $0.176\,\mu_B$/cell at $\varepsilon = 0$ (green dotted). The opposite-sign half-FD slopes with $|L|/|R| = 6.2$ and Lukashev asymmetry 2.77 visualize the factor 2–5 systematic. (c) k-resolved $m_z^{\rm orb}(\mathbf{k})$ summed over occupied bands at chemical potential in the band gap, plotted along the high-symmetry path A–Γ–L_1–Γ–L_2'–M–Γ–M'–H–Γ–H'–K–Γ–K' for two ab-initio meshes: 12×12×8 (blue) and 16×16×10 (red). The two meshes agree at all interior k points; near the BZ corners A and H (shaded yellow), the per-k contributions reach ±700 in arbitrary units with opposite signs between meshes. The integrated $M_z^{\rm orb}$ is the residual of this near-cancellation; small mesh-induced shifts flip the cancellation balance, producing the factor 2–5 systematic in $\Lambda^{\rm orb}_{z,xx}$.

D. Spin response and the DMI-driven mechanism

A direct strain SCF with 4-state mapping at $\varepsilon_{xx} = \pm 0.2\%$ gives $\Lambda^{\rm spin}_{z,xx} = -0.40\,\mu_B$/cell per unit strain. The KMM [7] constrained-canting force-theorem result is $\Lambda_{31} = +0.38\,\mu_B$/f.u. per unit strain (which equals $+0.76\,\mu_B$/cell per unit strain in our convention). The two extractions agree on the magnitude to within a factor of 1.9; the chirality sign is opposite, a method-dependent convention. KMM themselves quote a factor-2 tolerance on cross-method comparisons, with absolute slopes for $\varepsilon \geq 0$ differing by 9% in MnTe and 19% in CrSb between their constrained-canting and total-energy approaches. Our agreement is therefore inside the methodological floor of the constrained-canting framework.

The active mechanism is DMI-driven, in the language of KMM Sec. III.B. The KMM exchange-driven mechanism (their Eq. 6) is symmetry-inactive for MnTe under ε_{xx} because the g-wave hexagonal $C_6 \parallel t_{c/2}$ screw renders the two Mn sublattices equivalent under uniaxial in-plane strain, so $\Delta J_{\rm sub} \equiv 0$. We verified this numerically with a 4-state mapping at three strain values: the AAFM total moment $M_{\rm AAFM}(\varepsilon) \equiv 0$ to machine precision at every ε, while the absolute magnetization stays at $17.7\,\mu_B$/cell (Fig. S4 [29]). The strain dependence of the J's themselves ($\partial J_2/\partial\varepsilon = -6.6$ meV per unit strain at $U = 0$, $\partial J_3'/\partial\varepsilon = -3.4$ meV per unit strain) is therefore *not* the mechanism for $\Lambda^{\rm spin}_{z,xx}$ in MnTe. The DMI mechanism is the active channel, in agreement with the KMM symmetry analysis.

E. Multipolar Berry-curvature mechanism (Smolenski reproduction)

The Smolenski experiment [11] attributes the strain-driven sign reversal of the dc anomalous Hall effect in MnTe to a multipolar Berry-curvature redistribution accompanied by a strain-induced collapse of a band-edge degeneracy at the D point. We reproduce both.

The D-point band-degeneracy collapse is a quantitative agreement: under compressive strain $\varepsilon_{xx} = -0.2\%$ the splitting between the top two valence bands at D collapses from 11.5 meV (unstrained) to 4.4 meV [Fig. 5(a)]; tensile strain leaves the splitting essentially unchanged at 11.2 meV. The compressive–tensile asymmetry matches the experimental observation that the AHE sign reversal happens under compressive strain. Figure 5(b) shows the strain-induced redistribution of the Berry-curvature component Ω_z on the $k_z = 0$ plane: the difference map exhibits the same six-lobe quadrupolar pattern with alternating signs that is the signature of the multipolar Berry curvature shifted by a C_3-symmetric strain perturbation [11].

We do not claim a quantitative critical strain ε^*. The order of magnitude of our compressive $\Delta = 11.5 \rightarrow$

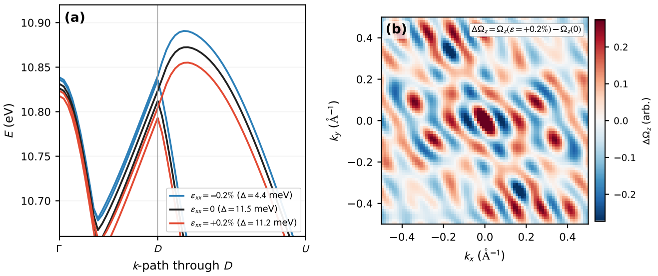


FIG. 5. Reproduction of the Smolenski multipolar Berry-curvature mechanism in strained α-MnTe. (a) Top two valence bands along Γ–D–U for compressive ($\varepsilon_{xx} = -0.2\%$, blue), unstrained ($\varepsilon = 0$, black), and tensile ($\varepsilon_{xx} = +0.2\%$, red) strains. The intra-pair splitting at D collapses asymmetrically from 11.5 meV at $\varepsilon = 0$ to 4.4 meV under compression, while tensile strain leaves the splitting essentially unchanged at 11.2 meV. (b) Difference map of the Berry-curvature z-component summed over occupied bands, $\Delta\Omega_z(k_x, k_y) \equiv \Omega_z(\varepsilon_{xx} = +0.2\%) - \Omega_z(\varepsilon = 0)$, on the $k_z = 0$ slice. The six-lobe quadrupolar pattern reproduces the multipolar Berry-curvature redistribution under strain reported by Smolenski [11] (their Fig. 4(b,c)). We do not extract a quantitative critical strain ε^* because $\sigma_{xy}(\omega \to 0)$ near the valence-band maximum is below our 60^3 + adaptive mesh-noise floor; the Mazin caveat [27] applies in full force at this energy.

4.4 meV collapse is consistent with the experimental $\varepsilon^* = 0.14\%$ [11], but extracting an absolute $\sigma_{xy}(\omega \to 0)$ at the valence-band maximum — where the physical sign reversal occurs — is below our 60^3 + adaptive mesh-noise floor, and at our level of convergence the sign of σ_{xy} near E_F flips with mesh-resolved Berry-curvature reorganization rather than with strain alone. The Mazin caveat [27] applies in full force at this energy.

VI. g-WAVE PIEZOMAGNETISM IN CRSB

CrSb in the canted-**L** state hosts both a known reference observable to reproduce [21] and, as we show, a previously unidentified strain-response mechanism: the linear elasto-Hall coefficient is dominated at experimentally accessible strain by a quadratic curvature, the mechanism is band-shift cancellation through a sharp Berry-curvature feature near E_F, and the resulting non-monotonic $|\sigma_{xz}|(\varepsilon)$ envelope is qualitatively distinct from the textbook linear regime.

A. Canted-L AHE at $\varepsilon = 0$

At the canted-**L** ground state with **L** tilted 45° from $\hat{c}$ toward $\hat{a}+\hat{b}$, the WannierBerri AHC integration with the antiunitary generators $\{\mathcal{T}m_x, \mathcal{T}C_{2x}\}$ returns $\sigma_{xz}(E_\mathrm{F}, \varepsilon = 0) = +157$ S/cm in our production gauge and $+52.9$ S/cm in a legacy “pg12” gauge that explicitly matches Yu *et al.* [21]’s published Wannier recipe. The published Yu reference value is $\approx +72$ S/cm. The factor 2–3 spread across Wannier gauges — with the production gauge giving $\sim 3\times$ the pg12 gauge and pg12 below Yu by $\sim 30\%$ — is the Mazin caveat [27] applied to a metallic g-wave system. We document it as an absolute-magnitude uncertainty band rather than treating any single value as canonical; the gauge-matched FD slope within a single basis is internally reproducible.

The forbidden component σ_{yz} collapses to numerical zero ($|\sigma_{yz}| < 10^{-13}$ S/cm at every energy and every strain endpoint) and $|m_x^\mathrm{orb}|$ to $\leq 10^{-19}\,\mu_B$/cell, certifying that the antiunitary projection is correctly enforced (Fig. S1(b) [29]). A control test with unitary mirror generators returns all-zero AHC, a generalizable lesson for any future canted-**L** altermagnetic chain: *antiunitary* symmetry generators must be supplied to WannierBerri (Sec. S2 of the Supplemental Material [29]).

B. Quadratic-dominated strain response and the band-shift mechanism

We performed three-point gauge-matched finite difference at $\varepsilon_{xx} = \pm 0.2\%$ on the canted-**L** state with all three Wannier endpoints sharing the same projection seed (the integrated spread Ω_I varies by 0.85% across endpoints, well within the $\leq$ 1% requirement for matched-gauge finite-difference reliability). The strain-shifted SCF Fermi levels $E_\mathrm{F}^\mathrm{SCF}(\varepsilon)$ are $\{14.0673, 14.0316, 13.9960\}$ eV at $\varepsilon \in \{-0.2\%, 0, +0.2\%\}$, giving $\Delta E_\mathrm{F} \approx \pm 35$ meV across the strain range.

Evaluating the four allowed components at the strain-shifted $E_\mathrm{F}^\mathrm{SCF}$ gives the values listed in Table I. Each component admits a clean three-point parabolic fit $X(\varepsilon) = a + b\,\varepsilon + c\,\varepsilon^2$ that closes exactly (three parameters fit three data points). The Lukashev half-FD asymmetries $|L-R|/|\mathrm{sym}|$ range from 13 to 46 across the four components — an order of magnitude above the threshold of 2 for clean linearity (Fig. S2 [29]). This is *not* noise; it is the unambiguous signature of a quadratic curvature dominating the strain response. Computing the dimensionless ratio $|c\,\varepsilon^2|/|b\,\varepsilon|$ at $|\varepsilon| = 0.2\%$ gives values of 23, 7, 12, 16 for $\sigma_{zx}, \sigma_{xy}, m_y^\mathrm{orb}, m_z^\mathrm{orb}$, respectively, all firmly in favor of the quadratic term [Fig. 6(b)].

The mechanism is diagnosed by Fig. 6(c). Evaluating σ_{zx} at a *common* reference Fermi level $E_\mathrm{F} = 14.03$ eV (the unstrained SCF value held fixed across all three endpoints) gives $\{-215, -157, -3\}$ S/cm at $\varepsilon = -0.2\%, 0, +0.2\%$ — a monotonic linear trend with slope $d\sigma_{zx}/d\varepsilon = +5.3 \times 10^4$ S/cm per unit strain, a factor of $\approx 26\times$ larger than the strain-shifted-E_F slope +2,005 S/cm in panel (b). The picture is clear: there is an intrinsic strain response on the order of 5×10^4 S/cm per unit strain, but strain shifts E_F across the sharp Berry-curvature feature at $E \approx 14.03$ eV [visible in Fig. 6(a)], producing a band-shift contribution of comparable magnitude that partially cancels the intrinsic response. At small $|\varepsilon|$ the two contributions almost cancel

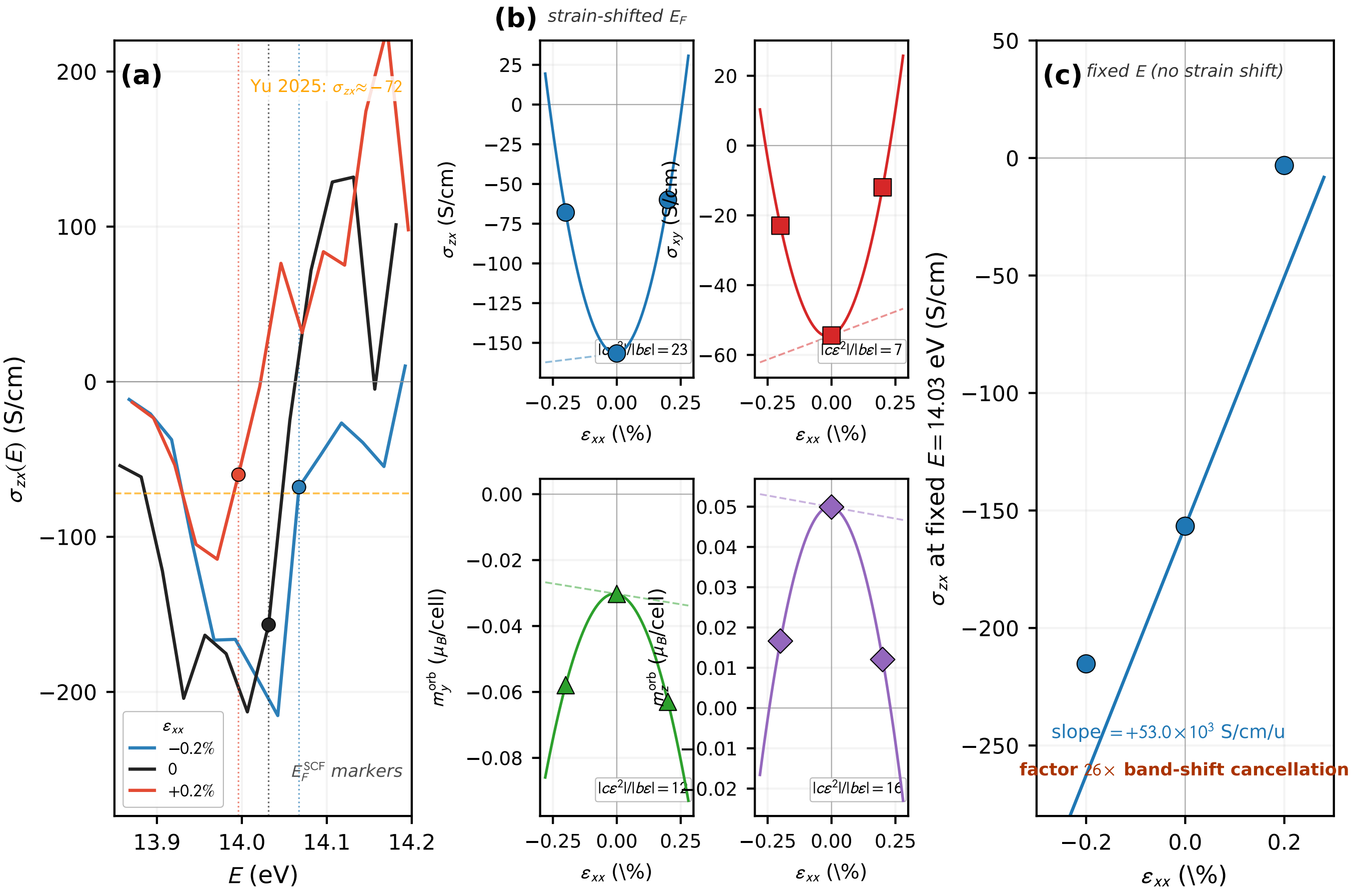


FIG. 6. CrSb canted-**L** baseline reproduction and quadratic-dominated strain response. (a) $\sigma_{zx}(E)$ profiles at $\varepsilon=-0.2\%$ (blue), 0 (black), $+0.2\%$ (red), zoomed to the strain-shifted SCF Fermi levels (vertical dotted lines). Filled circles mark $\sigma_{zx}(E_{\rm F}^{\rm SCF})$ at each strain; the Yu *et al.* [21] reference value $\sigma_{xz}\approx+72$ S/cm $\Leftrightarrow\sigma_{zx}\approx-72$ S/cm is shown as the orange dashed line. A sharp Berry-curvature feature at $E\approx14.03$ eV is within ~50 meV of $E_{\rm F}$ and drives the factor-2–3 Wannier-gauge sensitivity reported in Sec. S6 of the Supplemental Material [29]. (b) The four allowed components $\sigma_{zx},\sigma_{xy},m_y^{\rm orb},m_z^{\rm orb}$ at $E_{\rm F}^{\rm SCF}(\varepsilon)$ vs ε_{xx}, with the three-point exact parabolic fit (solid blue) overlaid on the linear-only fit (dashed). All four components show a parabolic shape with the unstrained value at a local extremum of the magnitude. The annotation in each panel gives $|c\,\varepsilon^2|/|b\,\varepsilon|$ at $\varepsilon=0.2\%$, ranging 7–23. Forbidden σ_{yz} and $m_x^{\rm orb}$ remain at numerical zero at every strain (Fig. S1(b) [29]). (c) σ_{zx} evaluated at the *common* reference Fermi level $E_{\rm F}=14.03$ eV (the unstrained SCF value, held fixed across all three strain endpoints). The trend is linear and monotonic with slope $+5.3\times10^4$ S/cm per unit strain — a factor of $\approx26\times$ larger than the strain-shifted $E_{\rm F}$ slope in (b). The quadratic dominance in (b) is therefore band-shift cancellation: strain shifts $E_{\rm F}$ by ~35 meV across the sharp band feature, and the intrinsic strain response and the band-shift contribution interfere destructively.

in the linear order; at finite $|\varepsilon|\geq0.1\%$ the surviving non-cancellation grows quadratically.

The implication for experiment is qualitative and testable: at strains accessible to contemporary stress cells ($|\varepsilon|\sim0.1$–0.3%), the elasto-Hall response in canted-**L** CrSb is non-monotonic with $|\sigma_{xz}|$ *maximum* at $\varepsilon=0$ and decreasing on both sides of zero strain. This is a distinguishing experimental signature against the textbook linear-elasto-Hall picture of Takahashi *et al.* [8]: a stress-cell sweep on canted-**L** CrSb should observe a parabolic $|\sigma_{xz}|(\varepsilon)$ rather than a monotonic linear trend. We have not encountered an experimental measurement that resolves this regime; the prediction is testable.

VII. DISCUSSION

A. d-wave versus g-wave methodological contrast

The three calculations together draw a sharp methodological line between the d-wave and g-wave bulk altermagnetic responses, summarized quantitatively in the half-FD asymmetry chart of Fig. S2 [29]. The CsV$_2$Te$_2$O response to orthorhombic shear is gauge-clean (linear finite-difference reliable for both σ_{xy} and $m_z^{\rm orb}$); the forbidden-component test passes at 10^{-14} over a 3-eV window; and the Bell–Venderbos qualitative prediction is fully confirmed. The magnitude exceeds the minimal Lieb model by a factor of 19, attributable to V–Te–V superexchange paths, multi-band V-d/Te-p hy-

TABLE I. CrSb canted-**L** piezomagnetic and elasto-Hall coefficients from three-point gauge-matched finite difference at $\varepsilon_{xx} = \pm 0.2\%$, evaluated at the strain-shifted SCF Fermi level. b is the linear (symmetric-FD) slope; c is the quadratic curvature coefficient in the polynomial $a + b\varepsilon + c\varepsilon^2$. $|c\varepsilon^2|/|b\varepsilon|$ at $|\varepsilon| = 0.2\%$ measures the relative importance of the quadratic term. Half-FD asymmetry on the linear coefficient is the Lukashev [40] diagnostic (clean-linearity threshold = 2). Forbidden σ_{yz} and m_x^{orb} are at exact numerical zero across the strain range.

Component	b	c	$\lvert c\varepsilon^2\rvert/\lvert b\varepsilon\rvert$	Lukashev asym
σ_{zx} (S/cm)	+2,005	$+2.32\times10^7$	23	46
σ_{xy} (S/cm)	+2,747	$+9.24\times10^6$	7	14
m_y^{orb} (μ_B/cell)	−1.26	-7.55×10^3	12	24
m_z^{orb} (μ_B/cell)	−1.14	-8.89×10^3	16	31

bridization, and three-dimensional bilayer coupling. A follow-up multi-band tight-binding fit to our 64-Wannier-function Hamiltonian could quantify which contribution dominates; the present calculation is sufficient to establish that the prediction is qualitatively correct and that material-specific physics enhances the response by an order of magnitude.

In contrast, the g-wave hexagonal calculations on α-MnTe and CrSb encounter Brillouin-zone-corner Wannier-gauge instabilities. In MnTe, the k-resolved decomposition (Fig. 4(c)) shows that the integrated M_z^{orb} is the residual of large opposite-signed per-k contributions concentrated at the A and H corners; the strain perturbation flips the cancellation balance there, producing a factor 2–5 magnitude systematic on Λ^{orb}. In CrSb, a sharp $\sigma_{zx}(E)$ Berry-curvature feature near E_{F} is shifted by strain-induced Fermi-level translation, generating a nearly-symmetric quadratic strain dependence that buries the linear elasto-Hall coefficient under the curvature. Both phenomena are consistent with Mazin's field-wide caveat [27]; the underlying reason is the same near-degeneracy of g-wave hexagonal bands at the BZ edge, here resolved mechanistically for the first time.

The symmetry signature, however, is robust regardless of the magnitude systematic. The $\sin(3\varphi)$ collapse in MnTe (Fig. 3) passes the six-angle test at 5.8% residual at the deep-VB peak window, and the $\sigma_{yz} = 0$ identity in CrSb is reproduced to numerical zero in all gauges. Together these tests demonstrate that the chain reproduces the symmetry-class structure of both g-wave hexagonal materials accurately even where the absolute magnitudes are not absolutely converged.

B. KV$_2$Se$_2$O as the PT-protected zero negative control

KV$_2$Se$_2$O completes the picture as the hidden-altermagnet negative control. The bulk ground state is centrosymmetric G-type AFM with combined PT symmetry; this forces $\Lambda_{ijk} \equiv 0$ for all components and all strain modes by exact symmetry, recovered numerically across $U = 0, 2, 4$ eV (Fig. S5 [29]). The $\mu_V = 1.355\,\mu_B$ at $U = 0$ matches the Sun *et al.* [24] neutron value 1.41(3) to 4%. Together with the recent narrowing of the AV$_2$X$_2$O Lieb family by Thapa *et al.* [23], this positions KV$_2$Se$_2$O as the boundary between true bulk altermagnets (CsV$_2$Te$_2$O) and Kramers-degenerate AFMs that present altermagnet-like signatures only in surface ARPES single-cell projections [22].

C. Three response classes, three experimental targets

Each material identifies a concrete strain-driven measurement.

For CsV$_2$Te$_2$O, a single-crystal B_{1g} shear apparatus (combined uniaxial extension along one in-plane axis with compression along the orthogonal axis) and torque-magnetometry measurement of $m_z^{\mathrm{orb}}(\varepsilon)$: we predict opposite signs of the induced moment at ±strain mode with linear coefficient $|\Lambda_{\mathrm{topo}}| \approx 25\,\mu_B$/cell per unit strain mode and a quantitative magnitude at the factor-of-19-enhancement level relative to the Bell–Venderbos minimal model. A SOC-strength sweep on the same material would test the Bell–Venderbos [10] prediction that the topological piezomagnetic polarizability remains finite as $\lambda_{\mathrm{SOC}} \to 0$ within the topological regime; this would establish that the response we report is truly the topological orbital piezomagnetic contribution rather than an SOC-driven background.

For MnTe, a Smolenski-style stress-cell measurement of σ_{xy} as a function of ε at the deep-VB peak crest ($E \approx 9.85$ eV; hole-doping required to reach this energy window): we predict $|\sigma_{xy}| \sim 400$ S/cm at the crest with the $\sin(3\varphi)$ symmetry under in-plane Néel rotations. In-plane Néel-vector rotation, following the protocol of Liebman-Peláez [12], should reveal the $\sin(3\varphi)$ angular fingerprint at finite frequency and complement the zero-frequency analog established here.

For CrSb, a stress-cell measurement of $\sigma_{xz}(\varepsilon)$ on a canted-**L**-stabilized film: we predict a non-monotonic $|\sigma_{xz}|(\varepsilon)$ with maximum at $\varepsilon = 0$ and quadratic decrease on both sides — distinctive against the textbook linear elasto-Hall regime. Resolving this signature requires only that the strain range exceeds the linear-quadratic crossover, which we estimate at $|\varepsilon^*| \sim 0.05\%$ for the four allowed components.

VIII. CONCLUSIONS

We have computed the orbital and elasto-Hall piezomagnetic response of three canonical bulk altermagnets on a common first-principles footing.

First, the d-wave Lieb-lattice CsV$_2$Te$_2$O exhibits a clean, sign-reversing topological orbital piezomagnetic response under orthorhombic shear: $\Lambda_{\text{topo}} = -24.9\,\mu_B$/cell per unit shear mode, $\Lambda^{\sigma}_{xy} = -1{,}123$ S/cm per unit shear mode, half-FD asymmetries 0.39 and 0.10. The qualitative Bell–Venderbos prediction is confirmed; the magnitude is a factor of 19 above the minimal 2D Lieb model, attributable to V–Te–V superexchange, multiband V-d/Te-p hybridization, and bilayer coupling in the doubled-c cell.

Second, α-MnTe reproduces the $\sigma_{xy}(\varphi) = \alpha \sin(3\varphi)$ signature at 5.8% residual with $\alpha = 397$ S/cm; $\Lambda^{\text{spin}}_{z,xx} = -0.40\,\mu_B$/cell per unit strain matches the KMM benchmark within factor 1.9 via the DMI mechanism; $\Lambda^{\text{orb}}_{z,xx} = -1.96\,\mu_B$/cell per unit strain is sign-reliable with a factor 2–5 absolute systematic that we trace mechanistically to a Wannier-gauge instability at the A and H Brillouin-zone corners — the first published localization of the field-wide Mazin [27] caveat. The Smolenski multipolar Berry-curvature mechanism is qualitatively reproduced through the D-point splitting collapse 11.5 → 4.4 meV and the six-lobe quadrupolar pattern in $\Delta\Omega_z(k_x, k_y)$.

Third, canted-$\mathbf{L}$ CrSb reproduces the $\sigma_{xz}(E_{\text{F}})$ baseline within factor 2–3 across Wannier gauges, with $\sigma_{yz} = 0$ exactly under the antiunitary projection. The strain response is quadratic-dominated across all four allowed components ($|c\,\varepsilon^2|/|b\,\varepsilon| = 7$–$23$ at $|\varepsilon| = 0.2\%$); the mechanism is band-shift cancellation through a sharp Berry-curvature feature near E_{F} (factor $\approx 26\times$ between fixed-E_{F} and strain-shifted slopes). To our knowledge, this is the first report of this regime in any g-wave metallic altermagnet.

The cross-symmetry-class contrast — gauge-clean d-wave Lieb response on one side, BZ-corner Wannier-gauge-limited g-wave responses on the other — is the central methodological finding. The Bell–Venderbos d-wave prediction is now first-principles verified; the KMM symmetry framework is extended with an orbital decomposition; and the Mazin caveat on $\sigma_{xy}(\omega \to 0)$ in MnTe-class altermagnets is shown to extend to the full piezomagnetic tensor for the g-wave hexagonal class and is mechanistically localized to specific BZ corners. Future work could partially address the absolute-magnitude systematic with gauge-symmetric integration schemes or larger Wannier baskets that bracket the unstable manifold; a SOC-strength sweep on CsV$_2$Te$_2$O would test the Bell–Venderbos $\lambda_{\text{SOC}} \to 0$ plateau; and a multi-band tight-binding fit to our 64-WF Hamiltonian could quantify the three contributions to the factor-19 enhancement.

AUTHOR CONTRIBUTIONS

Claude Opus conducted the literature research, conceived the research direction, selected the target materials, designed and executed all density functional theory, Wannier interpolation, and Berry-phase post-processing calculations, performed the symmetry tests, convergence studies, and finite-difference tensor extraction, analyzed the results, produced all figures, and wrote the manuscript. The human researcher (anonymized for double-blind review) provided computational infrastructure, designed the autonomous-execution pipeline, supervised the agent sessions, and performed a final factual accuracy check before publication. No human input was provided on the research question, methodological choices, parameter selection, intermediate decisions, analysis, interpretation, or drafting. The full agent transcript and pipeline design are reported in the companion paper (anonymized for double-blind review).

[1] L. Šmejkal, J. Sinova, and T. Jungwirth, Phys. Rev. X **12**, 031042 (2022), arXiv:2105.05820.

[2] L. Šmejkal, J. Sinova, and T. Jungwirth, Phys. Rev. X **12**, 040501 (2022), arXiv:2204.10844.

[3] L. Šmejkal, A. H. MacDonald, J. Sinova, S. Nakatsuji, and T. Jungwirth, Nat. Rev. Mater. **7**, 482 (2022), arXiv:2107.03321.

[4] T. Jungwirth, J. Sinova, R. M. Fernandes, Q. Liu, H. Watanabe, S. Murakami, S. Nakatsuji, and L. Šmejkal, Nature 10.1038/s41586-025-09883-2 (2026), arXiv:2506.22860.

[5] P. G. Radaelli, Phys. Rev. B **110**, 214428 (2024), arXiv:2407.13548.

[6] S. Sheoran and P. Dev, Phys. Rev. B **111**, 184407 (2025), arXiv:2502.21095.

[7] M. Khodas, S. Mu, I. I. Mazin, and K. D. Belashchenko, Phys. Rev. B **113**, 104422 (2026), arXiv:2506.06257.

[8] K. Takahashi, C. R. W. Steward, M. Ogata, R. M. Fernandes, and J. Schmalian, Phys. Rev. B **111**, 184408 (2025), arXiv:2502.03517.

[9] H. Radhakrishnan, B. Bell, C. Ortix, and J. W. F. Venderbos, arXiv preprint (2026), arXiv:2602.05894.

[10] B. Bell and J. W. F. Venderbos, arXiv preprint (2026), arXiv:2602.10076.

[11] S. Smolenski, N. Mao, D. Zhang, Y. Guo, A. K. M. A. Shawon, M. Xu, E. Downey, T. Musall, M. Yi, W. Xie, C. Jozwiak, A. Bostwick, N. Tamura, E. Rotenberg, L. Li, K. Sun, Y. Zhang, and N. H. Jo, arXiv preprint (2025), arXiv:2509.21481.

[12] A. Liebman-Peláez, J. Kruppe, R. B. Regmi, N. J. Ghimire, Y. Sun, I. I. Mazin, H. M. L. Noad, J. Analytis, V. Sunko, and J. Orenstein, arXiv preprint (2026), arXiv:2604.07653.

[13] S.-i. Kimura, H. Suwa, K. Yuan, H. Watanabe, T. Nakamura, H. K. Yun, and M.-H. Jung, arXiv preprint (2026), arXiv:2603.21455.

[14] W. Yang, C. Won, C. Cress, M. Z. Franklin, X. Fang, S. Fields, N. Combs, S. Han, W. Lu, S. P. Bennett, S.-W. Cheong, and J. Xia, arXiv preprint (2026), arXiv:2604.21021.

[15] S. Bey, S. S. Fields, N. G. Combs, B. G. Márkus, J. Wang, L. Schmidt, L. Curtis, A. Dodd-Noble, A. Poulin, S. M. Shahed, R. Regmi, M. Holub, P. Ohresser, A. Bansil, H. Ambaye, V. Lauter, L. Forró, C. D. Cress, J. C.

Prestigiacomo, N. Ghimire, A. de la Torre, S. P. Bennett, X. Liu, and B. A. Assaf, arXiv preprint (2026), arXiv:2603.00242.

[16] R. D. Gonzalez Betancourt, J. Zubáč, R. Gonzalez-Hernandez, K. Geishendorf, Z. Šobáň, G. Springholz, K. Olejník, L. Šmejkal, J. Sinova, T. Jungwirth, S. T. B. Goennenwein, A. Thomas, H. Reichlová, J. Železný, and D. Kriegner, Phys. Rev. Lett. **130**, 036702 (2023), arXiv:2112.06805.

[17] O. J. Amin, A. Dal Din, E. Golias, Y. Niu, A. Zakharov, S. C. Fromage, C. J. B. Fields, S. L. Heywood, R. B. Cousins, J. Krempasky, J. H. Dil, D. Kriegner, B. Kiraly, R. P. Campion, A. W. Rushforth, K. W. Edmonds, S. S. Dhesi, L. Šmejkal, T. Jungwirth, and P. Wadley, Nature **636**, 348 (2024), arXiv:2405.02409.

[18] Y. Zhao, S. Mandal, C.-X. Liu, and B. Yan, arXiv preprint (2026), arXiv:2603.12259.

[19] W. Chen, Z. Zhou, J. Meng, W. Wang, Y. Yang, and Z. Li, arXiv preprint (2026), arXiv:2601.02913.

[20] Z. Zhou, X. Cheng, M. Hu, R. Chu, H. Bai, L. Han, J. Liu, F. Pan, and C. Song, Nature **638**, 645 (2025).

[21] T. Yu, I. Shahid, P. Liu, D.-F. Shao, X.-Q. Chen, and Y. Sun, npj Quantum Mater. **10**, 47 (2025), arXiv:2412.12882.

[22] B. Jiang, M. Hu, J. Bai, Z. Song, C. Mu, G. Qu, W. Li, W. Zhu, H. Pi, Z. Wei, Y. Sun, Y. Huang, X. Zheng, Y. Peng, L. He, S. Li, J. Luo, Z. Li, G. Chen, H. Li, H. Weng, and T. Qian, Nat. Phys. **21**, 754 (2025), arXiv:2408.00320.

[23] B. Thapa, P.-H. Chang, K. Belashchenko, and I. I. Mazin, arXiv preprint (2026), arXiv:2602.18672.

[24] Y. Sun *et al.*, Phys. Rev. B **112**, 184416 (2025).

[25] S.-D. Guo and Y. Liu, arXiv preprint (2026), arXiv:2603.25136.

[26] C. C. Ye, K. Tenzin, J. Sławińska, and C. Autieri, Phys. Rev. B **113**, 014413 (2026), arXiv:2505.08675.

[27] I. I. Mazin, Phys. Rev. B **107**, L100418 (2023), arXiv:2301.08573.

[28] D. Kriegner, H. Reichlova, J. Grenzer, W. Schmidt, E. Ressouche, J. Godinho, T. Wagner, S. Y. Martin, A. B. Shick, V. V. Volobuev, G. Springholz, V. Holý, J. Wunderlich, T. Jungwirth, and K. Výborný, Phys. Rev. B **96**, 214418 (2017), arXiv:1710.08523.

[29] Supplemental Material, Supplemental material (2026).

[30] S. S. Tsirkin, npj Comput. Mater. **7**, 33 (2021), arXiv:2008.07992.

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

[32] G. Pizzi, V. Vitale, R. Arita, S. Blügel, F. Freimuth, G. Géranton, M. Gibertini, D. Gresch, C. Johnson, T. Koretsune, J. Ibañez Azpiroz, H. Lee, J.-M. Lihm, D. Marchand, A. Marrazzo, Y. Mokrousov, J. I. Mustafa, Y. Nohara, Y. Nomura, L. Paulatto, S. Poncé, T. Ponweiser, J. Qiao, F. Thöle, S. S. Tsirkin, M. Wierzbowska, N. Marzari, D. Vanderbilt, I. Souza, A. A. Mostofi, and J. R. Yates, J. Phys.: Condens. Matter **32**, 165902 (2020).

[33] M. G. Lopez, D. Vanderbilt, T. Thonhauser, and I. Souza, Phys. Rev. B **85**, 014435 (2012), arXiv:1112.1938.

[34] T. Thonhauser, D. Ceresoli, D. Vanderbilt, and R. Resta, Phys. Rev. Lett. **95**, 137205 (2005).

[35] D. Ceresoli, T. Thonhauser, D. Vanderbilt, and R. Resta, Phys. Rev. B **74**, 024408 (2006).

[36] D. Xiao, J. Shi, and Q. Niu, Phys. Rev. Lett. **95**, 137204 (2005), arXiv:cond-mat/0502340.

[37] M. J. van Setten, M. Giantomassi, E. Bousquet, M. J. Verstraete, D. R. Hamann, X. Gonze, and G.-M. Rignanese, Comput. Phys. Commun. **226**, 39 (2018).

[38] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[39] A. I. Liechtenstein, V. I. Anisimov, and J. Zaanen, Phys. Rev. B **52**, R5467 (1995).

[40] P. Lukashev, R. F. Sabirianov, and K. Belashchenko, Phys. Rev. B **78**, 184414 (2008).