Machine-Checked Triple Correctness of Scientific Simulations

Xinrun Wang[1], Chang Yang[2],
[1]Singapore Management University, [2]The Hong Kong Polytechnic University
xrwang@smu.edu.sg

When a simulated number disagrees with experiment, one cannot tell today whether the model, the numerics or the code is at fault. We introduce *machine-checked triple correctness*: mathematical and software correctness are proved, with the proofs checked by machine; physical correctness, which no proof can establish, is tested against observation under protocols and data frozen in advance, and the comparison is recomputed by machine. The method, *verify the checker, not the solver*, comes from certifying algorithms and proof checking: untrusted solvers emit certificates that Lean programs, proved sound, re-check in exact arithmetic; what is new is its reach to nonlinear self-consistent problems of materials science at near-production settings. In our realisation, FORMAL IN SILICO, 668 of 859 recorded claims are machine-proved, 305 of them down to the executed code. For silicon, the band gap of a discretised Kohn–Sham LDA model at pre-registered settings is bounded over the whole Brillouin zone to $[0.4146, 0.4856]$ eV, 0.741–0.847 eV below experiment, so that the deviation contains no algebraic, rounding or software error. Certified answers also serve as exact references for other programs and for published data. Given certified discrete problems under protocols frozen in advance, ABINIT, Quantum ESPRESSO and DFTK at default or tutorial settings returned numbers outside the certified intervals, about 10^{-9} Ha wide, in all 33 runs that gave numbers, 29 of them without a warning; in a post hoc reading, 11 of these runs solved a different discrete problem, and the other 22 solved the certified one, within the code's own tolerance if an energy criterion t is taken to allow eigenvalue errors of order $\sqrt{t}$ (24 of the 33 under this reading, 11 if every criterion is read literally); on silicon and aluminium phosphide no quantity a code reports as converged changes at thresholds fixed in advance. Of 47 comparable binary CALPHAD database files (32 papers), 21 (16 papers) define a model proven not to reproduce, beyond the printed digits, at least one invariant reaction of the table printed in their own paper, a test of reproducibility, not of accuracy; most of the 34 such differences with a closed certified interval are below 1 K, and post hoc, 28 of them lie within what rounding of the printed parameters to their last digit can produce, while 6, in 4 databases, exceed a linearised estimate of that range. In 21 of 63 databases (7 on a uniform grid, 16 of 52 databases at points near invariant reactions), pycalphad at its default settings returns an equilibrium proven not to be the global equilibrium of the defined model. No gradient-corrected functional is certified in three dimensions.

1 Introduction

Computer simulation has become an instrument of science and engineering in its own right. Density functional theory (DFT) predicts the stability of materials that have never been synthesised; molecular dynamics (MD) samples the conformations of drug candidates; the finite element method (FEM) certifies that bridges, aircraft and offshore platforms can carry their loads; computational thermodynamics (CALPHAD) predicts which phases an alloy will form. In each case a decision is taken on the strength of a number printed by a program.

That number is the end of a long chain. It starts from the real world, which a physical model is meant to describe; the model is posed as a problem in a function space, discretised, solved by an iterative algorithm, executed by a parallel program and rounded into IEEE 754 arith-

metic. Each link can fail independently: an uncontrolled discretisation sank the Sleipner A platform (Jakobsen and Rosendahl, 1994), 15 solid-state codes agree on equations of state only with tightly converged numerical settings and recent pseudopotential libraries (Lejaeghere et al., 2016), a parallel MD code that mishandles a subdomain boundary can lose an atom, an error of protocol rather than physics, and the LDA band-gap error persists however accurate the numerics (Perdew and Levy, 1983; Sham and Schlüter, 1983) (Appendix C). Today the chain is vouched for almost exclusively by comparison, with analytic solutions, with experiment and with other codes. Comparisons are indispensable, but they cover only what was compared, and when a result disagrees with experiment they rarely say which link failed.

FORMAL IN SILICO starts from the thesis that a simulated result is trustworthy only if it is correct in three senses at once, which we call *triple correctness*. *Physical correctness*: the model (equations, functional, force field, constitutive law, thermodynamic database) describes the real world in a stated range. *Mathematical correctness*: what is solved is the model's equations; the solution exists, is unique where claimed, and the discretisation and iteration errors have rigorous bounds. *Software correctness*: the program faithfully executes the algorithm, its parallel protocols lose and duplicate nothing, and its floating-point error is bounded. A result is only as trustworthy as the weakest of the three.

We require the evidence for each of the three to be checked by machine as far as its nature allows (Table 1); the name *machine-checked triple correctness* refers to this, not to a proof of physics. Mathematical correctness is a set of theorems, proved in Lean 4 and checked by its kernel. Software correctness is established for small codes directly, by Frama-C proofs of C code and TLA$^+$ model checking of parallel protocols on small finite instances, and for large solvers by a checker whose soundness is a Lean theorem about the very function that is compiled and run. Physical correctness cannot be proved or machine-checked, only tested against observation; we test it under protocols registered in advance, and what a machine checks is the comparison: that the protocol and the hashed data set are unchanged since their timestamped registration, and that the deviation interval follows from the certified enclosure and the measurement in exact arithmetic.

The verification-and-validation literature (Oberkampf and Roy, 2010) already separates code verification, solution verification and validation, a three-way split that triple correctness inherits; but there the first two rest on testing and convergence studies, so the numerical and code errors of each comparison with experiment are estimated, not bounded. Formal proofs have covered both the mathematics of a numerical scheme and the C program that implements it (Boldo et al., 2013), without a comparison with experiment. AI-generated proofs now reach compiled production libraries (Lucas et al., 2026), but, as their authors stress, a proof checker cannot tell whether a specification matches reality; for a simulation that question is physical correctness, and only observation can answer it. To our knowledge FORMAL IN SILICO is the first programme in which, for the same computation, the mathematics and the software are machine-checked and the comparison with experiment is pre-registered and recomputed by machine, all recorded in one ledger that states for every claim how far the evidence reaches.

The three belong together because of what each gains from the others. Today a discrepancy between a computed and a measured quantity mixes three error sources: the model, the numerics and the code. If the numerical and software errors are enclosed by a machine-checked certificate, the discrepancy, minus that known interval, is the error of the certified model plus measurement error only, where the certified model includes any discretisation that the certificate does not bound; that error, which validation studies today can only estimate after assuming that the numerics have converged, becomes rigorously bounded up to the measurement uncertainty (§2.2). The same certificates, used as references, make the combined numerical and software error of other programs, and the agreement of published data with the models they define, measurable from outside against exact references. The attempt is feasible now because Lean 4 (de Moura and Ullrich, 2021) and Mathlib (The mathlib Community, 2020) provide one library of real and functional analysis, linear algebra and spectral theory; because certifying computation (McConnell et al., 2011; Hales et al., 2017) lets one verify a small checker instead of a large solver; and because AI-assisted

theorem proving, most mature for Lean, can carry a mathematical layer of many thousands of theorems.

Contributions.

1. **Machine-checked triple correctness** (§2): a chain of seven layers with an explicit obligation at each transition, and an observation-error decomposition that turns certified numerics into a rigorous bound on the error of the certified model. The mathematical and software obligations are discharged by machine (verify the checker, not the solver, with fingerprinted soundness statements and an axiom audit; a software layer model-checked on finite instances and proved at code level for a small kernel), the physical one by pre-registered validation recomputed on every full check; and a ledger of trust levels (T0–T4) for what is proved and validation grades (V0–V3) for what is observed.
2. **A certified deviation from experiment** (§3.1): for the LDA band gaps of silicon, aluminium phosphide and gallium phosphide, the whole-zone gap of a discretised Kohn–Sham model at pre-registered settings is bounded from both sides, so the deviation from experiment contains no algebraic, rounding or software error; as registered, it is not called a model error: the discretisation is part of what is certified.
3. **An audit instrument** (§3.2): under frozen protocols, certified answers locate numbers that ABINIT, Quantum ESPRESSO and DFTK return outside certified intervals without a warning (post hoc, by cause: a different discrete problem or the code's own tolerance under one reading), proven disagreements between published CALPHAD database files and the tables printed in their own papers and, in 21 of 63 databases, equilibria from default-setting pycalphad proved not to be global.
4. **Proofs that floating-point software cannot give** (§3.3), such as that a converged SCF solution is not the ground state, across plane-wave DFT, FEM, MD, CALPHAD and structure search, with 859 ledger claims.

§4 states the limits; the appendices give the full programme, every domain and the evaluation (guide in Appendix A).

2 Machine-Checked Triple Correctness

This section defines the obligations of triple correctness and the levels in which progress on them is recorded (§2.1–§2.3), how each is discharged by machine (§2.4–§2.7), and how certified answers are used for audits (§2.8); Table 1 summarises what is checked, by which machine, and what remains trusted.

2.1 Seven Layers, Six Obligations

We split the path from the real world to a floating-point number into seven layers (Fig. 1; Table 3 in Appendix D): the real world (L0), the physical model (L1), the continuous problem (L2), the discrete problem (L3), the abstract algorithm (L4), the parallel program (L5) and the floating-point execution (L6). Between each pair sits an explicit obligation: validation against observation, which is physical and can only be discharged by observation, never by proof; well-posedness, a discretisation error estimate and algorithmic convergence, which are mathematical; and refinement of the abstract algorithm by the parallel program and a bound on rounding, which belong to software. For a continuous solution u, its discrete approximation u_h, the algebraic approximation $u_h^{(k)}$ after k iterations and the computed floating-point value $\tilde{u}_h^{(k)}$,

$$\|u - \tilde{u}_h^{(k)}\| \;\le\; \underbrace{\|u - u_h\|}_{\text{discretisation (PDE theory)}} + \underbrace{\|u_h - u_h^{(k)}\|}_{\text{algebraic (algorithm theory)}} + \underbrace{\|u_h^{(k)} - \tilde{u}_h^{(k)}\|}_{\text{rounding (floating-point analysis)}} . \tag{1}$$

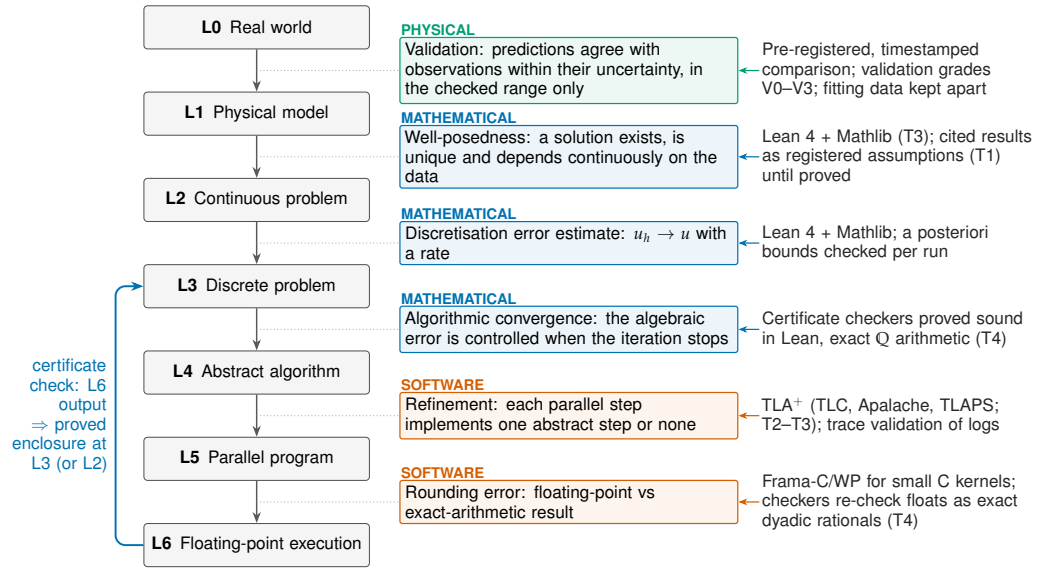


Figure 1: **Triple correctness on seven layers.** A simulation passes from the observed world (L0) to the number computed on IEEE 754 hardware (L6); each transition carries one obligation (boxes), colour-coded by the kind of correctness it establishes: physical (green, discharged only by observation), mathematical (blue, machine proof) or software (vermillion, specification, model checking and code-level proof). The right column names the tool that discharges each obligation. For the large solvers the chain L3–L6 is not proved step by step: a checker proved sound in Lean re-checks the solver's floating-point output in exact rational arithmetic and returns an enclosure of the discrete (L3) or, in some cases, the continuum (L2) quantity (blue arrow on the left; Fig. 2). Bottom: the observation-error decomposition; once the numerical and software term is enclosed, the difference between experiment and the certified interval is model error plus measurement error, the first bounded rigorously, the second known only statistically and reported separately.

Each term must be backed by a theorem, and the inequality has a premise that is easy to overlook: the program really computes $\tilde{u}_h^{(k)}$, that is, the parallel program faithfully executes the abstract algorithm, which the software layer must guarantee. Physical correctness adds one more term, outside (1). Instantiated for each domain, the obligations form a matrix that is the programme's work list (Table 4 in Appendix D).

2.2 The Observation Error Decomposition

Let y be an observable quantity (a band gap, a lattice constant, a phase-boundary temperature), y_{obs} its measured value, y_{true} its true value in the real world, $y(u)$ its value in the exact solution u of the model, and $\tilde{y}$ the number the program reports. Then

$$y_{\text{obs}} - \tilde{y} = \underbrace{(y_{\text{obs}} - y_{\text{true}})}_{\text{measurement (statistical)}} + \underbrace{(y_{\text{true}} - y(u))}_{\text{model (physical)}} + \underbrace{(y(u) - \tilde{y})}_{\text{numerical and software (certified)}} . \quad (2)$$

In a comparison with experiment today, the three terms are mixed: a deviation can come from the model, or from the cutoff, the k-point grid, SCF convergence, the pseudopotential or the code. When a certificate proves $y(u) \in [\text{lo}, \text{hi}]$, the last term is enclosed rigorously,

Table 1: **What is machine-checked.** For each correctness, the obligation (layers of Fig. 1), what a machine checks on every full run of `make verify-all` in strict mode, and what is trusted or open (trust levels T0–T4: §2.3). The trusted base is listed in full in Appendix T.

Correctness	Obligation	Checked by machine	Trusted or open
Physical	L0→L1: the model agrees with observation	the registered part of each claim, and its protocol text outside a header block, are unchanged since the registration commit, an ancestor of the checked head; the data set matches its SHA-256; recorded dates are in order; every deviation interval is recomputed in exact arithmetic from the hashed data set and the checker outputs	that the registration preceded the computation (commit timestamps in our own repository) and blinding; the measurement and its corrections; the model itself, which is only confirmed or refuted
Mathematical	L1→L4: well-posedness, discretisation, convergence	every theorem by the Lean kernel when its module is built; the statements of all checker soundness theorems against registered fingerprints; every constant against Lean's three standard axioms	registered assumptions, each listed in the ledger; that the Lean definitions say what is intended
Software	L4→L6: refinement, rounding	the compiled checker is the proved function (T4); parallel protocols by TLA$^+$ model checking on finite instances and trace validation; the C interval kernel by Frama-C/WP; every certificate test by a tampered copy or a wrong problem that must be rejected	the Lean kernel, compiler and runtime, parsers and printer, operating system and hardware; TLC (the TLA$^+$ model checker) and Frama-C for the software-layer claims only

and

$$y_{\text{true}} - y(u) \;\in\; [\,y_{\text{obs}} - \text{hi},\; y_{\text{obs}} - \text{lo}\,] \;-\; (y_{\text{obs}} - y_{\text{true}}), \tag{3}$$

a rigorous interval shifted by the measurement error, which is known only statistically, with uncertainty σ_{exp}. What remains of the discrepancy is the error of the certified object and measurement error only. When that object is the model fixed at L1 (the equations, the functional or force field, the pseudopotential, and the identification of the computed quantity with the measured one, for example a Kohn–Sham gap standing in for an optical gap), model error is then bounded rigorously, not estimated; when it is a discretisation of the model, as for three-dimensional DFT (§3.1), the bound is on the deviation of that discrete model, discretisation included, which we do not call model error. This is the point of putting the three correctnesses into one framework, and what distinguishes the programme from pure verification and from pure validation. The two kinds of uncertainty differ in nature, one a proved enclosure and the other a statistical statement about measurements, and reports keep them apart.

2.3 Trust Levels, Validation Grades and the Ledger

Not everything can be proved at once, so every claim states what it has reached (details in Appendices D.2 and E). Trust levels measure mathematical and software correctness: T0 tested, T1 paper proof, T2 model-checked on finite instances, T3 machine proof, T4 machine proof down to executable code and floating-point rounding (Table 5). A theorem may first enter as a registered assumption with a literature reference (T1) and is later replaced by a proof. Physical correctness cannot be proved and is measured on a separate axis, validation grades: V0 not validated, V1 a single comparison with a sourced measurement and its uncertainty, V2 a systematic validation on a protocol fixed in advance, V3 a blind prediction archived before the measurement (Table 6). A conclusion carries both, for example "band-gap interval T4, comparison with experiment V1"; the decomposition (3) needs a high T level for the computed interval and at least V1 for the comparison. Every claim is registered in the *trust ledger* with a project-unique identifier (e.g. DFT-B08), its layer, its current and

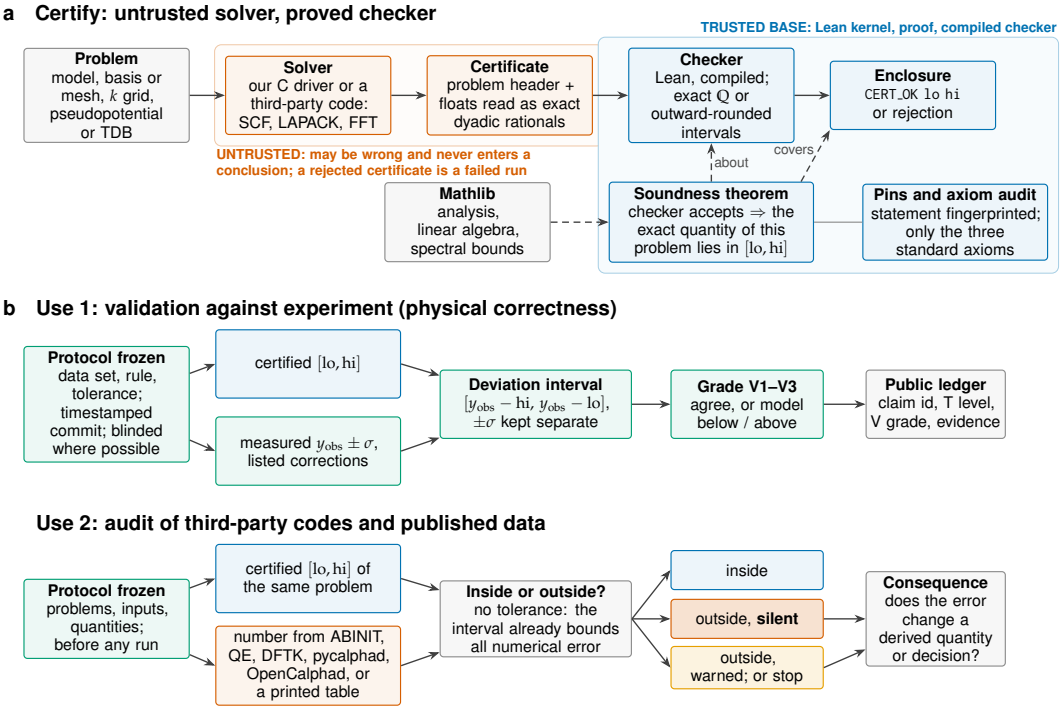


Figure 2: **Verify the checker, not the solver. a,** The solver (our C drivers or a third-party code) runs in floating point and is not trusted; it writes a certificate whose floating-point numbers the checker reads as exact dyadic rationals. The checker, a compiled Lean program, re-checks it in exact rational (or outward-rounded interval) arithmetic and either rejects it or prints an enclosure; its soundness theorem, built on Mathlib, states that acceptance implies that the exact quantity of the stated discrete (or continuum) problem lies in the printed interval, and is pinned and axiom-audited (§2.5). The trusted base is the Lean kernel, compiler and runtime, the certificate parser and the decimal printer, not the solver. **b,** Two uses of an enclosure. Validation: a protocol timestamped in advance turns it and a corrected measurement into a deviation interval, graded V1–V3, with the measurement uncertainty kept separate. Audit: a frozen protocol compares a third-party number with the certified interval of the same problem without tolerance, or a printed table at its printed precision, and classifies outside results as silent or warned; a further protocol asks whether a deviation changes a derived quantity or decision (Figs. 4 and 5).

target level, its evidence and its dependencies on other claims and on registered assumptions (A-DFT-01, . . .). One rule carries the programme: *every output has a theorem*; a number a program prints without a machine-checked guarantee marks an unfinished feature. Every program run prints a trust report whose lines cite ledger identifiers (Appendix U). The ledger now holds 859 mathematical and software claims, of which 668 (78%) are machine-proved and 305 at T4, and 24 validation claims: 14 at V2, four at V1 and six at V0 (Table 7). The T4 claims are of three kinds: 255 are theorems about checker functions that are compiled and run (soundness, problem binding, equality of fast and reference implementations; seven of them also state a public benchmark instance); 47 record the conclusions of particular certificate runs or archived sets of certificates, each accepted by the compiled checker (such as K29, W66 and R18); and 3 are Frama-C proofs about the C kernel, two of them under the registered assumption A-DFT-06, on which no checker output depends.

2.4 Verify the Checker, Not the Solver

Verifying a production solver line by line, as Boldo et al. (2013) did for a wave-equation program, is a landmark, but it does not scale to codes that call LAPACK, run SCF iterations with heuristic mixing and reduce across threads. We follow the tradition of certifying

algorithms (McConnell et al., 2011) and proof checking instead (Fig. 2), in which checkers have been formally verified before, for graph algorithms (Alkassar et al., 2014), SAT proofs (Tan et al., 2021), the inequalities of the Kepler conjecture (Hales et al., 2017) and interval bounds (Melquiond, 2008); what is new here is the target class, nonlinear self-consistent problems of materials science at near-production settings, the devices that make exact checking feasible there (below), and the use of the certificates to audit widely used codes and data (Appendix R). An untrusted driver (one of our C or Python programs, or a third-party code) computes a candidate and writes a *certificate*: the problem data and approximate solution data, with floating-point numbers entering only as exact dyadic rationals. The checker is an ordinary Lean function `check : Cert -> Option Output`, and its soundness theorem states that if `check c = some o` then the mathematical object defined by the data in `c` lies within the bounds in `o`. All arithmetic is in $\mathbb{Q}$; transcendental functions appear only through proven rational enclosures, and decimal output rounds lower bounds down and upper bounds up. The checker is compiled to a native executable, and since the soundness theorem is proved about the very function that is compiled, it is recorded at T4, with the Lean compiler, the certificate parser and the decimal printer in the trusted base (Appendix T); it is a proof about the function that is compiled, not a verified compilation. A wrong driver can only make the check fail or the interval wider; the solver can be replaced without touching a proof.

Crucially, the checker can target a *stronger* statement than the solver's own correctness. In all five domains it bounds the distance between the computed output and a mathematically defined target that the solver never computes exactly: the infimum of the discrete Kohn–Sham energy and its infinite-cutoff limit, the unique self-consistent solution in a ball (by a Newton–Kantorovich argument), the classical solution of a boundary-value problem, the true trajectory of Newton's equations, the global minimum of a Gibbs energy, an exact local minimum of a cluster energy. Where the target is a continuum object (one-dimensional DFT, FEM, MD), the three terms of (1) are discharged together, by one certified enclosure; for three-dimensional DFT the target is the discrete problem, so the enclosure discharges the algebraic and rounding terms and the discretisation term remains open (§4). Exact checking is slow for large problems, and three devices make it reach production settings in three dimensions: splitting the Newton–Kantorovich Jacobian into an exactly checked low-frequency block and a proved bound on the high-frequency tail; proved Bloch symmetry (time reversal and point groups, with conclusions about the full k mesh); and arithmetic modulo machine-word primes, proved in Lean to compute the same numbers as big integers. With them the silicon certificate of §3.1, 253 MB, checks in 196 CPU hours (Appendix K).

2.5 What a Soundness Theorem Guarantees

A proof that passes shows only that *some* statement was proved, about *some* problem. Four mechanisms, all re-run by machine, make it the intended statement about the intended problem. *Pinned statements*: the statements of all 515 checker soundness theorems, with the transitive closure of the project definitions they use, and the ACSL contracts (the C specification language) of the trusted C kernel are fingerprinted, and the full check fails on any change to a pinned statement, whether an added hypothesis, a looser bound or a changed definition, unless the statement is re-registered explicitly, which shows in the commit diff. *Axiom audit*: every constant, 36,091 in the library and those of the entry modules of all eleven checkers, may depend only on Lean's three standard axioms; any `sorry` (Lean's placeholder for a missing proof) or project axiom, any use of `native_decide`, and any compiled function replaced by other code, is reported. *Problem binding*: a checker certifies the problem written in its certificate, so the largest pieces of problem data are bound by the checker itself: the one-dimensional DFT checker compares the external potential with a Lean definition of the potential named on the command line, the CALPHAD checker builds its model from the bytes of the TDB file, and for S1 the registered problem is a Lean definition stated with Mathlib alone, whose pseudopotential coefficients are proved to be what a total parser reads from the bytes of the hashed file; the S1 checker mode decides that the certificate states exactly this problem and checks the SCF and zone certificates in one process, so its soundness theorem has no premise about another run (DFT-K30–K36, T3; Appendix K). It has bound the archived silicon certificates, the two-sided result by a composition theorem over separate runs whose glue lies outside Lean (K38, T4); the other

checkers' remaining parameters are echoed exactly and compared with the command line (T0). *Adversarial tests*: every certificate test case needs a tampered copy or a wrong problem that is caught (255 in all); 162 carry known answers computed independently of the driver, which catch a driver that consistently solves the wrong problem; and adversarial reviews of the checkers left a regression harness of 686 forged certificates. One command, `make verify-all`, re-checks all of this: it brings the 776 Lean modules up to date, re-checks the test and showcase certificates, recomputes every audit verdict from the archived checker outputs and re-checks a fixed sample of the audit certificates, in 99 minutes with 32 parallel jobs on one workstation in a *cold* run, which reads no result cache but reuses the Lean build output; a build from scratch takes longer (Appendices I and P). The three-dimensional production certificates (0.17–2.1 GB, up to about 1 800 CPU hours per check) were checked once or twice; full runs verify their archived checker outputs by SHA-256 and recompute every verdict from them.

2.6 The Software Layer

Verifying the checker makes most driver bugs harmless, and the rest of the software layer covers what remains. Concurrent control is specified in TLA$^+$ (Lamport, 2002): the SCF controller and the shared-memory k-point protocol are model-checked; the one-dimensional DFT driver logs its controller state and thread events, trace validation (Cirstea et al., 2024) replays the log against the specification (two runs on every full check), and a benchmark test (T0) finds its output bitwise identical for 1, 2, 4 and 7 threads. For showcase S3, an MPI driver of the three-dimensional point-group LDA problem distributes k-points, bands and FFT planes as the model-checked protocol `PWParallel` prescribes; its logs pass trace validation against that protocol, and its output, certificate included, is bitwise identical in all 25 tested layouts with between 1 and 32 processes (T0; Appendix I.2). The other three-dimensional drivers are covered by the checker alone. The small C kernel that computes residual bounds and inertia counts in outward-rounded interval arithmetic is proved with Frama-C/WP (Kirchner et al., 2015) on the C code, both free of run-time errors and correct under an IEEE 754 model (DFT-F03, F01, F02, T4), under two registered assumptions: that WP's floating-point model and the C-library contracts describe the compiled program (A-DFT-06), and that the ACSL statements say what the Lean ones say (A-DFT-07). The drivers are fuzzed under sanitizers; the fuzzer found inputs for which a driver wrote a certificate for a different problem, such as `3*T**1.5` read as $3T + 0.5$ in a TDB file; for the largest pieces of problem data, problem binding now removes this class of error (Appendix I.2).

2.7 Pre-Registered Validation, Recomputed by Machine

A certified interval removes the numerical part of a comparison with experiment; pre-registration removes the freedom to choose the comparison after seeing the number. A protocol is committed to the project's repository before the certified quantity exists: the quantities, the data set (frozen and hashed, each value with source, conditions, uncertainty and corrections), the separation of fitting from validation data, the rule for which computed interval may be used, and the agreement criterion. The machine then checks the comparison, not the physics: the registered part of a claim and its protocol must be unchanged since their registration commit, the data set must match its hash, and `make verify-all` recomputes every deviation interval in exact arithmetic from the archived checker outputs and the hashed data set. Two rules complete the protocol: a disagreement is a result and not a reason to re-run, and a stronger rule that becomes available later is recorded as a supplement next to the first verdict.

2.8 Certificates as an Audit Instrument

A certified answer to a precisely stated discrete problem is a reference against which any other program, or any published table, can be judged with no tolerance beyond the precision the table states. We use it in three steps. First, a protocol is fixed before any run: the problems, the programs and versions, the inputs (aligned with the certified problem, at default settings, or as in official tutorials), and the decision rules. Second, every number is judged *inside* or

outside the certified interval, and outside results are classified by what the program told its user: *warned* (an error, a warning or a non-zero exit) or *silent* (convergence reported, exit code 0, no warning). Third, a consequence protocol, also frozen before computation, asks whether a deviation changes a quantity or decision a user would derive from it, with thresholds fixed in advance. For databases the reference is internal: a database file against the calculated table printed in its own paper, and a solver's equilibrium against the certified global equilibrium. Any choice made after the results is labelled post hoc and reported with its alternatives. Verdicts say that a program or file disagrees with a certified answer, not why; mechanisms are established, where at all, by separate diagnostic runs.

3 Results

The realisation comprises about 219k lines of Lean (some 9.1k theorems), about 14.8k lines of C (the Python drivers aside), four model-checked TLA$^+$ specifications, three of them with trace validation, and a Frama-C-verified interval kernel, across one- to three-dimensional DFT, one- and two-dimensional FEM, one- and three-dimensional MD, binary CALPHAD and Lennard-Jones structure search (Appendices A–O). Its drivers agree with DFTK (Herbst et al., 2021), scikit-fem (Gustafsson and McBain, 2020), ASE (Larsen et al., 2017) and pycalphad (Otis and Liu, 2017a) on the cross-check problems of Appendix P, whose values lie inside the certified intervals. The three subsections report a certified deviation from experiment, audits of other programs and of published data, and what is proved; every interval attributed to our checkers is a machine-checked enclosure, while numbers from other programs, and their distances from the enclosures, are floating point (T0). Table 2 in Appendix B lists all results with their ledger identifiers.

3.1 Physical Correctness: A Certified Deviation from Experiment

The Kohn–Sham LDA band gaps of silicon, carbon and germanium were pre-registered on 10 October 2026, before any certified LDA gap of a real material existed (O04–O06): Slater exchange and PW92 correlation, GTH-LDA q4 pseudopotentials with file hashes, a static lattice, minimum cutoffs and k meshes, the corrections to the measured gaps, and three rules (two-sided, discrete two-sided, one-sided) for which computed interval may be used. For silicon at the protocol's settings (lattice constant 5.416 Å, cutoff about 10 Ha, $8 \times 8 \times 8$ mesh), the checker proves that the discrete zero-temperature KS-LDA equations have exactly one solution within $2.1 \cdot 10^{-10}$ of the certificate's centre (under the chemical-potential convention of Appendix K) and encloses all 8192 band energies in intervals at most $8.6 \cdot 10^{-9}$ Ha wide (K29); this is local uniqueness, not a proof that the solution is the ground state of the discrete model. Octrees over the Brillouin zone, with a Rayleigh–Ritz test for the valence band and a counting theorem of Lehmann type for the conduction band, turn this into a bound on the gap over the whole zone: at most 0.530902 eV under the one-sided rule (W51), and $[0.4146, 0.4856]$ eV under the two-sided one (939 CPU hours, W52–W66). DFTK 0.8, given the identical discrete problem, lies inside the certified gap over the $8 \times 8 \times 8$ mesh (K29, W217). Against the frozen data set, the comparable observation, the low-temperature gap with zero-point renormalisation and spin–orbit splitting removed, is 1.2437 ± 0.017 eV, so the discrete model lies $[0.741, 0.847]$ eV below the observation, measurement uncertainty included (grade V1; the first, one-sided verdict, at least 0.695 eV, stays in the record; Fig. 3). That LDA underestimates band gaps is textbook knowledge; what is new is that the model side is a proved bound, so no part of the deviation is algebraic, rounding or software error. Two further materials were registered later under rules of the same form, with the data entries blinded until their certified gaps existed. Aluminium phosphide (O20; lattice constant taken at 300 K, without the zero-point correction used for silicon, as the protocol fixed; about 10 meV per 0.01 Å in floating point) has a certified gap in $[1.4071, 1.4781]$ eV under the two-sided rule (705 CPU hours, W300), $[1.075, 1.293]$ eV below the observed 2.6263 ± 0.073 eV (V1; the first, one-sided verdict, at least 1.075 eV from the bound 1.478051 eV of W96, stays in the record). Gallium phosphide (O24), with the full GTH kernel for Ga, has a certified gap in $[1.4668, 1.5378]$ eV (1 842 CPU hours, W192), $[0.873, 1.045]$ eV below the observed 2.4617 ± 0.05 eV (V1; the one-sided verdict, at least 0.825 eV from W191, stays in the record).

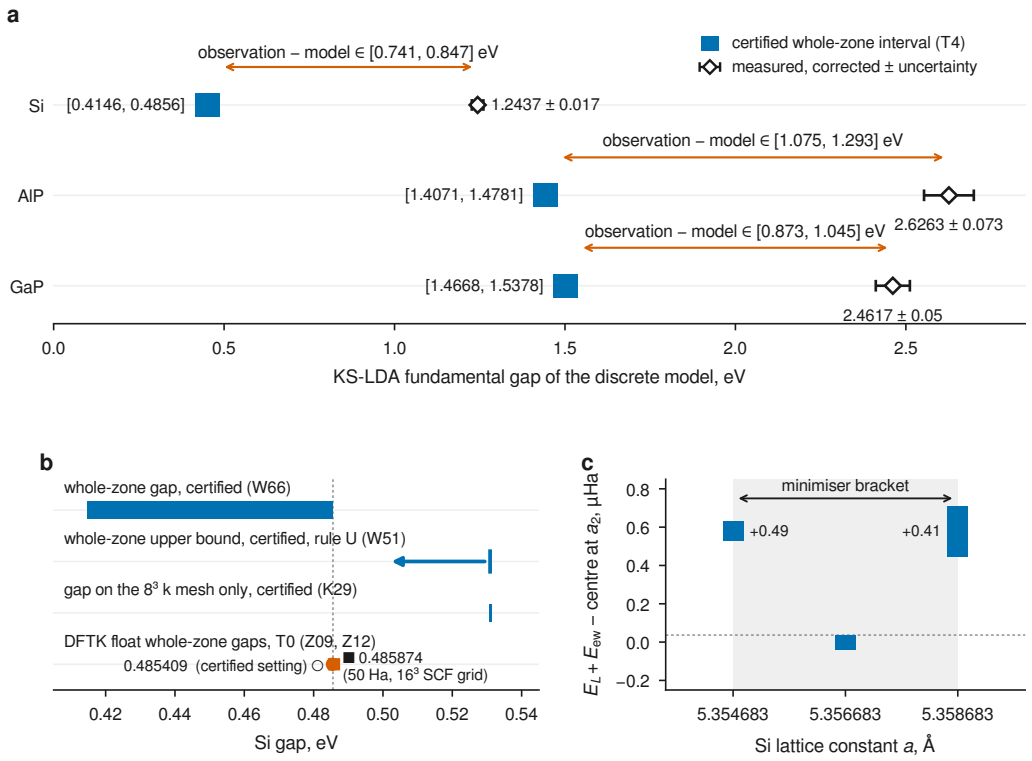


Figure 3: **Certified KS-LDA results and their comparison with experiment.** All intervals are proved (T4) for the *discrete* Kohn–Sham LDA model (GTH pseudopotentials, fixed cut-off, 8^3 Monkhorst–Pack mesh, fixed exchange–correlation grid, zero temperature, no spin–orbit; locally unique solution), not for converged LDA and not for the material; displayed bounds are rounded outward. **a,** Whole-zone gap, eV, under the two-sided rule, against the corrected measured gap (V1); each is a supplement to the pre-registered one-sided comparison, whose verdict stays in the record. Si ($a = 5.416$ Å): $[0.4146, 0.4856]$ eV (W66); observation minus model $[0.741, 0.847]$ eV (O04). AlP (blinded entry): $[1.4071, 1.4781]$ eV (W300); $[1.075, 1.293]$ eV (O20). GaP (blinded entry): $[1.4668, 1.5378]$ eV (W192); $[0.873, 1.045]$ eV (O24). The deviations include the discretisation, the pseudopotential and the difference between the Kohn–Sham and the fundamental gap, not the functional alone, and are therefore not called model errors. **b,** Si in detail: the one-sided bound (W51), the gap on the mesh alone (K29), and floating-point DFTK values (T0) at the certified setting and at a converged reference, a different discrete problem (Z09, Z12). **c,** Si lattice bracket: certified intervals (bar heights) of $E_L + E_{\mathrm{ew}}$ at three lattice constants with Q fixed, relative to the centre at a_2 (W105–W107); every minimiser lies in $(5.354683, 5.358683)$ Å, assuming that the end-point solutions are ground states; no comparison with experiment.

In all three materials the discretised model lies far below the observation, as expected for LDA. Each interval encloses, up to the measurement uncertainty, the deviation of the discretised model from experiment: it contains the error of the functional, the difference between the Kohn–Sham and the quasiparticle gap, the pseudopotential and the fixed discretisation (for silicon, continuum effects of a few meV, estimated in floating point; §4), and it excludes algebraic, rounding and software error, without saying which of the included parts is responsible. S1 is a demonstration of feasibility at near-production settings: its value is this exclusion, not a narrower number, since the certified whole-zone intervals, all 0.071 eV wide because the driver leaves fixed margins at the band edges (Appendix K), are wider than the meV-level uncertainty of a converged floating-point gap and, for silicon, than the measurement uncertainty. For silicon the checker also encloses bands 1–8 at 73

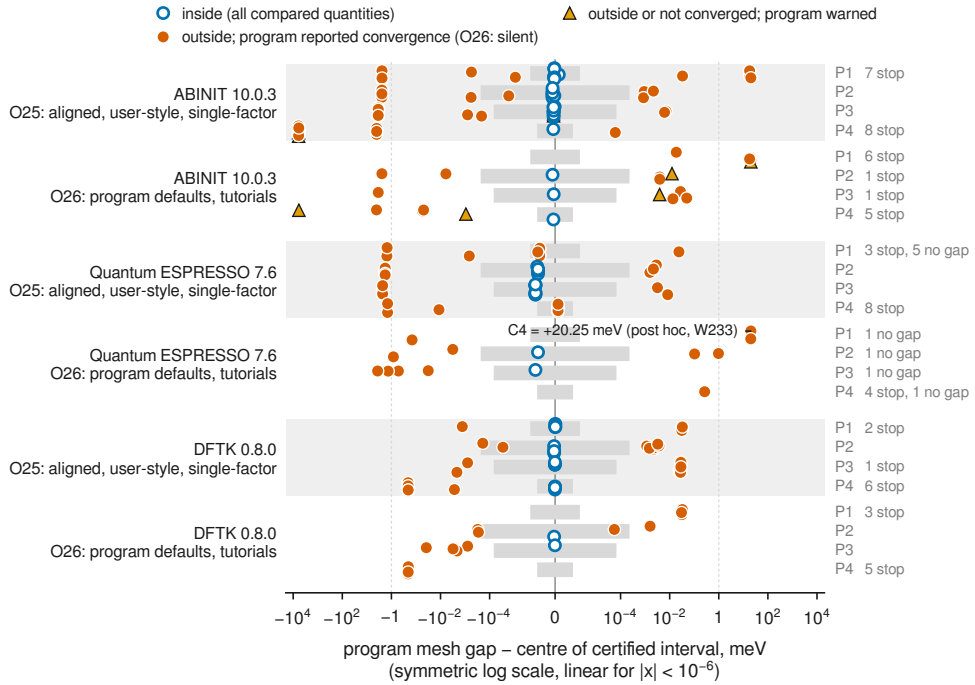


Figure 4: **ABINIT, Quantum ESPRESSO and DFTK against certified answers.** Each point is one run on one of four certified discrete problems P1–P4 (lanes; vertical jitter only separates points); horizontal: the program's mesh band gap minus the centre of the certified interval (grey bars; half-widths $1.2 \cdot 10^{-6}$ to $2.3 \cdot 10^{-4}$ meV), meV, symmetric logarithmic axis. Marker: the registered verdict on every compared quantity, with no tolerance, so a run can be outside through its eigenvalues or energies while its gap is near 0; right margin: runs that stopped or reported no gap. Shaded bands: protocol O25 (219 runs: aligned inputs, typical user inputs and single-factor variants; W201–W204); white bands: protocol O26 (86 runs: program defaults and official tutorials; W230–W232). For O26, "silent" means exit code 0, convergence claimed and no warning; O25 did not register this classification, so its filled points mean only that the program reported convergence. Labelled P1 points: the shifted-grid case (registered runs $+20.3$ to $+21.5$ meV); its mechanism and the size $+20.25$ meV of the same-FFT-grid pair come from a diagnostic run after the results, post hoc, and change no verdict (W233). Statements hold for these inputs and versions (ABINIT 10.0.3, QE 7.6, DFTK 0.8.0).

points along a high-symmetry path to about 0.17 meV (W282, Fig. 6), the total energy of the certified solution to $8.1 \cdot 10^{-8}$ Ha (W89) and, provided the end-point solutions are ground states, every energy minimum of the fixed-basis discrete model in lattice constant in $(5.354683, 5.358683)$ Å, not compared with experiment (W103–W107).

In an earlier, linear example all three correctnesses were assessed together. Certified band energies of the Cohen–Bergstresser empirical pseudopotentials, compared once under a protocol committed before any certified number existed, put silicon within the 0.3 eV measurement uncertainty, while germanium and GaAs disagree by 0.54 and 0.38 eV, in regions where the fit has no data (O01, O02, V2); a pre-registered extension to nine more semiconductors against one photoemission study (Ley et al., 1974) finds 37 of 50 valence-level comparisons in disagreement, the deep X_3 level in every material (O07–O19; Appendix Q.1).

3.2 Auditing Widely Used Codes and Data

DFT Codes. Under a protocol frozen before the runs (O25), ABINIT 10.0.3, Quantum ESPRESSO 7.6 and DFTK 0.8.0 were run 219 times on four discrete problems whose answers are certified intervals (W201; Fig. 4). With inputs aligned to the certified problem the codes reproduce it wherever they converge, apart from energies at the 10^{-8} Ha level and below (W202). With convergence set by an energy tolerance of 10^{-10}, ABINIT and QE reported convergence with exit code 0 and no warning while unoccupied-band eigenvalues lay up to $8 \cdot 10^{-3}$ Ha outside the intervals (W204), in the top one or two computed bands, for which ABINIT provides buffer bands (`nbdbuf`) and QE documents a looser threshold; our protocol requested no buffer bands, and without those bands the runs lie within each code's own tolerance under the $\sqrt{t}$ reading below (read literally, 1 of the 4; W318, post hoc). A consequence protocol, frozen before any computation (W208), found that 28 of the 131 runs judged outside or not converged change at least one derived quantity at thresholds fixed in advance (ABINIT 21, QE 7, DFTK 0; W209); on silicon and aluminium phosphide no quantity a code reports as converged changes, here or at the default settings below (W319). A third protocol left every numerical setting at the program defaults or as in the official tutorials, with only the physical problem aligned (W230): all 33 such runs that gave numbers lie outside the certified intervals, about 10^{-9} Ha wide, 29 of them without a warning (W231, W232). A post hoc reading, defined before its numbers were computed but after the results were known (W316, W317), finds that 11 solved a different discrete problem: six, on a silicon toy problem (cutoff 1.6 Ha, $2 \times 2 \times 2$ grid shifted by $\frac{1}{2}$), because QE by default, and ABINIT once its refusal is overridden, symmetrise the density with all 48 cubic operations, 36 of which do not preserve the grid (and use a different FFT grid; gap $+20.3$ to $+21.5$ meV; W233), and five because DFTK used its default FFT grid (gap shifts of at most 0.21 meV). The other 22 solved the certified problem, within the code's own tolerance if an energy criterion t is taken to allow eigenvalue errors of order $\sqrt{t}$, an order-of-magnitude sensitivity reading, not a bound (24 of the 33 under this reading, 11 read literally). On the toy problem the certified problem lies 29–46 meV from a k-converged reference, the symmetrised one 5–44 meV, closer in gap, direct gap and energy (W320): a symmetrised run does not solve the certified discrete problem, which does not make it further from the physics. A fourth protocol (O28) used Krack's GTH-PBE pseudopotential for silicon, whose h_{12} violates the HGH relation by which diagonal-only formats rebuild it: paths that read the full h matrix are inside the certified intervals, while ABINIT given a diagonal-only file is outside by up to 0.51 Ha without a warning about h_{12} (Appendix K.1). These statements are about these inputs and versions, and about certified problems written in DFTK's conventions, one discretisation among several (Appendix K). The same instrument shows that a classic table of nonlocal empirical pseudopotential band energies (Chelikowsky and Cohen, 1976) is not the spectrum of its printed model to the printed digits: of 111 registered values, 28 agree to the printed digits and 64 within the stated accuracy (O21, W158).

CALPHAD Databases. The frame of the audit was fixed in advance: every binary, non-kinetic database published as *Calphad* supplementary material in the TDBDB index (161 files, 99 papers; CAL-B39; Fig. 5). Metric 1 asks whether the published file regenerates the published table, a question of reproducibility, not of the accuracy of either. Against the invariant tables printed in their own papers, the model defined by the file disagrees with at least one printed reaction robustly in 21 of 47 comparable databases (16 of their 32 papers), that is also beyond one unit of the last printed digit and for either value of the gas constant (25 counting all disagreements; 41,788 certificates, CAL-B40). Of the 46 robust disagreements, 34 have a closed certified interval, 8 less than 0.1 K beyond the printed precision, 13 within 0.1–1 K and 13 at least 1 K (10 from one Ca–Ga paper; 4 databases, 3 papers), the largest registered one 7.53 K; of the other 12, supplementary certificates close 5, up to 16.85 K, 2 stay open and 5 are rows of doubtful printed composition (CAL-B70–B72, post hoc). Propagating the rounding of the printed parameters, read as rounded in their last digit (CAL-B76–B78, post hoc), 28 of the 34 have a certified file within half a unit of every last printed digit whose invariant temperature meets the printed one; 6, in 4 databases, exceed the linearised rounding range (the largest 2.31 K beyond the printed precision), as does a supplementary 9.58 K Hf–Sn case. Under the same frozen protocol, pycalphad 0.11.2 at its documented

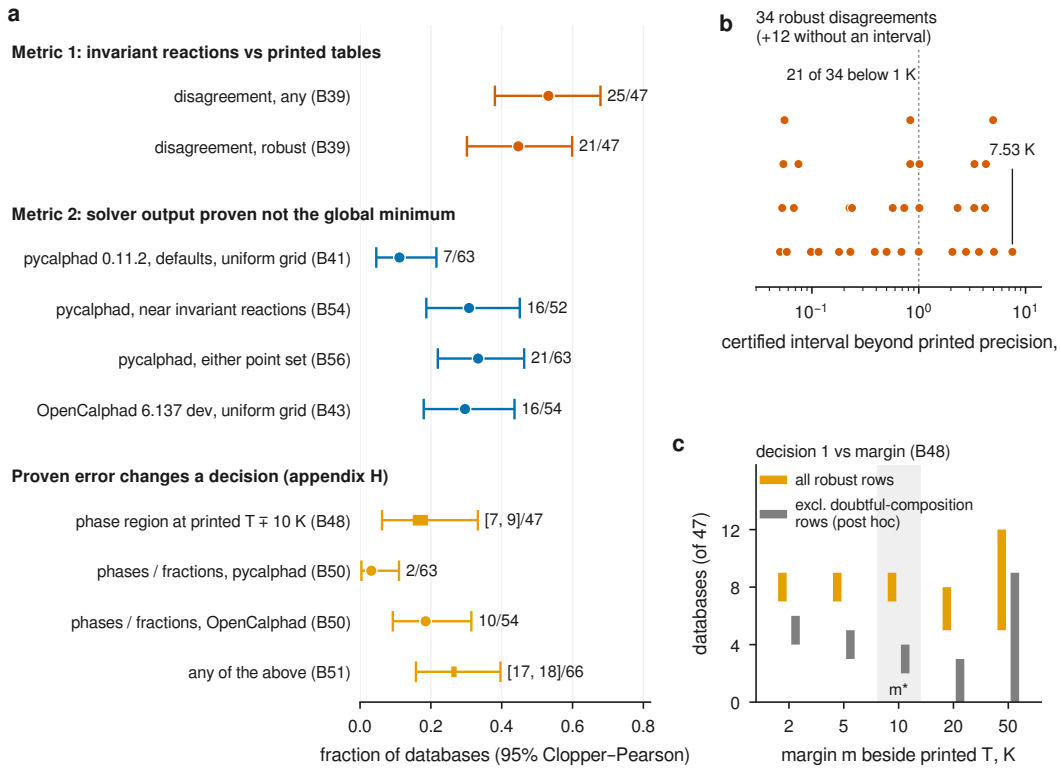


Figure 5: **An audit of published CALPHAD databases.** Frame: binary databases published as *Calphad* supplementary files (pre-registered; statements hold for the database files, not for the alloys). **a,** Fraction of databases with at least one proven disagreement (metric 1) or proven non-global equilibrium (metric 2), with exact 95% Clopper–Pearson intervals. Metric 1: a certified invariant reaction disagrees with the table printed in the same paper (25 of 47; 21 robustly, i.e. beyond the printed precision and independent of the gas constant used; CAL-B39, CAL-B40). Metric 2: a solver at its defaults returns a state certified not to be the global equilibrium of the model the file defines, on a fixed (T, x) grid (pycalphad 0.11.2, CAL-B41, CAL-B42; OpenCalphad 6.137 development build, CAL-B43) or, for pycalphad, at points within 2 K of the printed invariant reactions (CAL-B54, CAL-B55); the registered quantity counts a database if either point set has such a point (CAL-B56). Consequences: addendum H (CAL-B48–CAL-B51; for the solvers, errors on the grid only); ranges arise from unclassifiable cases. **b,** Distance by which the certified invariant temperature lies outside the printed value's precision interval, for the 34 robust disagreements with a certified interval (12 further have none): mostly below 1 K, at most 7.53 K. **c,** Decision 1 against the margin m: $m^* = 10$ K was chosen after the disagreements had been seen, so all five registered margins are shown; grey: without table rows of doubtful composition (post hoc sensitivity).

default settings returns an equilibrium proven by a Lean-checked certificate not to be the global equilibrium of the model the file defines in 21 of 63 databases (95% interval 0.22–0.46; CAL-B56): in 7 on a fixed (T, x) grid (4 without points counted only through a phase-fraction filter; CAL-B41, CAL-B42) and in 16 of 52, 14 of them only there, at points within 2 K of the printed invariant reactions or just outside their certified intervals, fixed mechanically and committed before any pycalphad call, with Gibbs energies up to 1.2×10^4 J/mol above the certified global minimum (CAL-B54, CAL-B55). A denser starting grid removes 49 of these 50 near-invariant non-global equilibria (CAL-B58); recomputed at the failing points only, 2 of 63 databases keep one that does not depend on the filter (CAL-B74), so these counts quantify a known limitation of grid-based minimisation near phase transitions at the

documented defaults. On the grid, a development build of OpenCalphad 6.137 returns a state proven not to be the global equilibrium in 16 of 54 databases, with a reading check revised after the results and labelled post hoc (CAL-B43); the two rarely do so at the same point. A further addendum, frozen after the metric-1 and grid results but before any new certificate and before the near-invariant runs, asked whether these proven disagreements and the grid's non-global equilibria (pycalphad and OpenCalphad) change two engineering decisions (CAL-B48–CAL-B51): in 17–18 of 66 databases at least one does (12–13 without rows of doubtful printed composition, a post hoc sensitivity); the process margin of 10 K was chosen after the disagreements had been seen, so all five margins (2–50 K) are reported. Post hoc, the 17 are 7 databases from metric 1 (5 only through rows of doubtful composition, 2 through open-interval temperatures, none through the shift of a closed interval, at most 8.05 K) and 10 from the OpenCalphad development build (pycalphad's 2 lie among the 10); without rows of doubtful composition and without the development build, 4–6 of 66 (CAL-B75). Earlier audits, of the NIST solder database (CAL-B35, CAL-B36), of published binary assessments (CAL-B37, CAL-B38) and of the 17 databases shipped with pycalphad, are reported in Appendices N and Q.

3.3 Mathematical Correctness: What Is Proved

Beyond the scale of the ledger, the certificates make statements that floating-point software cannot. In one dimension they enclose the discrete Kohn–Sham ground-state energy and its infinite-cutoff limit, proved equal to the ground energy of the continuum model, and the ground-state energy of $-\mathrm{d}^2/\mathrm{d}x^2 + x^4$ on the whole line in an interval $1.1 \cdot 10^{-16}$ wide; the classical solutions of boundary-value problems and two-sided bounds for membrane eigenvalues; the unique true trajectories of Newton's equations; and the global equilibria of thermodynamic databases (Appendices J–O). For the reduced Hartree–Fock silicon model the infimum of the discrete energy over all finite plane-wave sets lies in an interval $2.2 \cdot 10^{-4}$ Ha wide (C29–C33); for LDA the argument does not close, since the density nearly vanishes at the ion sites, and a Lean theorem proves the functional side of this obstruction: the constant the LDA checkers use for the second-order term has no finite bound as the density bound tends to zero (Y05, Y06). Showcase S2 proves that a converged spin-restricted LDA solution of stretched hydrogen is not the ground state of its discrete model, with a certified margin of 330 meV (M20–M26, Z06–Z08); the ordering is textbook physics, so this is a first in rigour, not a discovery. And disjoint balls of local uniqueness prove that a discrete LDA problem has at least five exact solutions with pairwise distinct densities, whose free-energy intervals separate them into three ordered groups (DFT-X09, X13; Appendix Q).

4 Discussion and Limitations

The Weakest Correctness. By the programme's own thesis, a result is only as trustworthy as its weakest correctness, and that is physical correctness. All certified three-dimensional exchange–correlation results are for LDA: PBE certification is about 300-fold short of the Newton–Kantorovich margin at S1 scale with the present proof route and is not pursued (Appendix K.1). Silicon, aluminium phosphide and gallium phosphide are bounded from both sides; carbon and germanium, registered in the same protocol, are out of reach of the present checker (cost and a nearly closed gap) and stay registered but not compared (O05, O06). Every one-dimensional result and every FEM and MD result is at V0, and the experimental values of the Cohen–Bergstresser comparisons come from compilations and await review by an experimentalist. Physical correctness is the one correctness that is not machine-checked: it is tested against observation, and the machine guarantees only how the comparison was made, not that the measurement is right; the registration timestamps are commits to our own repository, not entries in an independent registry. Most results should therefore be read as statements about models, proved to be solved correctly, and not about materials (Appendix S).

The Discrete Model and Continuous KS-LDA. What S1 certifies is the discrete problem: GTH-LDA at the protocol's cutoff, k mesh and integration grid. For silicon, floating-point

convergence puts the effect of the cutoff on the gap over the mesh at $+1.21$ meV against an 80 Ha reference (W140, T0), the difference of two larger parts: with the potential frozen at the certified solution, the shift is proved to lie in $[4.17, 6.32]$ meV (W146, T4; floating point $+5.25$ meV), while the self-consistent relaxation of the potential, -4.04 meV in floating point, is not bounded. A rigorous bound on the self-consistent continuum gap remains open, since norm-based bounds cannot see this cancellation (W141, C46). These effects of a few meV are two to three orders of magnitude below the $[0.741, 0.847]$ eV by which the discrete model lies below the observation, so the S1 verdict does not depend on them; formally, it is a statement about the discrete model, with floating-point evidence of convergence (Appendix K.2).

Relation to A Posteriori Estimates. Guaranteed a posteriori bounds for the linear Kohn–Sham operator of three-dimensional silicon (Herbst et al., 2020) and for eigenvalues (Cancès et al., 2017), a posteriori analysis of nonlinear eigenvalue problems (Dusson and Maday, 2017), practical error estimates (Cancès et al., 2022) and computer-assisted existence proofs for nonlinear elliptic problems (Plum, 2001; Nakao et al., 2019) provide the mathematics on which our three-dimensional checkers build. Measured on our certified problems, DFTK 0.8.0's estimates are accurate on the discrete SCF error (within 2.2%) and narrower than our intervals, which contain them; they do not cover rounding or establish existence and uniqueness (W214–W219). We therefore do not claim tighter bounds than a posteriori estimates, only bounds of a different kind: machine-checked, about the true discrete object, with no unverified constants (Appendix R).

The Trusted Base. The Lean kernel, compiler and runtime, the certificate parsers, and the operating system and hardware that run the checkers are trusted; TLC, Frama-C with its floating-point model, the SMT solvers, the C compiler and the hand-checked agreement of the ACSL contracts with the Lean statements only for the software-layer claims, on which no checker output depends (Appendix T). Where checkers bind the problem, what is trusted is that the Lean definitions of the DFT potentials, of the S1 problem and of reading a TDB file (930 lines) say what is intended (A-DFT-05, A-CAL-01); the remaining problem data are echoed but not recomputed. Twenty-three claims are below their target level. The software layer adds assurance about the drivers, not about the headline numbers, which are checker outputs and hold whatever the drivers do: the TLA$^+$ protocols are model-checked only on small finite instances, with only a paper argument for any size (T1); that the S3 driver follows its protocol and is bitwise reproducible is shown by tests (T0) of the executions that were run (Appendix I.2); and the Frama-C-proved kernel enters no checker output.

External Confirmation. The audit verdicts say that a program run or a database file disagrees with a certified answer, not which party is wrong; they hold for the stated inputs and versions. They have not yet been confirmed or disputed by the developers and authors concerned; the frozen protocols, the archived checker outputs and the certificates, the largest identified by SHA-256, will be released with the code for independent checks.

Conclusion. Machine-checked triple correctness asks of a simulated number that its mathematics and its software be proved, with the proofs checked by machine, and that its comparison with observation be frozen in advance and recomputed by machine. FORMAL IN SILICO realises this on seven layers with an explicit obligation at each transition, trust levels for what is proved and validation grades for what is observed, in one ledger. Verifying small checkers rather than large solvers makes the rigorous part attainable at production settings; once numerics and software are certified, the residual against experiment is the deviation of the certified model, discretisation included, plus measurement error, as for the discretised LDA band gaps of silicon, aluminium phosphide and gallium phosphide, and the behaviour of widely used codes, and the agreement of published data with the models they define, become measurable against exact references. We hope that the ledger, separating what is tested, what is proved, what is observed and what is still open, will prove useful beyond this project for reporting the reliability of scientific computation.

Reproducibility. Code, data, certificates, ledgers, protocols and the archived showcase and audit runs will be released at `https://github.com/SMU-DIGA/formal-in-silico` upon

publication; the programme platform, with the trust ledger and news of each result, is at `https://smu-diga.github.io/formal-in-silico-programme/`. `make verify-all` re-checks the repository (§2.5); all verification runs cited here are cold strict runs with 32 parallel jobs (no result cache read, Lean build output reused, a missing tool fails the run).

References

Alkassar, E., Böhme, S., Mehlhorn, K., and Rizkallah, C. (2014). A framework for the verification of certifying computations. *Journal of Automated Reasoning*, 52(3):241–273.

Anderson, D. G. (1965). Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560.

Ansara, I., Dinsdale, A. T., and Rand, M. H., editors (1998). *COST 507: Thermochemical Database for Light Metal Alloys, Volume 2*. Office for Official Publications of the European Communities, Luxembourg.

Appel, A. W. (2011). Verified software toolchain: (invited talk). In *European Symposium on Programming*, pages 1–17. Springer.

Bauer, F. L. and Fike, C. T. (1960). Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141.

Boldo, S., Clément, F., Faissole, F., Martin, V., and Mayero, M. (2017). A Coq formal proof of the Lax–Milgram theorem. In *Proceedings of the 6th ACM SIGPLAN Conference on Certified Programs and Proofs (CPP)*, pages 79–89.

Boldo, S., Clément, F., Filliâtre, J.-C., Mayero, M., Melquiond, G., and Weis, P. (2013). Wave equation numerical resolution: a comprehensive mechanized proof of a C program. *Journal of Automated Reasoning*, 50(4):423–456.

Boldo, S. and Melquiond, G. (2011). Flocq: A unified library for proving floating-point algorithms in Coq. In *IEEE 20th Symposium on Computer Arithmetic (ARITH)*, pages 243–252.

Bordignon, A., Cancès, E., Dusson, G., Kemlin, G., Lainez Reyes, R. A., and Stamm, B. (2025). Fully guaranteed and computable error bounds on the energy for periodic Kohn–Sham equations with convex density functionals. *SIAM Journal on Scientific Computing*, 47(5):A2881–A2905.

Bosoni, E., Beal, L., Bercx, M., Blaha, P., Blügel, S., Bröder, J., Callsen, M., Cottenier, S., Degomme, A., Dikan, V., et al. (2024). How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. *Nature Reviews Physics*, 6(1):45–58.

Cancès, E., Chakir, R., and Maday, Y. (2012). Numerical analysis of the planewave discretization of some orbital-free and Kohn–Sham models. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(2):341–388.

Cancès, E., Deleurme, A., and Lewin, M. (2008). A new approach to the modeling of local defects in crystals: the reduced Hartree–Fock case. *Communications in Mathematical Physics*, 281(1):129–177.

Cancès, E., Dusson, G., Kemlin, G., and Levitt, A. (2022). Practical error bounds for properties in plane-wave electronic structure calculations. *SIAM Journal on Scientific Computing*, 44(5):B1312–B1340.

Cancès, E., Dusson, G., Maday, Y., Stamm, B., and Vohralík, M. (2017). Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations. *SIAM Journal on Numerical Analysis*, 55(5):2228–2254.

Cancès, É., Ehrlacher, V., Gontier, D., Levitt, A., and Lombardi, D. (2020). Numerical quadrature in the Brillouin zone for periodic Schrödinger operators. *Numerische Mathematik*, 144(3):479–526.

Céa, J. (1964). Approximation variationnelle des problèmes aux limites. *Annales de l'Institut Fourier*, 14(2):345–444.

Chelikowsky, J. R. and Cohen, M. L. (1976). Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors. *Physical Review B*, 14:556–582.

Cirstea, H., Kuppe, M. A., Loillier, B., and Merz, S. (2024). Validating traces of distributed programs against tla+ specifications. In *International Conference on Software Engineering and Formal Methods*, pages 126–143. Springer.

de Moura, L. and Ullrich, S. (2021). The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *LNCS*, pages 625–635.

Dusson, G. and Maday, Y. (2017). A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem. *IMA Journal of Numerical Analysis*, 37(1):94–137.

Dusson, G., Sigal, I., and Stamm, B. (2023). Analysis of the Feshbach–Schur method for the Fourier spectral discretizations of Schrödinger operators. *Mathematics of Computation*, 92(339):217–249.

Filliâtre, J.-C. and Paskevich, A. (2013). Why3 – where programs meet provers. In *Programming Languages and Systems (ESOP 2013)*, volume 7792 of *Lecture Notes in Computer Science*, pages 125–128.

Fumex, C., Marché, C., and Moy, Y. (2017). Automating the verification of floating-point programs. In *Verified Software: Theories, Tools, and Experiments (VSTTE 2017)*, volume 10712 of *Lecture Notes in Computer Science*, pages 102–119.

Goedecker, S., Teter, M., and Hutter, J. (1996). Separable dual-space Gaussian pseudopotentials. *Physical Review B*, 54:1703–1710.

Gómez-Serrano, J. (2019). Computer-assisted proofs in PDE: a survey. *SeMA Journal*, 76(3):459–484.

Gustafsson, T. and McBain, G. D. (2020). scikit-fem: A Python package for finite element assembly. *Journal of Open Source Software*, 5(52):2369.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric Numerical Integration*. Springer, 2nd edition.

Hales, T., Adams, M., Bauer, G., Dang, T. D., Harrison, J., Hoang, L. T., Kaliszyk, C., Magron, V., McLaughlin, S., Nguyen, T. T., et al. (2017). A formal proof of the Kepler conjecture. In *Forum of mathematics, Pi*, volume 5, page e2. Cambridge University Press.

Hartwigsen, C., Goedecker, S., and Hutter, J. (1998). Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B*, 58:3641–3662.

Herbst, M. F., Levitt, A., and Cancès, E. (2020). A posteriori error estimation for the non-self-consistent Kohn–Sham equations. *Faraday Discussions*, 224:227–246.

Herbst, M. F., Levitt, A., and Cancès, E. (2021). DFTK: A Julian approach for simulating electrons in solids. *Proceedings of the JuliaCon Conferences*, 3(26):69.

Hillert, M. (2001). The compound energy formalism. *Journal of Alloys and Compounds*, 320(2):161–176.

Hillert, M. and Jarl, M. (1978). A model for alloying in ferromagnetic metals. *Calphad*, 2(3):227–238.

Immler, F. (2014). Formally verified computation of enclosures of solutions of ordinary differential equations. In *NASA Formal Methods Symposium*, pages 113–127.

Immler, F. (2018). A verified ODE solver and the Lorenz attractor. *Journal of Automated Reasoning*, 61(1–4):73–111.

Inden, G. (1981). The role of magnetism in the calculation of phase diagrams. *Physica B+C*, 103(1):82–100.

Jakobsen, B. and Rosendahl, F. (1994). The Sleipner platform accident. *Structural Engineering International*, 4(3):190–193.

Kato, T. (1949). On the upper and lower bounds of eigenvalues. *Journal of the Physical Society of Japan*, 4(4–6):334–339.

Kirchner, F., Kosmatov, N., Prevosto, V., Signoles, J., and Yakobowski, B. (2015). Frama-C: A software analysis perspective. *Formal Aspects of Computing*, 27(3):573–609.

Krack, M. (2005). Pseudopotentials for H to Kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts*, 114:145–152.

Lamport, L. (2002). *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley.

Larsen, A. H. et al. (2017). The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002.

Lejaeghere, K. et al. (2016). Reproducibility in density functional theory calculations of solids. *Science*, 351(6280):aad3000.

Leroy, X. (2009). Formal verification of a realistic compiler. *Communications of the ACM*, 52(7):107–115.

Ley, L., Pollak, R. A., McFeely, F. R., Kowalczyk, S. P., and Shirley, D. A. (1974). Total valence-band densities of states of III–V and II–VI compounds from x-ray photoemission spectroscopy. *Physical Review B*, 9:600.

Lucas, K., Lynch, A., Tong, L., Champine, L., Straznickas, Z., Distefano, D., Agrawal, H., Ben Asher, E., Howell, A., Jawhar, S., Rohlf, C., Zeldovich, N., Carlini, N., Agrawal, R., and Gross, J. (2026). LLMs are making formal software verification feasible at a much larger scale. Anthropic research post. 22 October 2026.

Lukas, H. L., Fries, S. G., and Sundman, B. (2007). *Computational Thermodynamics: The Calphad Method*. Cambridge University Press.

Martin-Dorel, É. and Melquiond, G. (2016). Proving tight bounds on univariate expressions with elementary functions in Coq. *Journal of Automated Reasoning*, 57(3):187–217.

Martin-Dorel, É. and Roux, P. (2017). A reflexive tactic for polynomial positivity using numerical solvers and floating-point computations. In *Proceedings of the 6th ACM SIGPLAN Conference on Certified Programs and Proofs (CPP 2017)*, pages 90–99. ACM.

Marzari, N., Vanderbilt, D., De Vita, A., and Payne, M. C. (1999). Thermal contraction and disordering of the Al(110) surface. *Physical Review Letters*, 82(16):3296–3299.

McConnell, R. M., Mehlhorn, K., Näher, S., and Schweitzer, P. (2011). Certifying algorithms. *Computer Science Review*, 5(2):119–161.

McDonald, C. M. and Floudas, C. A. (1995). Global optimization for the phase stability problem. *AIChE Journal*, 41(7):1798–1814.

Melquiond, G. (2008). Proving bounds on real-valued functions with computations. In *Automated Reasoning (IJCAR 2008)*, Lecture Notes in Computer Science, pages 2–17. Springer.

Methfessel, M. and Paxton, A. T. (1989). High-precision sampling for Brillouin-zone integration in metals. *Physical Review B*, 40(6):3616–3621.

Monkhorst, H. J. and Pack, J. D. (1976). Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192.

Moore, R. E., Kearfott, R. B., and Cloud, M. J. (2009). *Introduction to Interval Analysis*. SIAM.

Nakao, M. T., Plum, M., and Watanabe, Y. (2019). *Numerical Verification Methods and Computer-Assisted Proofs for Partial Differential Equations*. Springer Series in Computational Mathematics. Springer, Singapore.

Newcombe, C., Rath, T., Zhang, F., Munteanu, B., Brooker, M., and Deardeuff, M. (2015). How Amazon Web Services uses formal methods. *Communications of the ACM*, 58(4):66–73.

Oberkampf, W. L. and Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press, Cambridge.

Otis, R. and Liu, Z.-K. (2017a). pycalphad: CALPHAD-based computational thermodynamics in Python. *Journal of Open Research Software*, 5(1):1.

Otis, R. A. and Liu, Z.-K. (2017b). High-throughput thermodynamic modeling and uncertainty quantification for ICME. *JOM*, 69(5):886–892.

Paulson, N. H., Bocklund, B. J., Otis, R. A., Liu, Z.-K., and Stan, M. (2019). Quantified uncertainty in thermodynamic modeling for materials design. *Acta Materialia*, 174:9–15.

Perdew, J. P. and Levy, M. (1983). Physical content of the exact Kohn–Sham orbital energies: Band gaps and derivative discontinuities. *Physical Review Letters*, 51(20):1884–1887.

Perdew, J. P. and Wang, Y. (1992). Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244–13249.

Plum, M. (2001). Computer-assisted enclosure methods for elliptic differential equations. *Linear Algebra and its Applications*, 324(1–3):147–187.

Prager, W. and Synge, J. L. (1947). Approximations in elasticity based on the concept of function space. *Quarterly of Applied Mathematics*, 5(3):241–269.

Rump, S. M. (2010). Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449.

Schmid-Fetzer, R., Andersson, D., Chevalier, P. Y., Eleno, L., Fabrichnaya, O., Kattner, U. R., Sundman, B., Wang, C., Watson, A., Zabdyr, L., and Zinkevich, M. (2007). Assessment techniques, database design and software facilities for thermodynamics and diffusion. *Calphad*, 31(1):38–52.

Sham, L. J. and Schlüter, M. (1983). Density-functional theory of the energy gap. *Physical Review Letters*, 51(20):1888–1891.

Slater, J. C. (1951). A simplification of the Hartree–Fock method. *Physical Review*, 81(3):385–390.

Solovej, J. P. (1991). Proof of the ionization conjecture in a reduced Hartree–Fock model. *Inventiones Mathematicae*, 104:291–311.

Swope, W. C., Andersen, H. C., Berens, P. H., and Wilson, K. R. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649.

Tan, Y. K., Heule, M. J. H., and Myreen, M. O. (2021). cake_lpr: Verified propagation redundancy checking in CakeML. In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2021)*, Lecture Notes in Computer Science, pages 223–241. Springer.

Temple, G. (1928). The theory of Rayleigh's principle as applied to continuous systems. *Proceedings of the Royal Society of London. Series A*, 119(782):276–293.

Tessier, S. R., Brennecke, J. F., and Stadtherr, M. A. (2000). Reliable phase stability analysis for excess Gibbs energy models. *Chemical Engineering Science*, 55(10):1785–1796.

The mathlib Community (2020). The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pages 367–381.

Tucker, W. (2002). A rigorous ODE solver and Smale's 14th problem. *Foundations of Computational Mathematics*, 2(1):53–117.

Tucker, W. (2011). *Validated Numerics: A Short Introduction to Rigorous Computations*. Princeton University Press.

van de Walle, A., Nataraj, C., and Liu, Z.-K. (2018). The Thermodynamic Database Database. *Calphad*, 61:173–178.

A Overview of the Programme and the First Realisation

The main text condenses the programme to its idea, its method and its headline results. This appendix gives the fuller overview from which it was condensed, Appendix B lists all results in one table, and the remaining appendices give the details: the motivating failures (Appendix C); the seven layers, the kinds of evidence and the obligation matrix (Appendix D); trust levels, validation grades and the ledger (Appendix E); principles, technical decisions and the roadmap (Appendices F–H); the shared infrastructure (Appendix I); the domains, DFT in one and three dimensions, FEM, MD, CALPHAD and structure search (Appendices J–O); the evaluation (Appendix P); certificates as research instruments and the physical results (Appendix Q); related work (Appendix R); open questions and limitations (Appendix S); and the trusted base, a trust-report excerpt, a planned TLA$^+$ specification and the parity milestones (Appendices T–W).

The Chain and Its Ingredients. Each link of the chain from the real world to a printed number can fail independently. The model may miss the physics, as the local-density approximation misses the band gaps of semiconductors. The problem may be ill-posed or the discretisation too coarse. An iteration may stop before it has converged, or a parallel protocol may lose data. Rounding can accumulate, and a reduction order can change with the process count. Mathematics already supplies many of the links: well-posedness theorems, a priori and a posteriori error estimates, convergence theorems for iterations; they are rarely connected to the code that prints the number. Formal methods supply tools for the software: interactive theorem provers, model checkers for concurrent protocols, deductive verifiers for C code; they have rarely been applied to scientific computing as a whole. A result is only as trustworthy as the weakest of these, so each needs its own kind of evidence and its own entry in a ledger.

What Makes This Feasible Now. Three developments make the attempt realistic. First, Lean 4 (de Moura and Ullrich, 2021) and Mathlib (The mathlib Community, 2020) now provide a unified, actively maintained library of real and functional analysis, linear algebra, spectral theory and measure theory, so the mathematics of numerical analysis can be stated and proved against a single foundation. Second, the pattern of certifying computation (McConnell et al., 2011; Hales et al., 2017) has matured. Instead of verifying a large solver, one verifies a small checker for its output. This turns the hardest part of the problem, verifying LAPACK, FFTs and heuristic nonlinear solvers, into a much smaller one. Third, AI-assisted theorem proving is advancing quickly and is most mature for Lean. This matters for a programme whose mathematical layer will eventually contain many thousands of theorems.

The Programme. FORMAL IN SILICO is organised around a small number of commitments, developed in Appendices C–H:

1. **Triple correctness on a layered chain** (Appendix D). A simulation is split into seven layers, from the observed real world (L0) to floating-point execution (L6). The first obligation, validation against observation, is physical; the next three are mathematical; the last two belong to software. The error of the printed number is bounded by a sum of discretisation, algebraic and rounding terms under the premise that the parallel program faithfully implements the abstract algorithm, and the residual against experiment decomposes into measurement, model and certified numerical errors. Every term and the premise become obligations, collected in a matrix that serves as the programme's work list.
2. **Three kinds of evidence, one bridge, one foundation** (Appendix D). PDE theory and numerical analysis are formalised in Lean; parallel algorithms are specified in TLA$^+$ and C code is proved with Frama-C; models are confirmed against traceable experimental data; refinement connects algorithm and program; floating point is part of the specification, not an afterthought.
3. **Honest levels and grades in one ledger** (Appendix E). Every claim is registered with its trust level (T0–T4) for mathematical and software correctness, and, for comparisons with experiment, a validation grade (V0–V3) on a separate axis; with

its evidence and assumptions. Every program run produces a report citing those claims. Unproved claims are listed, not hidden.

4. **Every output has a theorem** (Appendix F). A feature counts as finished only when each number it prints is covered by a machine-checked theorem. If the number is an approximation, the theorem bounds its error.
5. **Verify the checker, not the solver** (Appendices F and I). Large numerical libraries stay untrusted; their outputs are certified by small checkers proved correct in Lean and executed in exact rational arithmetic.
6. **A phased roadmap with external yardsticks and showcases** (Appendix H). The programme started with the smallest DFT program that still exercises the mathematical and software layers, and grew towards feature parity with established open-source codes (DFTK, scikit-fem, pycalphad, ASE), each used as an independent reference on the same discrete problem. Its current goal is three-dimensional DFT on real materials with one showcase for each correctness: a certified deviation from experiment (physical), a machine-checked proof that a converged SCF solution is not the ground state of its discrete model (mathematical; a first in rigour, since the orderings are known physics), and parallel correctness with bitwise reproducibility (software).

First Results. Appendices I–Q report the first realisation of this plan in detail. We state its balance up front: it realises mathematical and software correctness in depth, and physical correctness in a first example and a pre-registered protocol (Appendix Q.1). The charter was revised to triple correctness on 8 October 2026; the earlier version explicitly excluded validation, and most of the one-dimensional work reported here was done under it. The realisation is a code base, to be released with this paper, with four domains, built first in one dimension and now reaching three-dimensional DFT on three real materials, two-dimensional finite elements and three-dimensional periodic molecular dynamics, together with a structure-search track (Appendix O):

- `mini-DFT`, one-dimensional plane-wave reduced Hartree–Fock (rHF) DFT. It has certified enclosures of the discrete ground-state energy and its infinite-cutoff limit, of finite-temperature free energies, of solutions of the non-variational Methfessel–Paxton (of any order) and Marzari–Vanderbilt SCF equations, including the nonlinear SCF equations of LDA with Slater exchange and PW92 correlation under any finite-temperature smearing, at the Γ point and with k-points, of k-point sampled energies via an exact supercell equivalence, of densities and eigenvalues, and of band structures and densities of states for both ground states and finite-temperature SCF solutions, at the Γ point and, on the primitive cell, with k-points. Collinear-spin LSDA, with spin-polarised Slater exchange or with spin-interpolated PW92 correlation in a uniform Zeeman field, has certified SCF solutions at the Γ point and with k-points, at finite temperature and, under a common gap, at zero temperature, with enclosed magnetisation, Fermi level, energy and free energy, and per-spin bands and densities of states. Closed-shell Hartree–Fock (exact exchange) has certified SCF solutions at finite temperature at the Γ point, forces on Gaussian sites have certified intervals, and the infinite-cutoff limit is proved equal to the ground energy of the continuum model. The checker rejects every certificate that does not name its external potential, and verifies the potential against a Lean definition of the one named on the command line. Outside the periodic model, the ground-state energy of the continuous operator $-\mathrm{d}^2/\mathrm{d}x^2 + x^4$ on the whole line is enclosed in an interval $1.1 \cdot 10^{-16}$ wide (Appendix J).
- Three-dimensional DFT, built on the one-dimensional checkers (Appendix K). Its certificates cover the band energies of the Cohen–Bergstresser empirical pseudopotential for Si, Ge and GaAs, and the self-consistent solutions of reduced Hartree–Fock and of Kohn–Sham LDA (Slater exchange and PW92 correlation) for fcc crystals with GTH pseudopotentials, at finite and zero temperature, on Monkhorst–Pack meshes reduced by time reversal and by a proved point-group symmetry. At the settings of the pre-registered band-gap protocol S1 they certify the KS-LDA solu-

tion of silicon on an $8 \times 8 \times 8$ mesh, bound its band gap over the whole Brillouin zone from both sides and its total energy, enclose its bands along a high-symmetry path, and bracket the energy minimum of the discrete model in lattice constant; aluminium phosphide and gallium phosphide (with the full GTH kernel for Ga) are certified at their own pre-registered settings, with two-sided gap bounds as well. A first instance of showcase S2 proves that a converged spin-restricted solution of stretched hydrogen is not the ground state of its discrete model. The registered S1 problem is written in Lean, and a checker mode whose soundness theorems conclude about it binds the archived silicon certificates to it: the SCF and one-sided zone certificates in one process, the two-sided result by a composition theorem over separate runs.

- `mini-FEM`, P1 and cubic Hermite elements in one dimension and P1 triangles in two. Its certificates enclose the *classical solution* of the continuous boundary-value problem at every node, including pure Neumann problems, whose compatibility condition the checker decides, bound $\sup |u - u_h|$ and the energy-norm error, drive adaptive refinement by proven element-wise bounds, and bound the eigenvalues of the continuous Sturm–Liouville problem from both sides. In two dimensions they give guaranteed Prager–Synge error bounds for the Poisson problem and two-sided bounds for the first Dirichlet eigenvalue of the square and of an L-shaped membrane (Appendix L).
- `mini-MD`, a one-dimensional Lennard-Jones chain and a three-dimensional periodic box of Lennard-Jones atoms with a smooth cutoff, integrated with velocity Verlet. It has machine-checked conservation laws, symplecticity and reversibility, proofs that Newton's equations have a unique global solution in both settings, and certified distances between the computed trajectory and that true trajectory. Further checkers enclose the period of a pendulum, the distance between exact-arithmetic Verlet and the true solution, and the normal-mode energies of the Fermi–Pasta–Ulam–Tsingou (FPUT) chain (Appendix M).
- `mini-CALPHAD`, binary CALPHAD equilibria from TDB databases, including magnetic phases, sublattice phases of the compound energy formalism with up to two mixing sublattices, interstitial solutions and magnetic compounds. Its certificates cover the *global* Gibbs-energy minimum, the chemical potentials, the driving forces and the composition windows of every phase in any equilibrium, at a given overall composition or a given chemical potential, and the Gibbs energy, entropy, enthalpy and heat capacity of each phase at fixed composition. They assemble into phase diagrams whose partition into regions is certified complete at every grid temperature, and further certificates hold at every temperature of an interval, which gives certified intervals for eutectic and eutectoid temperatures. The checker reads the TDB file itself, and a further kind of certificate proves that a phase has no miscibility gap (Appendix N).

The code base comprises about 219k lines of Lean (some 9.1k theorems; all 36,091 constants of the library depend only on Lean's three standard axioms), about 14.8k lines of C (including the ACSL contracts of the trusted kernel), four model-checked TLA$^+$ specifications, three of them with trace validation, black-box validation of the run reports, sanitizer-instrumented fuzzing of the drivers, and a Frama-C/WP-verified interval kernel. The statements of all 515 checker soundness theorems, and the ACSL contracts of the trusted C kernel, are fingerprinted, so that a change of a pinned statement fails verification unless it is re-registered explicitly, and certificate tests include planted known answers that catch a driver solving the wrong problem (Appendices I.1 and P). Results agree with DFTK (Herbst et al., 2021), scikit-fem (Gustafsson and McBain, 2020), ASE (Larsen et al., 2017) and pycalphad (Otis and Liu, 2017a) on identical problems, and the reference values lie inside our certified intervals (Appendix P). Finally, the certificates are used as research instruments: they prove that a discrete LDA problem has several exact solutions, and they audit the 17 thermodynamic databases shipped with pycalphad point by point (Appendix Q). Certified answers make widely used codes and published data auditable: under protocols frozen before the runs, ABINIT, Quantum ESPRESSO and DFTK return numbers outside certified intervals without a warning, also at their default settings, where a post hoc reading finds a different discrete

problem or a deviation within the code's own tolerance under one reading (Appendix K); a classic table of empirical pseudopotential band energies is not the spectrum of its printed model; in 21 of 47 comparable binary CALPHAD databases published with their assessments, the model the file defines disagrees, beyond the printed digits, with at least one invariant reaction printed in the same paper, a question of reproducibility, and for 28 of the 34 such differences with a closed certified interval within the rounding of the printed parameters (post hoc); pycalphad at its default settings returns equilibria proved not to be global in 21 of 63 databases, as does a development build of OpenCalphad in 16 of 54; and some of these disagreements and non-global equilibria change two stylised engineering decisions (Appendix N). The audit is also where the realisation first touches physical correctness: CALPHAD databases are fitted to experiments, and the anomalous liquid miscibility gaps it certifies lie where the parameters are extrapolated beyond the data (Appendix Q.1). Appendix S collects the open questions and the limits of what has been achieved.

B Results at a Glance

Table 2 lists the results of the main text and the appendices in one place, grouped by the correctness they serve. The column *Kind* names it: physical (P), mathematical (M) or software (S); audits are mathematical results about other programs and published data (A). The column *Level* gives the trust level of the cited claims and, for comparisons with experiment, the validation grade. In audits the certified side is T4, while the verdict about another program is T0 or T1 in the ledger, because the other program's number is floating point; the table gives both. Every row is stated for the discrete problem, inputs and versions named in the cited appendix.

Table 2: Results at a glance. Ledger identifiers as in the trust ledger; the appendix of each row gives the protocol and the evidence.

Result	Kind	Ledger	Level
Physical correctness: certified deviations from experiment (Appendices K, Q.1)			
Si, KS-LDA with GTH-LDA q4 at the pre-registered S1 settings ($a = 5.416$ Å, cutoff about 10 Ha, $8 \times 8 \times 8$ mesh): exactly one solution within $2.1 \cdot 10^{-10}$ of the certificate's centre (under the chemical-potential convention of Appendix K); band gap over the whole zone in $[0.4146, 0.4856]$ eV (939 CPU hours); $[0.741, 0.847]$ eV below the observed 1.2437 ± 0.017 eV	P, M	DFT-O04, K29, W52–W66	T4; V1
AlP at its pre-registered settings: gap in $[1.4071, 1.4781]$ eV (705 CPU hours), $[1.075, 1.293]$ eV below the observed 2.6263 ± 0.073 eV	P, M	DFT-O20, W300	T4; V1
GaP with the full GTH kernel for Ga: gap in $[1.4668, 1.5378]$ eV (1 842 CPU hours), $[0.873, 1.045]$ eV below the observed 2.4617 ± 0.05 eV	P, M	DFT-O24, W192	T4; V1
Cohen–Bergstresser empirical pseudopotentials, protocol committed before any certified number: Si within the 0.3 eV measurement uncertainty; Ge and GaAs disagree by 0.54 and 0.38 eV	P	DFT-O01, O02	V2
Pre-registered extension to nine more semiconductors against one photoemission study: 37 of 50 valence-level comparisons disagree, the deep X_3 level in every material	P	DFT-O07–O19	V2
Si, S1 solution (not compared with experiment): bands 1–8 at 73 points of a high-symmetry path to about 0.17 meV; total energy to $8.1 \cdot 10^{-8}$ Ha; every energy minimum of the fixed-basis model in lattice constant in $(5.354683, 5.358683)$ Å, provided the end-point solutions are ground states	M	DFT-W282, W89, W103–W107	T4 (bracket theorem T3)
Audits of widely used DFT codes and pseudopotential data (Appendices K, K.1)			

Result	Kind	Ledger	Level
ABINIT 10.0.3, Quantum ESPRESSO 7.6 and DFTK 0.8.0, 219 runs on four certified discrete problems under a frozen protocol: with an energy tolerance of 10^{-10}, ABINIT and QE report convergence without warning while unoccupied-band eigenvalues lie up to $8 \cdot 10^{-3}$ Ha outside the certified intervals, in the top one or two computed bands, without which they lie within each code's own tolerance (post hoc, under the $\sqrt{t}$ reading below; read literally, 1 of the 4; the protocol had no buffer bands)	A	DFT-O25, W201–W204, W318	T4 intervals; T0 verdicts
Consequences, protocol frozen before any computation: 28 of 131 runs judged outside or not converged change at least one derived quantity (ABINIT 21, QE 7, DFTK 0); on Si and AlP, here and at default settings, no quantity a code reports as converged changes	A	DFT-W208, W209, W319	T0
Default or tutorial settings: all 33 runs that gave numbers lie outside, 29 of them without a warning; post hoc, 11 solved a different discrete problem (on a k grid shifted by $\frac{1}{2}$, symmetrisation with all 48 cubic operations, 36 of which do not preserve the grid, makes QE, and ABINIT once its refusal is overridden, shift the gap by $+20.3$ to $+21.5$ meV; or DFTK's default FFT grid) and 22 the certified one, within each code's own tolerance if an energy criterion t is taken to allow eigenvalue errors of order $\sqrt{t}$ (24 of the 33 under this reading, 11 read literally)	A	DFT-W230–W233, W316, W317	T0
The shifted-grid toy problem: against a k-converged floating-point reference, the certified discrete problem lies 29–46 meV away, the symmetrised one 5–44 meV, closer on all three quantities (gap 8.5 against 28.8 meV below it; post hoc)	A	DFT-W320	T0
Krack's GTH-PBE pseudopotential for Si: h_{12}, like that of all 211 multi-projector channels of its family, violates the HGH relation; paths reading the full h matrix are inside, ABINIT given a diagonal-only file is outside by up to 0.51 Ha without warning	A	DFT-O28	T0
Census of 2821 public GTH/HGH files from seven sources (113 775 parameters): 932 of 946 groups of copies agree; CP2K v2026.2 and cp2k-data give two different B `GTH-PBESOL-q3`; against the papers 1856 of 1926 HGH and 1126 of 1186 Krack entries agree; all 7040 valence charges agree	A	DFT-O29, W240–W252	T0
A classic table of nonlocal empirical pseudopotential band energies is not the spectrum of its printed model: of 111 registered values, 28 agree to the printed digits and 64 within the stated accuracy	A	DFT-O21, W158	T4
Audits of CALPHAD databases and solvers (Appendices N, Q)			
Frame fixed in advance (161 files, 99 papers): in 21 of 47 comparable binary databases the model the file defines disagrees robustly with at least one invariant reaction printed in the same paper (41,788 certificates); of 34 robust disagreements with a closed certified interval, most lie less than 1 K beyond the printed precision and the largest registered one 7.53 K; 12 more have no closed interval	A	CAL-B39, B40	T4
Post hoc: 16 of 32 papers; the 34 lie 8 below 0.1 K, 13 within 0.1–1 K and 13 at 1 K or more (10 from one Ca–Ga paper), so 4 of 47 databases (3 papers) differ by 1 K or more; of the other 12, supplementary certificates close 5 (up to 16.85 K), 2 stay open and 5 are rows of doubtful composition	A	CAL-B70–B72	T1; T0

Result	Kind	Ledger	Level
Rounding of the printed parameters (post hoc): for 28 of the 34, a file within half a unit of every last printed digit has a certified invariant temperature that meets the printed one; 6, in 4 databases, and a supplementary Hf–Sn case exceed the linearised rounding range (7339 certificates)	A	CAL-B76–B78	T4 certificates; T1, T0
pycalphad 0.11.2 at its default settings returns an equilibrium proven not to be the global equilibrium of the model the file defines in 21 of 63 databases (95% interval 0.22–0.46): 7 on a fixed grid (4 without points counted only through a phase-fraction filter) and 16 of 52 at points fixed in advance near the printed invariant reactions, up to 1.2×10^4 J/mol above the certified minimum	A	CAL-B41, B42, B54–B56, B73	T4 certificates; T1 verdicts
These verdicts do not depend on Python's hash seed: all 32 grid and 50 near-invariant non-global points are non-global for each of eight seeds, and none of 200 and 100 agreeing controls becomes non-global; a denser starting grid removes 49 of the 50 near-invariant non-global equilibria, and recomputed at the failing points leaves 2 of 63 databases with one outside the filter	A	CAL-B64, B65, B58, B74	T0
A development build of OpenCalphad 6.137 on the grid: a state proven not to be the global equilibrium in 16 of 54 databases (reading check revised post hoc)	A	CAL-B43	T1
Engineering decisions, addendum frozen before its runs: in 17–18 of 66 databases at least one of two decisions changes (all five margins, 2–50 K, reported); post hoc, 7 through metric 1 (5 only through rows of doubtful composition, none through a closed-interval shift) and 10 through the OpenCalphad development build (pycalphad's 2 lie among the 10); without rows of doubtful composition and without the development build, 4–6 of 66	A	CAL-B48–B51, B75	T1
NIST solder database: 4 of 26 entries of its published invariant table disagree with the database; a Pt–Pb assessment whose printed parameters do not reproduce its printed table	A	CAL-B35–B38	T4 certificates; T1 verdicts
The 17 databases shipped with pycalphad: 246,650 equilibria in 316 binary systems certified; in five systems liquid miscibility gaps that open on heating, onsets, each bracketed to 1 K, between 2177 and 5054 K	A	Appendix Q	T4
Mathematical correctness: statements floating-point software cannot give (§3.3)			
Showcase S2: a converged spin-restricted LDA solution of stretched hydrogen is not the ground state of its discrete model, certified margin 330 meV	M	DFT-M20–M26, Z06–Z08	T4
A discrete LDA problem has at least five exact solutions with pairwise distinct densities, in three groups ordered by free energy	M	DFT-X09, X13	T4
Reduced Hartree–Fock Si: the infimum of the discrete energy over all finite plane-wave sets lies in an interval $2.2 \cdot 10^{-4}$ Ha wide	M	DFT-C29–C33	T4
LDA continuum: the constant the LDA checkers use for the second-order term has no finite bound as the density bound tends to zero	M	DFT-Y05, Y06	T3
One dimension: the infinite-cutoff limit of the discrete rHF energy is the ground energy of the continuum model; the ground energy of $-\mathrm{d}^2/\mathrm{d}x^2 + x^4$ on the whole line in an interval $1.1 \cdot 10^{-16}$ wide	M	DFT-C02; Y21–Y28	T3; T4
FEM: first Dirichlet eigenvalue of the unit square in $[19.739208, 19.740200]$ and of an L-shaped membrane in $[9.632674, 9.645136]$; guaranteed Prager–Synge error bounds for the 2D Poisson problem (Appendix L)	M	FEM-B25, B26, B12	T4

Result	Kind	Ledger	Level
MD: certified distance between the computed and the unique true trajectory of Newton's equations; global existence for the 3D periodic box; FPUT solutions on the checked interval (Appendix M)	M	MD-B05, C10–C13, B33	T4; T3
Structure search: exact local minima of LJ_3 to LJ_{13}, unique up to symmetry, whose intervals contain the Cambridge Cluster Database energies (Appendix O)	M	CSP-B01, B04	T4
Software correctness and the repository (Appendices I, P)			
Frama-C/WP proves absence of run-time errors and correctness of the interval arithmetic of the trusted C kernel under IEEE 754 (159/159 and 1468/1468 goals)	S	DFT-F01–F03	T4
Four TLA$^+$ specifications model-checked, four planted parallel bugs caught; the 1D driver's logs pass trace validation and its output is bitwise identical for 1, 2, 4 and 7 threads	S	DFT-P02	T0 (target T2)
Showcase S3: on a small Si case, the MPI-parallel 3D LDA driver is bitwise identical in 25 tested layouts with between 1 and 32 processes, its 305,619 logged events pass trace validation against `PWParallel`, six tampered logs are rejected and four bugs planted in the code are caught	S	DFT-R13–R16	T0 (R14, R15 target T2)
Its certificates pass the Lean checker: a small Si case, and the S1 setting from six layouts with between 1 and 32 processes (one 253 MB certificate, checked in 7 h 53 min; mesh gap in $[0.5307364, 0.5307366]$ eV, inside the interval of the serial certificate)	S	DFT-R17, R18	T4
The S1 problem in Lean: a Mathlib-only specification of the registered problem; its pseudopotential coefficients proved to be what a total parser reads from the hashed file; a one-process checker mode whose soundness theorems conclude about this problem with no premise about another run; 13 tampered copies and a wrong problem rejected; on the archived silicon certificates, the SCF and one-sided zone certificates bound in one 8 h 42 min run and the two-sided result by a composition theorem over the 32 ranges of W66, checked by two further processes, and their combination (glue outside Lean: same file bytes, same executable, comparison of printed numbers)	M, S	DFT-K30–K38	T3; T0 (K37); T4 (K38)
859 ledger claims, 668 machine-proved, 305 down to the executed code; 36,091 library constants depend only on Lean's three standard axioms; 515 soundness statements fingerprinted; 213 test certificates (162 with known answers) accepted, 255 tampered copies and 686 adversarial cases caught; one cold run of `make verify-all` re-checks this, with fixed samples of the archived research certificates, in 99 minutes (§2.5)	S	Appendix I	–

C Motivation: How Simulations Earn Trust Today

Scientific software usually earns trust in three ways: comparison with analytic solutions, comparison with experiment, and comparison with other programs. All three work, and none of them will be replaced. But they share two weaknesses. They only cover the cases that were tested. And when a result deviates, they rarely tell us *which layer* the deviation came from. Four examples illustrate the gap: two mathematical failures, one software failure and one physical failure.

Discretisation Error That Nobody Checked (FEM, 1991). The concrete gravity base of the Sleipner A platform in the North Sea failed during a ballast test and sank. The investigation found that poorly shaped elements in the linear-elastic finite element analysis had

underestimated the shear stress in the tricell walls by about 47%, so the reinforcement was insufficient (Jakobsen and Rosendahl, 1994). The program ran normally and reported no error. *Lesson:* mesh quality and discretisation error should be checkable preconditions with explicit error bounds, not a matter of engineering judgement.

Numerical Settings That Nobody Specified (DFT, 2016). Lejaeghere et al. (2016) compared the equations of state of 71 elemental crystals computed with 15 solid-state codes and 40 pseudopotential or basis-set schemes. Recent methods agree well. Older schemes deviate noticeably, and the source of each difference could only be located by item-by-item comparison. *Lesson:* results depend on many implicit numerical settings, such as cutoffs, k-grids, smearing, convergence thresholds and mixing parameters. These settings belong in the specification, tied to an error estimate.

Concurrency Errors That Are Not Physics (MD and Beyond). A production MD code decomposes space over thousands of MPI ranks. If an atom is lost or duplicated as it crosses a subdomain boundary, the energy curve may show only a barely visible jump. The order of floating-point reductions changes with the number of ranks, so the same input gives bitwise-different trajectories on machines of different size. These errors have nothing to do with physics. They are errors in concurrent protocols, and concurrent protocols are exactly what TLA$^+$ (Lamport, 2002) and model checking are good at; industrial practice in distributed systems confirms this (Newcombe et al., 2015).

A Model Error That No Numerics Can Remove (DFT, since 1983). Kohn–Sham band gaps computed with LDA and GGA systematically underestimate measured gaps: about 0.5 eV for silicon against an experimental 1.17 eV, and germanium comes out almost gapless in LDA. Perdew and Levy (1983) and Sham and Schlüter (1983) showed that the cause is the derivative discontinuity of the exact exchange–correlation potential, which these functionals lack, and not the numerics: a computation carried out to arbitrary precision keeps the discrepancy. *Lesson:* the numerics can be entirely correct and the conclusion still wrong. Physical correctness can only be tested by observation; and only when the mathematical and software errors are rigorously bounded can a discrepancy with experiment be attributed to the model with certainty.

Three Kinds of Error, One Kind of Check. Sleipner was a mathematical error (uncontrolled discretisation), the Δ-gauge spread largely a mathematical one (unspecified numerical settings), lost atoms a software error, and the band gap a physical error in the model itself. Common practice checks all of them the same way, by comparing with experiment or with other programs, which is why a deviation cannot be located. An earlier version of this programme accordingly restricted itself to *verification*, solving the stated equations correctly, and left *validation*, whether those equations describe reality (Oberkampf and Roy, 2010), to others. We have withdrawn that restriction: the three are separated, but all three are in scope, each with its own kind of evidence and its own entries in the ledger.

D Triple Correctness in Detail

D.1 Seven Layers, Six Obligations

We split the path from the real world to a floating-point number into the seven layers of Table 3 and write an explicit obligation between each pair. The first obligation is physical: it can only be discharged by observation, never by proof. The next three belong to PDE theory and numerical analysis, the fifth to software specification and the sixth to floating-point analysis. The error budget (1) of §2 is the backbone of mathematical and software correctness: each of its terms, discretisation, algebraic and rounding error, must be backed by a theorem, and its premise, that the parallel program faithfully executes the abstract algorithm, must be guaranteed by the software-layer specification. Physical correctness adds one more layer outside (1), treated by the observation error decomposition of §2.2.

Layer	Object	Obligation to the next layer	Correctness
L0	Real world	Validation: the model's predictions agree with observations within their uncertainty, and agreement is claimed only in the range where it was checked	Physical
L1	Physical model	Well-posedness: a solution exists, is unique and depends continuously on the data	Mathematical
L2	Continuous problem	Discretisation error estimate: $u_h \to u$, with a rate	Mathematical
L3	Discrete problem	Algorithmic convergence: when the iteration stops at step k, the algebraic error is controlled	Mathematical
L4	Abstract algorithm	Refinement: every step of the parallel program implements one abstract step or leaves the abstract state unchanged	Software (TLA$^+$)
L5	Parallel program	Rounding error: distance between the floating-point and the exact-arithmetic result	Software (floating point)
L6	Floating-point execution	— (the number actually computed on IEEE 754 hardware)	

Table 3: The seven layers of a simulation and the obligation connecting each layer to the next. L0: measured band gaps, lattice constants, magnetic moments, phase diagrams, elastic moduli, diffusion coefficients, with their measurement conditions and uncertainties. L1: Kohn–Sham, Newton or Langevin equations, linear elasticity, a thermodynamic database. L2: a variational form in a function space, e.g. find $u \in H^1$ with $a(u, v) = \ell(v)$ for all v. L3: stiffness matrices, plane-wave coefficients, particle coordinates. L4: SCF, velocity Verlet or conjugate gradients in exact arithmetic. L5: MPI ranks, messages, threads, GPU streams, checkpoint and restart.

D.2 Three Kinds of Evidence, One Bridge, One Foundation

Each correctness has its own tool and its own kind of evidence: machine proof for the mathematics, specification, model checking and code-level proof for the software, and traceable observation for the physics.

The Mathematics Layer: PDE Theory and Numerical Analysis. This layer proves well-posedness (Lax–Milgram, inf–sup conditions, existence of minimisers of Kohn–Sham energies), discrete convergence (Céa's lemma, interpolation estimates, plane-wave truncation; a posteriori estimates), structure preservation (symplecticity, time reversibility, conservation of momentum and charge, crystal symmetry) and iterative convergence (SCF fixed points, Newton and Krylov methods, stopping criteria). The tool is Lean 4 with Mathlib, which already contains Hilbert spaces, measure theory and the Lax–Milgram theorem. The strategy is incremental. A theorem may first be stated with a literature reference as a registered assumption, so that upper layers can depend on it at once, and is later replaced by a machine proof. Every unproved statement is registered in the trust ledger.

The Software Layer: TLA$^+$ and Model Checking. This layer proves safety properties: atoms are never lost or duplicated, every degree of freedom has exactly one owner, parallel assembly has no write conflicts, there is no deadlock. It also proves liveness: messages are eventually delivered, and the SCF loop either converges or reports failure in finitely many steps. Its subjects are protocols: domain decomposition and halo exchange, distributed FFT transposes, checkpoint and restart, adaptive refinement and load rebalancing. Algorithms are written in TLA$^+$ or PlusCal; TLC exhausts small instances, Apalache checks symbolically, and TLAPS proves properties at any scale. The link to the code is *trace validation* (Cirstea et al., 2024): the implementation emits an event log at run time, and the log is checked

to be a behaviour the specification allows. Sequential C code that computes guaranteed quantities is proved with Frama-C/WP, and the large untrusted solvers are covered by proved certificate checkers (Appendix F).

The Physical Layer: Validation against Traceable Observations. This layer confirms that a model's predictions of observable quantities (band gaps, lattice constants, bulk moduli, magnetic moments, phase boundaries, elastic responses, transport coefficients) agree with experiment within a stated range and uncertainty. Where they do not, it reports a certified deviation, the observed value minus the rigorous computed interval, together with the measurement uncertainty (3); it is a model error only if the computed interval is for the model itself, not for a discretisation of it. It also records the model's range of validity: the compositions, temperatures, pressures and structures where the model was checked, and those where there are no data. Every experimental value used is registered with its source, measurement conditions, uncertainty and the corrections needed before it can be compared with the model (for band gaps, zero-point renormalisation and temperature; for lattice constants, zero-point expansion), and data files are hashed and versioned with the repository. The rules are fixed before any comparison is made: the quantities, tolerances and data sets are committed in advance; data used to fit a model may not be used to validate it; and a conclusion holds only in the range that was validated, with "not validated" stated outside it. Physical correctness cannot be proved, only corroborated or refuted. Measurement uncertainty is statistical, unlike the rigorous intervals of the other layers, and the two are reported separately.

The Bridge: Refinement. The abstract algorithm is a TLA$^+$ specification A whose state is, say, the global density ρ or the global phase-space point (q, p), and whose steps are one mathematical iteration each. The parallel implementation is a specification P whose state is spread over processes. A refinement mapping π projects the state of P onto that of A, for example by stitching local densities into the global ρ. Once $P \Rightarrow A$ is proved, the convergence theorems proved in Lean about A hold for the parallel program. Performance work is handled in the same way: every optimisation is a new refinement and must be proved to implement the level above.

The Foundation: Floating Point as a First-Class Citizen. IEEE 754 addition is not associative, and naive rounding-error analysis of large solvers is hopeless. The programme therefore never lets a floating-point result enter a conclusion directly. The output of a large solver goes to a proved checker that uses either exact rational arithmetic or outward-rounded interval arithmetic. The only floating-point fact we then have to trust is that round-to-nearest returns the nearest double. Global reductions use a deterministic summation order, so results do not depend on the number of processes or threads. The rounding model is part of the specification and sits in the same error budget as the discretisation and algebraic errors.

D.3 The Proof-Obligation Matrix

Table 4 instantiates the seven layers for the three domains the programme started with. It is the programme's work list: every cell must eventually correspond to a specification, a theorem, a registered assumption or a registered validation in the repository. The first row is new in this version of the programme; its first DFT entries are the Cohen–Bergstresser validation and the pre-registered LDA band-gap protocol (Appendix Q.1), while the MD and FEM cells are still empty. A fourth domain, CALPHAD, was added later (Appendix H). Its column is shorter but follows the same pattern. Its observations are measured phase boundaries, invariant temperatures and thermochemical data, to which the databases are fitted by their assessors. The model is a Gibbs-energy description per phase. Continuum theory asks for the existence and characterisation of equilibria as global minimisers. Discrete theory asks for global, not local, minimisation over compositions and, for sublattice phases, over internal states; floating point enters through logarithms and polynomials in T. The first realisation covers the cells most needed to walk one complete path from L1 to L6 in

Layer	DFT (electronic structure)	MD (atomic trajectories)	FEM (continuum mechanics)
Observation (physical)	Measured band gaps, lattice constants, bulk moduli, magnetic moments and formation energies with uncertainties (including zero-point and temperature corrections); compared with rigorous computed intervals, giving certified deviations of the model from experiment	Measured equations of state, radial distribution functions, diffusion coefficients and heat capacities; force-field fitting sets kept apart from validation sets	Measured material constants and loads; displacements, strains and failure loads in structural tests compared with computed intervals
Model	Kohn–Sham equations, a nonlinear eigenvalue problem $H[\rho]\psi_i = \varepsilon_i\psi_i$, $\rho = \sum_i f_i\lvert\psi_i\rvert^2$	Hamiltonian dynamics $\dot{q} = \partial H/\partial p$, $\dot{p} = -\partial H/\partial q$; Langevin dynamics and its Fokker–Planck equation for thermostatted runs	Elliptic, parabolic and hyperbolic PDEs, e.g. linear elasticity $-\nabla\cdot\sigma(u) = f$ and its weak form $a(u,v) = \ell(v)$
Continuum	Lieb variational principle; existence of minimisers for LDA-type Kohn–Sham models; self-adjointness of the Hamiltonian	Existence and uniqueness for Lipschitz forces; conservation of energy, momentum and angular momentum; Liouville; invariant measure and ergodicity of Langevin dynamics	Well-posedness via Lax–Milgram; inf–sup conditions for mixed problems (Stokes, nearly incompressible elasticity); regularity sets the attainable rate
Discrete	A priori error of the plane-wave cutoff; Brillouin-zone k-point integration error; pseudopotential model error registered separately	Symplecticity and time reversibility of velocity Verlet; backward error analysis: a modified Hamiltonian keeps the energy error $O(\Delta t^2)$ for exponentially long times	Céa's lemma plus interpolation: $\lVert u - u_h\rVert_{H^1} \le Ch^k\lvert u\rvert_{H^{k+1}}$; mesh shape regularity as an explicit premise; reliability and efficiency of a posteriori estimators
Algorithm (TLA$^+$)	Local convergence of SCF with Anderson/Pulay mixing; orthogonality in LOBPCG and Davidson; the SCF state machine: converge, oscillate, restart	Neighbour lists: no pair is missed while the maximum displacement since the last rebuild is below half the skin; convergence of SHAKE/RATTLE	Convergence and stopping criteria of CG/GMRES with preconditioners; convergence of the SOLVE $\to$ ESTIMATE $\to$ MARK $\to$ REFINE loop
Parallel (TLA$^+$)	k-point/band/plane-wave data distribution; distributed 3D FFT transposes lose and duplicate nothing; every process sees the same ρ after reduction	Atom migration is conserved under domain decomposition; ghost exchange is consistent; no deadlock; state is equivalent across checkpoint and restart	Parallel assembly is write-conflict free; every degree of freedom has exactly one owner; the mesh stays conforming after distributed refinement; rebalancing loses no cells
Floating point	Accumulated rounding in orthogonalisation; a bound on the deviation of the computed charge from $\int\rho = N$	Rounding accumulation over long integrations; bitwise irreproducibility from reduction order	The limit a condition number $\kappa \sim h^{-2}$ puts on attainable accuracy; quadrature error

Table 4: The obligation matrix: what each of the three original domains must validate (first row) or prove (other rows) at each layer. Every cell must eventually correspond to a specification, a theorem, a registered assumption or, in the first row, a registered validation against experimental data in the repository.

Level	Meaning	Scope
T0	Tested	Regression tests, analytic solutions, comparison with established codes; holds only for the inputs tested
T1	Paper proof	A proof on paper with a literature source and explicit assumptions, not yet machine-checked
T2	Model checked	TLC or Apalache has exhausted every behaviour of a finite instance
T3	Machine proof	A Lean, Rocq or TLAPS proof valid at any size, for an abstract model, exact arithmetic or a line-by-line model of the code
T4	End to end	Machine proof about the code that is executed, including floating-point rounding where it occurs (for Lean checkers, about the function that is compiled, with the compiler trusted); precedent: Boldo et al. (2013)

Table 5: Trust levels. Levels only go up and are never overstated: a Lean theorem containing `sorry`, or a TLA$^+$ property not yet run through TLC, is at most T1.

Grade	Meaning	Scope
V0	Not validated	No comparison with observation; the claim is only that the model gives this number
V1	Single comparison	Compared with experimental data that have a source and an uncertainty; the deviation is reported together with the computed interval and the measurement uncertainty
V2	Systematic validation	Compared on a data set and protocol fixed in advance, separate from any data used for fitting; reports the distribution of model errors and the range of validity
V3	Blind prediction	The computed interval was archived, with a time stamp and its certificate, before the measurement, and later confirmed or refuted by an independent experiment

Table 6: Validation grades for physical correctness, an axis separate from the trust levels of Table 5. A conclusion carries both, e.g. "band-gap interval T4, comparison with experiment V2".

each domain; the step from L0 to L1 is the physical track of the roadmap (Appendix H), and Appendix S lists what remains.

E Trust Levels, Validation Grades and the Trust Ledger

Not everything can be proved at once, and a programme of this kind is only credible if it says exactly what has and has not been proved. Every component therefore states which trust level (Table 5) it has reached, and lists the assumptions it depends on, in the *trust ledger*. We regard the ledger as the most important file of the project.

T0–T4 measure mathematical and software correctness. Physical correctness cannot be proved, so it is measured on a separate axis, the *validation grades* V0–V3 of Table 6. A conclusion carries both: a certified band-gap interval may be at T4 and its comparison with experiment at V2. The two axes are independent in both directions: a T4 interval can be V0 (never compared), and a widely validated model can be computed at T0 (as most production simulations are today). The combination that the observation decomposition (3) needs is a high T level for the computed interval and a V grade of at least V1 for the comparison.

The ledger is a set of Markdown tables: a root file with rules and the trusted base, and one file per domain. Each row registers a claim with a project-unique identifier (e.g. DFT-B08: domain, category C/D/A/P/I/F/…, number), the layer it belongs to, its current and target level, the evidence (a Lean theorem name, a TLA$^+$ module, a Frama-C run) and its dependencies on other claims and on registered assumptions (A-DFT-01, …). Six rules govern it:

1. **Numbering.** Identifiers are unique across the project and are cited verbatim in reports.

2. **Honest labelling.** The *current* column records only what has been achieved.
3. **Synchronised updates.** A level change is committed together with its evidence.
4. **Convention changes.** When a domain convention changes, affected claims fall back to T1 until the evidence is re-checked.
5. **Every output has a Lean guarantee.** Every number a program outputs must correspond to a Lean theorem. Where the result is an approximation (SCF iteration, discretisation, k-point sampling, finite temperature, floating point), what is proved is an error bound, not merely a test. Benchmarks, model checking and sanitizers are supporting evidence only. A feature without such a theorem is unfinished and is listed in the domain ledger's *gaps* section.
6. **Pinned statements.** The statement of every checker soundness theorem, together with the transitive closure of the project definitions it uses (stopping at the checker function, which may change freely), is fingerprinted, and `make verify-all` fails on any change. A passing proof only shows that *some* statement was proved; the fingerprint shows that it is still the statement the ledger cites, not one weakened by an added hypothesis, a looser bound or a changed definition. Intentional changes regenerate the fingerprints and state the reason in the commit message.

Rule 5 turns the ledger into a to-do list, and rule 6 keeps its entries from being met by moving the goalposts. Validation claims follow the same pattern: each states the compared quantity, the data set (frozen and hashed), the protocol and its V grade, and cites the T-level claims of the computation it rests on, and `make verify-all` re-runs the checker on the archived certificate and recomputes every deviation interval in exact arithmetic before comparing it with the ledger. Protocols can be pre-registered: a registered claim is machine-checked to be unchanged since its registration commit until it is compared (Appendix Q.1). The ledger now contains 24 validation claims (Appendix Q.1): 14 at V2 (the pre-registered Cohen–Bergstresser comparisons), four at V1 (the silicon, aluminium phosphide and gallium phosphide band gaps of showcase S1, and the entropy of aluminium in a CALPHAD database against a reference value that the database was itself fitted to) and six at V0, four of them registered and not yet compared (two of these, GaAs and AlAs, will not be computed); all other claims in Table 7 are mathematical or software claims, and implicitly at V0. Each domain ledger ends with a gaps section, a list of implementation notes (T0) and a list of assumptions together with how each is backstopped, for example by a run-time check or a cross-check. A script renders all ledgers into one HTML reading version and checks that it is in sync. Every program run prints a *trust report* whose lines cite ledger identifiers and which ends with the assumptions the run used (Appendix U). T4 claims are of three kinds: 255 theorems about checker functions that are compiled and run, among them the soundness theorems, including those that bind a certificate to its problem (exact rational arithmetic, so no floating point is involved), and seven benchmark bridge theorems that also state their public instance; 47 conclusions of particular certificate runs or archived sets of certificates (for example the S1 certificates DFT-K29, W51, W66, W300, W192, the parallel runs R17, R18, the S1 problem binding K38, the re-checked CALPHAD audit certificates and the certificates of the rounding propagation, CAL-B77), each accepted by the compiled checker on one machine; and 3 Frama-C proofs that the C interval kernel has no run-time errors and that its interval arithmetic is correct under an IEEE 754 model, two of them (DFT-F01, F02) under assumption A-DFT-06; no checker output depends on these three. The Lean/Rocq interface claim DFT-I01 was withdrawn when decision D1 chose Frama-C (Appendix G). T0 claims include the Lean definition of how a TDB file is read, a specification that can only be cross-checked (Appendix N), with its extension to interstitial and magnetic sublattice phases and the pressure-free limit of an Fe–C database, the threaded DFT implementation and the tests of the MPI-parallel three-dimensional driver (bitwise identity, trace validation, planted bugs), the reading and validation of user potentials, and, since the reliability audit of Appendix P, the unverified input and output code of the checkers: problem echoes, input gates, output escaping and a SHA-256 of the TDB file, and since the second adversarial review the common exit-code rule, complete phase tables and exact headers; the Hartree–Fock driver, its DFTK fixed-point check and its known answers; and in `mini-CSP` the searcher's known answers and the planned proved particle-swarm kernel; the problem echoes of the two-dimensional FEM and three-dimensional MD

Domain	Claims	T4	T3	T2	T1	T0	below target
`mini-DFT`	649	235	268	6	14	125	19
`mini-FEM`	56	14	37	–	–	5	–
`mini-MD`	60	26	31	–	–	3	–
`mini-CALPHAD`	83	27	22	–	17	17	3
`mini-CSP`	11	3	5	–	–	3	1
Total	859	305	363	6	31	153	23

Table 7: The trust ledger at the time of writing, by current level. The 859 claims concern mathematical or software correctness; the 24 validation claims (V grades) are listed separately in Appendix Q.1. 668 of 859 claims (78%) are machine-proved; the remaining DFT claims are at T1 (fourteen, among them the paper proofs of the continuum bounds, of the local LDA continuum argument and of the potential-space LDA continuum theorem, and the two-line corollary that bounds the minimum of the lowest conduction band of silicon over the zone) or not yet reached (one: the continuous-band output of `certify-pw`, not started), and the seventeen CALPHAD claims at T1 are the verdicts of the pre-registered database audits, of a supplementary sample, of one case outside the audit frame and of five post hoc reanalyses of the systematic audit (CAL-B70, B72, B73, B75, B78), combined from certified checker outputs by scripts outside Lean. Claims below target are listed in Appendix S.

checkers; a floating-point comparison of lower-bound methods for the quartic oscillator (Appendix J); the timing measurements of the faster two-sided zone checker, the shell of the band-location mode and its DFTK comparison; the descriptive checks of the systematic CALPHAD audit (a denser starting grid, the gas constant and the hash seed) and the floating-point parts of its post hoc reanalysis (intervals clustered by paper, the denser-grid recount, the linearised rounding ranges); the recorded runs and verdicts of the DFT code audits, whose reference intervals are T4 certificates, and their post hoc reading by each code's own tolerance (Appendix K); and the shell and tests of the checker mode that binds S1 to its Lean problem.

F Principles and Non-Goals

The thesis of Appendix D leaves much room for interpretation. The following principles fix how the programme proceeds; each was adopted because a weaker version would have let unverified numbers slip into results.

Specification before code. Every module has a mathematical statement and, where it is concurrent, a TLA$^+$ specification before it has an implementation.

Assumptions are visible. Regularity, mesh conditions, step-size limits and convergence thresholds are written into the specification, not hidden in input-file defaults. The Sleipner and Δ-gauge lessons are about exactly this.

Honest levels. Better to mark T0 than to overstate T3.

A small trusted kernel. Only a few kernels need T4; upper layers inherit their guarantees through composition theorems.

Floating point enters the specification. The rounding model and its error bounds join the discretisation error in one error budget.

Reproducible by default. The same input gives bitwise-identical results at any process or thread count, unless the user explicitly gives this up.

Correct first, fast later. First obtain a provable version, then optimise without breaking the proof; every optimisation is a new refinement.

Verify the checker, not the solver. Large numerical libraries such as LAPACK or FFT libraries are not proved. Their outputs are checked by proved checkers. Solvers are replaceable; the guarantee comes from the checker.

Every run comes with a report. Each line of a run report cites a ledger claim, from which its level and assumptions can be looked up.

Every output has a theorem. Ledger rule 5, elevated to a principle: an output without a theorem is unfinished.

The three correctnesses are kept apart. Mathematical and software conclusions carry T levels, physical conclusions V grades, and reports state them separately. Rigorous intervals and statistical uncertainties are never mixed.

Experimental data are traceable. Every experimental value used for validation has a source, measurement conditions, an uncertainty and documented corrections; fitting and validation sets are kept apart, and the comparison protocol is fixed before the comparison.

Open. Specifications, proofs, benchmarks, experimental data sets and the trust ledger are to be public (they will be released with the code upon publication), so anyone can re-check them.

The checker principle deserves emphasis because it shapes everything else. Verifying a production solver line by line, as Boldo et al. (2013) did for a wave-equation program, is a landmark, but it does not scale to codes that call LAPACK, run SCF iterations with heuristic mixing and reduce across threads. A checker, by contrast, is small and can be written for exact arithmetic. Crucially, it can target a *stronger* statement than the solver's own correctness. In all five domains of the first realisation, the checker proves a bound on the distance between the computed output and the mathematically defined target (an infimum, a classical solution, every trajectory, a global minimum, an exact local minimum). Where the target is a continuum object, the three terms of (1) are therefore discharged together, by one certified enclosure; in three-dimensional DFT the target is the discrete problem, and the discretisation term remains open.

Non-Goals. Three things the programme deliberately does not do. It does not *claim to prove a physical model*: physical correctness can only be confirmed by observation, its conclusions always carry measurement uncertainty, and they hold only in the validated range (the V grades of Table 6). The programme measures model error but does not aim to develop new functionals, force fields or constitutive laws. Versions of the charter before 8 October 2026 excluded validation altogether ("we do not judge whether the physical model is correct"); that non-goal is withdrawn, and validation is now one of the three correctnesses. It does not *replace existing software at the outset*: Quantum ESPRESSO, VASP, LAMMPS, GROMACS, deal.II and FEniCS are benchmarks and references, not targets for replacement. And it does not *prove all the code*: input parsing, I/O and visualisation may remain at T0 as long as they do not enter the error budget.

G Technical Decisions

Two decisions fix the toolchain. Both are recorded with their rationale, rejected alternatives, cost and the condition under which they will be revisited.

D1: Lean 4 for Mathematics and the Floating-Point Model; Frama-C's Floating-Point Model at the Code Level. Lean 4 with Mathlib carries L1–L4: well-posedness, discretisation error, structure preservation and iterative convergence. It also carries two parts of L6: the certificate checkers, which run in exact rationals and are compiled to executables, and a line-by-line binary64 model of the interval-arithmetic kernel. Frama-C/WP (Kirchner et al., 2015) carries the code level: the absence of run-time errors in the C kernel and, since the code-level tool was chosen on 6 October 2026, the floating-point correctness of its interval arithmetic. WP translates every `double` operation into the IEEE 754 binary64 theory (Fumex et al., 2017) of Why3 (Filliâtre and Paskevich, 2013), the ACSL contracts state over the reals what the Lean theorems state, and Alt-Ergo, CVC5 and Z3 prove every goal on the actual C code (Appendix I.2).

Rationale. The mathematics layer is the bulk of the work. Mathlib's analysis and measure theory form one unified library, while real analysis in Rocq is spread across several partly incompatible libraries; and AI-assisted proving tools mostly support Lean first. The decision was revised once and completed once. The original plan used Rocq for all of L6 and stated every discrete-layer claim twice, in Lean and in Rocq, linked by paired interface statements. Adopting "verify the checker, not the solver" shrank the floating-point obligations so much that a small binary64 model in Lean suffices. The only property used is that the rounded result is the nearest double, after which the checker steps one ulp outward. This removed the cost of double formalisation, and the code-level tool was left open until that step began. When it began, Frama-C's floating-point model was chosen over Rocq with Flocq (Boldo and Melquiond, 2011), VST (Appel, 2011) and CompCert (Leroy, 2009): Frama-C already proved the kernel's run-time safety, so one annotation file and one proof run extend to floating point without a third prover or a second C semantics; the rounding obligations are discharged automatically by SMT solvers with floating-point theories; and without Rocq the interface rule and its single claim (DFT-I01) are withdrawn. Rocq's toolchain and the T4 precedent of Boldo et al. (2013) remain the route for forward error analysis beyond the reach of SMT solvers, and for a verified compiler. *Cost.* That the ACSL contracts state the same as the Lean theorems is checked by hand and registered as a ledger assumption. The trusted base grows by WP's floating-point model, a contract for `nextafter`, Frama-C's library contracts for `sqrt`, `fabs` and `isfinite`, the SMT solvers, and one `admit` for the value of `INFINITY`, a literal that WP cannot translate. The SMT proofs depend on hand-written invariants and instantiation hints.

D2: C11, Split into a Proved Kernel and an Untrusted Driver. The trusted kernel is a few hundred lines of restricted, ACSL-annotated C: roots of unity, grid density and charge intervals, interval residual bounds and LDL^{H} inertia counting. It is proved free of run-time errors and, under an IEEE 754 model, correct on the C code (both T4). The drivers (SCF loop, LAPACK calls, FEM assembly, Verlet integration, equilibrium search, threads, I/O, reports, event logs, certificate output) stay at T0. Every number they compute must pass the kernel or a Lean checker before it appears as a guaranteed result. Python and Julia appear only in benchmarks and trace-validation tooling, outside the trust chain. Compilation forbids fused multiply–add contraction (`-ffp-contract=off`) and any value-changing optimisation such as `-ffast-math`, so the compiled arithmetic matches the proofs. The code is strict C11 (`-std=c11 -Wpedantic`) and requests the few POSIX functions it uses explicitly, so it never relies on implicit declarations, which recent GCC and clang reject.

Rationale. The mature code-level verification routes (Frama-C/WP with Why3, VST, CompCert) all start from C, and LAPACK and FFTW have C interfaces. The main risk of C, memory safety, is addressed by proof in the kernel and by AddressSanitizer, UBSan and fuzzing in the drivers (Appendix I.2). *Not chosen:* Rust, whose verifiers either have limited floating-point support or translate to Lean, and which does not connect to the code-level tools above; Julia and Python, which have no route to code-level proof; Fortran, which has few verification tools; and Lean itself as the solver language, since its `Float` has no formal semantics and no numerical ecosystem. Lean is used for the checkers precisely because they avoid `Float`.

H Roadmap

H.1 Current Goal: A Three-Dimensional Showcase of Triple Correctness

The current goal, set on 8 October 2026, is to extend DFT to three-dimensional real materials and to give, on the same systems, evidence for all three correctnesses. Each showcase puts one correctness in front while satisfying the other two (Table 8). The results so far are reported in Appendices K and Q.1; the plan, its phases and its open decisions are in the repository (`docs/plan-dft3d.md`).

The plan has phases: a pilot with a literature review, measurements of the arithmetic route at silicon scale and a floating-point census of systems with several SCF solutions; linear 3D band structures with the Cohen–Bergstresser empirical pseudopotentials, where the full

3D infrastructure is built without SCF; 3D SCF for reduced Hartree–Fock and LDA with local pseudopotentials, where a Newton–Kantorovich certificate must combine an exact low-frequency block with analytic tail estimates; nonlocal (GTH/HGH) pseudopotentials, Ewald sums, total energies and forces on real materials; and 3D LSDA with the real-material S2 example (NiO); the first S2 example, simple-cubic hydrogen, needed only phase 2 and is done. Scale is the main risk: the one-dimensional checkers use exact rationals, which do not scale to 3D, so the plan compares exact rationals, dyadic interval arithmetic proved in Lean, and the Frama-C-proved floating-point interval kernel. In the event, exact arithmetic with structure (split modes, symmetry reductions and, finally, arithmetic modulo machine-word primes) rather than floating-point intervals carried the S1 certificate for Si; the results are reported in Appendix K.

Alongside these phases runs the *physical track* P: the V grades are written into the ledger rules, each domain ledger gets a validation section, and `make verify-all` recomputes the deviation intervals; a versioned experimental data set is assembled (band gaps of common semiconductors and insulators with zero-point renormalisation, zero-kelvin lattice constants and bulk moduli, and saturation magnetisations of bcc Fe, fcc Ni and hcp Co), each value traced to its original publication; comparison protocols are committed before comparisons; and run reports get three columns, mathematical (T-level intervals), software (specifications and traces) and physical (V grades and deviations). In the one-dimensional domains, which have no direct experimental counterpart, validation is proposed only where real data exist: experimental phase-diagram points for CALPHAD, argon for MD and structural tests for FEM.

H.2 Strategy

Three strategic choices shaped the roadmap that led to the first realisation.

Walk One Complete Path First. Before broadening, the programme builds the *smallest* program that still exercises the mathematical and software layers and walks through L1 to L6 once. That program is a 1D DFT code. It is small enough that every obligation can be written down, yet it keeps the full skeleton of a Kohn–Sham calculation: a nonlinear eigenvalue problem, self-consistent iteration and plane-wave discretisation. The infrastructure built on the way (trust ledger, certificate checkers, run-report format, trace validation) is reused by every later domain.

One Dimension First, Then d Dimensions. In 1D many deep obstacles disappear or shrink. Sobolev spaces can be bypassed with Green's functions and classical solutions, plane-wave matrices are small enough for exact rational re-checking, and global minimisation over a composition interval is a branch-and-bound over a line. Each domain is therefore completed in 1D before the Lean model is generalised to d dimensions.

External Yardsticks. "Feature-complete" is meaningless without a reference. Each domain therefore adopts an established open-source code as its yardstick and independent reference on the same discrete problem: DFTK.jl (Herbst et al., 2021) for DFT, scikit-fem (Gustafsson and McBain, 2020) for FEM, pycalphad (Otis and Liu, 2017a) for CALPHAD and, for the basic MD version, ASE (Larsen et al., 2017). A feature counts as **done** only when three conditions hold: (1) it is *implemented* in a driver or checker; (2) it *agrees* with the reference code on the same discrete problem, in an automated benchmark; and (3) every number it outputs is *certified* by a Lean theorem, where possible a theorem about the true solution of the continuous problem or the true global optimum. The standard was tightened during the project. Initially conditions (1) and (2) sufficed and certificates were tracked separately; this was changed so that a feature that agrees but is not certified counts only as *aligned*, not done.

H.3 Phases

Table 9 lists the five phases. Phase 0 is the minimal DFT program. Phase 1 extends it in physics (exchange–correlation), sampling (k-points) and parallelism (MPI), with bitwise

reproducibility across process counts as exit criterion. Phase 2 transfers the framework to MD and FEM. Phase 3 closes the gap between T3 and T4 for a few performance-critical kernels. Phase 4 benchmarks against established codes on community test sets.

The phases are not executed strictly in sequence. Several Phase 1 items (k-points, a model-checked thread-parallel k-point protocol, certified LDA at finite temperature with k-points) were completed early, along with many unplanned items: finite-temperature occupations with free-energy, entropy and energy certificates; Anderson mixing; non-self-consistent bands and densities of states; and the truncation-error certificate. Phase 2 met its exit criteria before Phase 0 closed. Phase 0's literature checks are done: no published result covers the one-dimensional periodic Hartree kernel directly, so they reached their target T1 with paper proofs (Appendix S). With the interval kernel's floating-point correctness proved on the C code and the interface claim withdrawn, every Phase 0 obligation is at its target and Phase 0 meets its exit criteria.

H.4 Parity Goals

To make progress measurable, the roadmap is complemented by four parallel parity goals, each tracked feature by feature in a public table that maps every feature of the reference code to a status and the ledger claims that certify it.

G1: DFT Parity with DFTK. DFTK is a natural yardstick for DFT. It is written by mathematicians with a focus on numerical analysis, has a clean modular term structure (kinetic, external, Hartree, exchange–correlation, nonlocal, …), and can be configured to solve exactly the same 1D discrete problem. The order is 1D first (M1–M3), where Lean models change little and certificates can keep pace; then a d-dimensional Lean model for 2D and 3D (M4–M5); and infrastructure last (M6). The hardest scientific step on this path is LDA exchange–correlation. It destroys convexity, so the lower-bound certificates fail, and uniqueness becomes local; certificates must then rely on local uniqueness and a posteriori estimates, in the same way as the certificates for non-variational smearing. The first step is done: for Slater exchange and PW92 correlation the nonlinear SCF equations now have certificates of local existence and uniqueness at any finite temperature, at the Γ point and with k-points, with enclosed energy, free energy, band structure and density of states (Appendix J). Collinear spin is complete (DFTK's `Magnetic` vector-potential term aside): LSDA with spin-polarised Slater exchange or spin-interpolated PW92 correlation in a Zeeman field is certified at the Γ point and with Monkhorst–Pack k-points, at finite temperature and at zero temperature under a common gap, with enclosed (at zero temperature, exact integer) magnetisation, energy and free energy, and per-spin bands and densities of states. Exact exchange (closed-shell Hartree–Fock) and forces have started, at finite temperature at the Γ point and on Gaussian sites for rHF respectively. That uniqueness really is only local is itself certified: one discrete LDA problem has at least five distinct exact solutions (Appendix Q).

G2: Certified Finite Elements against scikit-fem. The FEM goal goes further than reproducing scikit-fem. Wherever possible the certificate must speak about the *true solution of the continuous problem*, with its discretisation error proved. In 1D this is achieved with Green's functions and classical solutions (Appendix L). The step to 2D triangular meshes (F2) was expected to require a Lean model of H^1, which Mathlib lacks, and guaranteed a posteriori upper bounds without unknown constants, for instance by equilibrated fluxes in the spirit of Prager and Synge (Prager and Synge, 1947). The first 2D results avoid the Sobolev space: the error bound is stated in dual form against smooth compactly supported test functions, and eigenvalues are defined by Rayleigh quotients of Lipschitz functions (Appendix L). The validity of the mesh, the lesson of Sleipner, is a premise verified by the checker; its shape regularity affects only the size of the bound, not its validity.

G3: Certified CALPHAD against pycalphad. CALPHAD was not among the original three domains. It was added because it shows a different kind of certified statement. CALPHAD solvers find local minima of non-convex Gibbs energies, and missing a phase is a classic failure mode. A certificate can instead establish the *global* equilibrium of the database model, which no conventional solver claims. The binary case (C0–C1) can be

certified by branch and bound on an interval. Multicomponent and sublattice models (C2) require branch and bound over products of simplices. Sublattice phases with two mixing sublattices are the first step: the Gibbs energy of such a phase at a given composition is a minimum over internal states, and the checker already runs its branch and bound over the square of internal states (Appendix N).

G4: Certified MD against ASE. For MD a parity goal analogous to G1–G3 was set on 13 October 2026, with ASE 3.29.0 as the reference on the same potential, cell, initial state and step. Its first two milestones are the one-dimensional chain of `mini-MD` and a cubic periodic box with the minimum-image convention and a smoothly switched Lennard-Jones cutoff (Appendix M). ASE's own smooth cutoff is a cubic polynomial in r^2 and only C^1, which the defect bounds of the MD certificates cannot use, so our C^4 potential is wrapped as an ASE calculator for the comparison and the item counts as aligned. The next MD steps follow Table 4: neighbour lists with the half-skin criterion, general cells, thermostats, and the TLA$^+$ modules for atom migration and halo exchange.

I First Realisation: Shared Infrastructure

This and the following appendices report the first realisation of the programme: Phase 0, the parts of Phases 1 and 2 completed so far, the first milestones of the four parity goals and the three-dimensional DFT work towards the showcases. The repository mirrors the layers. `spec/math` holds the Lean 4 development: discrete models, error-bound theorems and the certificate checkers of all five domains. `spec/tla` holds the TLA$^+$ specifications, and `trace` the trace-validation tooling. `kernels` holds the C11 code, split into the trusted kernel and the untrusted drivers. `bench` holds benchmarks and cross-checks (T0, outside the trust chain), and `domains/*` holds each domain's specification, conventions, ledger and parity table. One command, `make verify-all`, re-checks the whole repository. A complete cold run with 32 parallel jobs on one workstation (192 cores, shared with other jobs) took 99 minutes on the final state of the repository; cold means that no result cache is read, while the Lean build output is reused, so Lean modules are recompiled only where their sources changed. From a fresh clone of the repository one day earlier (before the K38 archive and its check were added), on 32 cores of the same machine, cloning, compiling the drivers, setting up the MPI environment and unpacking Mathlib's published build cache took 11 minutes, building the library (755 modules; Mathlib's modules come from its cache and are not re-elaborated) 2 h 40 min, and the full check, which also builds the remaining modules and the eleven checker executables (compiling Mathlib's cached C code to object files for linking), 2 h 04 min: 4 h 54 min in all, without the download of the compressed Mathlib cache (under 2 GB), which was already on disk; the build output takes 17 GB. It compiles the drivers and brings the 776 Lean modules and the eleven checker executables up to date. It audits axioms: every constant of the library (36,091) and of the entry modules of all eleven checker executables (975) may depend only on `propext`, `Classical.choice` and `Quot.sound`, any `sorry` or project axiom is reported, and no compiled function may be replaced by other code (`implemented_by`, `extern`, `unsafe`); the step fails if an executable declared in the build file is missing from the audit list.

It checks that the ledger is in sync and its cross-references resolve, that the fingerprints of the 515 pinned statements are unchanged (ledger rule 6), and runs TLC on every specification. For every domain it generates small certificates (213 in all), requires the Lean checkers to accept them, and requires each to have a tampered copy or a wrong problem that is caught (255 in all); 162 of them carry known answers (Appendix P). The `PROBLEM` lines of the DFT, FEM, MD and CALPHAD checkers must echo the problem the command line asked for, a driver that solved for a different external potential must be rejected by the problem binding, and every DFT certificate must be rejected once its `pot` line is removed (Appendix J); a convexity case must also be rejected at a temperature where the phase has a miscibility gap. It re-checks the archived certificates of the showcases (Appendix P) and of the research results (Appendix Q) with the current checkers. It also runs trace validation, report validation, a one-minute sanitizer fuzzing run of all four drivers together

with fixed regression inputs from earlier fuzzing, differential tests of the trusted kernel and Frama-C/WP.

I.1 Certificate Checkers

Figure 2 shows the architecture shared by all five domains, in the tradition of certifying algorithms (McConnell et al., 2011). The checker is an ordinary Lean function `check : Cert -> Option Output`, and its soundness theorem states that if `check c = some o` then the mathematical object defined by the data in `c` lies within the bounds in `o`. Floating-point numbers from the driver enter only as exact dyadic rationals; all arithmetic is in $\mathbb{Q}$. Transcendental constants and functions appear only through proven rational enclosures: π through Mathlib's 20-digit bounds, log through an artanh series with range reduction, exp, sin and cos through truncated series whose remainders are bounded by Mathlib's `Complex.exp_bound'`, and the error function through alternating Taylor bounds. Decimal output rounds lower bounds down and upper bounds up. The checker is compiled to a native executable (`certify`, `certify-pw`, `certify-pwnl`, `certify-pw3d` and `certify-osc` for DFT, `certify-fem`, `certify-md`, `certify-md3d`, `certify-mdlad`, `certify-calphad` and `certify-csp`). Soundness theorems are proved about the very function that is compiled, so they are recorded as T4, with the Lean compiler, the certificate parser and the decimal printer in the trusted base (Appendix T). Their statements are pinned by fingerprints (ledger rule 6): a checker may be rewritten or optimised freely, but the theorem it must satisfy cannot change without a recorded decision.

The solver is therefore replaceable: LAPACK's `zheev`, a Thomas or banded LU solve, a Newton iteration or a convex-hull heuristic can be swapped without touching any proof. As argued in Appendix F, the certified statement is also stronger than "the floating-point computation is close to its exact-arithmetic counterpart". It concerns a mathematically defined target that the solver never computes exactly.

I.2 The Software Layer and the Trusted Kernel

TLA$^+$ and Trace Validation. The SCF controller of `mini-DFT` is specified in TLA$^+$ (Lamport, 2002) as `SCFLoop`. It terminates, and it reports convergence only if the residual is below tolerance, the gap is open and the certificate check passed. TLC explores all behaviours for iteration limits $\{1, 4, 10\}$ (97 states for limit 4). Mutation testing confirms that dropping the certificate check violates `ConvergedIsJustified` and that removing the iteration limit deadlocks. The shared-memory k-point parallel protocol `KPointParallel` is model-checked for 1–3 threads and 1, 3 and 4 k-points; it guarantees that every k-point is claimed once per phase and that reductions happen in a fixed order after all contributions are written. The driver logs its controller state, and each logged run is replayed against the specification via trace validation (Cirstea et al., 2024). Runs ending in all four outcomes (`converged`, `gap_closed`, `max_iter`, `cert_failed`) are accepted, while tampered traces are rejected. The threads log their events as well: every claim, completed k-point, exit and join, the chemical-potential step and each step of the ordered reduction is one action of `KPointParallel`, and TLC replays the log while checking the protocol's safety properties. Logs from 1, 4 and 7 threads are accepted. Double claims, claims out of counter order, an out-of-order reduction, a join before all threads have exited and a missing chemical-potential step are rejected. A claim is logged under the same lock as its atomic increment, so the log order is the linearisation order; the atomicity of the counter itself is checked only by ThreadSanitizer. The output is bitwise identical for 1, 2, 4 and 7 threads.

Showcase S3: A Parallel Three-Dimensional Driver. The three-dimensional parallel protocol `PWParallel` (pools of k-points claimed through an atomic counter, band groups within a pool, plane-wave processes within a group) and the slab transpose `FFTTranspose` of the distributed FFT are model-checked on finite instances (DFT-R08–R11, T2): on every full check `PWParallel` with two pools, three k-points, three bands and two FFT planes, and `FFTTranspose` with $N = 2$ and at most two processes per direction, and by hand up to three pools or $N = 3$; for any number of processes there is only a paper argument (R12, T1). An MPI driver now implements `PWParallel` and the slab case of `FFTTranspose` for the

zero-temperature point-group LDA problem of S1 (R13). Pools claim irreducible k-points with `MPI_Fetch_and_op` on a counter at rank 0, bands are dealt to groups cyclically, and the FFT processes of a group hold slabs of planes and transpose with one all-to-all. Every reduction runs in the order the specification fixes: a k-point's density by band index, the total density by k index and the energy by plane index, with no `MPI_Reduce`; one process runs the same algorithm, not a separate serial path. For every protocol event a process takes a ticket from a second atomic counter on rank 0 before the event's effect becomes visible to other processes, a claim in the same exclusive lock epoch as the claim itself, so ordering the merged log by ticket gives a linear extension of the causal order, and TLC replays it through a trace specification that instantiates `PWParallel` and checks the protocol's safety properties along the way (R14). On real Si (GTH-LDA q4, $Q = 17/2$, a $4 \times 4 \times 4$ mesh, 12^3 grid), all 25 tested layouts with 1, 2, 3, 4, 7, 8, 16, 29, 30 and 32 processes, uneven slabs included, give the same SHA-256 hashes of the eigenvalues, density, energy, certificate centre, the trajectory of all iterations and the certificate file (R15); their 305,619 logged events are accepted, and six tampered logs (a duplicated or swapped claim, two reduction orders, a k-point released early, a missing chemical-potential step) are rejected. Four bugs planted in the code are each caught (R16): a non-atomic claim only by trace validation, since the duplicated k-point is merely recomputed; a reduction in arrival order and a fold by band group by both tests, although the SCF converges as usual ($\|G\|$ about $3\text{–}4 \cdot 10^{-12}$), so the hash comparison sees rounding-level differences; and a transpose that loses its own block only by the hash comparison, a coarse control in which the SCF no longer converges, since the FFT is not in the trace. The certificate of this case passes the Lean checker (R17, T4). At the S1 setting itself (the problem of the serial S1 certificate, $8 \times 8 \times 8$ mesh, 29 irreducible k-points, 24^3 grid), six layouts with 1 to 32 processes give the same 253 MB certificate in 8 to 27 minutes of wall time, and their 223,690 logged events are accepted; the certificate's problem header is byte-identical to that of the serial certificate, and its Lean check passes in 7 h 53 min on 32 cores (184 CPU hours, peak 85.5 GiB), with the mesh gap in $[0.0195042035, 0.0195042094]$ Ha ($[0.5307364, 0.5307366]$ eV), inside the serial certificate's interval (R18, T4); each certificate proves uniqueness only in its own ball, and we do not claim that the two solutions coincide. The certificates are T4 whatever the driver does; the parallel evidence itself is tests (T0) and covers the executions that were run, on one machine and one library build. The pool root runs the dense eigensolver, so the band groups validate a residual loop, not a distributed eigensolver, and the transpose, Jacobian and refinement phases are outside the trace.

The Interval Kernel. Eigenpair residual bounds and inertia counts during SCF are computed by a small C kernel (about 500 lines) in outward-rounded interval arithmetic. Frama-C/WP (Kirchner et al., 2015) proves, on the C code, both the absence of run-time errors and the correctness of the interval arithmetic under WP's IEEE 754 model (159/159 and 1468/1468 goals for the two source files, about five minutes on a 384-core machine; `DFT-F03`, `F01`, `F02`, T4). The contracts are stated over the reals: each interval operation returns an interval containing the exact sum, difference, product, quotient or square of any reals in its inputs, or the whole line when a step overflows; the charge interval contains $2w \sum_i f_i \|c_i\|^2$ for the floating-point orbitals; and the residual bound is at least $\|Hc - \varepsilon c\|_2$. For the inertia count, Frama-C proves that at every step the intervals enclose an exact symmetric-pivoting Schur-complement elimination of $H - \sigma I$, written as an ACSL logic function, that every exact pivot is non-zero, and that the returned count is the number of negative exact pivots; the Lean theorem `exactElim_inertia` turns this into the eigenvalue count via Sylvester's law of inertia (A04). The proof assumes a contract for `nextafter` (the next double towards $\pm\infty$, as in C11) and the value of `INFINITY`; that the ACSL and Lean statements agree is checked by hand. The same properties are also proved in Lean against a line-by-line binary64 model of the kernel (T3), and a differential test compares that model with the C code bit for bit on 900 random inputs.

The Untrusted Drivers. Verifying the checker rather than the solver makes most driver bugs harmless: they make a certificate fail, not pass. Two pieces of driver code nevertheless stay on the path from input to claim. The first is the report. With `--verify` a driver runs its checker and copies the output into the report, prefixed with `Lean`. A black-box test checks, for every domain, that these lines are verbatim checker output, that no checker token

appears elsewhere, that a rejecting, crashing or missing checker makes the driver exit with an error, and that an unconverged run writes no certificate. It found that two drivers exited successfully when the checker failed or could not be started, and that all four split checker lines longer than 511 bytes and inserted a second prefix in the middle. The second is the problem statement. A checker certifies the problem written in the certificate, and the driver produces that problem from the command line or from a TDB database. We fuzzed the four drivers under AddressSanitizer and UBSan with about 43,000 mutated inputs. Besides a heap overflow when a TDB file is read from a pipe, an unbounded allocation for very large k-point grids and non-termination on NaN chemical potentials or sub-ulp temperature steps (and, in a longer run on 8 October 2026, an out-of-bounds write in the CALPHAD phase-diagram mode, now a fixed regression input), the fuzzer found inputs for which a driver wrote a certificate for a different problem, which the checker then accepted. In CALPHAD, `3*T**1.5` was read as $3T + 0.5$, coefficient tokens were truncated in fixed-size buffers (-30000 became -300), overflow in exact site-ratio arithmetic dropped a phase, and command-line tokens such as `0.25*2` meant one number to C and another to the Lean parser. In FEM, a long right-hand side was truncated and NaN parameters were written as zero. All drivers now parse numbers strictly and reject every input they cannot translate exactly, and their output for valid inputs is byte-identical to before. The stricter TDB parser also showed that three of pycalphad's test databases had been misread, for example through a parameter written with an abbreviated keyword, while four that cannot be translated exactly are now rejected. Later, the database audit of Appendix Q found one more such case: molecular species made only of A and B atoms, such as Cu_2 in COST 507 (Ansara et al., 1998), were dropped like unrelated elements, so one phase was certified as a stoichiometric compound it is not. These checks are T0 and run in `make verify-all`.

The structural remedy is for the checker to rebuild the problem from the raw input instead of trusting the driver's translation. This is now done for the two largest pieces of problem data. The DFT checker compares the external-potential coefficients with a Lean definition of the potential named in the certificate (Appendix J), and the CALPHAD checker builds the model from the TDB file itself (Appendix N); both soundness theorems are then statements about the named problem. For S1 the whole registered problem is a Lean definition that the checker compares with the certificate exactly (Appendix K); on the archived S1 certificates the SCF and one-sided zone checks are bound to it in one process, and the two-sided result by a composition theorem whose glue (same file bytes, same executable, comparison of printed numbers) lies outside Lean (DFT-K38). What remains trusted is that these Lean definitions say what is intended. The remaining DFT parameters (cell length, cutoff, electron count, temperature, smearing, functional, grid) are echoed exactly and compared with the command line, which is again T0; FEM and MD checkers echo their problems in the same way (Appendix P).

J Density Functional Theory in One Dimension

This appendix treats the one-dimensional code `mini-DFT`; the three-dimensional checkers built on it follow in Appendix K.

Model. `mini-DFT` solves the one-dimensional periodic reduced Hartree–Fock model on $[0, L)$ with N spin-paired electrons. For a one-body density matrix $0 \le \gamma \le 1$ with $\operatorname{Tr}\gamma = N/2$ and density $\rho_\gamma(x) = 2\gamma(x, x)$,

$$E(\gamma) = \operatorname{Tr}\big(-\tfrac{1}{2}\Delta\gamma\big) + \int V_{\mathrm{ext}}\,\rho_\gamma + \tfrac{1}{2}D(\rho_\gamma, \rho_\gamma), \qquad D(\rho, \rho) = \sum_{q \neq 0} \frac{4\pi}{q^2}\,|\hat{\rho}_q|^2. \tag{4}$$

Orbitals are expanded in plane waves $e^{i(q+\kappa)x}$, $q = 2\pi k/L$, $|k| \le K$, at the Γ point or on Γ-centred or Monkhorst–Pack (Monkhorst and Pack, 1976) k-grids. E_K denotes the discrete ground-state energy (an infimum). rHF is the cleanest member of the Kohn–Sham family: E is convex in γ and strictly convex in the nonzero Fourier modes of the density, so the ground-state density is unique. The solver is a self-consistent field (SCF) iteration with linear, Kerker-preconditioned or Anderson (Anderson, 1965) mixing and dense LAPACK diagonalisation.

Discrete Theory (T3). In Lean we prove that a discrete ground state exists (D01) and that its density is unique (D02). We also prove Galerkin monotonicity, $K_1 \le K_2 \Rightarrow E_{K_2} \le E_{K_1}$ (D03), and absence of aliasing: with $M \ge 4K+1$ grid points, densities and Hamiltonian applications computed on the grid coincide with their Galerkin definitions (D04, D05). For eigenpairs, a Hermitian residual bound places an eigenvalue within $\|Hv - \varepsilon v\| / \|v\|$ of ε (A03), and Sylvester's law of inertia (A04), proved from the spectral theorem, identifies *which* eigenvalue it is from the signs of an LDL^{H} factorisation of $H - \sigma I$.

Energy Certificate. The lower bound combines convexity with weak duality of a semidefinite relaxation. If $Z \succeq 0$ and $H[\gamma] - \mu + Z \succeq 0$, then

$$E(\gamma) - 2\,\mathrm{Re}\,\mathrm{Tr}(H[\gamma]\gamma) + 2(\mu N_{\mathrm{pair}} - \mathrm{Tr}\,Z) \;\le\; E_K \qquad \text{(DFT-B02)}. \tag{5}$$

Positive semidefiniteness is certified from a floating-point Cholesky factor R through $\|A - \eta - R^{\mathsf{H}}R\|_F \le \eta \Rightarrow A \succeq 0$ (B05). The upper bound rounds the approximately idempotent computed G to an admissible projector Q with $\|Q - G\|_F \le 2\|G^2 - G\|_F$ (B03, B04). The checker `checkFast` is proved equal to a specification `check` and sound (B08, T4). Intervals for E_K have widths of $6 \cdot 10^{-10}$ to $9 \cdot 10^{-9}$ and take about 0.8 s to check at $K = 16$.

Discretisation Error. Upper bounds are monotone in K, so $E_\infty := \lim_K E_K \le E_K$. For a lower bound on E_∞ valid for every $K' \ge K$, the high-frequency block satisfies $H - \mu \succeq \delta$, with δ computed from $\frac{1}{2}q_{K+1}^2$ and the ℓ^1 norm of the potential coefficients, and its coupling to the low block is absorbed by a Schur complement into a diagonal shift (L02–L04). The certificate therefore bounds the truncation error $0 \le E_K - E_\infty$ rigorously. For example, $E_K - E_\infty \le 5.9 \cdot 10^{-10}$ at $K = 8$ for a double-well potential. For the Gaussian double well the external potential's Fourier tail is also enclosed, so the bounds hold for the true Gaussian potential rather than its truncation (V05). That E_∞ is the ground energy of the continuous problem is a Lean theorem (C02, T3). We write the continuum model in Fourier space, as density matrices γ_{kl} on $\ell^2(\mathbb{Z})$ with $0 \le \gamma \le 1$, $\mathrm{tr}\,\gamma = N_{\mathrm{pair}}$ and finite kinetic energy, which avoids periodic Sobolev spaces; the energy is bounded below for absolutely summable $\hat{V}$ (C03), and zero-padding embeds every discrete state with the same energy, so $E \le E_K$ (C04). Conversely, for any continuum state γ the trial state $(1-c)P_{K'}\gamma P_{K'} + cP_{K_0}$, with c at most the trace τ outside $|k| \le K'$, is admissible for K', and its density coefficients differ from those of γ by at most $(2/L)(\tau/\eta + \eta N_{\mathrm{pair}} + \dots)$, an entrywise AM–GM estimate $|\gamma_{kl}| \le (\eta\gamma_{kk} + \gamma_{ll}/\eta)/2$ that replaces the trace-norm argument of the paper proof (C05–C07). For $N_{\mathrm{pair}} \le 2K+1$ this gives $E_{K'} \to E$ without the existence of a continuum minimiser (C08), so $\inf_{K'} E_{K'}$ equals the continuum infimum, and the `LIMIT_OK` interval also contains the continuum energy (C09, T3). That this matrix model is the usual operator model on H^1 is an assumption (A-DFT-08); the named Gaussian and site potentials are not yet linked to the summability hypothesis, and finite temperature and k-points remain.

Finite Temperature. For Fermi–Dirac and Gaussian smearing the checker encloses the free energy $F_{T,K} = \inf_\gamma E(\gamma) - 2T\,S(\gamma)$. The lower bound relies on a trace inequality whose proof uses the doubly stochastic overlap matrix between two eigenbases (T03), together with an approximate-unitary lemma proved via the polar decomposition (T05). The upper bound uses Jensen's inequality for the concave entropy. Interval widths are $4 \cdot 10^{-11}$ to $2 \cdot 10^{-10}$. Entropy and energy follow from three free-energy certificates at T and $T \pm h$ through proven slope bounds. We also prove existence of finite-temperature ground states and uniqueness of their density.

Non-Variational Smearing. Methfessel–Paxton (Methfessel and Paxton, 1989) and Marzari–Vanderbilt (Marzari et al., 1999) occupations can exceed 1, so the problem has no variational principle and "ground state" is undefined. For these schemes the certificate instead proves, Newton–Kantorovich style, that the discrete SCF equations have exactly one solution in a ball of radius r around the driver's approximation. Matrix-function differences are expressed through mixed-basis divided-difference operators of Daleckii–Krein type, which are Lipschitz in both spectral decompositions (T20–T25). Typical radii

are $r \approx 10^{-12}$. The energy and free energy of that solution are then enclosed, the latter to second order using stationarity: about $5 \cdot 10^{-14}$ wide at $K = 8$. The argument only needs bounds on f', on the Lipschitz constant of f' and on $|xf'(x)|$, and an entropy with $s' = xf'$. For Methfessel–Paxton of order N all of these are Hermite polynomials times e^{-x^2}; since $(H_n e^{-x^2})' = -H_{n+1}e^{-x^2}$, the derivative identities hold term by term for every N at once, and the checker computes the bounds from the Hermite coefficients using $x^{2m}e^{-x^2} \le (m/e)^m$ (T30, T31). For orders 2 to 4 the radii and free-energy widths stay at 10^{-12} and $4 \cdot 10^{-14}$–$5 \cdot 10^{-13}$.

LDA Exchange and Correlation. With the local density approximation the energy gains $E_{\mathrm{xc}} = \int e(\rho_\gamma)$ and the Hamiltonian the potential $v(\rho_\gamma) = e'(\rho_\gamma)$. `mini-DFT` evaluates both pointwise on the grid $x_j = jL/M$ with $M = 4K + 1$, the discrete problem DFTK solves with an FFT grid of size M. With Slater exchange (Slater, 1951) and PW92 correlation (Perdew and Wang, 1992), energies, all components including exchange–correlation, and all eigenvalues agree with DFTK at zero temperature, for Fermi–Dirac, Marzari–Vanderbilt and Gaussian smearing, on Monkhorst–Pack grids and for bands and densities of states. The energy is no longer convex, so there is no variational certificate. Instead, the SCF-solution certificate of the previous paragraph is generalised to a residual that is nonlinear in the unknown: the grid Parseval identity bounds the exchange operator by grid averages of v (X01), the grid density is Lipschitz in the unknown (X02), and Newton–Kantorovich is applied around the linearisation, with first- and second-order bounds that use $|v'|$ and the Lipschitz constant of v' (X03–X05). For Slater exchange these constants have closed forms whose cubes are rational, e.g. $|v'|^3 = 1/(9\pi\rho^2)$ (X07). PW92 correlation contains $\log(1 + 1/Q)$, with Q a polynomial in $s = \sqrt{r_s} = (3/(4\pi\rho))^{1/6}$. In the variable s, every derivative of the logarithm multiplied by a power of s is a rational function that is regular at $s = 0$, so that $v_c' = -(2\pi/9)\, s^6\, (sW)$ and v_c'' are polynomials in s, sL (with $0 \le sL \le 1/q(s)$) and these rational functions (X14). The checker encloses v_c, v_c' and e_c at the grid points by rational interval arithmetic with checked bounds for sixth roots and logarithms, and bounds $|v_c'|$ and $|v_c''|$ for all $\rho \ge \mathrm{lo}$ by splitting $[0, s(\mathrm{lo})]$ into 64 pieces (X15); the bounds exceed the true suprema by 10–25%. At the Γ point with any finite-temperature smearing (Fermi–Dirac, Gaussian, Marzari–Vanderbilt, or Methfessel–Paxton of any order), with Slater exchange alone or with PW92 correlation, the checker proves that the LDA SCF equations have exactly one solution within $r \approx 10^{-12}$–$8 \cdot 10^{-12}$ of the computed one (X09, T4) and encloses its energy, exchange–correlation included, and its free energy (X13, T4). At $K = 8$ the energy interval is $3 \cdot 10^{-9}$ to $2 \cdot 10^{-8}$ wide, tens of times wider than for rHF because of the extra first-order term; the free-energy interval is 10^{-15} to 10^{-12} wide, and checking takes 6–14 s (Fermi–Dirac is slowest, because of the rational bounds on $e^{-|x|}$ and log). DFTK's Fermi level, energy and free energy lie inside the Lean intervals, for both functionals. For Fermi–Dirac smearing the occupation $1/(1 + e^x)$ has its own derivative bounds, and its entropy is proved equal to the binary entropy used by the variational certificates (X16, X17). With k-points the same certificate applies to the equivalent supercell (next paragraph). At zero temperature the occupation is a step and the derivative bounds fail. Under a gap, however, a smooth step $f = p \circ \mathrm{clamp}_{[-1,1]}$ with a quintic p whose first two derivatives vanish at ± 1 equals the step outside $(-1, 1)$ (X21); the checker runs the same SCF-solution test with f, the gap half-width as T and μ_0 mid-gap, and also checks $|\tilde{a}_i| \ge 1 + h_0 + n_X' r$ for the eigenvalues of the centre. By Weyl counting no $X(y)$ in the ball then has an eigenvalue in $(-1, 1)$, so on the ball the smooth problem *is* the zero-temperature problem: the unique solution is the spectral projector of $H_{\mathrm{L}}[\gamma^*]$ below the gap, with trace N_{pair}, and its energy is enclosed (X22, X23; checker X24, T4; k-points X25, T4). For a double well at $K = 8$ the radius is 2.5–$5 \cdot 10^{-13}$, the energy interval 10^{-15} wide, and checking takes 4–6 s; a metal at zero temperature has no such certificate. Since uniqueness is only local, several certificates for the same problem can certify several solutions; Appendix Q uses this to prove that one problem has at least five.

***k*-Points.** Pinching a density matrix to its Bloch-block-diagonal part keeps it admissible, does not increase the energy and does not decrease the entropy (K01). This yields an exact

supercell equivalence: $E_{\text{supercell}}/n_k$ equals the k-sampled ground-state energy, for odd Γ-centred grids and, using half-integer frequencies, for even Monkhorst–Pack grids (K03, K06, K07). The energy, free-energy and SCF-solution certificates therefore extend to k-point runs, and for odd grids so does the infinite-cutoff limit (L06). For LDA the energy is not convex and pinching gives nothing, so the equivalence is proved differently. The certificate describes the supercell on a grid of $n_k M$ points, on which a shift by one cell is a shift of the grid. The exchange–correlation potential is therefore equivariant, $V_{\text{xc}}(\tau y) = S\, V_{\text{xc}}(y)\, S^{\mathsf{H}}$, so the shift of a solution is again a solution; the checker verifies that the computed centre is itself cell-periodic, so the shifted solution lies in the same ball, and local uniqueness forces the certified solution to be cell-periodic. Its Bloch blocks then satisfy the k-point LDA SCF equations, and the supercell energy is n_k times the per-cell energy with exchange–correlation integrated on the cell grid, as in DFTK (X18, X19). The checker `checkLDAK` adds these structural checks to the Γ-point checker and divides by n_k (X20, T4). Because the supercell needs $n_k M \geq 4K_s + 1$ grid points, certified runs use $M = 4K + 3$ rather than DFTK's default $4K + 1$; DFTK on the same grid lands inside the intervals. For $n_k = 2$–4 and $K = 4$ the radius is 2–$9 \cdot 10^{-12}$, the per-cell energy interval 10^{-9}–$2 \cdot 10^{-8}$ wide and the free-energy interval $3 \cdot 10^{-15}$–$8 \cdot 10^{-14}$ wide; checking takes 9–53 s.

Collinear Spin. In collinear-spin LSDA, DFTK's `spin_polarization = :collinear`, each spin $s \in \{\uparrow, \downarrow\}$ has its own density matrix $0 \leq \gamma_s \leq 1$, without the spin factor 2, and grid density ρ_s, and the total density $\rho_\uparrow + \rho_\downarrow$ enters the Hartree term. The first, simple version has spin-polarised Slater exchange only, $v_s = -(6\rho_s/\pi)^{1/3} = v_x(2\rho_s)$, which is separable in the two spins, and a uniform Zeeman field B that lowers the energy of spin ↑ by B and raises that of spin ↓ by B. The two spins share one chemical potential, the total electron number N is fixed and the magnetisation $\nu = \operatorname{Tr}\gamma_\uparrow - \operatorname{Tr}\gamma_\downarrow$ is free:

$$\gamma_s = f\big((H[\rho_\uparrow + \rho_\downarrow] + V_{\text{xc},s}[\rho_\uparrow, \rho_\downarrow] - \mu - \operatorname{sgn}(s)\, B)/T\big), \qquad \operatorname{Tr}\gamma_\uparrow + \operatorname{Tr}\gamma_\downarrow = N. \tag{6}$$

With the unknown $y = (u_\uparrow, \nu, u_\downarrow, \mu)$, where u_s collects the Fourier coefficients of the spin potentials, the SCF map $X(y) = \operatorname{diag}(X_\uparrow(y), X_\downarrow(y))$ is block diagonal. Matrix functions and Schur products of block-diagonal Hermitian matrices act block by block, and the Newton–Kantorovich argument of X03 holds for any index type with a linear projection and an affine right-hand side (`exists_unique_scf_abs`), so the two-spin problem reduces to the existing one-spin machinery. The new estimates are the first- and second-order bounds of $X(y)$: a Parseval bound over both spins, in which the zero-frequency components $\pm\nu/(2L)$ of the spin densities move with ν, and Lipschitz and Taylor bounds of the exchange potential on $[\text{lo}, \infty)^2$ (X26, T3). In a ball in which both spin densities stay above lo, the LSDA SCF equations then have exactly one solution, with common μ and free magnetisation (X27, T3). The checker for `fis-lsda 1` certificates treats each spin as an LDA certificate with potential coefficients $u_\uparrow + u_\downarrow$ and chemical potential $\mu_0 + \operatorname{sgn}(s)\, B$, uses a preconditioner of order $2(4K + 1)$, and prints intervals of radius r for ν and μ (X28, T4). On four systems, cosine potentials and free electrons at $K = 6$–8 with Fermi–Dirac smearing, the radius is $3.3 \cdot 10^{-14}$–$9.9 \cdot 10^{-13}$ and checking takes 2.7–6.1 s. DFTK 0.8 with collinear spin, Slater exchange and the Zeeman field written as a spin-dependent external potential gives ν and ε_F inside the Lean intervals, at most $8.2 \cdot 10^{-14}$ from the driver's values. For free electrons the density is uniform, and an independent arbitrary-precision solve of the two occupation equations gives ν and μ inside the intervals; at $B = 0$ the ν interval contains 0 and, since $v_s(\rho/2) = v_x(\rho)$, the μ interval meets that of the unpolarised LDA certificate. With Monkhorst–Pack k-points the driver solves the equivalent supercell: the supercell LSDA map commutes with translation by one primitive cell for any exchange-correlation potential, so the unique solution in the ball is cell-periodic (X35, T3), and each spin's density matrix is block-diagonal over the Bloch groups, $\gamma_{s,j} = f(X_{s,j})$, with the supercell density the k-average of the primitive-cell densities (X36, T3). The k-point checker adds the conditions that the supercell comes from the primitive-cell problem, that both spins' centres are cell-periodic and that n_k divides M and N, echoes the primitive cell's L, K, N and M, and prints the per-cell magnetisation (X37, T4); for a cosine potential at $L = 4$, $K = 4$, $N = 2$, $T = 0.05$, $B = 0.1$ the per-cell ν interval has width $1.6 \cdot 10^{-13}$ ($n_k = 3$, 20 s) and $5.3 \cdot 10^{-13}$ ($n_k = 4$, 49 s), and DFTK's ν and ε_F on

the same k-grid lie inside. At zero temperature both spins are filled by aufbau over the union of their levels; under a common gap the smooth step agrees with aufbau occupation, so the zero-temperature LSDA SCF solution exists near the certificate (X40, T3) and is unique there (X41, T3), and the `fis-lsda0 1` checker prints the exact integers $N_\uparrow$, $N_\downarrow$ and a gap interval containing no eigenvalue of either spin (X42, T4); DFTK with `Smearing.None` gives the same integer ν, its ε_F in the Lean gap, and HOMO and LUMO within 10^{-14} of the driver. For the energy, $E_S(\gamma_\uparrow, \gamma_\downarrow) = E((\gamma_\uparrow + \gamma_\downarrow)/2) + E_{xc}[\rho_\uparrow, \rho_\downarrow] - B(N_\uparrow - N_\downarrow)$ and $F_S = E_S - T\sum_s \sum_i s(x_{s,i})$; with the exchange-correlation energy entering only through a second-order Taylor interface (X45, T3), the energy and free energy of the SCF solution are bounded by quantities the checker can compute, the free energy to second order since it is stationary at the solution (X46, T3); the checker computes these bounds and prints intervals for E_S and F_S of the certified solution, per cell with k-points (X47, T4); their widths are $1.5 \cdot 10^{-10}$–$1.0 \cdot 10^{-8}$ for E_S and $4 \cdot 10^{-15}$–$6.6 \cdot 10^{-14}$ for F_S, and DFTK's E and F lie inside on six systems. With k-points at zero temperature the uniqueness is among zero-temperature solutions only, so the translated solution, which has the same gap, is the solution itself and is cell-periodic (X43, T3); the checker prints the supercell's integer occupations and, per primitive cell, exact rational ones (X44, T4): for free electrons with $n_k = 3$ and $B = 0.2$, $N_\uparrow = 5$, $N_\downarrow = 1$ in the supercell and $\nu = 4/3$ per cell, with one gap for all k-points and both spins. DFTK's zero-temperature occupation needs the same number of occupied bands at every k-point, which this system violates, so there the comparison uses Fermi–Dirac at $T = 0.002$, far below the gap; the per-cell ν agrees and ε_F lies in the Lean gap. Each spin has its own Bloch Hamiltonian $H_{L,\kappa,s}[\gamma_\uparrow^*, \gamma_\downarrow^*]$, Zeeman term included; the error split of the LDA band certificate carries over to one diagonal block of the block-diagonal $X(y)$ (X38, T4), so the checker bounds the per-spin bands, the per-spin DOS (without spin factor) and the total DOS at Γ and with k-points (S09–S11, T4), with band half-widths $1.6 \cdot 10^{-11}$–$3.0 \cdot 10^{-10}$; all 192 band and 252 DOS values from DFTK's `compute_bands` and `compute_dos` lie inside. For a Gaussian potential at Γ, changing the potential shifts each spin's X_s by $\delta V / T$ and the energy only through the external term (V15, T3), so the Newton–Kantorovich condition is re-checked with a larger h_0 and the solution, magnetisation, μ, E_S and F_S hold for the named potential (V16, V17, T4); for a double well with $V_0 = 2$, $L = 4$, $K = 6$, $T = 0.1$ the named-potential E_S interval has width $7.4 \cdot 10^{-9}$, while at $T = 0.05$ the condition fails for the named potential although it holds for the certificate, and the checker says so. Spin-interpolated PW92 correlation (DFTK's `Xc([:lda_x, :lda_c_pw])`) couples the spins: $v_\uparrow$ depends on $\rho_\downarrow$ and the Jacobian has cross terms. We prove closed forms for the potentials, Jacobian and third derivatives of the interpolated functional along lines (X29, T3) and derive the first- and second-order bounds from entrywise bounds (X30, T3), but uniform bounds on $[\mathrm{lo}, \infty)^2$ are about $100\times$ the true sup $|J|$ and fail the Newton–Kantorovich condition on two of four test systems. The checker therefore uses a local version of the LSDA argument, with Lipschitz and Taylor bounds on a small box around each grid point's spin densities, of half-width $2C_\infty r_0$ (X31, T3), and rational interval balls for v_s and J_{st} on those boxes (X32, T3); a generic LSDA checker over any exchange-correlation interface is instantiated with PW92 at Γ (X33, T4) and with k-points (X34, T4). For a cosine potential at $L = 6$, $K = 6$, $N = 2$, $T = 0.05$, $B = 0.1$ the radius is $5.4 \cdot 10^{-14}$ and the ν interval has width $1.1 \cdot 10^{-13}$ (10 s, three to four times the Slater-only time); DFTK's ν and ε_F lie inside on four systems, within $4.5 \cdot 10^{-14}$ of the driver. Since 8 October 2026 PW92 has every result Slater exchange has. At zero temperature the generic checker uses the smooth step under a common gap with the local box constants (X48–X50; radii $1.1 \cdot 10^{-15}$–$8.4 \cdot 10^{-15}$ at Γ in 7–12 s, $\leq 1.1 \cdot 10^{-13}$ with k-points in about 60 s). The energy and free energy follow from Lipschitz bounds of the potentials (X51–X53); the energy interval still uses the uniform $[\mathrm{lo}, \infty)^2$ constants and is one to two orders of magnitude wider than for Slater exchange (E: $1.3 \cdot 10^{-8}$–$5.9 \cdot 10^{-7}$, F: $3.6 \cdot 10^{-14}$–$1.6 \cdot 10^{-9}$, 24–116 s). Per-spin bands and DOS (X54, X55; half-widths $\leq 1.4 \cdot 10^{-10}$ and $\leq 1.9 \cdot 10^{-8}$, all 192 band and 252 DOS values of DFTK inside) and the named Gaussian potential at Γ (X56–X58, with the box re-centred for the larger h_0; radius $7.5 \cdot 10^{-13}$, E width $6.6 \cdot 10^{-7}$) complete the item; DFTK lies inside every interval. Physical uniqueness (T34, T37) is proved for Slater exchange only. Not yet certified are bands and densities of states at zero temperature, the named Gaussian potential with k-points (such certificates are rejected)

and its LSDA bands, and DFTK's `Magnetic` term, a vector potential rather than a Zeeman field.

Spectra and Densities. At zero temperature and the Γ point, the checker also encloses, for *any* discrete ground state γ^*: the Coulomb-norm and pointwise distance between the computed density and ρ^*; the Hartree potential; every eigenvalue of the ground-state Hamiltonian H^* and its gap; the kinetic, external and Hartree energies; and the non-self-consistent band structure and Gaussian-broadened density of states (E01–E03, S02, S03). At finite temperature there is no ground-state energy to argue from, but the SCF-solution certificate already pins down the Hamiltonian of the unique nearby solution: $H[\gamma^*] = T\,X(y^*) + \mu^*$ with $\|X(y^*) - X(y_0)\|_F \le n_X r$ and $|\mu^* - \mu_0| \le r$. Replacing the kinetic diagonal of the rational matrix $X(y_0)$ by its value at κ gives the Bloch sector, so $H_\kappa[\gamma^*]$ differs from a diagonalised rational matrix by a scalar shift and a perturbation of bounded Frobenius norm. Weyl-type eigenvalue counting and the density-of-states argument of the zero-temperature case then enclose every band energy and DOS value, for rHF with any of the four smearings and for LDA, whose potential adds one more bounded term (S04, S05, T4). For a double well at $L = 10$, $K = 8$, $T = 0.01$ the band half-widths are about 10^{-11} (rHF, MV) and 10^{-9} (LDA with PW92, FD), and DFTK's non-self-consistent bands and DOS lie inside.

With k-points the certificate describes the equivalent supercell, while bands are wanted on the primitive cell. The cell-periodic density of the solution, the density DFTK reports, is that of the block average $\bar\gamma^* = n_k^{-1}\sum_j \gamma_j^*$ of the supercell density matrix over its Bloch blocks. For any supercell density matrix γ, the j-th diagonal block of the supercell Hamiltonian is the primitive Bloch Hamiltonian $H_{\kappa_j}[\bar\gamma]$; for LDA, if γ is block diagonal, the exchange–correlation potential on the supercell grid has the potential of $\rho_{\bar\gamma}$ on the cell grid as its diagonal blocks. For any primitive k-point κ, $H_\kappa[\bar\gamma^*]$ is therefore a principal submatrix of a Bloch sector of the supercell Hamiltonian plus an exact kinetic difference, and since principal submatrices do not increase Frobenius norms or quadratic-form bounds, the supercell error bounds carry over unchanged. This certifies bands and DOS on the primitive cell for rHF with any of the four smearings (S06), for LDA (S07) and for zero-temperature rHF, where the density bound of the ground-state certificate takes the place of the Newton–Kantorovich radius (S08), all at T4. For a double well with $n_k = 3$ the band half-widths are at most $5.6 \cdot 10^{-12}$ (rHF, MV, $K = 4$), $3.9 \cdot 10^{-11}$ (LDA with PW92, FD, $K = 4$) and $3.5 \cdot 10^{-4}$ (zero temperature, $K = 6$, as at Γ), checking takes about 9, 50 and 2 s, and DFTK's `compute_bands` and `compute_dos` lie inside the Lean intervals. The zone-boundary point X is certified too but not compared: there DFTK's spherical cutoff keeps $2K + 2$ plane waves and ours $2K + 1$, a difference in the discretisation only. Not yet certified are the bands of zero-temperature LDA with k-points, k-point bands for the named Gaussian potential, and the bands of finite-temperature runs certified by the variational free-energy certificate.

Binding to the Named Potential. Every theorem above is about the problem written in the certificate, and the driver writes it. To close the largest part of this gap, `mini-DFT` writes a `pot` line into every certificate: the family of the external potential (free, cosine, or a Gaussian double well with R copies per cell) and the exact doubles of its parameters. In Lean, `ExtPot` defines these potentials in real space together with their Fourier coefficients, and we prove that the Fourier series converges to the real-space potential, so the coefficients are not a further convention (DFT-V07). For the free and cosine families the coefficients are rational, and the checker requires every coefficient in the certificate to equal its Lean value. The certificate's problem is then *equal* to the problem of the named potential, at the Γ point and, for k-points, for the primitive cell (I02, I03, T4), so every soundness theorem of this section becomes a statement about the named potential. Gaussian coefficients contain $\sqrt{2\pi}$ and exponentials; rational enclosures give an ℓ^1 bound d on the deviation of the certificate's coefficients from the exact ones, about $5 \cdot 10^{-16}$ in practice. The perturbation bound for the ground-state energy (V01) and its finite-temperature analogue $|F'_{T,K} - F_{T,K}| \le 2N_{\text{pair}} \sum_{|m|\le 2K} |\hat V'_m - \hat V_m|$ (V08) then turn the variational certificates into intervals for the named Gaussian potential, widened by $2N_{\text{pair}}d$: at zero and finite temperature, at Γ and

with k-points, for any number of copies R (V09, T4). The remaining problem data (cell length, cutoff, electron count, k-point count and, depending on the certificate, temperature, smearing, functional, grid and Zeeman field) are parameters rather than derived quantities; the checker prints their exact values in `PROBLEM` lines, and `make verify-all` compares them with the command line. The binding is mandatory: `certify` runs it first for every kind of certificate and rejects a certificate without a `pot` line with `POT_FAIL` and no conclusion, so every conclusion it prints is about the named potential (DFT-I04, T4). When the driver solves for a different potential but names the requested one, or when one coefficient of a cosine potential is changed by one ulp, the certificate is rejected in the same way. The archived showcase and research certificates were rebound by rerunning the driver with the archived command, checking that every problem datum matches line by line and inserting its `pot` line; their candidate solutions and conclusions are unchanged. For SCF-solution, spectral and band certificates a perturbation estimate is needed instead: replacing the potential changes H_{ext}, $H[\gamma]$, its Bloch sectors, $H_{\mathrm{L}}[\gamma]$ and the SCF maps $X(y)$, $X_{\mathrm{L}}(y)$ by one and the same Toeplitz matrix δV, with $|x^{\mathsf{H}}\delta V x| \le d\,\|x\|^2$ (V10). The checker then repeats the Newton–Kantorovich test with the residual widened by $\lceil\sqrt{n}\rceil d/T$, and on success the named Gaussian problem has a unique SCF solution near the certificate, rHF or LDA, at Γ or with k-points; spectra, bands, DOS, entropy and energy are widened accordingly (V11–V14, T4). The price is modest but visible: for a double well with $d \approx 1.3 \cdot 10^{-15}$ the radius of uniqueness grows from $1.3 \cdot 10^{-12}$ to $2.6 \cdot 10^{-11}$ and the energy interval from $3.7 \cdot 10^{-10}$ to $7.6 \cdot 10^{-9}$. Zero-temperature LDA and the zero-temperature energy components still speak about the certificate's coefficients, within d of the named potential.

User Potentials. Since 7 October a user can bring a potential of their own: a file of rational Fourier coefficients (`--pot file`) or periodised Gaussian sites with rational positions, depths and widths (`--pot sites`). The certificate carries the data as exact fractions and the checker echoes every value; given the same input file or site list, the checker reads it itself, with the driver's rules, and requires it to equal the certificate's data, so a run is bound to the user's input as a `--tdb` run is to its database (DFT-V18–V21; reading the file is unverified I/O, V23, T0). Dyadic coefficients bind exactly, as for the cosine potential; for non-dyadic coefficients the checker computes the ℓ^1 deviation d exactly in rationals, and for sites it encloses envelope and phase in rational complex balls ($d \approx 5 \cdot 10^{-16}$), so all named-potential results of V09–V14 carry over (V22). DFTK's energies for a complex non-dyadic coefficient file and for three asymmetric sites with one negative depth lie inside the Lean intervals (widths $2.0 \cdot 10^{-9}$ at zero temperature, $4.4 \cdot 10^{-11}$ with Fermi–Dirac smearing, $2.0 \cdot 10^{-13}$ for an LDA SCF solution; V24). The infinite-cutoff limit for sites (no Fourier tail bound yet) and LSDA or zero-temperature LDA with a non-exact user potential are not covered. Because sites are a family of potentials depending smoothly on the site positions, `bindStrict_sound` now states exactness only for families other than `gauss`, `sites` and `file`; this is the one pinned statement the change touched.

Forces. For rHF at any temperature and with any entropy function, the ground (free) energy is concave in the external potential, and the strong convexity of the Hartree term in the density bounds the second-order remainder: $\langle\delta,\gamma_0\rangle - L\sum_{m\neq 0}|\delta_{-m}|^2/c_m \le F[\hat V + \delta] - F[\hat V] \le \langle\delta,\gamma_0\rangle$ (DFT-N01, T3). Hence, if the coefficients $\hat V_m(\theta)$ are differentiable with derivatives w_m, the energy is differentiable with derivative $\langle w,\gamma^*\rangle = \mathrm{Re}\,L\sum_m \overline{\hat\rho^*_m}\,w_m$ for any minimiser γ^*, the Hellmann–Feynman theorem, with no continuity of the minimiser in θ needed and no Pulay term, since the plane-wave basis does not move (N02, T3); a rigid translation leaves F unchanged, so forces on sites sum to zero (N03, T3). A computable checker encloses $\langle w,\gamma^*\rangle$ from the certificate's Coulomb-norm density error (N04, T4). Since 8 October 2026 it is run by `certify` for Gaussian sites (`--pot sites`): the site-derivative interface supplies rational envelopes of $w_m = \partial\hat V_m/\partial x_j$ with an ℓ^1 error (N05), and for rHF Γ-point certificates at zero temperature (N06) and with Fermi–Dirac or Gaussian smearing (N07) the checker prints `FORCE_OK j lo hi` for the force $-\partial E/\partial x_j$ of the named problem and `FORCE_SUM`, whose interval contains 0 by theorem (N08). The half-width is set by the density error and is about the square root of the energy width: $1.8 \cdot 10^{-4}$ at zero temperature and

4.5–4.7 $\cdot 10^{-5}$ at $T = 0.05$ for $L = 4$, $K = 16$; central differences of certified energies and DFTK's `compute_forces` (`ElementGaussian` with `AtomicLocal`; six forces within $2.7 \cdot 10^{-13}$ of the driver) lie inside. Forces with k-points, for SCF-solution certificates (MV, MP), for LDA and LSDA (where the derivative along a locally unique SCF branch needs the implicit function theorem, N09) and on the wells of `--pot gauss` remain.

Exact Exchange. Closed-shell Hartree–Fock, DFTK's `ExactExchange` in 1D, adds $E_x(\gamma) = -\operatorname{Re}\operatorname{Tr}(K[\gamma]\gamma)$ with $K[\gamma]_{ab} = L^{-1}\sum_{|m|\le 2K} v_m \gamma_{a-m,b-m}$ to rHF; the SCF equations are $\gamma = f((F[\gamma]-\mu)/T)$, $\operatorname{tr}\gamma = N_{\text{pair}}$, with the Fock matrix $F = H - K$ (H01, T3). In 1D the Coulomb kernel $-2\pi|x|$ is fixed only up to a constant; the Hartree energy does not depend on it, but the exchange energy does (H05, T3), so the kernel $v_m = 4\pi/q_m^2$ ($1 \le |m| \le 2K$), $v_0 = \pi L^2/6$, is part of the model and echoed by the checker. Exchange is nonlocal, so the unknowns are the n^2 coordinates of the mean-field matrix $W = J[\gamma] - K[\gamma]$ and μ; with Frobenius bounds on K and J (H02) the Newton–Kantorovich argument of the SCF-solution certificate carries over (H06, H07), and the checker proves that the HF SCF equations have exactly one solution near the certificate and encloses its μ (H08, T4): radii $4.5 \cdot 10^{-15}$–$9.3 \cdot 10^{-14}$ in 0.4–3.1 s on seven systems, with the free-electron μ from arbitrary-precision arithmetic inside. For one electron pair the exchange cancels the self-interaction, $E_x = -E_H/2 - v_0/L$ (H04), which the driver checks to $6 \cdot 10^{-15}$. DFTK's own HF SCF does not converge on these 1D systems, so the comparison is a fixed-point check: at the certificate's centre DFTK's Hamiltonian equals the Fock matrix to $1.7 \cdot 10^{-14}$ and the energy terms agree to $9 \cdot 10^{-14}$ (H11). The HF energy interval (H09), zero temperature, k-points and hybrids are not done; checking time grows like n^5, and gapped systems at small T fail honestly.

A Continuum Ground State on the Whole Line. The two questions of this section, the ground state of a discretised problem and that of the continuous one, can be answered for a linear Schrödinger operator $H = -\mathrm{d}^2/\mathrm{d}x^2 + V$ on all of $\mathbb{R}$ with a polynomial V, the quartic oscillator $V = x^4$ first; a separate checker, `certify-osc`, handles both. In the basis of the first n Hermite functions the Galerkin matrix H_n contains square roots $\sqrt{(i+1)/2}$, but a diagonal similarity removes them: with $\Delta_0 = 1$, $\Delta_{i+1} = \Delta_i\sqrt{(i+1)/2}$ and $v = \Delta w$, $v^{\mathsf{T}} H_n v = w^{\mathsf{T}} B w$ and $\|v\|^2 = \sum_i m_i w_i^2$ with a rational symmetric B and $m_i = i!/2^i$ (DFT-Y22, T3). The checker proves $\lambda_{\min}(H_n) \ge \text{lo}$ from positive pivots of an exact rational LDL^{T} factorisation of $B - \text{lo}\,\operatorname{diag}(m)$ (Y21) and $\lambda_{\min}(H_n) \le \text{hi}$ from the exact Rayleigh quotient of a rational vector (Y23, T4); for $n = 40$ the interval is $8.7 \cdot 10^{-19}$ wide and contains an independently computed high-precision value. The continuum ground-state energy is defined without Sobolev spaces, $E_0 = \inf\{\int(u'^2 + Vu^2)/\int u^2\}$ over Lipschitz, compactly supported $u \ne 0$ with u' from Rademacher's theorem. The upper bound is the exact Rayleigh quotient of a continuous piecewise-polynomial trial function (Y27). The lower bound uses the Riccati (Barta-type) inequality rather than a spectral argument: for any piecewise-smooth g that jumps only downwards at its nodes, $\int u'^2 + Vu^2 \ge \int (V - g' - g^2)u^2$, since the difference is $\int(u' - gu)^2$ after one integration by parts, so $V - g' - g^2 \ge \text{lo}$ everywhere gives $E_0 \ge \text{lo}$ (Y26, T3). The checker verifies $V - g' - g^2 - \text{lo} \ge 0$ on each piece by a centred form with bisection, the tails $|x| \ge 5$, where $g = \pm(x^2 - 1)$, by non-negative coefficients, and the downward jumps, all in exact rationals (Y25, Y28, T4). The quality of the bound depends only on how well g follows the Riccati equation $g' = V - e - g^2$, with no h^2 discretisation error and no need for the second eigenvalue; the driver integrates it at $e = E_0 - \delta$, where the solution turns smoothly onto the growing branch, while for $e > E_0$ it blows up, so the method cannot produce a wrong lower bound. For x^4, 284 pieces of degree 12 for the lower and 64 for the upper bound enclose E_0 in $[1.06036209048418278, 1.06036209048418290]$, an interval $1.1 \cdot 10^{-16}$ wide that contains the value 1.06036209048418289964704603 from two independent floating-point routes (a Hermite basis with $n = 120$ and inward shooting, agreeing to $3 \cdot 10^{-24}$); checking takes about a second. The width is a choice: with $\delta = 10^{-20}$ the interval is $1.1 \cdot 10^{-20}$ wide (5626 pieces, 11.5 s; that certificate is not archived). On the same problem, floating-point trials of the alternatives reach $2.1 \cdot 10^{-8}$ with Temple's

inequality ($n = 40$, assuming the exact second eigenvalue, which itself would need a rigorous lower bound and a spectral theorem for unbounded operators that Mathlib lacks) and about 10^{-7} with Dirichlet–Neumann bracketing and Liu–Oishi bounds for 8000 P1 elements (Y30, T0). Only the ground state is covered, since the Riccati bound needs a ground state without zeros; V must have even degree and a positive leading coefficient; and that E_0 is the bottom of the spectrum of the self-adjoint operator is classical but not proved in Lean.

Usability. The driver now ends its report with machine-readable `STATUS`, `FLOAT` and `HINT` lines, whose names never end in `_OK` or `_FAIL`; `fis` reads only these, not the human report. A failed SCF prints the options to add (Anderson mixing, a smaller step, or smearing when a zero-temperature gap closes). `fis multiplicity` packages the argument of Appendix Q: SCF runs from several starting densities, one checked certificate per distinct solution, and an exact comparison of the problem data and of the uniqueness balls, which splits certificate lines exactly as the checker's parser does. Ten checker theorems behind output lines whose names do not end in `_sound` are now pinned too.

K Three-Dimensional DFT and Showcase S1

The three-dimensional work follows the phases of Appendix H.1: arithmetic and the model first, then SCF certificates of growing size, nonlocal pseudopotentials, zero temperature, k-point meshes and symmetry, and finally the certificate that showcase S1 needs for silicon. The empirical Cohen–Bergstresser band structures, a linear three-dimensional problem, and their validation are described with the other physical results in Appendix Q.1.

Arithmetic and Model. The first step towards three-dimensional SCF was arithmetic. A dyadic interval library proved in Lean (mantissa $\times 2^e$ with outward rounding to p bits) gives `certify-pw` a second arithmetic route, `--dyadic p`, whose soundness theorems state exactly what the exact-rational ones state (J08, J09, T4); band indices come from interval LDL^{H} inertia counts at gap midpoints, and a cluster bound proved by a dimension argument and a Gershgorin bound on the Gram matrix places each cluster of bands (J01–J07, T3). On the Si, Ge and GaAs certificates of Appendix Q.1 every interval of the dyadic route intersects the exact one (83 of 83 output lines), the five lowest bands pass from $p = 26$ and all nine bands need $p = 32$–36, because interval widths grow during elimination near a shift with a small gap at X. The dyadic route is not yet faster than exact rationals at $n \approx 200$ (11–22 s against 6–14 s per k-point for nine bands): with unbounded integer mantissas the gain needs $p \leq 31$, so preconditioning or 2×2 pivots come before machine-word mantissas. Second, the model of phase 2 was put into Lean: reduced Hartree–Fock on any Bravais lattice with the Hartree kernel $4\pi/|G|^2$ and local pseudopotentials summed over atoms, with existence of minimisers at zero and finite temperature, uniqueness of the ground-state density, monotonicity in the cutoff, the Euler–Lagrange and Aufbau characterisation that the SCF certificate will check, and bridges showing that the one-dimensional model is the case $d = 1$ and that without Hartree the Hamiltonian is the Cohen–Bergstresser one (M01–M06, T3); for LDA, minimisers exist but the density need not be unique, since exchange is concave (M16).

First SCF Certificates. The first certificate for a three-dimensional SCF solution followed: for fcc Si with a model local pseudopotential (not a fitted one), reduced Hartree–Fock at the Γ point and Fermi–Dirac smearing $T = 0.01$ Ha, a new checker `certify-pw3d` proves that the discrete SCF equations have exactly one solution in a ball of radius $7 \cdot 10^{-12}$ to $6 \cdot 10^{-10}$ around the certificate and gives intervals for the chemical potential and the lowest 16 bands, for 15 to 89 plane waves (M07–M12, the checker theorems at T4; abstract Newton–Kantorovich theorem U01); DFTK's solution of the same discrete problem lies 6–$8 \cdot 10^{-14}$ from the certificate's centre, inside the ball. Three limits are measured rather than hidden. The dense exact Jacobian costs about n^4 (38 CPU hours at $n = 89$, against an estimated 1.5 CPU years per k-point at the $n \approx 400$ that meV convergence needs); the Newton–Kantorovich condition fails below $T \approx 1/300$ Ha for this gapped system, so zero

temperature needs the Aufbau route of M05; and the split into an exactly checked low-frequency block and an analytic high-frequency tail (U02–U05) closes numerically with the true tail ($\kappa \approx 0.1$–0.45) but not yet with the provable Coulomb-only tail bound, which is 300 to 10^4 times too large: a kinetic-dominance bound for the tail was the open step on which the scale of phase 2 depended. That step is now proved (U06–U10, T3): low-energy pair densities are computed exactly, the divided differences between occupied and high-energy states are bounded by the kinetic energy rather than by the global maximum, and a symmetry of the Jacobian makes the low-row, high-column block exact. With it, a split mode of the checker (U11–U13, T4, same conclusions as the dense mode) computes only the 169 low-frequency columns exactly and certifies the same Si model up to $n = 411$ plane waves ($d = 2951$ unknowns, cutoff about 10 Ha, where this model converges to the meV level) in 56 minutes of wall time; the contraction constant stays near 0.5 from $n = 89$ to $n = 411$, and at $n = 89$ the split mode needs 25 times less CPU time than the dense one. The provable constants are still 18 to 90 times looser than the true ones, so there is room to tighten further; the dominant remaining cost is big-integer arithmetic in the low-frequency columns.

Showcase S2: Not the Ground State. The first instance of showcase S2 followed (M20–M26, Z06–Z08). For simple-cubic hydrogen (lattice constant 7 bohr, two atoms in the cell, the local part of the GTH-LDA pseudopotential, Slater exchange, Γ point, $T = 0.01$ Ha, 59 plane waves), the checker proves that the spin-restricted LDA equations have exactly one solution in a ball of radius $3.7 \cdot 10^{-10}$, that this solution is also a self-consistent solution of the collinear LSDA problem, and that its free energy is at least -0.3623446 Ha, while an antiferromagnetic trial state built from rational vectors has LSDA energy at most -0.3744735 Ha; by the variational principle the converged restricted solution is therefore not the ground state, with a certified margin of 330 meV. DFTK's restricted and broken-symmetry solutions of the same discrete problem lie inside the respective bounds. The ordering itself is textbook physics (symmetry breaking in stretched hydrogen), so this is a first in rigour, not a discovery; the cutoff (about 1.3 Ha) is far from converged, at a smaller cutoff ($n = 27$) the broken-symmetry state does not exist and the S2 check correctly fails, and the checker binds only fcc cells.

Pseudopotentials, Zero Temperature and k Points. Three pieces of showcase S1 followed, each first in a checker mode of its own. First, real silicon: the model and checker take separable GTH/HGH pseudopotentials with nonlocal projectors, written exactly as DFTK implements them (Q01–Q06, Q12, T3; checker modes Q07–Q09, T4). For Si with the GTH-LDA q4 pseudopotential from DFTK's pseudopotential library (local part, two s projectors and the p channel, with file hash and provenance recorded, Q11), reduced Hartree–Fock at the Γ point and $T = 0.01$ Ha, the SCF solution is certified up to $n = 411$ plane waves (cutoff about 10 Ha, 6155 s of wall time), with band intervals $1.3 \cdot 10^{-7}$ Ha wide, and DFTK's solution lies 10^{-14}–10^{-13} from the certificate's centre at every size checked (Q10); the tail theorem holds unchanged, since the nonlocal term does not depend on the density. Second, zero temperature: with an occupation that is exactly the Aufbau step outside the gap and the chemical potential fixed at mid-gap, the Newton–Kantorovich argument closes at zero temperature, and because reduced Hartree–Fock is convex the conclusion is global: the certified density matrix is the unique ground state of the discrete model, with an interval for the Γ-point gap (M30–M34; the checker theorems at T4). For the Si model the gap is enclosed to about 5 μeV at $n = 411$; the zero-temperature LDA equation is proved (M35, T3) but has no checker yet. Third, k points: a block-diagonal multi-sector model with Monkhorst–Pack weights, in which the sectors couple only through the total density, reduces to the Γ model for one k point and has a weighted Newton–Kantorovich theorem whose constants do not grow with the number of k points (M13–M15, M40, T3); the checker certifies meshes up to $4 \times 4 \times 4$ (64 k points, 87 s) with band intervals at each k point, and all 2864 DFTK eigenvalues checked lie inside (M41, M42, T4). The three pieces were then combined, and the two remaining gaps of S1 were narrowed. A unified LDA mode of the checker covers Slater exchange with or without PW92 correlation, GTH nonlocal projectors at every k point, zero temperature or Fermi–Dirac occupation and Monkhorst–Pack meshes under one soundness theorem (M50–M54, T3; M56, M57, T4). The first combined certificates, real Si with GTH-LDA q4 at zero temperature on $2 \times 2 \times 2$ and $4 \times 4 \times 4$ meshes, enclose the gap in intervals under 10^{-9} Ha wide ($4 \times 4 \times 4$ in 228 s), and all 1023 DFTK eigenvalues checked

lie inside (M58); at a cutoff of about 1.6 Ha they test the pipeline and are not S1 numbers. Time reversal makes the SCF maps of the blocks at κ and $-\kappa$ equal (K08–K10, T3), and a split mode per k point checks one block of each pair, exactly only in the low frequencies, with a conclusion about the full mesh (U21–U24, T4): the Si model is certified with up to 180 plane waves per k point on a $4 \times 4 \times 4$ mesh (73 min) and 410 on $2 \times 2 \times 2$ (2.5 h), at about a sixteenth of the dense mode's CPU time; the measured cost per block grows as about n^3, so a $4 \times 4 \times 4$ mesh at 410 plane waves would take about 1100 CPU hours, and point-group symmetry is not yet used. For the cutoff error, reduced Hartree–Fock's convexity gives a bound with no continuum Jacobian: the distance from the continuum SCF density to any trial density is at most the residual of the SCF map (C20–C23, T3), and the resulting intervals for the continuum gap and energy have a paper proof (C24–C26, T1); in a floating-point trial on the Si model they contain the gap of a 5577-plane-wave reference with widths from 126 down to 9 meV. The argument needs convexity, so LDA, the functional of S1, needs another one, and the rHF Si model turned out to have no gap over the Brillouin zone in floating point.

LDA at the Protocol's Cutoff, and Symmetry. The next round pushed towards the protocol's thresholds. For LDA the exchange–correlation kernel is local, so the high-frequency block lacks the Hartree $1/|G|^2$ decay and the earlier tail bound does not close; a pointwise Cauchy–Schwarz bound on the response density with the exact energy denominator of every occupied–unoccupied pair and the operator norm of a pair Gram matrix does (U25–U31, T4). With it, an LDA split mode with time reversal and GTH projectors at every k point has a soundness theorem whose conclusion is word for word that of the unified mode, about the full mesh (M59–M67, T4); for real Si at zero temperature on a $2 \times 2 \times 2$ mesh with 24–27 plane waves per k point it passes with $\kappa_0 = 0.24$ in 32 s, and DFTK's 128 eigenvalues on the full mesh lie inside (M69). At about 410 plane waves per k point (10 Ha) a first check passed the density-positivity gate but was stopped unfinished after six hours on a shared machine; the protocol's cutoff needed the sharper constants described below. Point-group symmetry is now proved usable: the block SCF maps are equivariant under the space-group operations of the crystal, which act on the unknowns as signed permutations, so the unique solution in the certificate's ball is symmetric and the equation on the irreducible k points has the same zero as the full one (K11–K14, T3); the checker verifies the operations itself and its conclusion is again about the full mesh (K15–K18, T4). The Si model passes on the $8 \times 8 \times 8$ mesh the protocol requires (29 irreducible of 512 k points, 80 CPU hours at 180 plane waves) and on $4 \times 4 \times 4$ at 410 plane waves (276 CPU hours), with all 8192 DFTK eigenvalues of the $8 \times 8 \times 8$ case inside; the extrapolated cost of the S1 certificate is about 10^3 CPU hours, against $8.6 \cdot 10^3$ with time reversal alone (K20).

The Whole Brillouin Zone and the Continuum. Band bounds over the whole Brillouin zone close the third gap: the effective Hamiltonian at an arbitrary k point, with its own k-dependent basis, is bounded from a passing SCF certificate, and a checker gives two-sided band intervals at listed k points, off the mesh included, and an upper bound on the highest valence band over the zone from an octree of cubes with a gradient bound for the nonlocal kernel (W30–W40; the checker theorems at T4). On a coarse real-Si certificate this yields a certified upper bound on the gap, the quantity of the protocol's one-sided rule, with 48 DFTK band values inside; the matching lower bound needs about 10^6 cubes at the first-order leaf test and is left for a second-order one. Finally, the first continuum result is a T4 certificate: for the rHF Si model at 411 plane waves the infimum of the discrete ground-state energy over all finite plane-wave sets lies in an interval $2.2 \cdot 10^{-4}$ Ha wide (C29–C33). For LDA the local continuum argument does not close (C34, T1): the density of the pseudopotential model nearly vanishes at the ion sites (10^{-5}–10^{-4}), where the exchange–correlation kernel and its derivative blow up, the same assumption the literature makes without proof. The functional side of this obstruction is now a Lean theorem: any Lipschitz constant N of the Slater kernel $f_x = v_x'$ on $[\mathrm{lo}, \infty)$ satisfies $N^3 \geq 8/(243\pi\,\mathrm{lo}^5)$, and the rational constant the LDA checkers use for the second-order Newton–Kantorovich term has no finite bound as $\mathrm{lo} \to 0$, for Slater exchange alone and with PW92 correlation (Y05, Y06, T3). A potential-space route to the LDA continuum limit, which weights the potential by the density to absorb the blow-up at

low density, is proved in its abstract parts (C44, T3) but closes on real Si only with estimated constants, so it gives no certificate (C40–C49).

The S1 Certificate for Silicon. With these pieces the protocol's certificate for silicon exists. Sharper zero-temperature constants for the exchange–correlation term and a preflight that rejects a certificate before the expensive stages when the Newton–Kantorovich condition cannot hold (M70–M74, T4) made 410 plane waves closable, and an LDA point-group mode combines them with the proved symmetry reduction, again with a conclusion about the full mesh (M75–M77, K21–K28, T4). The exact rational and big-integer arithmetic of its inner products is replaced by arithmetic modulo several machine-word primes, proved in Lean to compute the same numbers, so the soundness theorem carries over word for word (J10–J17, T4). For real Si with GTH-LDA q4 at the protocol's lattice constant (exactly 5.416 Å), cutoff ($Q = 107/2$, about 10 Ha, 399–419 plane waves per k point) and $8 \times 8 \times 8$ Monkhorst–Pack mesh (29 irreducible of 512 k points), the check passes in 7 h 57 min on 32 cores (196 CPU hours, against an estimated 1150 with the big-integer route): the discrete zero-temperature KS-LDA equations have exactly one solution within $2.1 \cdot 10^{-10}$ of the certificate's centre (under the chemical-potential convention of K32, below), all 8192 band energies lie in intervals at most $8.6 \cdot 10^{-9}$ Ha wide, and the band gap on the mesh lies in $[0.5307362, 0.5307368]$ eV (K29). The mesh gap is computed inside the verified checker: a Lean function takes the certified intervals of bands N and $N+1$ at every k point of the mesh, and a proved theorem (`runUGR_bandgap_sound`, W314, chained to the SCF theorem) guarantees that $\min_k \varepsilon_{N+1}(k) - \max_k \varepsilon_N(k)$ of the certified self-consistent Hamiltonian lies in the printed interval; the archived production lines for Si, AlP and GaP are exactly this function's values (W315). The mesh gap bounds the gap of this self-consistent Hamiltonian over the whole Brillouin zone from above, but it is not the protocol's quantity: the one-sided rule of S1 also requires an upper bound on the highest valence band over the whole zone, and the two-sided rule a lower bound on the lowest conduction band. A band mode of the same checker supplies the first of these from this certificate (W42–W51; checker T4). It covers the zone by an octree of about 7500 cubes; on each cube a Rayleigh–Ritz test in a few trial vectors, with the kinetic part expanded exactly, the nonlocal part bounded by a mean-value form of the separable GTH kernel and the potential taken from the SCF certificate's frequency differences, shows that the highest valence band stays below a bound everywhere in the cube. At two listed mesh points, Γ and $(0, 5/8, 5/8)$, full-basis interval counts bound the fourth and fifth bands from both sides to about $6 \cdot 10^{-6}$ Ha. Together they give the S1 record for silicon under the one-sided rule: the KS-LDA band gap at the protocol's settings is at most 0.530902 eV (W51); DFTK 0.8 agrees at every listed point. The check takes 44 min on 32 cores. A Rayleigh–Ritz point off the mesh, near 0.84 of the way from Γ to X, gives the sharper bound 0.4856 eV, but the protocol's listed points must be bounded from both sides, so it is not used in the record. The zone-wide valence bound of this mode is still loose (0.03 Ha above the band maximum), which does not enter the one-sided gap. The lower bound on the conduction band that the two-sided rule needs comes from a further mode with a second octree (W52–W64; checker T4): a counting theorem of Lehmann type, proved in Lean, shows on every cube that the effective Hamiltonian has no more eigenvalues below a bound than there are occupied bands, from a gap condition on patches of cubes that share trial vectors and a residual condition on each leaf; together with a sharper upper octree and the listed points it gives a two-sided bound on the band gap over the whole zone, and the range-split form lets its parts be checked and resumed separately. The silicon result and its comparison with experiment follow below.

The Total Energy of the Certified Solution. A further mode of the checker bounds the total LDA energy of a certified SCF solution (W80–W89). The electronic part uses an identity at the solution: for any trial density matrix, the energy of the true solution equals a band sum plus a correction that is first order in the band-sum error and second order in the distance to the trial matrix, with the Hartree remainder bounded through the same class counting as the SCF certificate (W81–W83). The ion–ion energy is an Ewald sum at a fixed splitting parameter η, bounded by rational bounds on erfc and exp and explicit bounds on both lattice-sum tails (W86, W87). Soundness of the mode is proved in Lean (W85, W87, T4). For the S1 silicon certificate it gives $E_L + E_{\mathrm{ew}} \in [-7.928721861, -7.928721780]$ Ha per cell, a

width of $8.0 \cdot 10^{-8}$ Ha (2.2 μeV); the Ewald part alone is 10^{-17} Ha wide (W89), and DFTK 0.8 on the same discrete problem lies inside the interval (W88). That the Ewald sum does not depend on η is classical but not proved here; the checker bounds the sum at $\eta = 7/2$, the formula DFTK evaluates.

A Second Material: AlP. Aluminium phosphide was pre-registered before any calculation (O20: zinc blende at $a = 5.4672$ Å, the 300 K value with neither thermal nor zero-point expansion removed, unlike silicon's static-lattice value, GTH-LDA with Al and P pseudopotentials, cutoff 10.08 Ha, an unshifted $8 \times 8 \times 8$ mesh, zero temperature) and certified at those settings (W90–W97). The checker now carries one pseudopotential and one structure factor per species, and checks the symmetry of the two-species crystal itself (W90, W91, T4); the zone-wide valence bound carries over with the same proof (W93). The SCF certificate has radius $6.05 \cdot 10^{-11}$ and bounds all 8192 band energies to within $2.6 \cdot 10^{-9}$ Ha; the gap on the mesh lies in $[1.4778801, 1.4778803]$ eV, from Γ to X (W95). The zone-wide check gives the S1 record under the one-sided rule: the KS-LDA band gap at the protocol's settings is at most 1.478051 eV (W96). The SCF check takes 14.5 h on 32 cores (362 CPU-hours, 103 GB of memory), the zone-wide check 1 h. Before the two-sided runs for AlP and GaP, the dominant step of the two-species check, the lower-bound leaf test, was made 14–69 times faster (45 times averaged over the leaves of the GaP certificate): profiling showed the time going into recomputation and rational-number overhead, not precision, and exact-integer and memoised versions of the leaf test are each proved in Lean equal to the functions they replace (W260–W267, T4), so the soundness theorem and the trusted base are unchanged; the single-species checker used for silicon was not changed. The projected cost of the GaP check fell from about 32 600 to about 1 700 CPU hours (W268, T0), and the patch condition is now the largest remaining part. The two-sided zone check (rule B) with the two-species, full-kernel checker followed (`runBZBXD_of_ranges`, `runBZBXD_sound`): a 2.0 GB certificate with 362 lower-bound patches and 17 050 own lower-bound leaves, checked in 25 separately checked ranges, all accepted, 705 CPU hours on 32 cores (27 h wall clock, 51 GB); for every wave vector of the zone band 4 lies below -0.026555547 Ha and band 5 above 0.025155547 Ha relative to μ^*, so the gap of the discrete model lies in $[1.4071, 1.4781]$ eV (W300, T4), an interval 0.0709 eV wide, below the registered width limit of 0.10 eV. One driver setting differs from GaP: the floating-point margin left for the patch condition was lowered from 0.02 to 0.008, because the conduction-band minimum lies exactly at X, where subdividing a patch lowers the margin; the protocol leaves such checker parameters free, and patches with floating-point margins down to 0.008 all pass in Lean. DFTK 0.8 on the same discrete problem puts all 8192 bands inside the intervals (W97); its first run did not, because floating-point rounding at the cutoff sphere dropped one plane wave from one star of k points, which made it a different discrete problem; raising its cutoff by a relative $2 \cdot 10^{-11}$ restores the same basis. The comparison with experiment follows below.

A Third Material: GaP. Gallium phosphide was registered before any certificate (O24: zinc blende at $a = 5.4343$ Å, GTH-LDA with the full kernel for Ga, $Q = 59$ (11.04 Ha), an unshifted $8 \times 8 \times 8$ mesh, zero temperature) and certified at those settings. The SCF certificate has radius $4.5 \cdot 10^{-11}$; the gap on the mesh lies in $[1.5855484, 1.5855487]$ eV, from Γ to X (W190, T4). The zone-wide check bounds the highest valence band over the whole zone with 43 359 leaves and gives the one-sided S1 record: the gap of the discrete model is at most 1.585748 eV (W191, T4). Here the conduction-band minimum lies off the mesh, on the Γ–X line, so the bound exceeds the floating-point gap of the same model by 48 meV; the protocol predicted about 47 meV before any certificate. The SCF check takes 30 h on 32 cores (931 CPU-hours with the driver, 129 GB), the zone-wide check 6.3 h (174 CPU-hours); DFTK 0.8 puts all 8192 bands inside the intervals (W193). The two-sided zone check (rule B), with a 2.1 GB certificate (496 lower-bound patches, 51 519 own lower-bound leaves) that lists a point at the conduction-band minimum on Γ–X ($t = 0.861$), was checked in 34 ranges by four processes of 32 cores, all accepted: 1 842 CPU hours, as projected from a timing step before production, and 46 h wall clock (69 GB). For every wave vector band 4 lies below -0.028533917 Ha and band 5 above 0.025370475 Ha relative to μ^*, so the gap of the discrete model lies in $[1.4668, 1.5378]$ eV (W192, T4), 0.0709 eV wide; its upper end lies 48 meV below the one-sided bound because the listed off-mesh point is included, and the floating-point

gap of the same model, 1.5376 eV, lies inside. The registered GaAs and AlAs protocols (O22, O23) were not computed, a decision taken on cost alone before any of their certified numbers existed; their claims stay registered and not compared.

Comparison with Experiment. The three S1 records were compared with the frozen datasets `band-gaps-batch1` and `band-gaps-batch2` (version 1, unchanged since registration) by the pre-registered rules (O04, O20, O24). The comparable observation is the low-temperature gap with the zero-point renormalisation and the spin–orbit splitting removed, with the uncertainties added: 1.2437 ± 0.017 eV for Si, 2.6263 ± 0.073 eV for AlP and 2.4617 ± 0.05 eV for GaP. Against the certified bounds (≤ 0.530902, ≤ 1.478051 and ≤ 1.585748 eV), the observation exceeds the model by at least 0.695 eV (Si), 1.075 eV (AlP) and 0.825 eV (GaP): "model below observation", grade V1, for all three. The one-sided rule gives only this lower bound on the deviation, and the object is the discretised KS-LDA model at the protocol's cutoff and k mesh (its distance from continuous KS-LDA is discussed in Section K.2). That LDA underestimates band gaps is textbook knowledge; what is new is that the model side is a proved bound, so no part of the deviation is algebraic, rounding or software error. The comparison tool (`fis.s1val`) only copies numbers from the archived checker outputs and combines them by the registered rule; `verify_all` recomputes every comparison from those outputs. For silicon the two-sided zone check (rule B) has since run at the same settings, 939 CPU hours in 32 separately checked ranges (W52–W66, T4): the gap of the discretised model lies in $[0.4146, 0.4856]$ eV, and DFTK 0.8 lies inside at all 27 compared points. The deviation from the observation is then $[0.741, 0.847]$ eV, still "model below observation" at grade V1. It is recorded as a supplement comparison, the form the protocol fixed in advance for a stronger rule that becomes available after a comparison; the first, one-sided verdict stays in the record (O04). For aluminium phosphide and gallium phosphide the two-sided records (W300, W192) give, in the same way, deviations of $[1.075, 1.293]$ eV and $[0.873, 1.045]$ eV, "model below observation", V1 (O20, O24); for aluminium phosphide the floating-point gaps of the same model computed before the certificate, 1.4779 eV at the protocol's cutoff and 1.4738 eV at 50 Ha, lie inside the certified interval.

The S1 Problem in Lean. The soundness theorems above conclude about the problem written in the certificate. That this is the registered problem was first checked outside Lean, by comparing the echoed problem lines with the hashed pseudopotential file and the registered settings (Q11, T0; the problem-header comparison recorded with R18), and the zone-wide results compose two checker runs whose glue lies outside Lean as well. The registered problem is now itself a Lean definition (K30–K38). A 165-line specification, on top of the 321-line model specification of benchmark D4, which imports only Mathlib and writes out the discrete GTH-LDA plane-wave model (truncated basis, local and nonlocal GTH terms, Hartree term, Slater exchange and PW92 correlation on the grid, zero-temperature self-consistency), defines the problem `S1Si` ($a = 5.416/0.529177210903$ bohr, $Q = 107/2$, the decimals of the file `Si.gth`, two atoms, the unshifted $8 \times 8 \times 8$ Monkhorst–Pack mesh, a 24^3 grid, four occupied bands), the Hamiltonian and bands at any wave vector for any enumeration of its basis, and three conclusions: a self-consistent solution with its gap on the mesh, and the zone-wide bounds of the one-sided and the two-sided rule (`ClaimSCF`, `ClaimU`, `ClaimB`; K36). The 217 bytes of the hashed `Si.gth` are a Lean string; a total parser of the CP2K format gives its coefficients, proved by kernel evaluation without new axioms, and the checker's parameters ($\alpha = 625/242$, v_1, v_2 and the nonlocal record) are derived from this parse by theorems (K30); `make verify-all` checks that the checker prints exactly these bytes, with the registered SHA-256. A Boolean test compares every problem field of a certificate with that of `S1Si` exactly, and a theorem shows that a certificate passing it states exactly `S1Si` (K31). The SCF soundness theorem is restated in the specification's terms (K32): a zero-temperature self-consistent solution with gap at least T, unique among those whose chemical potential follows the convention $\mu' = \lambda_0 + \operatorname{Re} \hat{v}_{\mathrm{eff}}[\gamma'](0)$, with gap at least T and effective potential within r of the centre; uniqueness for an arbitrary chemical potential would need block gaps that the point-group proof does not expose. The bands of the specification at any wave vector are proved to be those the zone checkers bound (K34). In the mode `--expect s1-si` the checker decides the problem test after parsing and before any expensive stage; with `--with-scf` it then checks the SCF certificate, compares the zone certificate's claimed radius and bounds with that output and checks the zone certificate, all

in one evaluation, and the soundness theorems `runS1U_sound` and `runS1B_sound` conclude `ClaimU S1Si` and `ClaimB S1Si` with no premise about another run (K33, K35, T3). For checks split over separate runs, composition theorems state every premise and leave outside Lean that the processes evaluated the same certificate, from the same file bytes with the same executable, and the comparison of the numbers they print (rounded outward). A certificate of another problem and 13 tampered copies (lattice constant, Q, α, v_1, the sign of h_{12}, r_1, atoms, k shift, k mesh, grid, functional and number of occupied bands, and a finite temperature, caught at parsing) are rejected (K37, T0); on the four archived S1 certificates the problem test alone passes, in about six minutes each, which shows that they state this problem but does not bind their conclusions to it. That needs the full check in this mode, which has now run on the archived certificates (K38, T4). One run of `certify-pw3d` with `--expect s1-si --with-scf` on the SCF certificate of K29 and the one-sided zone certificate of W51 passed in 8 h 42 min on 32 cores (213 CPU hours, 92 GB of memory), so `runS1U_sound` gives `ClaimU S1Si` with no premise about another run: exactly one zero-temperature self-consistent solution in the sense of K32 within $r \leq 8.462 \cdot 10^{-11}$ of the certificate centre, with $\mu^* \in [0.2353261934, 0.2353331758]$ Ha, and the zone bounds of W51, whose zone output lines are identical to the archived ones apart from the explanatory text of the line that echoes the SCF claim. These SCF numbers are tighter than, and nested in, those of K29 ($r = 2.093 \cdot 10^{-10}$, $\mu^* \in [0.2353210487, 0.2353383208]$ Ha), which we attribute to the current checker approximating the kinetic prefactor $(2\pi/a)^2$ on a 2^{-96} rather than a 2^{-52} grid (W170), the only change between the two checkers that alters this computation; the run is a new and stronger check of the same certificate, not a reproduction of K29, whose uniqueness radius and mesh-gap interval remain results of the earlier checker. The two-sided result W66 follows by a further composition theorem, `s1_zoneB_of_runS1U`, from that run, the 32 ranges of W66's plan checked in this mode by two further processes (727 CPU hours, about 13 h on two 32-core workers) and their combination, with the glue (same file bytes, same executable, and the comparison of the printed radius and lower bound) outside Lean, as for the split check of W64; the combined gap bounds are identical to the archived ones of W66. `make verify-all` checks the archived outputs of these runs by SHA-256, that their inputs are the certificates recorded for K29, W51 and W66, that their problem, zone and combined lines equal the archived outputs of those claims, and the printed radius and lower-bound comparisons; it does not re-run the checks. `checkS1` and the one-process rule-B check `runS1B` have not run on production certificates, so K33 and K35 stay at T3.

A Certified Band Structure of Silicon. From the S1 SCF certificate of silicon (K29), band certificates at listed points enclose bands 1–8 at 73 rational points along L–Γ–X–W–K–Γ, refined on Γ–X near the conduction-band minimum: 584 enclosures 0.163–0.170 meV wide (W282, T4; Fig. 6). Relative to the valence-band top at Γ, band 5 lies in $[2.5569, 2.5574]$ eV at Γ, $[0.6265, 0.6269]$ eV at X and $[1.4631, 1.4635]$ eV at L. The statements hold at the listed points only; between them the bands of the discrete model are not even continuous, since plane waves enter and leave the cutoff sphere. A further mode of the checker with its own soundness theorem (`bz1_sound`, W280, T4) covers segments of the path by the lower-bound patches and leaves of rule B: everywhere on L–Γ–X–W–K–Γ except the window $t \in (211/256, 441/512) = (0.82421875, 0.861328125)$ of Γ–X, band 5 lies at or above a threshold that exceeds its value at $t = 0.843$, so its minimum along the path lies in that window (W283, T4). With the zone-wide lower bound of rule B (W66) the infimum of band 5 over the whole zone lies in $[0.006086340, 0.008089457]$ Ha relative to μ^* (W285, T1). Where the minimum over the whole zone lies is not certified: that it lies on the path, let alone on Γ–X, would need an exclusion over the whole zone, estimated at 700–1 000 CPU hours with the present checker. DFTK 0.8 on the same discrete problem puts all 584 values inside the intervals and the minimum of band 5 on a fine Γ–X line at $t = 0.845$, inside the window (W284, T0). The band certificates took 89 CPU hours to check, the location certificates 34.

The Full GTH Kernel. Gallium and arsenic need the complete GTH/HGH nonlocal part: angular momenta $l \leq 2$ with up to three projectors per channel. The LDA checker and the zone-wide band checker now take it (W110–W119): the projector Fourier transforms, the d-channel kernel on the fcc lattice and its behaviour under time reversal and the point group are proved in Lean, and without the new terms the checker reduces symbol for symbol

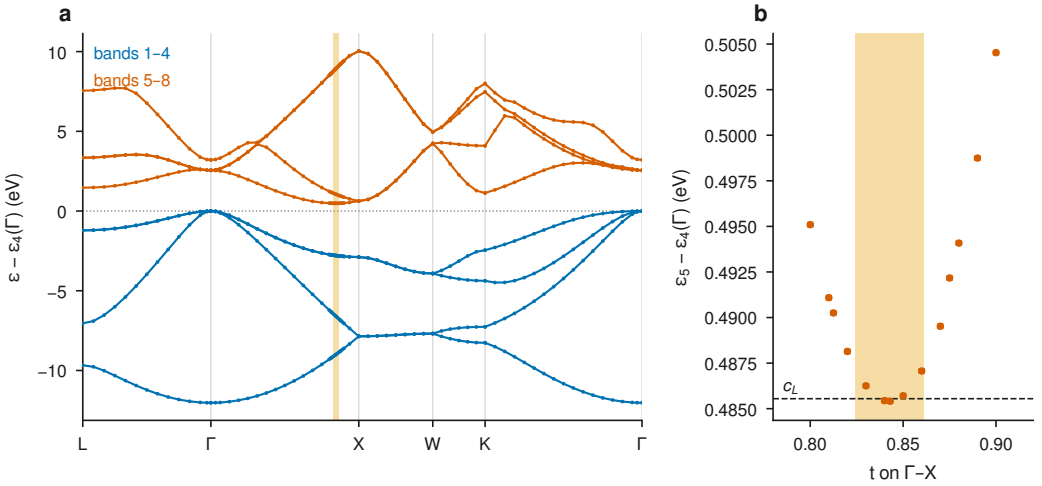


Figure 6: **The certified band structure of the S1 silicon model.** **a**, Bands 1–8 at the 73 listed points along L–Γ–X–W–K–Γ (dots), each enclosed by the Lean checker in an interval 0.163–0.170 meV wide, too narrow to draw (584 enclosures, W282, T4); energies relative to the midpoint of the enclosure of band 4 at Γ. The lines join interval midpoints; nothing is claimed between listed points. **b**, Band 5 on Γ–X, t the fraction of the way to X. Shaded (also in **a**): the window $t \in (211/256, 441/512)$; everywhere on the path outside it band 5 is proved to lie at or above the threshold c_L (dashed, shifted like the bands), which exceeds its value at $t = 0.843$, so the minimum along the path lies in the window (W283, T4). Where the minimum over the whole zone lies is not certified. The object is the discrete model of the protocol (cutoff $Q = 107/2$, $8 \times 8 \times 8$ mesh, the SCF solution of K29).

to the old one. On a GaAs test at a low cutoff ($a = 10.5$ bohr, about 1 Ha, $2 \times 2 \times 2$ mesh) the SCF certificate has radius $5.6 \cdot 10^{-12}$, all 104 DFTK eigenvalues lie inside its intervals, and the band check passes. (GaP at S1 settings is described above; GaAs and AlAs were registered but not computed.) The two-sided zone check (rule B) carries over to the full kernel, including the range-split form that lets parts be checked and resumed separately (W120–W129, T4); on the same GaAs test it proves the band gap over the whole zone to lie in $[0.188, 2.202]$ eV, the first positive lower bound with this kernel.

Extended Precision. For gallium arsenide at its registered settings the double-precision certificate cannot close: the Newton–Kantorovich quantity $4a\eta_T$ is 23.9 on a $2 \times 2 \times 2$ preflight. The k-point checker now rounds the kinetic prefactor to 2^{-96} (W170, T4; the soundness statements are unchanged), and the driver refines the centre in extended precision (W171); the same preflight gives $4a\eta_T \approx 0.011$ (W172, a preflight, not a certificate).

A Lattice-Constant Bracket. Certified SCF solutions and total energies of silicon at three lattice constants, with the basis fixed in reduced coordinates ($Q = 107/2$) and the other S1 settings unchanged, put the middle point at least $4.09 \cdot 10^{-7}$ Ha below both ends (by at least 0.49 and 0.41 μHa at the two ends) with intervals at most $2.61 \cdot 10^{-7}$ Ha wide (W105–W107). A bracket theorem proved in Lean (W103, W104), with the comparison decided by the kernel on exact numbers, then places every energy minimum of this fixed-Q discrete family in $(5.354683, 5.358683)$ Å, provided the certified solutions at the two end points are ground states, an assumption the checker does not provide. It is not the converged LDA lattice constant (Section K.2) and is not compared with experiment.

An Audit of a Classic Pseudopotential Table. Chelikowsky and Cohen's 1976 nonlocal empirical pseudopotentials are a standard reference for semiconductor band structures. Under a protocol registered before any band energy was computed (O21), a new checker, `certify-pwnl` (W155, T4), bounds the eigenvalues of the model exactly as printed, for nine

parameter sets at Γ, X and L, with intervals at most $5 \cdot 10^{-7}$ eV wide (W157). Of 111 registered printed values, 28 agree to the printed digits and 64 within the paper's stated accuracy (W158). The clearest case is silicon's local model, whose parameters are printed in full: all 11 printed levels lie 0.034–0.087 eV from the model's eigenvalues, a difference that neither the truncation (changes below $4 \cdot 10^{-4}$ eV between $Q = 33.5$ and 150) nor the paper's own basis explains (W159). The statement is that the printed table is not the spectrum of the printed model to the printed digits; it involves no experiment.

Three Codes against Certified Answers. Under a protocol frozen before the runs (O25), ABINIT 10.0.3, Quantum ESPRESSO 7.6 and DFTK 0.8.0 were run 219 times on four discrete problems whose answers are certified intervals, each number judged inside or outside with no tolerance (W201). With inputs aligned to the certified problem the codes reproduce it wherever they converge, the exceptions being energies at the 10^{-8} Ha level and below (W202); with default inputs none of the 12 (problem, code) pairs is inside, mainly because ABINIT and QE default to another LDA correlation, a gap shift of 1.8–4.0 meV (W203). With convergence set by an energy tolerance of 10^{-10}, ABINIT and QE reported convergence with exit code 0 and no warning while unoccupied-band eigenvalues lay up to $8 \cdot 10^{-3}$ Ha outside the intervals (occupied bands and the gap within 10^{-7} Ha); one aligned ABINIT run with as many bands as plane waves exited normally, with a single iteration warning, and a gap 0.22 Ha off (W204). The frozen protocol gave ABINIT no buffer bands (`nbdbuf`), so in two aligned runs (P2, P3) the highest computed bands did not reach the wave-function tolerance and were recorded as not converged; this was an oversight of the protocol, recorded in the audit notes, and its effect on the verdicts is read post hoc below. These statements are about these inputs and versions.

Consequences and Defaults. A second protocol, frozen before any computation (W208), asked which of these deviations change quantities users derive from band structures and energies, with thresholds fixed in advance (10 meV for energy differences, 1% for curvatures, 1 meV/atom, 0.001 Å). Of the 131 runs judged outside or not converged, 28 change at least one quantity (ABINIT 21, QE 7, DFTK 0; W209); with default-style inputs on the two S1 problems the gaps move by 1.8–3.4 meV and nothing changes at the main thresholds (W211), while ABINIT run as its own error message suggests (`chksymbreak 0`) silently shifts the gap of the small silicon problem by +18.0 meV. QE's documented Ewald precision uses 0.24 of the smallest margin of the lattice-constant bracket, which therefore stands (W213). A third frozen protocol left every numerical setting at the program defaults or as in the official tutorials, with only the physical problem aligned (W230): all 33 such runs that gave numbers are outside, 29 of them silently (W231, W232). The largest effect is on the small silicon problem, whose k-point grid is shifted by $\frac{1}{2}$: QE by default, and ABINIT once its refusal is overridden, symmetrise the density with all 48 cubic operations, 36 of which do not preserve the grid (and use a different FFT grid, 10^3 and 12^3 against 8^3), and so solve a different discrete problem (gap +20.3 to +21.5 meV, −2.50 meV/atom; a pair on the same 12^3 grid that differs only in symmetry gives +20.25 meV and −2.49 meV/atom, so symmetrisation dominates). The mechanism was established by a single-factor diagnostic run after the results, which also confirmed that the k points the programs used equal the certificate grid (W233); it changes no verdict. Fig. 4 plots for each run the quantity Q2 of both protocols, the mesh band gap $\min_k \varepsilon_{N+1} - \max_k \varepsilon_N$ over the program's k points, minus the centre of the certified interval (axis linear below 10^{-6} meV), with the marker giving the verdict on all compared quantities (plane-wave counts, eigenvalues, gap, energies).

Each Code's Own Tolerance (Post Hoc). To separate the causes of the outside verdicts, we read the archived outputs of both protocols once more against what each program itself promises (W316–W320, T0). This reading is post hoc: its definition was committed on its own before any of its numbers was computed, but its author had read the results of O25–O27, so it is neither pre-registered nor blind, and it changes no frozen verdict. For every run the widening follows from the input and the documented defaults: an ABINIT `toldfe` t widens eigenvalues and energies by t, `tolvrs` and `tolwfr` by $\sqrt{t}$; QE's `conv_thr` by half its value, with, for empty bands under the default `diago_full_acc = .false.`, half

of $\max(5\,\texttt{ethr}, 10^{-5})$ Ry more, and the documented Ewald precision for the total energy; DFTK's `tol` by its value, plus 10^{-3} for its band step. Only bands the program claims converged are compared, so DFTK's bands beyond `n_bands_converge` and ABINIT's buffer bands are left out, and different plane-wave sets or k points are never widened away. The definition read every criterion literally (nominal, its main reading) and, as a sensitivity, let an energy criterion t widen eigenvalues by $\sqrt{t}$ (scaled; an order of magnitude with an $O(1)$ constant, not a bound), since the total energy is second order in the density error and eigenvalues first order; it classified the causes of each outside run under the scaled reading, in this order: (i) a different discrete problem, read from the program's own output (plane-wave set, k points, functional, FFT grid, or more than the 12 grid-preserving operations on the shifted grid), (ii) inside the code's own tolerance, (iii) inside once the top one or two computed bands are dropped, and (iv) other. Among the runs that claim convergence, 42 (nominal) and 46 (scaled) of 138 in O25 lie within the code's own tolerance, against 33 inside the unwidened intervals (ABINIT 11 and 13 of 47, QE 7 and 9 of 46, DFTK 24 of 45), and 20 and 42 of 60 in O26, against 7 (ABINIT 6 and 16 of 23, QE 2 and 14 of 20, DFTK 12 of 17). Of the 33 default and tutorial runs that gave numbers, 11 solved a different discrete problem (10 of them without a warning): six symmetrised the density of the shifted grid with 48 operations (ABINIT after `chksymbreak 0`, and QE), and five were DFTK at its default FFT grid instead of the certified 8^3 (12^3 on P1, gap $+0.03$ meV; 9^3 on P4, -0.21 meV); the other 22 (19 without a warning) solved the certified problem and lie within tolerance under the scaled reading, so 24 of the 33 do in all, two of the DFTK runs included, and 11 under the nominal one, where the 33 split as 11, 9, 3 and 10 over the four classes. A tolerance check on the gap alone would have let the symmetrised runs pass, since $+20.3$ meV lies inside $2\sqrt{t}$ (54 meV for ABINIT's `toldfe`, 38 meV for QE's `conv_thr`); only their total energy, -2.50 meV/atom, is outside. Of the 125 outside runs of O25, 91 solved a different problem, 10 are within tolerance, 4 fall into class (iii) and 20 are other: 8 in which the program itself reports non-convergence, 5 QE runs on the toy problem whose E_L is outside, 6 QE runs on P4 outside by $9.4 \cdot 10^{-11}$ Ha, and one ABINIT run whose lowest band is off by 1.49 meV, inside the program's own reported $\sqrt{\texttt{residm}}$ of 3 meV; of the 54 of O26, 22, 28, 3 and 1 fall into classes (i)–(iv). Class (iii), 4 runs in O25 and 3 in O26, consists exactly of the silent empty-band deviations at tight energy tolerances, those of W204 and their counterparts in O26: their large errors sit in the top one or two computed bands, and without them the runs lie within the code's own tolerance (under the scaled reading, by which the classes are defined; read literally, 1 of the 4 in O25 and 1 of the 3 in O26); ABINIT's band step keeps two buffer bands by default, and QE documents its looser threshold for empty bands, so this reflects the missing buffer bands of our protocol, not behaviour that contradicts the programs' documentation (W318). On the two realistic problems, silicon (P2) and aluminium phosphide (P3), none of the 63 runs of the consequence protocol that gave numbers changes a derived quantity at the main thresholds (all 28 changes are on P1 and P4), and of the 34 default and tutorial runs only ABINIT's AlP tutorial band step does, by splitting the bands 7 and 8 at X by 4.7 meV: exactly the two buffer bands that ABINIT does not converge by default, with a warning; no quantity a program claims converged changes (W319). For the toy problem P1 (cutoff 1.594 Ha, $2 \times 2 \times 2$ grid shifted by $\frac{1}{2}$) a floating-point DFTK reference changes only the k quadrature of the density, up to an unshifted $48 \times 48 \times 48$ mesh, and compares the gap over the eight points of the certificate's grid (the minimum of band 5 minus the maximum of band 4), the smallest direct gap there and the total energy per atom (W320). The certified problem lies -28.8 meV, -34.3 meV and $+46.2$ meV/atom from the k-converged values, the symmetrised one -8.5 meV, -5.0 meV and $+43.7$ meV/atom, with k-convergence errors estimated at 0.075 meV, 0.043 meV and 0.55 meV/atom; as controls, DFTK on the certificate's problem lies inside the certified gap interval, and the symmetrised minus the certified gap, $+20.2828$ meV, equals the registered QE value to 10^{-6} meV. Two deviations from the definition are recorded: the reference runs use a 12^3 FFT grid, since DFTK's dense diagonalisation refuses 8^3 at points off the certificate's grid (0.03 meV on the certified problem), and the k meshes were extended to 48^3, since the defined $n \leq 20$ left an energy uncertainty of 8.0 meV/atom. From k-convergence the certified problem thus lies 29–46 meV and the symmetrised one 5–44 meV, closer in all three quantities; the audit's statement is that a number is not the solution of the certified discrete problem, not that

one discrete problem is better. Two limitations remain. The reading of a program's own tolerance is not unique: the nominal and scaled readings give 11 and 24 of the 33, and both are reported. And the certified problems are written in DFTK's conventions (cutoff sphere, FFT and integration grid, form of the GTH pseudopotential; Appendix K), so agreement with DFTK and disagreement with ABINIT and QE is partly a matter of convention; the aligned runs, in which ABINIT and QE reproduce the certified problem from its written definition wherever they converge (W202), are the evidence that the problem is not DFTK's alone.

K.1 The GTH-PBE Pseudopotential Layer

The Question. PBE itself is not certified (Appendix S), but the GTH mode of the checker takes the local part, the projectors and the coupling matrix h of a pseudopotential as parameters (Q07–Q09), so Krack's GTH-PBE pseudopotential for silicon needs no new mathematics. A fourth frozen protocol (O28) asked how ABINIT 10.0.3, Quantum ESPRESSO 7.6 and DFTK 0.8.0 read and use this file, in particular its off-diagonal element h_{12}. Before any run, an analysis of the files found that in Krack's GTH-PBE family, as distributed for CP2K and DFTK, all 211 channels with two or three projectors satisfy the back-transformation of the generating program's orthogonalised projectors (for silicon $h_{12} = -\sqrt{3/5}\,h_{22} = -2.70627082\,\mathrm{Ha}$) and none satisfies the HGH relation ($h_{12} = -\frac{1}{2}\sqrt{3/5}\,h_{22}$) by which diagonal-only formats (ABINIT's pspcod 3, a branch of QE's `hgh2qe`) rebuild the off-diagonal elements; in the GTH-LDA family of the earlier audit 158 of 205 such channels, silicon's among them, satisfy it, so that audit could not tell how a program treats h_{12}. Krack's h_{12} is not an error.

Pre-Registration. The protocol was frozen in commit `87486f14` and pushed as a timestamp branch on 22 October 2026 at 15:27:08 (UTC+8), before ABINIT, QE or DFTK ran any registered input (W220). It fixes six paths: Krack's CP2K-format file F0 and five conversions or distributions of it (Table 11), among them a diagonal-only pspcod 3 file as positive control (F3); the verdict rules of the code audit, unchanged and with no tolerance; four counterfactuals for h_{12} (H-a: the HGH relation; zero; opposite sign; doubled); and 14 predictions. ABINIT and QE were blind, their comparison table committed before any DFTK run on the PBE files; DFTK was not, since a feasibility study had read the file.

What Is Certified. The carrier is reduced Hartree–Fock, fcc silicon at the Γ point, Fermi–Dirac occupation at $T = 1/100\,\mathrm{Ha}$, on three discrete problems G2, G3 and G4 with 27, 169 and 259 plane waves. Six certificates, for F0 and for the GTH-LDA calibration file, pass the checker; for F0 the first 16 bands have half-widths at most $2.8 \cdot 10^{-10}$, $1.9 \cdot 10^{-8}$ and $8.1 \cdot 10^{-8}\,\mathrm{Ha}$, and the checker's echo equals the file digit by digit, $h_{12} = -2.70627082$ included (W221, T4). A dump mode of the checker prints every local form factor (11, 38 and 50 shells) and every nonlocal matrix element (729, 28,561 and 67,081) of the three problems as rational intervals with errors at most $8.8 \cdot 10^{-19}$ (W226, T4): none of DFTK's is outside beyond floating-point resolution (largest distances $1.7 \cdot 10^{-14}$ and $1.6 \cdot 10^{-16}$), while under counterfactual H-a every matrix element to which h_{12} contributes is outside (largest distance 0.114). Calibration with the GTH-LDA file put all three programs inside on all three problems.

Program	Path	h_{12}	Outcome on G2, G3, G4
ABINIT 10.0.3	F1, F2 (pspcod 10)	read	inside; deviation $\leq 1.2 \cdot 10^{-11}$ Ha
ABINIT 10.0.3	F3 (pspcod 3)	rebuilt	outside, matching H-a; on G2 exit code 0, convergence reported, no warning on h_{12}; G3, G4 not converged
QE 7.6	F1, F2 (`.gth`)	read	inside; deviation $\leq 1.8 \cdot 10^{-10}$ Ha; not converged
QE 7.6	F4 (`upfconv.x`)	–	cannot be read; exit code 0, no file written
QE 7.6	F5 (QE's UPF)	read	G2 outside by $1.0 \cdot 10^{-8}$ Ha (36 half-widths)
DFTK 0.8.0	F0 (CP2K format)	read	inside; deviation $\leq 1.6 \cdot 10^{-13}$ Ha (not blind)
DFTK 0.8.0	F5 (QE's UPF)	read	G2 outside by $1.7 \cdot 10^{-8}$ Ha (61 half-widths)
CP2K 2026.2	F0 (native)	read	not run; source code only

Table 11: The GTH-PBE pseudopotential audit (O28; W223–W225). F1 is cp2k-data's ABINIT conversion, F2 ours. F4 converts a CPMD file. "Read": the full h matrix is taken from the file; "rebuilt": from diagonal elements by the HGH relation. Inside and outside: eigenvalues against the certified intervals, with a common energy zero; F5 is inside on G3 and G4.

Results. No path that reads the full h matrix (ABINIT F1 and F2, QE F1 and F2, DFTK F0) shows an h_{12} convention error: ABINIT echoes $h_{12} = -2.7062708$ and DFTK's fields are the nearest doubles of the file's numbers (W222), every eigenvalue comparison is inside, and all four counterfactuals are at least 0.099 Ha away (W227). The positive control F3 shows that the audit can see such an error: it is outside on all three problems, and its eigenvalues agree with those of H-a band by band to $1.4 \cdot 10^{-10}$, $3.5 \cdot 10^{-11}$ and $5.1 \cdot 10^{-11}$ Ha, deviate by up to 0.51 Ha, and put the lowest band 0.084, 0.041 and 0.038 Ha above its certified interval. ABINIT echoes only the diagonal elements and gives no warning about the off-diagonal ones; on the smallest problem, G2, it exits with code 0 and reports convergence (on G3 and G4 it does not converge), and since the file it was given no longer carries F0's h_{12}, the error is attributed to the conversion (W223). The null result is that these programs correctly consume an h_{12} that the HGH relation cannot check; the risk lies in diagonal-only formats, which give another pseudopotential for every one of the 211 multi-projector channels of the family.

Side Findings. QE's `upfconv.x`, whose help text lists CPMD files, cannot read the CPMD file (F4) yet exits with code 0 (W222). QE's UPF file (F5, byte-identical to the one on QE's website on 22 October 2026) carries the ten significant digits of the original fit where the CP2K file has eight decimals; the difference in h_{12}, $3.0 \cdot 10^{-9}$, is within F0's printed rounding, so this is not a convention error but another pseudopotential, differing at the ninth significant digit, whose shift of about 10^{-8} Ha is outside on G2, the narrowest problem. Post hoc, in floating point and outside the verdict, QE's F5 eigenvalues lie within $8.6 \cdot 10^{-10}$ Ha of those of F0 with the ten-digit parameters (W224). ABINIT gives no warning that the file's functional (PBE) differs from the input's (none).

Predictions, Deviations and Limitations. Of the 14 predictions, 9 hit, 2 partly and 3 missed: the warning on the functional, and the F5 deviations of QE and DFTK, about 10^{-8} Ha against a predicted 10^{-7} or more and outside only on G2 (W228). Three dated deviations are recorded (a longer time limit for the G4 certificate checks, QE's spelling `'NOX-NOC'` of the empty functional, two script corrections with both versions reported), and `make verify-all` re-derives every verdict from the archive. The spectral carrier is reduced Hartree–Fock, not PBE: GTH-PBE is certified only at the pseudopotential layer, and the PBE functional is out of scope. No QE run reached its convergence threshold (10^{-24} Ry), though its numbers for the calibration file, F1 and F2 are inside; CP2K's handling of h rests on its source code; and the statements are about these inputs and versions.

A Census of GTH/HGH Parameter Files. A further protocol, frozen and timestamped before any comparison (O29, W240), compared the numbers in every publicly distributed GTH/HGH parameter file exactly, with no tolerance beyond the rounding of the printed digits: 2821 files from seven sources (CP2K, cp2k-data, PseudoLibrary as used by DFTK,

ABINIT, the Quantum ESPRESSO website, PySCF and this repository), 7040 entries and 113 775 parameters, across distributions, against the tables of the 1996, 1998 and 2005 papers (Goedecker et al., 1996; Hartwigsen et al., 1998; Krack, 2005) and against their internal relations (W241–W252, T0). Of 946 groups of copies, 932 agree in every member; CP2K v2026.2 and the cp2k-data repository it draws from give two different pseudopotentials under the name B `GTH-PBESOL-q3`, PySCF's UZH file gives other pseudopotentials for Ca, Cr, Tc, Ru, Rh and Pd, and its Na `GTH-*-q9` carries CP2K's `_old` numbers (W243, W244). Against the papers, 1856 of 1926 matched HGH entries and 1126 of 1186 Krack entries agree; the semicore Ce, Ga, Y, Zr and Na `GTH-PADE` entries of the CP2K lineage, which includes the LDA families DFTK uses, and the Krack Na and Ga entries differ from the printed tables, while ABINIT's copies follow them (W245, W247); no difference is a transcription error. All 7040 valence charges agree (W248). The census is about the numbers in the files; whether a difference changes a computed quantity is not part of this protocol.

K.2 From the Discrete Model to Continuous KS-LDA

The Object Certified. Every S1 statement is about a discrete problem: silicon with GTH-LDA q4, Slater exchange and PW92 correlation at $a = 5.416$ Å, the plane waves inside the cutoff $Q = 107/2$ (about 10 Ha), exchange–correlation integrated on a 24^3 grid, the unshifted $8 \times 8 \times 8$ Monkhorst–Pack mesh and zero temperature (K29, T4). The continuous KS-LDA model with the same pseudopotential, functional and k quadrature differs from it only by the cutoff. How far apart the two are is known from floating-point evidence and, for one part, from a certificate; a rigorous bound on the whole difference is open.

The Truncation Effect, Part by Part. A checker mode bounds eigenvalues of the continuum operator $T + V_{\mathrm{loc}} + V_{\mathrm{nl}} + w_D - \mu^*$, with no cutoff and the Hartree and exchange–correlation potential frozen at the certified SCF solution; continuum eigenvalues are infima over all finite plane-wave sets, so Lean needs no unbounded-operator theory. A Rayleigh–Ritz upper bound on a 45 Ha section and a Schur-type lower bound on a 35 Ha section, with closed-form pseudopotential tail sums, give two-sided intervals (W143, W144, T3; checker soundness W145, T4). Two production certificates, for the valence-band maximum at Γ and the lowest conduction band on the mesh at $X_m = (0, \frac{5}{8}, \frac{5}{8})$, took about 1.5 h each on 16 cores (W146); Table 12 sets them beside the floating-point decomposition of the full truncation effect on the mesh gap $\varepsilon_5(X_m) - \varepsilon_4(\Gamma)$ (W140). DFTK 0.8 diagonalising the frozen operator at 35, 45 and 80 Ha (2685–9257 plane waves) lies inside every certified interval, with a gap shift of $+5.25$ meV at 80 Ha (W147, T0). The certificate bounds only these two points: it does not prove that the band edges of the frozen continuum operator stay at Γ and X_m (in floating point they do).

Continuum minus discrete	Certified	Floating point	Ledger
$\varepsilon_4(\Gamma)$, frozen potential (Ha)	$[-1.377, -1.338] \cdot 10^{-3}$	$-1.352 \cdot 10^{-3}$	W146 T4; W140 T0
$\varepsilon_5(X_m)$, frozen potential (Ha)	$[-1.185, -1.145] \cdot 10^{-3}$	$-1.159 \cdot 10^{-3}$	W146 T4; W140 T0
Mesh gap, frozen potential	[4.17, 6.32] meV	$+5.25$ meV	W146 T4; W140 T0
Mesh gap, self-consistent relaxation	not bounded	-4.04 meV	W140 T0
Mesh gap, total	not bounded	$+1.21$ meV	W140 T0

Table 12: The effect of the plane-wave cutoff on the S1 silicon gap over the $8 \times 8 \times 8$ mesh. Certified: the continuum operator with the potential frozen at the certified SCF solution, minus the discrete model. Floating point: DFTK 0.8 against an 80 Ha reference (the 40 Ha reference gives the same frozen-potential shift to $3 \cdot 10^{-9}$ Ha).

Floating-Point Convergence. For silicon (Z09–Z12, T0), DFTK at the certified settings gives a mesh gap of 0.5307380 eV and a gap over the whole zone of 0.485409 eV. Raising the cutoff from 10.08 to 50 Ha changes the whole-zone gap by -0.76 meV and the mesh gap by $+1.21$ meV, not monotonically; refining the SCF mesh from 8^3 to 16^3 adds $+1.23$ and $+1.19$ meV. At the converged reference (50 Ha, 16^3) the whole-zone gap is 0.485874 eV, $+0.47$ meV from the certified settings, and the mesh gap $+2.41$ meV. For aluminium phos-

phide at its registered settings ($a = 5.4672$ Å, $Q = 109/2$, 8^3; W142, T0) the gap, 1.477880 eV between Γ and X, both on the mesh, agrees with the certified value; the cutoff raised to 40 Ha changes it by -4.10 meV and the SCF mesh refined to 12^3 by $+0.73$ meV, and the converged reference (40 Ha, 12^3) gives 1.474541 eV, -3.34 meV. All of these are at the meV level, against certified whole-zone intervals 0.0709 eV wide for silicon, aluminium phosphide and gallium phosphide (W66, W300, W192). The width is the same for all three because the driver places the zone-wide bounds at fixed margins from its floating-point band edges, about $6 \cdot 10^{-4}$ Ha above the valence-band maximum and $2 \cdot 10^{-3}$ Ha below the conduction-band minimum; the upper end of each interval comes from listed points, so the floating-point gap of the same model lies near it, and tighter margins would need smaller cubes and more leaves.

Why a Rigorous Continuum Bound Is Out of Reach. The difficulty is the cancellation in Table 12, which a bound built from norms cannot see. In the linear part, the residual of the discrete eigenvectors in the large basis, 0.105–0.107 Ha, lies almost entirely outside the cutoff sphere (W140), and classical residual bounds fail: Bauer–Fike (Bauer and Fike, 1960) gives ± 0.11 Ha per band edge, and Kato–Temple (Kato, 1949; Temple, 1928), whose denominator is the distance to the neighbouring levels (0.02–0.09 Ha here), gives 0.12–0.58 Ha, worse than no bound; only a Schur-type bound whose denominator is the kinetic energy outside the cutoff, about 10 Ha, has the right size, about $0.9 \cdot 10^{-3}$ Ha against measured shifts of $1.16–1.35 \cdot 10^{-3}$ Ha. In the self-consistent part, the fixed-point residual of the trial potential from the certified 10 Ha solution is $\eta_\infty = 3.4 \cdot 10^{-2}$ Ha in the sup norm (80 Ha reference), and the true change of the potential is of the same size ($r/\eta = 0.99–1.02$; W141). The gap, however, moves by only $-1.5 \cdot 10^{-4}$ Ha, about 470 times less than Weyl's bound $2r$, because the change of the potential is of high frequency and nearly orthogonal to the smooth band-edge densities; a Weyl-type transfer gives a gap interval of ± 1.9 eV even with r known. A bound must therefore start from a trial potential at a much higher cutoff; its residual falls by a factor 6–7 per doubling of the cutoff ($3.4 \cdot 10^{-2}$, $4.8 \cdot 10^{-3}$ and $7.9 \cdot 10^{-4}$ Ha at 10, 20 and 40 Ha), so a 1 meV gap bound would need about 300 Ha and a 10 meV bound about 130 Ha (extrapolated; W141, T0). More fundamentally, the Newton–Kantorovich argument in potential space does not close with the constants we can prove: on a Γ-point model the present rigorous-type bound on the second-order constant is $3.0–3.8 \cdot 10^3$, against directional values of at least 39 (sign-type directions normalised by their grid supremum; an earlier normalisation by the supremum of the trigonometric interpolant had given about 3), so that the closure quantity $4\beta^2 q\eta$, which must be at most 1, is 50–180 at 40 Ha (C46, T0). The blocking constant is the bulk second-order response (about three quarters Hartree, one quarter exchange–correlation), not the low density at the ions. In the most favourable extrapolation, with improved rigorous constants on the Γ-point model, the argument closes near 60 Ha and gives a gap bound about 80 meV wide, still two orders of magnitude above the true effect (W141). Guaranteed a posteriori bounds in the literature cover linear, non-self-consistent operators (Herbst et al., 2020) and the energy of convex functionals (Bordignon et al., 2025); for LDA only asymptotic estimates exist (Cancès et al., 2022), and we found no guaranteed bound on a self-consistent KS-LDA gap with computable constants (W214). On the frozen-potential shift, DFTK's residual bounds are $5.3 \cdot 10^3$ and $1.15 \cdot 10^4$ times wider than the certified interval, and both its second-order estimate (2.34 meV) and a convergence-study value at 20 Ha (6.52 meV) fall outside it (W218, T0; Appendix R). A rigorous upper bound on the bulk second-order response constants, within about a factor 10 of the measured values, is therefore the open problem on which a certified continuum gap depends.

What Is Proved about the Continuum. In density space the local continuum argument for LDA does not close on real Si (C34, T1; C38, T0), because the second-order constant grows like $\rho_{\min}^{-5/3}$ at the ion sites. Taking the local Kohn–Sham potential as the unknown removes this obstruction: the square root of a projected density is 1-Lipschitz in the projector difference, so $\rho \geq 0$ need not be proved (C42, T3); the linearisation remainder of Slater exchange is quadratic in $\sqrt{\rho}$ for all $\rho \geq 0$ with the optimal constant $2|v_x'(\rho')|$ (C41, T3); their combination bounds the exchange part of the remainder pointwise with a coefficient that grows only like $\rho'^{-2/3}$ (C43, T3); and the abstract fixed-point and Newton–Kantorovich

theorems are in Lean (C44, T3). On this basis Theorem P states that if $4\beta^2 q\eta \le 1$ and the gap is preserved, a ball of radius at most $2\beta\eta$ around the trial potential, in a weighted sup norm, contains a zero-temperature SCF solution of continuum KS-LDA, whose eigenvalues differ from the trial ones by at most $\varphi_{\max}$ times the radius (C40, T1); Theorem P′ adds local uniqueness under a positivity margin (C45, T1). The Schauder fixed-point theorem and the compactness argument are on paper only, as Mathlib has no Schauder theorem; with the directional estimate of the second-order constant (at least 39, a lower estimate) the closure quantity is below 1 from about 40 Ha, with rigorous-type constants it is not (C46). These are the only statements about continuum KS-LDA in this paper; for reduced Hartree–Fock, which is convex, the continuum limit is certified (C29–C33, T3–T4).

The Lattice Constant. The bracket $(5.354683, 5.358683)$ Å (W105–W107) is likewise a statement about the discrete model at fixed $Q = 107/2$. A floating-point fit (T0, not a ledger claim) puts the energy minimum of this model 0.0244 Å below that of a converged reference (5.3811 Å at 50 Ha), six times the width of the bracket; the bracket is therefore not a bound on the converged LDA lattice constant, and is not compared with experiment.

L Finite Elements

Problems. `mini-FEM` solves $-u'' = f$ on $(0, L)$ with boundary conditions $A_0 u(0) - B_0 u'(0) = g_0$, $A_L u(L) + B_L u'(L) = g_L$ (Dirichlet, Neumann and Robin, including Neumann at both ends) using P1 elements. It also solves the Euler–Bernoulli beam equation $u'''' = f$ with clamped, pinned, free and sliding ends using cubic Hermite elements. Right-hand sides are sums and products of polynomials, $e^{\lambda x}$, $\sin$ / $\cos(\omega x)$ and $\sin$ / $\cos(q\pi x)$.

Avoiding Sobolev Spaces. Mathlib has no Sobolev spaces yet. In one dimension we work with classical solutions and Green's functions instead. We prove that the Green representation $u = \ell + \int_0^L G(\cdot, y) f(y)\, dy$ is a classical solution and that it is the only one (FEM-C01). The key discrete fact is *nodal exactness*: on any mesh, the nodal values of the true solution satisfy all P1 equations, and every vector of nodal values admits a discrete Green representation (D01). Together with the element-wise interpolation bound $|u(t) - V_h(t)| \le Mh_k^2/8 + \max_{\text{ends}} |u(x_j) - V_j|$ (D02) and a discrete maximum principle (D04), this gives pointwise control of the error against the true solution. We also prove Céa's lemma (Céa, 1964) for bounded coercive bilinear forms on real normed spaces (D03) in preparation for 2D.

Certificate. The checker builds a Taylor model on each element for the right-hand side expression, proved valid by `FExpr.tm_valid`, and integrates its polynomial part exactly to enclose the load vector. From candidate error vectors $e^{\pm}$, it verifies exactly that the residual equations $R(e^{\pm}) = r^{\pm}$ hold; the discrete maximum principle then gives $e^- \le u - V \le e^+$ at the nodes. The output encloses the *continuous* solution at every node (`NODE_OK`) and bounds $\sup_t |u(t) - V_h(t)|$ over the whole interval (`LINF_OK`, FEM-B01). The beam checker replaces the maximum principle, which fails for fourth-order problems, with a discrete stability estimate obtained from Taylor transfer matrices (D05). It encloses nodal values and slopes, and bounds the interpolation error by $Mh^4/384$ (D06, B03).

Pure Neumann Problems. With Neumann conditions $-u'(0) = \hat{g}_0$, $u'(L) = \hat{g}_L$ at both ends, a solution exists only if $\int_0^L f + \hat{g}_0 + \hat{g}_L = 0$, and it is unique only up to a constant. We therefore certify the projected problem $-u'' = f - \bar{c}$, $\bar{c} = (\int_0^L f + \hat{g}_0 + \hat{g}_L)/L$, normalised by $\int_0^L u = 0$: its classical solution exists and is unique, and $\bar{c} = 0$ whenever the original problem has a classical solution, so for compatible data the two coincide (FEM-C03). Nodal exactness survives up to a computable shift: the true nodal values minus $s = -\sum_k \beta_k / L$, where β_k is the integral of the element bubble $(y - a)(b - y)/2$ against $f - \bar{c}$, solve the discrete Neumann system normalised by $\int W_h = 0$ (D07). The checker encloses $\bar{c}$ (`COMPAT_OK`) and s (`SHIFT_OK`) by integrating Taylor models exactly, reduces the error equation to a mixed

Dirichlet–Neumann problem so that the maximum principle applies again, and outputs nodal enclosures and the sup-norm bound as before (B04, T4). For polynomial loads $\bar{c}$ is exact and the checker requires $\bar{c} = 0$, an exact decision of compatibility; for transcendental loads it certifies the projected problem. scikit-fem's Lagrange-multiplier solution differs from the driver's by a constant, which lies inside `SHIFT_OK` up to scikit-fem's rounding.

Certified Adaptivity. With `--adapt tol`, the driver bisects every element whose *proven* element bound `ELEM_OK` exceeds the tolerance, rather than an error *estimate*. The refinement loop itself is untrusted; the conclusion rests on the certificate of the final mesh. For $f = 20e^{4x}$ (P1), 9 rounds reach $N = 731$; for a cantilever with $f = 100e^{10x}$, 7 rounds reach $N = 146$. In both cases the final certified sup-norm error is within tolerance.

Eigenvalues. With `--eigs n` the driver solves the P1 pencil $Kx = \lambda Mx$ for $-u'' = \lambda u$ with Dirichlet, Neumann or Robin ends and writes rational shifts σ with claimed counts. The checker evaluates the pivots of the tridiagonal $K - \sigma M$ in interval arithmetic; when none contains zero, Sylvester's law of inertia for the pencil (FEM-D08, reusing the argument of DFT-A04) turns the number of negative pivots into the exact number of eigenvalues below σ, and two counts bracket the k-th discrete eigenvalue (B05, T4). On six cases the intervals have relative width at most $2.1 \cdot 10^{-15}$ and contain a 50-digit reference; scikit-fem agrees to $3 \cdot 10^{-15}$ ($2.3 \cdot 10^{-13}$ for Neumann at both ends, solved by shift-invert), the observed convergence order is 2.000, and checking takes 0.45 s for $N = 1024$ with four eigenvalues and 2.6 s for $N = 2000$ with ten. For Dirichlet ends the continuous $(k\pi/L)^2$ lies below every Lean lower bound of these discrete eigenvalues; that this must be so is proved next.

Energy-Norm Errors. With `--energy` the checker also bounds the error in the energy norm, $\|u - V_h\|_E^2 = \sum_k \int_k (u' - s_k)^2 + \kappa_0 (u(0) - V_0)^2 + \kappa_L (u(L) - V_N)^2$, where s_k is the slope of V_h on element k and $\kappa_0, \kappa_L \geq 0$ are the Robin coefficients, zero at Dirichlet and Neumann ends (FEM-B20, T4). On an element, $U = u'$ satisfies $U' = -f$; with the Taylor model (p, δ) of f from B02 and an antiderivative q of p bounded by M, $\int_a^b (U - s)^2 \leq \int_a^b q^2 + M\delta(b-a)^2 + \delta^2(b-a)^3/3 + (\varepsilon_{k+1} - \varepsilon_k)^2/(b-a)$, where ε is the nodal error that the checker already encloses (D20, T3). Integration by parts for Lipschitz test functions and a pointwise AM–GM estimate turn this into the dual statement $|\int f\varphi - \int V_h'\varphi' + \ldots| \leq \eta \, \|\varphi\|_E$ for every Lipschitz φ vanishing at the Dirichlet ends (D21, D22, T3). For polynomial loads $\delta = 0$ and η^2 is exact up to the nodal error: on $-u'' = e^x$ with 16 elements the certified η exceeds the true error by $6 \cdot 10^{-19}$, and scikit-fem's nodal values written into the same certificate pass with the same η. Neumann conditions at both ends are not covered.

Eigenvalues of the Continuous Problem. For the Sturm–Liouville problem $-(pu')' + qu = \lambda w u$ with Dirichlet ends, λ_k is defined by min–max over Lipschitz functions vanishing at the ends, and `fis-fem-sl` certificates bound it from both sides. *Upper bound.* If on each element $p \leq P^+$, $q \leq Q^+$ and $w \geq W^- > 0$, and the LDL^T pivots of the tridiagonal $K^+ - \sigma M^-$ are non-zero with at least k of them negative, then completing the square and back-substitution give k P1 functions on which every non-zero combination has Rayleigh quotient at most σ, so $\lambda_k \leq \sigma$ (D23, D24, D26, T3): a conforming finite-element eigenvalue is an upper bound. *Lower bound*, of Liu–Oishi type. On an element, $\int_a^b (v - Iv)^2 \leq ((b-a)/\pi)^2 \int_a^b (v' - s)^2$ for the linear interpolant Iv (D25); nodal interpolation is the Ritz projection of the element-wise $\sum P^- \int v'^2$; and if $K^- - \sigma M^+$ has fewer than k negative pivots, any k admissible functions have a non-zero combination with $(\sigma/(1 + C^2\sigma) + \theta) \int wv^2 \leq \int pv'^2 + qv^2$, where $C^2 \geq \max W^+ h^2/(\pi^2 P^-)$ and $\theta = \min Q^-/W^+$, so $\lambda_k \geq \sigma/(1 + C^2\sigma) + \theta$ (D27, T3). The checker encloses the coefficients with the Taylor models of B02, counts pivots with the interval LDL^T of B05, takes the shifts from an untrusted search, and prints `SL_EIG_OK j lo hi` (B21, T4). For $-u'' = \lambda u$ on $(0, 1)$ with 512 elements, $\lambda_1 \in [9.869598, 9.869635]$, a relative width of $3.8 \cdot 10^{-6}$, which contains π^2. On uniform meshes Lean also proves $\lambda_1 \leq \lambda_{1,h}$ for every number of elements $N \geq 2$, and that the reverse inequality fails: at $N = 2$, $\lambda_{1,h} = 12$ while $\lambda_1 \leq 54/5$ (C12, T3). With variable coefficients the coefficients enter as element

constants, so the relative width is $O(h)$ rather than $O(h^2)$; Neumann and Robin ends and eigenvector bounds are not covered.

Two Dimensions: Guaranteed Error Bounds. A two-dimensional driver solves $-\Delta u = f$ with $u = 0$ on the boundary of a polygon, with P1 triangles, and writes an equilibrated Raviart–Thomas (RT0) flux $\sigma_k(x) = d_k x + c_k$ on every triangle. The conclusion `ETA_OK` η is a dual-norm bound: for every $\varphi \in C_c^1(\Omega)$, $|\int f\varphi - \int \nabla u_h \cdot \nabla\varphi| \le \eta \, \|\nabla\varphi\|_{L^2}$, where u_h is the P1 function of the certificate's nodal values (FEM-B12, T4). It is equivalent to $\|\nabla(u - u_h)\|_{L^2} \le \eta$ once the weak solution exists and C_c^1 is dense in H_0^1, neither of which is in Lean; the statement and its proof need no Sobolev space. The bound is that of Prager and Synge (Prager and Synge, 1947), $\eta = (\sum_k E_k)^{1/2} + C_F(\sum_k R_k)^{1/2}$ with $E_k \ge \int_{T_k} |\sigma_k - \nabla u_h|^2$ and $R_k \ge \int_{T_k} (f + \operatorname{div}\sigma_k)^2$, proved by two Cauchy–Schwarz steps without an L^2 space (D12). Lean proves the change of variables to the reference triangle and Fubini on it (D10), integration by parts on a triangle (D11; Mathlib's divergence theorem covers only boxes), the cancellation of the edge terms over the mesh by an involution on twin half-edges whose normal components agree (D13), and a Friedrichs inequality on the bounding rectangle, $(6/a^2 + 6/b^2)\int \varphi^2 \le \int |\nabla\varphi|^2$, which gives C_F (D15); all at T3. The checker computes E_k exactly by the edge-midpoint rule and computes the residual R_k exactly instead of requiring exact equilibration, which floating-point fluxes never satisfy. The mesh is a premise the checker verifies: a sorted sweep with a separating line for every pair of triangles shows that they overlap only in null sets, tile the domain and lie in its closure, P1 functions that vanish at the boundary nodes vanish outside Ω, and interpolants agree on shared edges (D16–D18, T3); clockwise, degenerate, overlapping or non-covering meshes and hanging nodes are rejected. The shape measure is only echoed (B13, T0), since the bound holds on every valid mesh; poorly shaped elements make η larger, not wrong. On the unit square with $f = 1$ and a 16×16 mesh the certified η is 0.0253 against a true error of 0.0210 (efficiency about 1.2 on uniform meshes), 8192 triangles are checked in about 4.5 s, and scikit-fem's nodal values on the same mesh agree with the driver's to $3 \cdot 10^{-16}$. Since 14 October 2026 any polynomial right-hand side with rational coefficients has a certified η as well: the residual integrals R_k are computed exactly through the Beta integral $\int_0^1 s^a(1-s)^n \, \mathrm{d}s = a!\,n!/(a+n+1)!$ on the reference triangle, with the same soundness statement (B12); for non-polynomial loads the checker prints `ETA_T0`. Two limits remain. The residual term uses the Friedrichs constant of the whole domain, not the local h_T/π of convex elements. And the hypothesis that the certificate's description of the domain matches Ω is proved in Lean only for the unit square, so for rectangles and L-shaped domains the conclusion holds for any Ω satisfying it.

Eigenvalues of Membranes. For a domain $\Omega \subset \mathbb{R}^2$, $\lambda_1(\Omega)$ is defined as the infimum of $\int |\nabla u|^2 / \int u^2$ over Lipschitz functions on $\mathbb{R}^2$ that vanish outside Ω, with partial derivatives from Rademacher's theorem (D30, D31, T3); this class lies in $H_0^1(\Omega)$ and is dense in it, so it is the usual first Dirichlet eigenvalue, but no Sobolev space appears. Every admissible function bounds λ_1 from above; the checker uses Q1 functions on a structured grid with rational coefficients from the certificate, whose two integrals equal nine-point stencil quadratic forms computed exactly (D33, B25, T4). For a rectangle the lower bound $\lambda_1 \ge \pi^2(1/a^2 + 1/b^2)$ is exact, by one-dimensional Wirtinger inequalities on every line and Fubini (D32). For the L-shaped domain, with its re-entrant corner, the lower bound uses the non-conforming rotated Q1 (Rannacher–Turek) element: its interpolant v has the same edge integrals as u, the energy splits as $\int_K |\nabla u|^2 = Q_K(v) + \int_K |\nabla(u - v)|^2$, and one-dimensional Wirtinger inequalities of Dirichlet and Neumann type give $\int_K (u - v)^2 \le (h_{\max}/\pi)^2 \int_K |\nabla(u - v)|^2$ (D34–D36, T3). If $K - \sigma M \succeq 0$ for the rotated-Q1 matrices, which the checker verifies by exact rational diagonal dominance of $K - \sigma M - LL^{\mathsf{T}}$ for an untrusted floating-point Cholesky factor L, then $\lambda_1 \ge \sigma/(1 + \sigma h_{\max}^2/\pi^2)$ (B26, T4). On 128×128 grids, λ_1 of the unit square lies in $[19.739208, 19.740200]$, which contains $2\pi^2$, and that of the L-shaped domain $(0,2)^2 \setminus [1,2] \times [0,1]$ in $[9.632674, 9.645136]$, a relative width of $1.3 \cdot 10^{-3}$, which contains the classical value 9.6397238 (a known answer, not a theorem). scikit-fem's discrete Q1 eigenvalue agrees with the exact Rayleigh quotient the checker computes to a relative

$2.1 \cdot 10^{-14}$; the certified statement, unlike scikit-fem's, is about the continuous eigenvalue. The upper-bound checker also refutes the belief that finite-element eigenvalues are lower bounds: P1 on the unit square divided by both diagonals into four triangles gives 24 (a hand computation, not formalised), while Lean proves $\lambda_1((0,1)^2) < 24$ by running the checker in the kernel on a 4×4 Q1 trial function (C15, T3). Only the first eigenvalue is certified, only for the L-shaped family with the re-entrant corner on grid lines, and the width is limited by the corner singularity.

M Molecular Dynamics

Problem. `mini-MD` integrates N atoms of a one-dimensional Lennard-Jones chain (free boundaries, all pairs, no cutoff) with velocity Verlet (Swope et al., 1982). A three-dimensional periodic version, the pendulum and the FPUT chain follow at the end of this section.

Structure (T3). We prove energy and momentum conservation for the Newton equations (MD-C01, C02). For the discrete map Φ_h, we prove time reversibility for arbitrary forces, momentum conservation, and symplecticity (D01–D03). Symplecticity follows by factoring Φ_h as kick $\circ$ drift $\circ$ kick and showing that each derivative preserves $\omega(u,w) = \sum_i (u_{q,i} w_{p,i} - u_{p,i} w_{q,i})$; it relies on the symmetry of the Lennard-Jones force Jacobian. We also prove that the Newton equations have a solution on all of $[a,\infty)$ from any initial configuration with the atoms in order, unique on every bounded interval (MD-C03). Since $\phi \geq -\varepsilon$, energy conservation bounds the kinetic energy and every pair potential, which yields a minimal distance $r_0 > 0$ and a momentum bound P that depend only on the initial data, not on the length of the interval (C04). Mathlib's Picard–Lindelöf theorem gives a solution of the vector field truncated at r_0 and P, which is bounded and globally Lipschitz, and a first-exit argument with the same energy bounds shows that this solution, and any solution of the true equations, stays strictly inside the truncation region; the first is therefore a true solution, and uniqueness follows from the Lipschitz bound.

Certificate. The central tool is a tube theorem (MD-A01). Suppose the vector field is K-Lipschitz on a set S containing the ρ-neighbourhood of an approximate curve z with defect $\|z' - v(z)\| \leq \varepsilon$, the initial error is at most δ, and the Grönwall bound $g(t) = \delta e^{Kt} + (\varepsilon/K)(e^{Kt} - 1)$ stays below ρ. Then *every* solution stays within g of z. The proof needs no prior knowledge of where the solution is: the set of escape times is closed, and its infimum leads to a contradiction via Mathlib's `dist_le_of_approx_trajectories_ODE_of_mem`. For each step the checker interpolates the data with a quartic polynomial, bounds its defect through a chain of derivatives of the pair potential (A03), and obtains a local bound for the solution starting from the data point (B01). Chaining these bounds gives a global bound from the initial value (B02). Energy drift, momentum drift and the residual of the Verlet equations are computed exactly (B03, B04). A passed check also shows that the data points are in order, so by C03 the true trajectory from the initial value exists and is unique, and the certified bounds are about it rather than vacuous statements about any solution there might be (B05, T4).

Honest Limits. Local bounds are nearly sharp: the ratio of the true local error, computed by a high-order reference integration, to the bound is at most 0.99. Global bounds grow like e^{Kt} and cover one or two vibration periods (e.g. 500 steps at $N = 8$, $T_0 = 0.01$, $h = 10^{-3}$). Many-body Lennard-Jones dynamics is chaotic, so long-time pointwise bounds do not exist. Long-time statements must take another form, such as shadowing, modified-Hamiltonian backward error analysis (Hairer et al., 2006) or statistics. Existence holds for all time, so the limit is the exponential growth of the bound, not the solution.

Exact Verlet. A second mode of the checker (`certify-md --exact`) bounds, step by step, the distance between the driver's floating-point trajectory and the velocity-Verlet iterates $\Phi_h^j(Q_0, P_0)$ computed in exact real arithmetic, by a discrete Grönwall step with row-wise Lipschitz constants of the force on boxes around the data (MD-A13, T3; B26, T4). It separates

the rounding error of a run from the discretisation error of the method; when the lower bound on some pair distance in a box is not positive, it stops and reports the horizon reached.

Three Dimensions with Periodic Boundaries. A three-dimensional driver and checker (`certify-md3d`) treat N atoms in a cubic box of side L with the minimum-image convention. The pair potential is written as a function of $u = r^2$: Lennard-Jones up to r_s, multiplied by a switch $S(x) = 1 - 126x^5 + 420x^6 - 540x^7 + 315x^8 - 70x^9$ of $x = (u - r_s^2)/(r_c^2 - r_s^2)$ up to r_c, and zero beyond, with $2r_c < L$; it is C^4 at both junctions and bounded below by $-\varepsilon$ (C09). Forces are then rational in the coordinates, so the checker needs no square roots, and C^4 lets the defect bound of the one-dimensional certificate, which uses the fourth derivative, carry over without a theory of piecewise-smooth ODEs. Lean proves energy and momentum conservation and, from any initial state with positive minimum-image distances, existence of a solution on $[a, \infty)$, unique on every bounded interval, with pair distances bounded below in terms of the initial data (C10–C13, T3). The minimum image is discontinuous where a displacement component equals $L/2$; the truncated pair force used for Picard–Lindelöf is shown to be globally Lipschitz by distinguishing nearby and distant points, and energy conservation across the discontinuity uses that a pair there is beyond the cutoff. In each step's box the checker classifies every pair as near (every displacement component strictly inside $(-L/2, L/2)$, so the image is fixed) or far (some component at least r_c from $L\mathbb{Z}$, so the force vanishes), rejects the certificate if a pair within the cutoff could cross $L/2$ inside the tube, and applies the tube theorem with a box Lipschitz constant and a defect bound for quartic interpolants (A06–A08, T3). Its soundness theorems state the local and global bounds for every solution, the existence and uniqueness of the true trajectory, exact energy and momentum drifts, the Verlet residuals and the validity of the parameters (B15–B21, T4); the problem is echoed exactly (B23, T0). For 32 atoms in an fcc 2^3 cell with displacements of 0.02 and $T_0 = 0.01$, the global tube covers 74 of 100 steps, up to $t = 0.148$, about half the shortest harmonic vibration period (0.29); local bounds are $1.5 \cdot 10^{-7}$, position errors are below 10^{-4} at $t = 0.1$, and checking takes 46 s on 16 cores. With our potential wrapped as an ASE calculator, ASE's velocity Verlet agrees with the driver step by step to $4.4 \cdot 10^{-16}$ in the positions, and where no pair lies in the switching region our forces agree with ASE's unswitched Lennard-Jones calculator to $2.8 \cdot 10^{-14}$; a high-order reference integration stays inside every interval. The limit is again e^{Kt}: the maximum-norm Lipschitz constant $K \approx 47$ is about twice the highest vibration frequency, and pair classification costs $O(N^2)$ per step without neighbour lists.

The Pendulum. For $\theta'' = -k \sin\theta$ from rest at $\theta_0 \in (0, \pi)$, Lean proves that a global solution exists and is unique (C17) and that its minimal positive period is $4t^*$, where t^* is the first zero (C18, T3): time reversal makes θ even, reflection about t^* with $\theta \mapsto -\theta$ gives $\theta(t) = -\theta(2t^* - t)$, and energy conservation with the intermediate value theorem gives $\theta \le \theta_0$, which excludes a shorter period. The checker `certify-mdlad` runs a tube in scaled coordinates $(\theta, \omega/c)$, with global Lipschitz constant $\max(c, k/c)$ and defect bounds from rational enclosures of sin and cos (A14); it requires $\theta > 0$ throughout the tube until the last step and a sign change inside it, and then encloses the first zero and the period (B28, T4). For $k = 1$, $\theta_0 = 1$ and $h = 10^{-3}$ the period lies in $[6.699975658, 6.699975670]$, which contains $4K(\sin\frac{1}{2}) = 6.6999756644$, with K the complete elliptic integral of the first kind (a floating-point comparison only); checking takes about 4 s. Only motion from rest with $\theta_0 < 3$ is accepted (the checker uses $3 < \pi$).

The FPUT Chain. The FPUT-α chain with fixed ends and 31 moving unit masses has $H = \sum_i p_i^2/2 + \sum_b (r_b^2/2 + (\alpha/3) r_b^3)$ with bond stretches r_b; we take $\alpha = 1/4$ and start with all energy in the first normal mode (rational roundings of $\sin(i\pi/32)$). The motion is close to integrable, but a Lipschitz tube would still fail by $t \approx 20$. In the inner product $\langle \cdot, \cdot \rangle_G$ given by twice the harmonic energy the harmonic part of the vector field is skew, and $\langle x - z, v(x) - v(z) \rangle_G \le |\alpha| S \, \|x - z\|_G^2$ whenever every bond satisfies $|\Delta_b x + \Delta_b z| \le S$ (C19, T3), so the growth rate is $|\alpha|$ times the amplitude, about 0.05–0.1, instead of a Lipschitz

constant near 2. A one-sided Lipschitz tube theorem (A15) and its per-step form for Taylor polynomials of degree 16 (A16) give the tube, and since $\sqrt{E_k}$ is a seminorm with $E_k(e) \le C_k^2 \|e\|_G^2$ (C20), the checker prints rigorous envelopes of the normal-mode energies E_k on blocks of time (B30–B32, T4). The FPUT potential is unbounded below, so global existence is not obvious; a passing check yields a solution on the checked interval by coordinate truncation and Picard–Lindelöf, and only there (B33, T4). With $h = 1/8$ the energies of the first five modes are enclosed up to $t = 200$, the first block containing $E_1(0) = 32\sin^2(\pi/64)$. A longer run reached $t = 1272.875$ before the tube grew to its radius $1/100$, about 12% of the time $t \approx 10530$ at which the energy nearly returns to the first mode; it encloses the energies of the first three modes on all of $[0, 1272.875]$, with E_1 falling from 0.0770 to $[0.04854, 0.04899]$ near $t = 1120$. That check reached the one-hour limit while printing the fourth mode; each mode's lines are printed only after its checks pass, so the printed envelopes are valid, but the run did not complete. Floating-point solutions (DOP853) lie inside every envelope. The tube grows by about 0.032 e-folds per unit time, far more than the true near-integrable dynamics, and following the chain to the recurrence would need an initial defect near 10^{-146}.

Verlet Does Not Conserve Energy. Lean also settles a common misstatement: for the harmonic oscillator with $m = k = 1$ from $(1, 0)$, one Verlet step changes the energy by $-h^4(4 - h^2)/32$, which is non-zero for every $h > 0$, $h \neq 2$; with $h = 1$ the energies are $1/2$, $13/32$, $13/32$ (D04, T3). What Verlet conserves exactly is the modified energy $p^2/(2m) + (k/2)(1 - h^2k/(4m))q^2$ (D05, T3), the discrete counterpart of backward error analysis (Hairer et al., 2006).

N Computational Thermodynamics

Problem. `mini-CALPHAD` reads a TDB database and computes the equilibrium of a binary system A–B at fixed temperature, pressure and overall composition x_0 (Lukas et al., 2007), or at a given chemical potential of B instead of x_0. Phases are substitutional solutions (ideal mixing plus Redlich–Kister excess terms), stoichiometric compounds, and sublattice phases of the compound energy formalism (Hillert, 2001) with one or two sublattices on which A and B mix, the other sublattices being filled by A, B or vacancies. Solutions may carry the Inden–Hillert–Jarl magnetic term (Inden, 1981; Hillert and Jarl, 1978) $RT\log(1 + \beta(x))\, g(T_c(x)/T)$, where the Curie temperature T_c and the magnetic moment β are Redlich–Kister sums divided by an antiferromagnetic factor where negative, and g is a piecewise function with polynomial and inverse-polynomial branches. With one mixing sublattice $(\mathrm{A}, \mathrm{B})_a$ and n_A, n_B sites filled by A and B, the composition $x = (ay + n_\mathrm{B})/m$, $m = a + n_\mathrm{A} + n_\mathrm{B}$, is an affine function of the site fraction y of B, and mG has the form of a solution in y with mixing weight a; an example is $\theta = (\mathrm{Al})_2(\mathrm{Al}, \mathrm{Cu})_1$. With two mixing sublattices, as in the Laves phase $\mathrm{C15} = (\mathrm{Cu}, \mathrm{Mg})_2(\mathrm{Cu}, \mathrm{Mg})_1$, the composition $x = (ay_1 + by_2 + n_\mathrm{B})/m$ only constrains a line in the square of internal states, and, as in pycalphad, the Gibbs energy of the phase is the minimum over that line,

$$G(x) = \min\left\{H(y_1, y_2)/m \;:\; y_1, y_2 \in [0,1],\; ay_1 + by_2 = mx - n_\mathrm{B}\right\}, \tag{7}$$

where H collects the four endmember energies, the ideal mixing $RT\,(aS(y_1) + bS(y_2))$ and Redlich–Kister interactions on either sublattice. The equilibrium Gibbs energy $G^*(x_0)$ is the infimum of the total Gibbs energy over feasible configurations, i.e. finite families of (phase, composition, amount) with the correct overall composition. An equilibrium is a feasible configuration together with a line that lies below all phases and touches the configuration; its end values are the chemical potentials. Interstitial solutions and magnetic sublattice phases with one mixing sublattice were added later (see below). Phases outside these classes (three or more mixing sublattices, reciprocal parameters, endmembers missing from the database, magnetic phases with two mixing sublattices, order–disorder models) are skipped and listed in the report; every conclusion holds for the phases taken into account, the set the checker echoes as `PROBLEM PHASES`. Statements that a state is not the global equilibrium remain true when such phases are added, since extra phases can only lower the convex hull; positive statements need not, and this short argument is on paper, not in Lean. Since 26 October 2026 the driver and `fis` refuse by default (exit code 4) a run in which

skipped phases would silently restrict the phase set, and accept it only with `--allow-skip` or an explicit `--phases`: on pycalphad's `alni_dupin_2001.tdb` (Al–Ni, 1000 K, $x_{\mathrm{Ni}} = 0.5$) the restricted set gives LIQUID + AL3NI5, true for those phases, while the full database gives BCC_B2. The gas constant is a convention, $R = 8.3145\,\mathrm{J\,mol^{-1}\,K^{-1}}$ as in pycalphad (8.31451 for the comparisons with OpenCalphad), written in the certificate and echoed as `PROBLEM R` (A-CAL-02).

Theory (T3). We prove a quadratic lower bound for the mixing entropy $S(x) = x\log x + (1-x)\log(1-x)$ around any point, by writing S minus its tangent as a sum of two Kullback–Leibler terms (CAL-C01). We characterise equilibria as global minimisers (C02). We also prove that an equilibrium exists (C03): the set of supporting slopes touching on either side of x_0 is closed, the two sets cover $\mathbb{R}$, and connectedness of $\mathbb{R}$ yields a line touching on both sides. The argument needs continuous Gibbs energies. For magnetic phases this rests on an identity, linear in $1/p$ and proved by `ring`, which makes the two branches of g agree at $T = T_c$ for every structure factor p. For phases with two mixing sublattices we parametrise the constraint segment by $t \in [0,1]$, so that (7) is an infimum over a compact set; it is attained, every internal state on the segment bounds G from above, and G is continuous by Mathlib's `IsCompact.continuous_sInf` (C07). With one mixing sublattice the change of variables $y \leftrightarrow x$ carries the solution bounds over exactly (C06). The magnetic term has no convenient second derivative, so it is bounded on intervals by *secant slopes* instead (C04): $\log(1+v)$, the antiferromagnetic convention, t^n, t^{-n} and both branches of g each have rational bounds on the slope between any two points of an interval. Composed, they bound the magnetic energy on an interval below by a polynomial in x minus an error of the order of the interval width squared, which the branch and bound absorbs.

Certificate. Here the checker certifies a *global* property, not the local solution found by the driver's grid-plus-Newton heuristic. It proves a lower bound on the total Gibbs energy of every feasible configuration by branch and bound over composition intervals. On each interval the polynomial parts are shifted exactly, the mixing term is bounded with C01, and quadratics are minimised exactly. The certificate's phase assemblage, combined by the lever rule, gives a matching upper bound. For every equilibrium the checker then outputs: the interval for $G^*(x_0)$ (`G_OK`); intervals for both chemical potentials (`MU_OK`); composition windows that contain every phase present (`WINDOW`); phases absent from every equilibrium (`ABSENT`); and each phase's driving force $\max_x(\ell^*(x) - G(x))$ for the equilibrium tangent ℓ^* (`DF_OK`) (CAL-B01, T4). The driving force is the global maximum over compositions; pycalphad's solver, which starts from the phase alone at x_0, stops at a local maximum in 6 of 141 benchmark cases. In single-phase regions the checker replaces the driver's double-precision tangent by its own rational tangent at x_0; soundness holds for any tangent, so this needs no new proof and narrows the chemical potentials from about 10^{-4} to 10^{-8}–10^{-7} J/mol. When x_0 is exactly the composition of a compound on the lower hull, the tangent is not unique, and the chemical-potential intervals have the width of the genuine non-uniqueness; terminal solubilities too small for a double, such as $1 - x \approx 3 \cdot 10^{-37}$, are written as exact decimals. Temperature-dependent parameters $c\,T^n(\log T)^l$ are enclosed with the rational logarithm bounds of Appendix I.1.

Sublattice Phases. With one mixing sublattice the branch and bound runs over $y \in [0,1]$: the tangent becomes an affine function of y, the solution bound of the previous paragraph applies to mG, and windows are mapped back to compositions. With two mixing sublattices the checker never computes the internal minimisation in (7). Since x and the tangent ℓ are affine in the internal state, $G(x) - \ell(x)$ is the minimum of $F(y) = H(y)/m - \ell(x(y))$ over the constraint segment, so any lower bound for F on the whole square $[0,1]^2$ bounds $G - \ell$ for every composition. On a box with centre c the two one-variable parts of H are bounded with C01, and the cross terms are expanded to second order at c with the remainder absorbed into a constant E, giving $mF(y) \geq Q(y-c) - E$ for a bivariate quadratic Q, whose minimum over the box is bounded below in closed form (C07). A two-dimensional branch and bound with this bound replaces the interval search; it splits a box only while its lower bound lies below a running estimate of the minimum found so far, since with a fixed target

every box along the curve where F crosses the target would be split to maximum depth. The upper bound is evaluated at the internal states the certificate supplies. Windows are harder, because the boundary of $\{F \le \kappa\}$ is a curve and the number of boxes covering it grows exponentially with depth. Each box therefore contributes the intersection of its composition range with the range of x on the ellipse $Q \le m\kappa + E$, computed by Cauchy–Schwarz with a rational upper bound for the square root, and boxes whose ellipse is already accurate, or whose range lies inside the composition range of known points with $F \le \kappa$, are not split further. Each pass splits at most 5000 boxes. Soundness holds for any splitting rule and any budget, so a degenerate certificate can only produce wider intervals, within about 30 s, never a wrong one; typical certificates take 1–4 s (B01).

Binding to the Database. The translation from TDB text to the certified model (choosing the temperature range of each piecewise parameter, expanding function references, classifying phases, site ratios, the sign of parameters written in the order B,A, magnetic constants, species) was originally done only by the untrusted driver. It is now a Lean definition, `modelOfTdb`, the model a TDB text defines at temperature T for two components and a list of phases (CAL-I01). In `--tdb` mode, which the driver uses whenever it verifies a run from a TDB file, the checker reads the file as bytes, builds this model, requires it to equal the certificate's model and x_0 to match, and only then runs the checker above; `checkTdb_sound` states the conclusions of CAL-B01 for the model the TDB file defines (B02, T4). The checker echoes the problem (file, components, T, x_0, phases) and lists every applicable phase left out of the model together with the reason. The trust that remains is in the definition itself, about 930 lines of Lean, which is a specification and therefore at T0. It is cross-checked against the driver's translation on every binary system of COST 507 (Ansara et al., 1998) at two temperatures (595 certificates, 6701 phases: identical models, and the same phases skipped for the same reasons) and against pycalphad's `calculate` on sampled systems and on every sublattice phase of COST 507, to a relative 10^{-12} or better. A model coefficient in a certificate changed by 0.01% is accepted without `--tdb`, as a statement about the changed model, and rejected with it.

No Miscibility Gap. A point certificate can prove that a phase *has* a miscibility gap at (T, x_0), when its windows exclude x_0, but not that it has none. A second kind of certificate supplies the converse for a non-magnetic substitutional solution. At fixed T the parameters split exactly into $aG = P_0 + (\log T)\,P_1 + aRT\,S$, with a the site count of the phase and exact rational polynomials P_0, P_1, so $aG'' = P_0'' + (\log T)\,P_1'' + aRT\,(1/x + 1/(1-x))$ with $\log T$ entering only as a coefficient between rational bounds. The checker runs its own branch and bound for $aG'' > 0$ on $(0,1)$. Lean then proves that G is strictly convex on $[0,1]$, that in any equilibrium of any model containing the phase all of its parts have one composition, and that a model with this phase alone has the single-phase equilibrium at every x_0 (C05, T3; checker and TDB binding B03, T4). Together with point certificates this brackets the temperature at which a gap opens from both sides; a certificate takes 0.1–0.2 s to check. It is a statement about one temperature: nothing is certified between two temperatures, and magnetic phases would need bounds on the second derivative of the magnetic term, which C04 does not provide.

Prescribed Chemical Potential. pycalphad also accepts the chemical potential μ_B as a condition in place of the overall composition. The equilibrium tangent then passes through $(1, \mu_B)$, and what is minimised is $G - \mu_B N_B$ per mole of A, with minimum $\mu_A^* = \inf_{i,\,x<1}(G_i(x) - \mu_B x)/(1-x)$. A contact point supplied by the certificate gives $\mu_A^* \le a_{hi}$. For the line through $(0, a_{hi} - gap)$ and $(1, \mu_B)$ the tangent is fixed, and for a fixed tangent the first pass of the point checker, which proves $G_i - \ell \ge 0$ for every phase, is already the certificate of the lower bound. Any equilibrium tangent differs from this line by at most $\kappa = gap$, so the second pass yields composition windows for every phase present in any equilibrium, and from them the checker bounds the overall composition and the total Gibbs energy. Two windows mean two-phase coexistence, where μ_B does not determine the overall composition and pycalphad reports one of the two phases. Existence follows from the intermediate value theorem for the continuous map $\mu \mapsto \inf_K(g - \ell_\mu)$ on the compact union K of the phase curves (C08; checker and TDB binding B04, T4). On 16 benchmark points,

with solutions, a compound, a magnetic phase and sublattice and Laves phases, pycalphad's μ_A, phase compositions, overall composition and G_M lie inside the Lean intervals up to pycalphad's convergence tolerance, and the μ_A intervals are about $2 \cdot 10^{-9}$ J/mol wide. Activities are the more common condition, but $\mu_B = G_{ref}(1) + RT \ln a_B$, with pure B in a reference phase at the same temperature as reference state, is in general irrational. The checker therefore computes $G_{ref}(1)$ from the model and encloses μ_B in a rational interval, and one certificate covers the whole interval, using Gibbs–Duhem monotonicity: raising μ_B does not raise μ_A^* and does not lower the overall composition (C10). Its conclusions therefore hold for the true, irrational μ_B (B06, T4). On 17 points, with the activity converted from pycalphad's `MU`, the reference μ, μ_A, phase compositions, overall composition and G_M of pycalphad lie inside the Lean intervals, which are about 10^{-9} J/mol wide.

Thermodynamic Properties. pycalphad's `calculate` also returns the molar entropy $S = -\partial G/\partial T$, enthalpy $H = G - T\,\partial G/\partial T$ and heat capacity $C_p = -T\,\partial^2 G/\partial T^2$ of a phase at fixed composition. For a phase with two mixing sublattices the internal state is held fixed too, so these are derivatives of $H(y_1, y_2)/m$, not of the minimum (7). At a fixed rational composition every phase has the form $G(T) = (P(T) + KT + \Lambda T\, h(t_c/T))/m$ on an open set around T, where P is a sum of terms $c\,T^n(\log T)^l$ with exact rational coefficients, K the coefficient of the ideal mixing, $\Lambda = R\log(1+\beta)$, and h the branch of the magnetic function g on the side of T_c where T lies. The derivative of such a term is again a sum of such terms, so S, H and C_p have closed forms: the ideal mixing cancels from H, and the magnetic term enters through h, h' and h'' at the rational point t_c/T (CAL-C11, T3). These are the derivatives of the model the TDB file defines only if its piecewise parameters do not change near T. If no breakpoint of a parameter of the listed phases, or of a function such a parameter uses, equals T, Lean proves that the TDB file defines the same parameters at every rational temperature between the nearest breakpoints on either side; the real function of T is the unique continuous extension of these values, so on that interval it *is* the TDB model, and so are its derivatives (I02, T3). The checker (`--calc n`) evaluates the closed forms with the rational bounds for $\log T$ and prints, for each point, rational intervals about 10^{-25} wide for G, S, H and C_p, and in `--tdb` mode the temperature range on which the model is unchanged (B07, T4). A temperature at a breakpoint, where pycalphad takes the upper piece, or at the Curie temperature, where C_p has its λ peak, is rejected. On 713 points in 21 systems, with solutions, compounds, magnetic phases on both sides of T_c, and sublattice and Laves phases at fixed internal states, pycalphad's `GM`, `SM`, `HM` and `CPM` lie inside the Lean intervals up to a relative distance below $3 \cdot 10^{-14}$, the rounding of a double; for magnetic C_p the distance is $3.3 \cdot 10^{-12}$, because pycalphad evaluates $\tau = T/(T_c + 10^{-9})$. Writing and checking a certificate takes 0.14–0.74 s. Equilibrium-level properties, pycalphad's `equilibrium` with `HM`, `SM` or `CPM`, which weights the phases present by their amounts, are not yet certified.

Phase Diagrams. On a temperature grid the driver splits each isothermal section along the lower convex hull into single-phase regions and common tangents and writes one point certificate per region. The windows of a common-tangent certificate are certified phase boundaries at that temperature. Figure 7 shows them for magnetic Cr–Fe, where the ferromagnetism of iron opens a bcc miscibility gap, and for Al–Fe; the benchmark also maps Al–Cu with the sublattice phases θ, η, ε and γ, and Cu–Mg with the Laves phases C14, C15 and C36.

Point certificates alone cannot show that the list of regions is complete: a narrow region between two certified compositions could be missed. A partition certificate per temperature closes this gap for every $x_0 \in [0,1]$ (CAL-C09, B05, T4). It rests on one observation: two equilibria at $x_1 < x_2$ have tangents ℓ_1, ℓ_2 that both lie below all phases, so $\ell_2(x_1) \le \ell_1(x_1)$ and $\ell_1(x_2) \le \ell_2(x_2)$, and adding the two shows that the slope $\mu_B - \mu_A$ of the equilibrium tangent is non-decreasing in x_0. Point certificates at anchor compositions $a_{k-1} < a_k$ therefore bracket the slope of every equilibrium tangent at any x_0 in between. Such a tangent lies below the phases' rational upper bounds at a few points, so it lies below the chords between them and, beyond the outermost points, below rays with the bracketing slopes. Every phase present in the equilibrium touches the tangent and so satisfies $G_j \le L$ for one of these lines L; the second pass of the point checker with $\kappa = 0$ turns each line into composition windows,

and no new lower bound is needed. From the windows the checker derives single-phase regions, two-phase regions (in which both sides are present, since a feasible configuration has a part on each side of x_0, and the phase boundaries lie in the windows), and narrow intervals at phase boundaries where only the set of phases is known, and it checks that the regions cover $[0, 1]$. Where the anchors and points are placed is an untrusted heuristic, so a poor choice can only make the check fail or the regions coarser. All partition certificates of the benchmark, seven systems at 67 temperatures, pass, and the unresolved intervals at phase boundaries are at most $5.4 \cdot 10^{-8}$ wide. A partition certificate says nothing between grid rows (the next paragraph treats fixed compositions there), nor about a miscibility gap inside the window of a single-phase region, which a convexity certificate can exclude for non-magnetic solutions.

Temperature Intervals. Interval certificates speak about every real temperature in $[T_1, T_2]$. If $|\partial^2 G/\partial T^2| \le C$ on the interval, $G(T, x)$ differs from the linear interpolation of its values at T_1 and T_2 by at most $C(T - T_1)(T_2 - T)/2$; the checker obtains C from the exact symbolic second derivatives of the parameters, the ideal mixing being linear in T (CAL-C12), and equilibria at real temperatures, with their existence, are defined and proved as at rational ones (C13, T3). The first pass, the second pass and the upper bound of the point checker run at both endpoints, and the interpolation error covers the interior. A `tbox` certificate concludes, at a fixed x_0 and for every T in the interval, that every equilibrium has chemical potentials within κ of the interpolated tangent and its phases in the printed windows, and that an equilibrium exists when the model contains a substitutional solution (B13, T4); a `tabsent` certificate concludes that a non-magnetic solution phase lies strictly above every line below all phases, so it is absent from every equilibrium at every overall composition and every T in the interval (B14, T4). In `--tdb` mode the checker also proves that the TDB file defines the certificate's model at T_1 and the same parameters at every rational temperature of an open interval containing $[T_1, T_2]$ (B14, I02). In single-phase regions the uncertainty of the tangent grows with the interval width, so the driver uses narrow intervals there.

Invariant Reactions. A eutectic certificate chains `tabsent` certificates below the reaction, in which the liquid is absent at every composition, a `tbox` across it, in which only α (left of x_0), β (right of x_0) and the liquid can appear at x_0, and `tbox` certificates above it, in which x_0 is liquid; the temperature at which the liquid first appears therefore lies in $[T_{\mathrm{lo}}, T_{\mathrm{hi}}]$, and the middle windows enclose the three phase compositions (B15, T4). The checker also proves that $T_a < T_{\mathrm{lo}} < T_{\mathrm{hi}} < T_b$ and that equilibria exist, so none of the conclusions is vacuous (C33, T3). For Pb–Sn (`pbsn.tdb`) it certifies the eutectic temperature in $[454.5521, 454.5721]$ K with windows $[0.25270, 0.27391]$ for fcc, $[0.72725, 0.74725]$ for the liquid and $[0.97292, 0.97802]$ for bct, from 34 segments in about 70 s for driver and checker together; pycalphad's temperature-by-temperature equilibria at x_0 place the eutectic at 454.5601–454.5621 K, inside. Not proved is that three-phase coexistence actually occurs at some temperature of $[T_{\mathrm{lo}}, T_{\mathrm{hi}}]$, which needs a connectedness argument; peritectics have no certificate, and neither do phases with two mixing sublattices. For the regular solution $G = \Omega x(1 - x) + RT\,S(x)$ with $\Omega > 0$, Lean proves that G is non-convex on $[0, 1]$, so that a miscibility gap exists, exactly when $T < \Omega/2R$, that it is strictly convex otherwise, and that it is convex on $[0, b]$ inside the spinodal, $2\Omega b(1 - b) \le RT$ (C14, T3).

Interstitial and Magnetic Phases. A phase with one mixing sublattice on which two species mix, one of them possibly vacancies, has $n_{\mathrm{A}0} + n_{\mathrm{A}1}y$ atoms of A and $n_{\mathrm{B}0} + n_{\mathrm{B}1}y$ of B per formula unit, so the composition $x(y)$ is a ratio of affine functions of the site fraction y. This covers interstitial solutions such as $(\mathrm{Fe})_1(\mathrm{C}, \mathrm{Va})_3$ and sublattice phases with an Inden–Hillert–Jarl magnetic term. When $x(y)$ is increasing, Lean proves that it is a homeomorphism onto the admissible compositions, that $G - \ell = a(G_{\mathrm{sol}} - \ell')/m(y)$ for an affine ℓ' in y, so the solution bounds of C01 and C04 carry over, and that upper bounds and windows map back (C22); stoichiometric compounds may now carry a magnetic term (C23). The soundness theorems of all point-temperature checkers hold for these phases with unchanged statements, while the activity reference, property and convexity checkers report them as unsupported and the driver has been tested with them only for point equilibria (B23, T4); the TDB reader recognises them (I09, T0), and on six equilibria in Fe–C, Al–Fe,

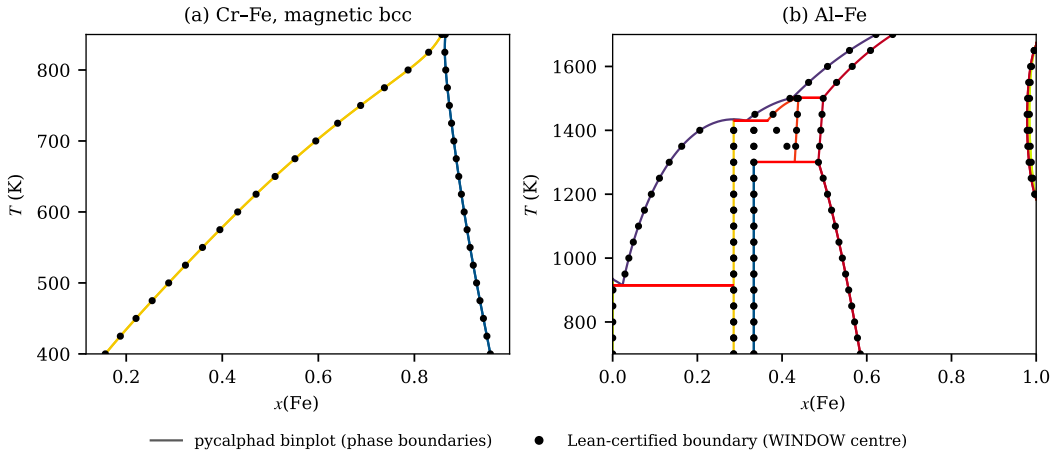


Figure 7: Certified phase boundaries. Each dot is the centre of a composition window, at most $2 \cdot 10^{-6}$ wide, from one common-tangent certificate; lines are pycalphad's `binplot` on the same database and phases. (a) Cr–Fe, 400–850 K: the miscibility gap of the magnetic bcc solution. (b) Al–Fe, 700–1700 K, with magnetic bcc, fcc and hcp. Between 1300 and 1430 K `binplot` omits the Al_2Fe region that pycalphad's own `equilibrium` and our certificates both contain.

Fe–Ni and Cr–Fe, pycalphad 0.11.2's G_M, chemical potentials and phase compositions lie inside the Lean intervals and windows. The interpolation bound extends to these phases, for magnetic terms whose T_c and β do not vary with the site fraction (C24, T3). With it, a eutectoid certificate for the metastable Fe–Fe_3C system, fcc $\rightarrow$ bcc + cementite, places the eutectoid temperature in $[999.6738, 999.6938]$ K, with fcc absent at every composition within 20 K below, and pycalphad's 999.6838 K inside (B24, T4). The database is `cfe_broshe.tdb`, shipped with pycalphad, in its exact pressure-free limit: its equation-of-state and Einstein terms vanish at $P = 0$ and are removed, and `M7C3_D101`, whose Gibbs energy keeps a non-SGTE term at $P = 0$, is left out (I10, T0); for seven phases at four temperatures, pycalphad's Gibbs energies from the reduced file differ from those of the original by at most $7 \cdot 10^{-8}$ J/mol at $P = 10^{-6}$ Pa and by 0.35–0.79 J/mol at atmospheric pressure, so the certificate is about the pressure-free model. Sublattices on which vacancies mix with both components, phases with more than one such sublattice and pressure-dependent TDB functions are not supported.

An Audit of the NIST Solder Database. As a first test of the checkers on a widely used reference, we compared the NIST solder database (`NIST-solder.tdb`) with the table of calculated invariant equilibria published with it, under a protocol registered before any certified calculation (CAL-B35, CAL-B36). All 26 entries received a verdict: 22 agree, 4 do not, each robustly (also under a widened tolerance). Ag–Sn L + ζ $\rightarrow$ Ag_3Sn lies at $[754.3137, 754.3337]$ K against the published 753.05 K; this entry was not computed before registration, and the 1999 file agrees with the table; the visible difference between the two files is the sublattice model of the ϵ phase. Ag–Pb L $\rightarrow$ (Ag) + (Pb) lies 1.20 K above the table; a floating-point spot check with pycalphad's default grid had put the difference at 0.5 K, which the certificate shows to be a numerical error of the grid. The two Cu–Pb reactions do not exist in the file: at the test compositions only FCC_A1 is stable, the liquid has a certified negative driving force even above the melting point of copper, and FCC_A1 has no Cu–Pb interaction parameter. Every verdict is phrased as "this file (dis)agrees with the table published with it"; it does not say which one is wrong and involves no experiment. The 2055 certificates are archived and re-checked by `verify_all`.

Published Binary Assessments. The same test applied to six published binary assessments that ship a database file and print a calculated invariant table (CAL-B37, CAL-B38) gives, for 56 registered reactions, 32 agreements (10 by point certificates), 3 disagreements and 21

reactions not compared by the registered rules; all 18 Al–Cu and 3 Mg–Si reactions agree. The disagreements are in Pt–Pb, by 4–17 K: the printed Redlich–Kister convention (Eq. 3) with the printed parameters does not reproduce the printed table, while the opposite sign of the first-order liquid parameter reproduces all four values within 0.08 K, which suggests, as a hypothesis, that the table was computed with the other sign. A further 49 reactions involve phase models the checker does not support (such as order–disorder BCC_B2) and are listed as not comparable.

A Systematic Audit. The frame of the next audit was fixed in advance: every binary, non-kinetic database published as *Calphad* supplementary material in the TDBDB index (161 files, 99 papers; CAL-B39). Against the invariant tables printed in their own papers, 25 of 47 comparable databases have at least one disagreeing reaction (0.53, exact 95% Clopper–Pearson interval $[0.38, 0.68]$), 21 of them robustly; of the 34 robust disagreements with a closed certified interval most are below 1 K beyond the printed precision and the largest registered one is 7.53 K, and 12 have none, characterised post hoc below (41,788 certificates, CAL-B40); a rule for broken certificate chains chosen after the verdicts were seen is reported only as a sensitivity analysis. Metric 2 asks whether pycalphad 0.11.2 at its documented default settings (one `equilibrium` call per database and temperature, nothing passed beyond conditions and phases) returns an equilibrium that a point certificate proves is not the global equilibrium of the model the file defines: its Gibbs energy exceeds the certified upper bound by more than $\max(0.01\,\mathrm{J/mol}, \delta_{\mathrm{num}})$, or its phase set contradicts a phase the certificate proves absent or present or a certified composition window. On a fixed grid ($x = 0.02, 0.06, \ldots, 0.98$, T every 50 K from 300 K) of 70,450 points, 67,797 have a certificate (the others exceed the 600 s limit per point in A14 Fe–C and A46 Si–Sr), pycalphad gives no result at 21 and returns such an equilibrium at 29 points in 7 of 63 databases (0.11, $[0.05, 0.22]$; CAL-B41, CAL-B42); 26 of the 29 are energy errors, 20 by more than 1 J/mol and 7 by more than 100 J/mol, and 3 count only because the registered rule filters out a phase present at a fraction below 10^{-6} (without them, 4 of 63). A pre-registered addendum with OpenCalphad 6.137 gives 16 of 54 (0.30, $[0.18, 0.44]$), with no answer at 14.9% of certified points; its single-phase reading check was revised after the results were seen and is labelled post hoc (CAL-B43). The two solvers rarely err at the same point: of 48,584 points both answered, 4 are non-global for both, 78 only for OpenCalphad and 4 only for pycalphad. The property checker now covers interstitial phases and magnetic compounds (CAL-B07, T4), which the reading check needed for six databases. A further addendum, frozen before any new certificate, asks whether these proven disagreements and non-global equilibria change two engineering decisions (CAL-B48–CAL-B51). Placing a process temperature 10 K beside a printed invariant temperature lands in another certified phase region in 7–9 of the 47 comparable databases, 2–4 if rows with a doubtful printed composition are excluded; the 10 K margin was chosen after the disagreements had been seen, so all five margins (2–50 K) are reported (678 new certificates, CAL-B49, T4). At the non-global solver points the user reads a different phase set or a phase fraction off by at least 0.05 in 2 of 63 databases for pycalphad and 10 of 54 for the OpenCalphad development version; in 17–18 of 66 databases at least one proven disagreement or non-global equilibrium changes one of the two decisions. A supplementary analysis, which leaves the registered results and denominators unchanged, reran a seeded random sample of the grid points left without a certificate in the three slowest databases (A14 Fe–C, A46 Si–Sr, A57 C–Nd) with the same driver, frozen checker and judge functions, only the per-point limit raised from 600 s wall clock to 7200 CPU s (CAL-B52, T1; CAL-B53, T4). None of the 45 points that received a certificate is a solver error (pycalphad agrees at all 35, OpenCalphad at 9 of 10 and gives no result at one), and no database verdict would change; in A14 every certified point lies at or below 1500 K, A57 was sampled at only five points, too few to extrapolate, and the full set of about 7000 points, an estimated 9000–11 000 CPU hours, was not run.

Near Invariant Reactions. The second point set of metric 2 was registered with the first but run only after the transcribed invariant tables and the metric-1 intervals were frozen (CAL-B54). For each of the 478 transcribed reactions of tier A it takes $T_{\mathrm{pub}} \pm 0.5$ and ± 2 K at the printed phase compositions, their midpoints and the metric-1 composition x_0, and for each certified interval $[T_{\mathrm{lo}}, T_{\mathrm{hi}}]$ the points $(T_{\mathrm{lo}} - 0.05\,\mathrm{K}, x_0)$ and $(T_{\mathrm{hi}} + 0.05\,\mathrm{K}, x_0)$; the

list, 6364 points in 53 databases (54 are listed; one has no point), follows mechanically from the transcriptions and was committed before any pycalphad call. Of these, 6110 have a certificate (CAL-B55, T4); 110 exceed the 600 s limit (A14 and A46) and 144 lie outside the temperature range on which the parameters are defined. pycalphad gives no result (GM is NaN) at 84 certified points, and nowhere does its energy fall below a certified lower bound. At 50 points in 16 of 52 databases (0.31, [0.19, 0.45]) it returns an equilibrium certified not to be global: 44 are energy errors (46 above δ_{num}, 20 above 1 J/mol, 4 above 100 J/mol), all 50 contradict the certified phase set (6 of them only the phase set), and all 50 change the phase assemblage a user would read; none depends on the 10^{-6} filter. In 4 of the 16 databases the largest energy error is below 1 J/mol (0.24–0.53 J/mol), and 14 of the 50 points are those just outside a certified transition interval. Typically, 0.5–2 K above an invariant temperature pycalphad still returns a low-temperature phase that the certificate proves absent, for example single-phase `AU4SC1` in Au–Sc at 1498.5 K and $x_{\mathrm{Sc}} = 0.2$, 6.22 J/mol above the certified bound, where the certificate proves liquid plus `AU3SC1`. The largest cases are single-phase HCP in Gd–Ti at 1150.5 and 1152 K and $x_{\mathrm{Ti}} = 0.499$ and single-phase DHCP in Nd–Ti at 1161 K, where the certificates prove two phases at least 1.02×10^4 and 1.23×10^4 J/mol lower. The rate per point ($8.3 \cdot 10^{-3}$) and the failure rate ($1.4 \cdot 10^{-2}$) are about 20 and 40 times those of the grid, but these points lie within 2 K of invariant temperatures by design, so the rates are not comparable. Combined as registered (a database counts if either point set has such a point), metric 2 is 21 of 63 (0.33, [0.22, 0.46]; CAL-B56, Fig. 5a): 14 databases count only near invariant reactions, 2 on both point sets and 5 only on the grid; without the three filter-only grid points it is 19 of 63. The combined count mixes two testing intensities: 10 of the 63 databases have no transcribed invariant table and one has no certified point near its reactions, so only the grid tests them; near invariant reactions alone the rate is the 16 of 52 above.

Invariant Reactions in a Temperature Scan. For each of the 331 reactions with a certified interval, pycalphad at its defaults scanned $[T_{\mathrm{pub}} - 20, T_{\mathrm{pub}} + 20]$ K at x_0 every 0.25 K, each change of phase assemblage bisected to 0.001 K, and the change was compared with the certified low- and high-temperature assemblages (CAL-B57). Of the 328 classifiable reactions, it reproduces 273 (0.83, [0.79, 0.87]) within the certified interval widened by 0.05 K, places 14 outside it (11 by less than 1 K, 3 by 1.0–2.4 K), misses 2, which it passes through an intermediate assemblage, and for 39 returns no result over a band of 0.002–5.6 K (median 0.14 K) between the certified low- and high-temperature assemblages; 3 are not classifiable because a point certificate leaves one side undetermined, and 2 of the 256 reactions with a certificate chain show an additional change inside a segment where the chain proves the assemblage constant. The class of the 39 is post hoc: it was defined after a first scan, run exactly as pre-specified, which stopped bisecting at the first NaN and gave 269 reproduced and 45 missed (archived with the results); after bisecting both edges of each failure band, read literally without the new class, 41 reactions are missed (0.125). A failure to return a result proves nothing about the model, so the 39 are counted neither as misplaced nor as missed.

Checks and Deviations. Three descriptive checks leave the verdicts unchanged (CAL-B58, T0). With a denser starting grid (`pdens` = 2000 instead of 60), 49 of the 50 near-invariant points and 27 of the 32 grid points (S1 and controls) are no longer non-global; the remaining near-invariant point, in Ga–Ti at 1347.5 K, is 0.455 J/mol above the bound, as at the default. pycalphad's own `calculate`, evaluated at the certified assemblage, lies inside the certified interval at every non-global point, so the difference arises in the search of `equilibrium`, not in a different reading of the file. Of 10 sampled non-global points near invariant reactions, the 8 energy errors remain energy errors with certificates rewritten for $R = 8.31451$ J/(mol K); the other 2 are phase-set errors only. Besides the post hoc failure class, the deviations from the protocol are readings of its wording (which certified intervals and compositions define the points, how a change is matched in the scan) and changes of environment (the cores, a bit-identical rebuild of the frozen checker); none changes a judging rule. Before this part was run, a proxy-user trial in another session had found a non-global pycalphad equilibrium near a Cr–Ti eutectoid in a database outside the frame (below); it prompted the timing of the run but read or computed no file of the frame.

Hash Seeds, and a Case outside the Frame. A supplementary registration, frozen before its runs, asks whether the verdicts depend on Python's string-hash seed, which the registered runs left unspecified (CAL-B64, T0). In a fresh process for each of PYTHONHASHSEED $= 0, \ldots, 7$, all 32 registered non-global points of the grid (29 in S1, 3 controls) are non-global for every seed, with $g -$ hi unchanged to $5 \cdot 10^{-10}$ J/mol, and none of 200 randomly drawn agreeing points becomes non-global (0 of 200, $[0, 0.018]$); at all 5225 points of these 209 calls the phase assemblages are the same for all seeds, and the energies differ in the last digits only. Near invariant reactions the same holds (CAL-B65, T0): all 50 registered non-global points are non-global for every seed, with $g -$ hi unchanged to $9 \cdot 10^{-11}$ J/mol, and none of 100 randomly drawn agreeing points becomes non-global (0 of 100, $[0, 0.036]$); so none of the registered verdicts of metric 2 depends on the seed. The same registration confirmed independently a case found by the proxy-user trial, `crtiv_ghosh.tdb` (Cr–Ti), a test database of pycalphad outside the audit frame; the case was not blind and enters no rate (CAL-B66). The driver refuses the file because two lines carry a double quote after a comment mark, which pycalphad ignores; with these two bytes removed, pycalphad's output is bit-identical on both files. Point certificates prove that pycalphad 0.11.2 at its defaults does not return the global equilibrium of this file's model at four points for all eight seeds, by 0.028 to 5791.5 J/mol (single-phase LAVES_C36 at 1073.84 K and $x_{\mathrm{Ti}} = 0.045$, where the certificate proves BCC_A2 plus LAVES_C15), and at 933 K and $x_{\mathrm{Ti}} = 0.84$ for five of the eight seeds, by 31.0 J/mol, while the other three return the certified LAVES_C15 plus HCP_A3. The two groups of seeds coincide with the two iteration orders of a Python set of components in pycalphad's Model, which leave the symbolic energy and the sample points unchanged and alter only the last digits of the energies on the starting grid; that this order sets the rounding through the order of additions is a hypothesis. Region certificates at 938.4805 K and 938.4923 K prove that at $x_{\mathrm{Ti}} = 0.84$ titanium-rich BCC_A2 is absent at the first temperature and present at the second; in the default scan, for every seed, it first appears at 932.45 K, about 6 K lower, after which the assemblage alternates between BCC_A2 and HCP_A3, each with LAVES_C15 (and, briefly, single-phase LAVES_C36), until its last change at 938.15 K, still below the certified bracket.

Reanalysis (Post Hoc). Everything in this and the next paragraph was decided after the registered results had been read, and leaves every registered verdict, rate and denominator unchanged; the methods were committed before any of the new numbers was computed, except that the size distribution and the list of open cases had been looked at while exploring the archive. Metric 1 measures whether the published file regenerates the published table, not whether either is accurate: measured invariant temperatures carry uncertainties of a kelvin or more. The 47 comparable files come from 32 papers, and the 21 with a robust disagreement from 16; clustered by paper, D_{robust} is 16 of 32 (0.50, Clopper–Pearson $[0.319, 0.681]$; a cluster bootstrap over papers gives $[0.286, 0.608]$ against the registered $[0.302, 0.599]$), and D_i is 20 of 32 ($[0.437, 0.789]$; CAL-B71, T0). Of the 46 robust disagreements, 34 have a closed certified interval; by their distance d beyond the printed precision, 8 lie below 0.1 K, 13 within $[0.1, 1)$ K and 13 at 1 K or more, 10 of which come from the two files of one Ca–Ga paper, the others from Nd–Sn (1.01 K, twice) and Ho–Sb (2.31 K); so $d \geq 1$ K occurs in 4 of 47 databases from 3 papers, $d \geq 2$ K in 3 from 2, and the largest registered distance is 7.53 K, in Ca–Ga ($\mathrm{L} \rightleftharpoons \mathrm{CaGa_2}$, 8.05 K from the printed value; CAL-B70, T1). Five databases count as robust only through cases without a closed interval. Of these 12 cases, supplementary certificates (outside the registered rule) close 5: one Ca–Ga reaction at $d = 16.85$ K, three at 5.75–6.12 K and one Hf–Sn reaction at 9.58 K; 2 stay open, Gd–Ti, where the same phase assemblage is certified up to 34.3 K above the printed temperature, and Cd–Pu, where the change is certified to lie between 1.0 and 11.2 K above it; and in 5, all rows of doubtful printed composition, the test composition lies in another phase field, so there is no comparable temperature (CAL-B72, T1). The largest certified distance is thus 7.53 K for the registered closed intervals and 16.85 K with the supplementary certificates. For metric 2, the 7 of 63 databases with a non-global pycalphad equilibrium on the uniform grid are 4 without the three that count only through the 10^{-6} phase-fraction filter; near invariant reactions it is 16 of 52, and 14 of the 21 databases of the combined count have such points only there, 14 of their 50 points being those 0.05 K outside a certified transition; by the largest energy difference, 4 of these 16 databases lie below 1 J/mol, 10 between 1 and 100 J/mol and

2 (Gd–Ti, Nd–Ti) above 10^4 J/mol (CAL-B73, T1). Recomputed with pdens = 2000 only at the failing points of the default runs, not as a full rerun, 49 of the 50 near-invariant and 24 of the 29 grid points are no longer non-global, and 5 of 63 databases keep one, 2 without filter-only points (CAL-B74, T0): the counts quantify the known sensitivity of grid-based minimisation near phase transitions to its starting grid. The stylised decisions of the addendum, $D_{\text{any}} = [17, 18]$ of 66, decompose as follows (CAL-B75, T1): metric 1 contributes $[7, 9]$ of 47, 5 only through rows of doubtful composition, 2 through temperatures of cases without a closed interval and none through the shift of a closed interval, whose largest absolute value, 8.05 K, is below the 10 K margin; pycalphad contributes 2 of 63, both also changed by OpenCalphad; and the OpenCalphad development build contributes 10 of 54. Without doubtful rows the count is $[12, 13]$, without the development build $[9, 11]$, and without both $[4, 6]$ of 66.

Rounding of the Printed Parameters (Post Hoc). The parameters in a file are the rounded constants printed in the paper, while the paper's table was usually computed with unrounded values; how far can rounding move an invariant temperature (CAL-B76–B78)? For the 34 closed-interval disagreements, and for information the 5 closed by supplementary certificates, every numerical coefficient of the assessed binary parameters of the phases involved (compound endmembers and interaction parameters, not the SGTE unary data, stoichiometric multipliers, exponents or range limits) is moved within half a unit of its last printed digit, trailing zeros included. A linearised range from central differences, widened by up to 18% where random vertices of the box move further, is computed in floating point (T0); the files at the two extreme vertices and witness files on the segment towards the printed value, 109 TDB files in all, are run through the registered pipeline unchanged, and their 7339 certificates are accepted by the current checker (CAL-B77, T4). For 28 of the 34 a file whose every assessed coefficient lies within half a unit of its last printed digit has a certified invariant temperature that meets the printed one: for 27 a witness file (25 agree by the registered rule; for two Ca–Ga rows of doubtful composition the interval meets the printed value but the reaction type differs) and for one the extreme-vertex file itself (CAL-B78, T1). These include all the 3–8 K differences of the Ca–Ga paper, whose compound entropy terms are printed as integers or with one decimal, giving rounding ranges of 4–206 K. Six, in 4 databases from 4 papers, exceed the rounding range, and both their extreme-vertex files are still certified to disagree: three Cd–Sr reactions ($d = 0.055$–0.068 K against ranges of 0.037–0.041 K), Ge–Lu (0.69 against 0.05 K), Nd–Ti (1.00 against 10^{-4} K) and Ho–Sb (2.31 against 0.96 K); the supplementary Hf–Sn case (9.58 against 0.04 K) does too (CAL-B76, T0). The "beyond rounding" class rests on the linearisation: no certificate covers the whole parameter box, which would need a new checker, and for Cd–Sr the certified margin beyond the extreme vertex is only 0.015–0.032 K. The "within rounding" class rests on reading every last printed digit as rounded: integer coefficients that were chosen rather than rounded would shrink the ranges, and rounding coarser than the last digit, which is not perturbed, would widen them. Seven deviations from the method are recorded in the audit notes, among them an implementation error that made the first witness search fall back to the unperturbed files; those runs were discarded and redone. With the printed precision as the yardstick most differences are thus within what rounding of the printed parameters can produce; under the rounding yardstick 6 of the 34, in 4 of the 47 databases, and the supplementary Hf–Sn case remain printed values that the published file is proven not to reproduce and that rounding of its printed parameters does not explain, the latter on a linearised range.

O Structure Search

Problem. mini-CSP (7 October 2026) starts a structure-search track with clusters of N identical atoms in three dimensions and the untruncated Lennard-Jones pair potential, with ε and σ the exact rationals in the certificate. The energy, gradient and Hessian are rational functions of the coordinates, invariant under orthogonal maps, translations and permutations (CSP-C01). The searcher, written in C from the published description of particle-swarm structure prediction, combines random generation, BFGS and Newton

relaxation, particle-swarm updates and fingerprint deduplication; it is untrusted and only proposes candidates.

Certificate. For each candidate x_0 the certificate gives a radius R, a convexity constant μ and a gradient bound g. The gauge is fixed by the Eckart slice $x_0 + W(x_0)$, the directions orthogonal to translations and rotations at x_0; Lean proves that every configuration can be moved into the slice by a rigid motion without moving further from x_0 (CSP-D01). A strong-convexity ball theorem (CSP-D02) then gives, from μ-convexity on the ball and $g < \mu R$, exactly one minimiser with quadratic growth, at distance at most g/μ, with energy in $[E(x_0) - g^2/2\mu, E(x_0)]$. The checker verifies the hypotheses in exact rational arithmetic, using explicit pair bounds on the Lennard-Jones second and third derivatives over the ball and a rounded Schur-complement test of positive semidefiniteness, and prints `MIN_OK`: an exact local minimum, unique up to rotation, reflection and translation, with $\nabla E = 0$ in all of $\mathbb{R}^{3N}$ (CSP-B01, T4). Two candidates are certified the same structure by a permutation and an exact rational orthogonal matrix from a Cayley parameterisation, or distinct by disjoint energy intervals (CSP-B04, T4). `BEST` names the lowest certified candidate of the run; it is not a claim of global optimality. For LJ_3 to LJ_{13} the searcher finds, and the checker certifies, intervals that contain the Cambridge Cluster Database energies, of width far below the printed fifteen decimals, in 0.1 to 17 s of checking; ASE's `LennardJones` agrees with the certified centres to $7.1 \cdot 10^{-15}$.

Search Mathematics (T3, Idealised). Under the usual stagnation model (fixed attractors, independent uniform random numbers), Lean characterises the first-order and mean-square stability regions of the particle-swarm update exactly, through Schur–Cohn criteria for real quadratics and cubics, and proves that the Clerc–Kennedy constants lie inside both (CSP-A01); independent uniform restarts reach any neighbourhood of the global minimum with probability tending to one (CSP-A02). These results are asymptotic and about an idealised search, not about the C code, which has no Frama-C proof yet (CSP-P01, target T4); the certified output lines do not depend on them.

P Evaluation

All experiments are scripted in `bench/` and can be re-run by anyone once the code is released. Cross-checks against reference codes are T0 evidence: they validate the drivers and the translation of inputs, while the guarantees come from the certificates. Two questions matter. Do our solvers solve the same discrete problems as established codes? Are the certified enclosures tight enough to be useful?

Agreement. Table 13 shows that the drivers reproduce the reference codes to near machine precision on identical discrete problems, and that the reference values lie inside our certified intervals. The cross-checks also exposed modelling subtleties. DFTK's default occupation threshold leaves an electron-count error of about $8 \cdot 10^{-9}$ at the Fermi level, which raises its free energy by roughly $2.6 \cdot 10^{-9}$: still above our certified lower bound, and consistent once the threshold is tightened. On grids containing the zone boundary $f = \frac{1}{2}$, DFTK's spherical cutoff yields $2K + 2$ plane waves, so the two codes solve different discrete problems; we neither compare nor certify those cases. For LDA, PW92 must use Libxc's original constant $A = 0.031091$; with 0.0310907 the energies differ from DFTK by $4 \cdot 10^{-7}$. In CALPHAD the certificates exposed two weaknesses of the reference: in 6 of 141 cases pycalphad's driving force is a local maximum, below our certified global value (its own `calculate` at our composition lands inside our interval); and at Al–Fe, 1000 K, $x = 0.9$ its multiphase equilibrium reports a G_M $6 \cdot 10^{-6}$ J/mol above its own single-phase value, just outside our interval. For phases with two mixing sublattices the benchmark also requires the driver's minimum over the constraint segment never to lie above pycalphad's minimum over 2001 points of the same segment; the audit of Appendix Q finds larger failures of the reference on whole databases.

Tightness. Tables 14 and 15 show that the certificates are close to the accuracy of the computation itself. The enclosures that matter physically, energies and true solutions, are between 10^{-15} and 10^{-8} wide. The FEM sup-norm bounds are within about 20% of the observed error, and the MD local bounds are nearly sharp. The weakest enclosures are where theory predicts weakness. Eigenvalue half-widths ($\sim 10^{-4}$) scale like the square root of a residual; single-phase chemical potentials did too, until the checker began to compute its own tangent. LDA energy intervals are tens of times wider than rHF ones because the nonlinear exchange term adds a first-order contribution. Global MD bounds grow exponentially in time. In three dimensions the widths stay small at production settings: the S1 certificate encloses every one of the 8192 band energies of silicon on its $8 \times 8 \times 8$ mesh to better than 10^{-8} Ha at a cutoff of 10 Ha; there the cost, hours on 32 cores and about 93 GB of memory, is the limiting factor (Appendix S).

Tampering. Every domain includes adversarial tests, and `make verify-all` requires every certificate test case to come with a tampered copy or a wrong problem that is caught. In DFT, zeroing a Cholesky factor is rejected. Perturbing an orbital either fails or yields a wider but still correct interval. Changing the smearing type, shifting μ_0 by 10^{-3}, mixing a k-point certificate with a Γ-point one, or perturbing an eigenvector of a finite-temperature band certificate is rejected. In FEM, altering a nodal value only enlarges `NODE_ERR`, while non-monotone meshes, negative boundary coefficients and incompatible pure Neumann data are rejected. In MD, a displaced position only enlarges the Verlet residual (from $2 \cdot 10^{-16}$ to $5 \cdot 10^{-10}$) and the local bounds, while negative masses and overlapping atoms are rejected. In two-dimensional FEM a flux whose normal component jumps across an interior edge is rejected, as are the invalid meshes of Appendix L; in three-dimensional MD, negative masses, a cutoff of at least half the box and a pair that could cross half the box inside the tube are rejected. The oscillator checker is conservative: the certificate for x^4 presented as one for $x^4 + x^2$, for which its lower bound would still hold, is rejected because the centred form cannot prove positivity near $x = 0$. For LDA, changing the grid size, shifting μ_0, presenting an LDA certificate as an rHF one or swapping the functional between Slater exchange and Slater exchange with PW92 correlation is rejected. For LSDA, doubling B, shifting ν_0 by 10^{-3}, adding an electron, shifting one coefficient of $u_\downarrow$, swapping the two spins' decompositions, swapping Slater exchange and PW92 correlation or presenting the certificate as an LDA one is rejected. For Hartree–Fock, a changed kernel, temperature, electron count or smearing, a shifted centre and a k-point or rHF header are rejected; for forces, sites relabelled or moved in the certificate only follow the echoed problem, and forged `FORCE_OK` lines are never conclusions. In CALPHAD, phase compositions outside $[0, 1]$, assemblages that do not contain x_0 and a positive antiferromagnetic factor are rejected, while a shifted tangent, another feasible assemblage or a poor driving-force point only widen the intervals. A chemical potential at or above the Gibbs energy of pure B, or a raised μ_B with unchanged contact points, is rejected, and so is a partition certificate whose first segment does not start at composition 0, and a property certificate with a site fraction above 1, a phase index out of range, or no internal state for a phase with two mixing sublattices. Problem binding catches tampering that is otherwise undetectable: a DFT external-potential coefficient changed by one ulp, or a CALPHAD model coefficient changed by 0.01%, is accepted as a statement about the changed problem without the binding and rejected with it, and a changed site count of θ or C15 is rejected by the binding. Cross-checks also catch mistakes in the model itself: a Methfessel–Paxton free energy lies outside the Gaussian interval, so using the wrong smearing formula is detected.

Adversarial Review. Tampering tests are written by the people who wrote the checkers. We therefore reviewed the checkers adversarially, with an attacker who controls the certificate and the driver but not the checker or the command line, and found seven flaws, none in a soundness theorem. Phase names containing a Unicode line separator or a tab are echoed verbatim, so a Python consumer splitting lines or fields read a forged `MODEL_OK` line or a forged composition window; certificates may now contain only printable ASCII outside comments, and the `fis` parser splits only at newlines and spaces. Exponents such as `3E-999999999` or 2^{10^7} made a checker hang; they are now bounded. The DOS theorem

sums over the k-points that passed, so a certificate with one broken k-point or with scaled weights still printed a DOS for the whole grid; `DOS_OK` is now printed only for a complete set with unit weight sum, and band labels are the certificate's k-point indices. Finally, `make verify-all` did not compare every DFT echo (smearing, functional, grid size) with the command line. A regression harness replays the forged certificates on every run (241 cases after the reliability audit below, including the zero-temperature LDA, activity, FEM eigenvalue, k-point band, thermodynamic-property and collinear-spin LSDA modes, the latter at finite and zero temperature, with k-points, bands, energies, the Gaussian potential and PW92 correlation, user coefficient files and sites, and `mini-CSP`); six weaker points were recorded, among them that a `--tdb` certificate could leave out phases of the database and that a failed secondary check did not change the exit code. A second review (7 October 2026) covered the entry points added since: collinear spin at finite and zero temperature with k-points, bands, energies and PW92, user potentials, `fis multiplicity`, the `mini-CSP` checker, and `fis` against the driver's machine-readable lines. It found no wrong soundness theorem and no checker printing a claim about a problem other than the one it echoes, and fixed eight points: one exit-code rule for all five checkers (0 only if no line starts with a tag ending in `_FAIL` or `_PARTIAL`, 3 when the main check passed and a secondary one failed); a complete phase table for `--tdb` certificates, or the user's `--phases`; exact CALPHAD headers; the input gate for the spin-down k-point lines, where a zero denominator had silently given $s = 0$; `fis multiplicity` using the named-potential ball rather than that of the certificate's rounded coefficients; `fis` hashing the user's input files before the driver runs; `make verify-all` comparing the LSDA band echo with the command line; and `fis` ignoring driver lines named like conclusions. Resource exhaustion by huge certificates, `fis` not comparing the echoed problem with the driver's options, and files changed while the checker runs remain recorded. The regression harness now replays 686 forged certificates.

Reliability Audit. A second audit (7 October 2026) of the checkers, their input and output code and the test harness found further holes, now fixed. The Lennard-Jones terms are written with r^{-1}, and Lean defines $0^{-1} = 0$; the MD checker verified the ordering of the atoms only for the states it stepped from, not for the final state or a trajectory of zero steps, and never checked σ, so certificates with coincident atoms or $\sigma = 0$ were accepted with an energy printed. The verified check now requires at least one step, positive ε, σ, h and masses, and every state strictly ordered, and the energy lines are proved to be the genuine Lennard-Jones energy. Phase names read from a TDB file, which the checker reads itself, could still inject output lines; all names are now escaped, and conclusions are printed by phase name only when the names are pairwise distinct, which is a theorem. The MD, FEM and CALPHAD checkers now echo their whole problem as exact `PROBLEM` lines and parse strictly (a missing MD line used to default to zero); the DFT checker also echoes the k-points, weights and DOS width of band and DOS runs and rejects an unknown header or key before checking; and the CALPHAD checker prints the SHA-256 of the TDB file it read. The SCF-solution theorems now state that the SCF solution itself is unique in the certified ball, not only the zero of the auxiliary map, since every SCF solution is shown to be such a zero. The ACSL contracts of the trusted kernel, the Frama-C command line and the floating-point compiler flags are now fingerprinted like the Lean statements, a skipped test case can no longer pass silently, and every `make verify-all` run leaves a record of the commit it checked. The echoes, input gates, escaping and hash are unverified input and output code (T0), cross-checked against the command line by `make verify-all` and by the regression harness.

Known Answers and Positive Controls. Except where it is bound to its problem, a checker certifies the problem written in its certificate, so tampering tests cannot detect a driver that consistently solves the *wrong* problem. Of the 213 certificate test cases, 162 therefore carry a known answer computed independently from the command line, never from the certificate: free electrons at Γ ($E = \pi^2/2$, $\varepsilon_n = \pi^2 n^2/8$) and with k-points ($\pi^2/54$ per cell); free electrons at finite temperature (Fermi–Dirac F, S, E); free electrons with LDA (PW92) at Γ, including the bands $\frac{1}{2}(2\pi(k+s)/L)^2 + v_{\mathrm{xc}}(\rho_0)$, and with k-points, where the density is uniform, the Hartree term vanishes and v_{xc} only shifts μ, so arbitrary-precision arithmetic gives μ, E and F, and at zero temperature E and the gap, at Γ and with k-

points; free electrons with spin-polarised Slater exchange in a Zeeman field, where the uniform spin densities make ν and μ the solution of two scalar occupation equations, solved in arbitrary precision, which also give E, F and the per-spin bands, and at zero temperature $N_\uparrow$, $N_\downarrow$ and the gap, at Γ and with k-points; the same with spin-interpolated PW92 correlation, whose potentials are obtained by numerically differentiating Libxc's $e_c(\rho_\uparrow, \rho_\downarrow)$ independently of the checker's closed forms, at finite and zero temperature, with energies and per-spin bands; free electrons with Hartree–Fock, whose diagonal HF equations $\varepsilon_a = q_a^2/2 - L^{-1}\sum_k v_{a-k} f_k$ are solved in arbitrary precision for μ; the exact solutions $\sin \pi x$ and, for Neumann data at both ends, $\cos \pi x$ in FEM, and the closed-form P1 eigenvalues of a uniform Dirichlet mesh; an MD chain at rest, whose energy is a sum of pair potentials; three CALPHAD cases with G, both chemical potentials and a zero driving force in closed form: a binary regular solution, a sublattice phase $(\mathrm{Cu})_1(\mathrm{Cu},\mathrm{Ni})_1$, and a phase $(\mathrm{Cu},\mathrm{Ni})_1(\mathrm{Cu},\mathrm{Ni})_1$ with two mixing sublattices whose symmetric endmembers put the internal minimum at $y_1 = y_2 = x$, so that $G = Dx(1-x) + RT\,S(x)$; the regular solution at a prescribed μ_B, where x solves $\mu_\mathrm{B} = RT \ln x + L_0(1-x)^2$, and at a prescribed activity, with the liquid itself or a pure-Ni compound as reference phase; the thermodynamic properties of the same two phases, $H = L_0 x(1-x)$, $S = -R\,S(x)$, $C_p = 0$ for the liquid and $H = -100 + 0.01\,T$, $S = 0.01(\ln T + 1)$, $C_p = 0.01$ for the compound with $G = -100 - 0.01\,T \ln T$; and the partition of a regular solution with a miscibility gap, whose binodal compositions x_b and $1 - x_b$ must lie in the certified windows; and the Lennard-Jones clusters LJ_3 to LJ_{13}, whose best certified energy interval must contain the Cambridge Cluster Database value. Later known answers include the exact Sturm–Liouville eigenvalues $(j\pi)^2$, the true energy-norm error for $-u'' = e^x$, $2\pi^2$ for the unit square and the classical value for the L-shaped membrane, exact rational Verlet iterates, a resting fcc lattice in the periodic box with energy $192\,\Phi(a^2/2)$, $E_1(0) = N \sin^2(\pi/2N)$ for the FPUT chain, the quartic-oscillator energy from two independent routes, and $\lambda_\mathrm{min} = 1$ for the harmonic oscillator. Each known answer must lie inside the Lean intervals. Positive controls show that this check has teeth: a driver run with a slightly wrong cell length, spacing or temperature produces a self-consistent certificate that the checker accepts, and only the known answer, now outside the intervals, reveals the error; for DFT the wrong cell length also shows in the echoed problem. A driver that solves free electrons in a weak cosine potential while its certificate names the free potential is self-consistent too, but the problem binding rejects it at Γ and with k-points, for rHF and LDA. This is direct evidence for the gap discussed in Appendix S: where the problem is neither bound nor echoed, a certificate proves the right answer to whatever problem the driver wrote down.

Showcases. Twelve complete runs, five in DFT, three in FEM, one in MD and three in CALPHAD, are archived with the driver report, the certificates, the checker's full output and an independent solution by the reference code. All 48 judged comparisons agree, and `make verify-all` re-checks the archived certificates with the current checkers, so a change that breaks an old certificate is noticed. The same data generate the showcase page of the project site.

Observed Numerical Pathologies. The certified pipeline separates convergence from correctness. In a symmetric double well with nearly degenerate frontier orbitals, SCF oscillates for every mixing parameter, and the controller ends with `max_iter` rather than reporting convergence. With only the Γ point, a system whose true potential period is $L/2$ appears gapped. On a k-grid, the interval inertia check correctly reports `gap_closed`: the system is a metal. A converged residual alone would have hidden both issues.

P.1 Archived Showcase Runs

Besides the benchmarks, 19 complete runs are archived in `showcases/` and shown on the project site, each with the same four steps: the driver computes a candidate and writes a certificate, the Lean checker checks it and prints the guaranteed conclusion, a reference code (DFTK, scikit-fem, ASE, pycalphad) solves the same discrete problem independently, and the comparison is recorded (T0). Every certificate is stored with its SHA-256 hash and

the checker's full output; `make verify-all` re-checks all of them with the current checkers and requires the output to be byte-identical to the archive (about half an hour with 32 parallel jobs). Table 16 lists them. Two silicon runs use a low cutoff ($Q = 17/2$) so that they can be re-checked in minutes; they are not converged silicon. The S1 production record (Appendix K) is attached to the zone-wide band case as an *archived record*: its checker output is stored and hash-checked, but not re-run by `make verify-all`, because one check takes about 196 CPU hours; the 253 MB certificate is referenced by its hash.

Table 16: The archived showcase runs. Each conclusion is printed by a checker with a Lean soundness theorem (T4 for the checker); the last column counts the cross-checks against the reference code that agree.

Run	Lean-checked conclusion	Reference
DFT		
Real Si, LDA, $4 \times 4 \times 4$ mesh	unique SCF solution within $6 \cdot 10^{-12}$; 1024 band intervals; mesh gap in $[1.567906704, 1.567906708]$ eV	DFTK 4/4
Si bands over the whole zone (+ S1 record)	$\varepsilon_4(\kappa) - \mu^* < 0.031140$ Ha for every κ; gap $\leq$ 1.3338 eV; archived S1 record: gap $\leq$ 0.530902 eV	DFTK 4/4
S2: not the ground state (H)	the spin-restricted SCF solution lies at least 330.0 meV above a spin-polarised trial state	DFTK 3/3
Cohen–Bergstresser vs. experiment	12 materials, 55 pre-registered comparisons: 14 agree, 39 disagree, 2 no verdict; intervals $\leq 1.1 \cdot 10^{-7}$ eV	DFTK 12/12
rHF electron pair in a double well	$E \in [-2.592240050, -2.592240045]$; 33 eigenvalues, 627 band values, 201 DOS values enclosed	DFTK 4/4
Metal at finite temperature	F to $3.5 \cdot 10^{-11}$; S, E enclosed	DFTK 3/3
LDA (Slater + PW92) SCF	exactly one SCF solution nearby; F to $2.5 \cdot 10^{-14}$	DFTK 3/3
LDA with order-3 Methfessel–Paxton	exactly one SCF solution nearby; μ to $4.1 \cdot 10^{-13}$	DFTK 3/3
k-point insulator	E per cell to $1.2 \cdot 10^{-8}$	DFTK 1/1
FEM		
Poisson, manufactured sin πx	$\sup \lvert u - V_h \rvert \leq 1.21 \cdot 10^{-3}$	scikit-fem 2/2
Neumann at both ends	compatibility decided; $\sup \lvert u - V_h \rvert \leq 1.60 \cdot 10^{-3}$	scikit-fem 3/3
Adaptive cantilever	146 elements, $\sup \lvert u - V_h \rvert \leq 9.8 \cdot 10^{-7}$	scikit-fem 2/2
L-shaped membrane, λ_1	$\lambda_1 \in [9.6205, 9.6553]$, both bounds proved	scikit-fem 3/3
MD		
Lennard-Jones chain, Verlet	energy drift $\leq 4.7 \cdot 10^{-7}$; position error $\leq 10^{-3}$ at $t = 0.5$	ASE 3/3
Lennard-Jones atoms, periodic 3D box	true solution exists and is unique; position error $\leq 9.5 \cdot 10^{-4}$ at $t = 0.144$	ASE 2/2
CALPHAD		
Al–Zn FCC miscibility gap	G^* to 10^{-9}; two-phase windows $\leq 3 \cdot 10^{-7}$	pycalphad 13/13
Cr–Fe magnetic BCC gap	G^* to 10^{-9}	pycalphad 9/9
Al–Zn phase diagram	35 regions, 12 common tangents, boundary windows $\leq 3.7 \cdot 10^{-6}$	pycalphad 2/2
Pb–Sn eutectic	$T_E \in [454.5521, 454.5721]$ K, for every real temperature in the interval	pycalphad 6/6

Q Certificates as Research Instruments

Floating-point software can report that it found a solution; it cannot show that no lower one exists, that a phase does not appear, or that two computed solutions are really different. The

certificates make exactly such statements, so they can answer questions of this kind. We report two results here; the audits of widely used codes, of a classic pseudopotential table and of published thermodynamic databases against certified answers are reported with their domains (Appendices K and N). Each is archived with its certificates and the checker's output, and `make verify-all` re-checks the archive with the current checkers. They are not ledger claims: the step that combines the checkers' conclusions, comparing problem data and intervals, is a short script in exact rational arithmetic outside Lean.

Several Exact Kohn–Sham Solutions. rHF has a unique ground-state density, but the exchange energy is concave, and LDA SCF iterations started from different densities may converge to different points. That alone proves nothing: the points may approximate one solution, or none. Consider four electrons per period $L = 0.6$ in the cosine potential $20\cos(8\pi x/L)$, whose period is $L/4$, with Slater exchange and PW92 correlation, Fermi–Dirac smearing at $T = 0.03$, $K = 6$ and $M = 25$. Five LDA SCF-solution certificates (DFT-X09, X13) for this problem are accepted, with radii between $2 \cdot 10^{-13}$ and $1.6 \cdot 10^{-12}$. Their problem data are byte-identical, so the five theorems are about one system of equations; their centres are 0.048–0.096 apart, far more than the sum of any two radii, so the five balls are disjoint, also in the density components. The discrete SCF equations therefore have at least five exact solutions with pairwise distinct densities: the symmetric one and two pairs of charge-density waves. Their free-energy intervals, at most $5 \cdot 10^{-11}$ wide, are disjoint between the groups, which orders them: one pair lies below the other, and both lie below the symmetric solution, about 0.039 Ha above the lowest. The mechanism is a half-filled shell at $k = \pm 2\pi/L$, which exchange turns into a density wave of frequency 2; the potential has only frequency 4 and pins its phase to discrete values. Without the potential the phase is a continuous zero mode, the Jacobian is singular, and no certificate can pass. The result is a lower bound for one discrete problem, not a count, and which solutions are stable is not proved. Floating-point runs find five solutions at $K = 8$, 12 and 16 but only two at $K = 24$, none of them certified, and the effect is an open-shell effect of a small cell at the Γ point, not a bulk instability. The archived certificates were regenerated for the mandatory potential binding of Appendix J: the current driver reproduces all five byte for byte apart from the added `pot` line, so the checker binds each of them to the named cosine potential, with unchanged radii and intervals.

A Certified Audit of CALPHAD Databases. Over what range of temperature and composition does a published database give physically sensible equilibria, and where does a general-purpose solver miss the global equilibrium? We took every binary system of the 17 databases shipped with pycalphad, COST 507 (Ansara et al., 1998) among them, on a grid $T = 300, 400, \ldots$ K up to the upper limit of the database's parameters (at most 6000 K) and $x = 0.02, 0.06, \ldots, 0.98$, with all phases the driver supports (3470 phase–system pairs; the others are listed with the reason). At every point the driver wrote a certificate that the checker verified in `--tdb` mode, and pycalphad computed the equilibrium on the same phases. All 246,650 points in 316 binary systems are certified for the model their TDB file defines; COST 507 alone takes about 30 minutes with 150 processes on a 384-core machine. Four kinds of finding stand out. *Anomalous liquid gaps.* In five systems the liquid loses convexity at high temperature, so a liquid miscibility gap opens as the temperature rises, far above the data the parameters were fitted to; the upper limits of up to 6000 K in these files follow the SGTE convention for the range of the Gibbs-energy functions and are not a claim of validity by the assessors. Convexity certificates below and point certificates above bracket each onset to 1 K (Table 17), and pycalphad's own G'' changes sign inside the same kelvin. *Equilibria pycalphad misses.* In the Fe–Ni database FeNi_deep_branching at 700 K and $x_{\mathrm{Ni}} = 0.98$, pycalphad returns one fcc phase; the certificate proves two fcc phases near $x_{\mathrm{Ni}} = 0.9774$ and 0.9881, a narrow gap, and pycalphad's G_M is at least 0.37 J/mol too high. In COST 507 Cr–Ti at 300 K and $x_{\mathrm{Ti}} = 0.34$, pycalphad returns a single hcp phase at $+812.7$ J/mol, while the certificate proves the equilibrium to be the C15 Laves phase plus the $\mathrm{D0}_{19}$ phase `ALM_D019` at -9006.96 J/mol, so pycalphad is at least 9.82 kJ/mol too high (32 grid points in this system; given only these two phases, pycalphad finds the same equilibrium). At four points of Mn–Ti pycalphad stops at a local internal state of the C14 Laves phase, at least 4.5 J/mol too high, and in COST 507 Fe–Mg at 1000 K it reports the right phases at least

1.47 J/mol too high. Everywhere else its G_M lies inside the Lean interval or less than 1 J/mol above it. *No model mismatch.* At no point does pycalphad's G_M fall below the certified lower bound, which would mean that the two read the database differently. This check is what exposed the dropped molecular species of Appendix I.2, where pycalphad lay up to 128 J/mol below the lower bound; after the fix, the driver and the Lean reader both mark such phases unsupported. *Solids above both melting points.* The certificates prove that the liquid is not the single-phase equilibrium where, for example, the C15 phase of COST 507 Cu–Mg coexists with liquid or is stable alone up to the top grid temperature of 2900 K, far above the melting point of copper, 1358 K; similar cases are `ALTI` in Al–V, `SNTI3` in Sn–Ti and `SI3TI5` in Si–Ti. Refractory compounds can be stable above both melting points, so whether such a case is a defect needs a person; the certificate settles only what the database says. Conclusions that something is *not* an equilibrium, or that pycalphad's G_M is too high, hold whatever phases were skipped, since adding phases can only lower the hull; conclusions that a phase *appears* hold only among the phases included, and all conclusions hold at the grid points only.

Q.1 Physical Correctness So Far

Measured against triple correctness, the first realisation is unbalanced, and we say so plainly. Mathematical correctness is realised in depth (668 machine-proved claims, certificates about true solutions and global minima), and software correctness substantially (model-checked TLA$^+$ specifications with trace validation for the one-dimensional code and for the parallel three-dimensional driver, a Frama-C-proved interval kernel, fingerprinted statements, adversarial review and fuzzing). Physical correctness was not, until the first three-dimensional example below; every one-dimensional result in this paper, and every two- and three-dimensional FEM and MD result, is at V0. The one-dimensional DFT, FEM and MD models have no direct experimental counterpart, and the Lennard-Jones cluster energies of Appendix O are compared with a database of computed, not measured, values.

Among the one-dimensional domains, the CALPHAD audit is the one place where the work touches the physical layer, and only indirectly. A CALPHAD database is a physical model fitted by its assessors to experimental phase-boundary and thermochemical data; its physical correctness is established, outside our ledger, in the region where those data lie. The audit certifies what the model says everywhere on its declared range, including where it was not fitted. The anomalous liquid miscibility gaps of Table 17 open between 2177 K and 5054 K (each onset bracketed to 1 K), above the melting points of both components of each system, within the temperature range of the files' functions (up to 6000 K by the SGTE convention), where the parameters are extrapolated beyond the experimental data rather than fitted to it; the solid phases certified stable far above the melting point of a component, such as C15 in Cu–Mg up to 2900 K, are of the same kind. Such certified features in data-free regions show what a model says where it was not fitted; they are not validation in the sense of Table 6, since no measured value is compared, and whether a given case is a defect needs a person and a measurement. Positive validation, a certified computed interval set against a measured value with its uncertainty, is the work of the physical track P and of showcase S1 (Appendix H.1).

A First Example with All Three Correctnesses. The empirical pseudopotential of Cohen and Bergstresser (CB, 1966) for 14 diamond and zinc-blende semiconductors is a linear, three-dimensional model whose six form factors per material were fitted to optical transition energies. Its band energies at a k-point are eigenvalues of a Hermitian plane-wave matrix, and a new checker, `certify-pw`, proves intervals for them: from the driver's approximate eigenvectors X it forms $X^{\mathsf{H}}(H - \sigma)X$ exactly in Gaussian integers and counts eigenvalues below each shift σ by diagonal dominance, with the cutoff basis, the gauge and the irrational $(2\pi/a)^2$ handled inside the Lean soundness theorem (W13, W14, T4). DFTK's values for the same discrete problem and the exact empty-lattice levels lie inside every interval, and a wrong form factor puts them outside. Separately, Lean proves that the discrete counts bound the eigenvalues of the continuous Bloch operator (W08, W24, T3, modulo the operator identification A-DFT-09) and that band edges over the whole Brillouin zone follow from finitely many certified k-points (W20–W25); the protocol layer of the planned

three-dimensional parallel code is model-checked in TLA$^+$, including four planted parallel bugs that TLC finds (R08–R11).

The validation was pre-registered. CB's Table I lists exactly the experimental splittings they fitted; we froze those 86 values as the fit set and ten other measured quantities as the validation set, and committed a protocol fixing the quantities, the label-to-band mapping, the computed-interval rule, the corrections and the agreement criterion before any certified CB number existed (a timestamped commit to the repository before the checker produced output). The comparison then ran once, with the final checker at a cutoff of $|k+G|^2 \le 67/2$ in units of $(2\pi/a)^2$:

Claim	Quantity	CB (certified)	Measured	Verdict
O01 (V2)	Si valence-band width	12.6134 eV	12.50 ± 0.3	agrees
	Ge valence-band width	11.9670 eV	12.6 ± 0.3	off by 0.54 eV
O02 (V2)	Si $L_{3'} - \Gamma_{25'}$	-1.2530 eV	-1.2 ± 0.3	agrees
	GaAs $L_3 - \Gamma_{15}$	-0.9137 eV	-1.30 ± 0.3	off by 0.38 eV
O03 (V0)	Si $\Gamma_{2'} - \Gamma_{25'}$	3.8892 eV	4.18 (no σ)	none

The measured values are shown before the spin–orbit corrections that the protocol applies; the deviations are after them. Because the computed intervals are about 10^{-7} eV wide, the two disagreements belong to the model at the protocol's cutoff and to the measurement, not to the computation; they lie where CB's fit has no data (the deep valence band, which CB state is "not tied down to any experimental data", and a zinc-blende L level). Even on its own fit set the converged model misses 5 of the 22 fitted values for these three materials by more than 0.2 eV, while CB's printed values, computed in a basis of about twenty plane waves, are reproduced within 0.18 eV. By the protocol a disagreement is a result, not a reason to re-run; the claims are graded V2 because two comparisons with stated uncertainties were made, independently of their outcome. The experimental values come from compilations rather than primary sources and are flagged as such in the data set, which awaits review by an experimentalist.

Extension to Nine More Semiconductors. The same comparison was extended, before any band energy of the new materials was computed, to nine more materials of the 1966 paper (GaP, GaSb, InP, InAs, InSb, ZnS, ZnSe, ZnTe, CdTe; O07–O19), against 50 valence-band energies and widths from one X-ray photoemission study (Ley et al., 1974), with values inside the span of CB's fitted data excluded; the registration was committed and pushed on 17 October 2026 before the certificates were made. `certify-pw` certifies the bands at Γ, X and L in intervals $2 \cdot 10^{-9}$ Ha wide (W70–W78, T4); all 243 DFTK eigenvalues of the same discrete problems lie inside, and raising the cutoff moves the compared quantities by at most 1.5 meV (W79, T0). Of the 50 entries 12 agree, 37 disagree and one, a UPS value without a stated error, has no verdict: InP agrees in four of six levels, InSb and ZnS in none. The pattern is systematic: the X_3 valence level is measured 0.45–2.4 eV deeper than the model in all nine materials, and the valence-band width is off by 0.9–1.7 eV in InAs, InSb, ZnS and ZnSe. Four claims disagree outright and eight agree in part (all V2); the CdTe claim stays at V0. That a local pseudopotential fitted to optical transitions misplaces the deep valence band is not new: the nonlocal pseudopotentials of Chelikowsky and Cohen (1976) were introduced in part to improve the valence-band widths of these materials. What is new is the form of the conclusion: the verdicts were fixed before the numbers existed, and the model's side is certified, so each disagreement belongs to the model and the measurement (XPS final-state effects included), not to the numerics. CB's printed values are reproduced within 0.42 eV (ZnS) and within 0.14 eV for the III–V compounds, and the converged model misses 31 of the 55 fitted values of these materials by more than 0.2 eV.

Towards Showcase S1. The second comparison, of self-consistent Kohn–Sham LDA band gaps, was pre-registered on 10 October 2026, before any certified LDA gap of a real material existed: functional, pseudopotential files with hashes, a static lattice, minimum cutoffs and meshes, the corrections to the measured gaps, the agreement rule and three rules (two-sided, discrete two-sided and one-sided) for which computed interval may be used, each tied

to what the certificates can show (O04–O06). The certified side now exists for silicon at the protocol's settings (Appendix K); its mesh gap is about $4 \cdot 10^{-7}$ eV wide, six orders of magnitude below the size of the band-gap problem; the whole-zone interval used in the two-sided comparison is 0.071 eV wide. The zone-wide bound that the one-sided rule requires now exists as well, and with it the S1 record for silicon: a gap of at most 0.530902 eV (W51). The two-sided rule followed, with the gap of the discrete model in $[0.4146, 0.4856]$ eV (W66), and both records have been compared with experiment by the registered rules (Appendix K): the model lies below the observation by at least 0.695 eV under the one-sided rule and by $[0.741, 0.847]$ eV under the two-sided one, grade V1; aluminium phosphide, registered later under the same rules, lies $[1.075, 1.293]$ eV below under the two-sided rule (O20, W300), and gallium phosphide $[0.873, 1.045]$ eV below (O24, W192). A floating-point study at the certified settings (Z09–Z12, T0) shows that raising the cutoff to 50 Ha and the SCF mesh to 16^3 changes the true gap by 0.47 meV and the mesh gap by 2.4 meV; the 45 meV between the one-sided bound and the true gap of the model come from the conduction-band minimum lying off the mesh, not from the discretisation. When the rule's requirement was reviewed after the silicon number existed, it was kept as frozen, since relaxing it then would have weakened the pre-registration.

R Related Work

Verified Numerical Programs. Boldo et al. (2013) proved a C program for the 1D wave equation correct down to rounding, using Coq, Frama-C and Flocq; Boldo et al. (2017) formalised Lax–Milgram in Coq. Floating-point verification relies on Flocq (Boldo and Melquiond, 2011), CompCert (Leroy, 2009) and VST (Appel, 2011). Isabelle/HOL contains verified enclosures of ODE solutions (Immler, 2014), with which Immler (2018) re-checked the numerical part of the proof that the Lorenz attractor exists. FORMAL IN SILICO differs in scope, spanning five domains and all layers from model to floating point, and in strategy: certificates are checked rather than solvers verified, so production-style solvers can be used unchanged. At a much larger scale, Lucas et al. (2026) had a language model write Rocq proofs that 239 entry points of 123 open-source libraries, compiled to RISC-V, cannot crash on any input up to 64 MB, by turning existing test suites into specifications; formalising the instruction semantics also exposed hundreds of errors in the reference documentation. Our audits (§3.2) rest on the same effect for scientific reference data: stating a model precisely enough to certify it exposes where published tables and files disagree with their own models. Their trusted base (the machine models, the exact theorem statements and the proof checker) has the same shape as ours (Appendix T), and their caveat that a proof does not show that the specification matches reality is the gap that the physical leg of triple correctness addresses.

Validated Numerics and Certifying Algorithms. Interval methods and verification methods (Moore et al., 2009; Rump, 2010; Tucker, 2011) compute rigorous enclosures, famously in the proof that the Lorenz attractor exists (Tucker, 2002); the enclosure software is often trusted, though its result for the Lorenz attractor has since been re-checked by a formally verified solver (Immler, 2018). Computer-assisted existence and enclosure proofs for nonlinear elliptic boundary-value problems, by Newton–Kantorovich-type fixed-point arguments around an approximate solution (Plum, 2001; Nakao et al., 2019; Gómez-Serrano, 2019), are the mathematical precedent of our SCF-solution certificates; we apply such an argument to self-consistent Kohn–Sham equations and have the check itself machine-verified. Certifying algorithms (McConnell et al., 2011) pair untrusted computation with checked witnesses, and their checkers have been formally verified, for LEDA (Alkassar et al., 2014), for SAT unsatisfiability proofs down to machine code with a verified compiler (`cake_lpr` (Tan et al., 2021)), for the nonlinear inequalities and linear programs of the Flyspeck project (Hales et al., 2017), and for interval and polynomial-positivity bounds in Coq (CoqInterval and ValidSDP (Melquiond, 2008; Martin-Dorel and Melquiond, 2016; Martin-Dorel and Roux, 2017)). We follow this pattern and prove the checkers in Lean 4 (de Moura and Ullrich, 2021) on top of Mathlib (The mathlib Community, 2020), with the Lean compiler trusted rather than verified. The idea of verifying the checker is therefore not ours; what is new is the target class, nonlinear and nonconvex self-consistent problems of materials science

checked at production-like settings, the devices that make exact checking feasible there (a split Newton–Kantorovich Jacobian, proved symmetry reduction, arithmetic modulo machine-word primes), and the use of certified answers to audit widely used codes and published data.

Error Control in Electronic Structure. A priori and a posteriori analyses of plane-wave discretisations of Kohn–Sham-type models (Cancès et al., 2012; 2022; Herbst et al., 2020) provide the mathematics that guaranteed bounds need; DFTK (Herbst et al., 2021) was designed to make such analyses practical. Our certificates are machine-checked counterparts, for one-dimensional models and, in three dimensions, for Kohn–Sham LDA of silicon with GTH pseudopotentials. Herbst et al. (2020) obtained guaranteed a posteriori bounds for eigenvalues of the linear (non-self-consistent) Kohn–Sham operator of three-dimensional silicon, accounting for discretisation and floating-point error; our three-dimensional checkers build on the same ideas and add machine checking, nonlinear SCF solutions with k-point meshes and symmetry, and a pre-registered comparison with experiment. Our strongest outputs, enclosures of the discrete ground-state energy, of its infinite-cutoff limit and of SCF solutions for non-variational smearing, carry no unverified constants. Measured on our certified problems (W214–W219), the error estimates of DFTK 0.8.0 (Cancès et al., 2022) are accurate on the discrete SCF error (within 2.2%) and narrower than our intervals, which contain them; they do not cover rounding or establish existence and uniqueness. For the cutoff shift of the silicon gap with the potential frozen, residual bounds are 10^3–10^4 times wider than the certified interval and the second-order estimate (2.34 meV) falls outside it. We therefore do not claim tighter bounds than a posteriori estimates, only bounds of a different kind. Guaranteed lower and upper eigenvalue bounds (Cancès et al., 2017), the a posteriori analysis of nonlinear Gross–Pitaevskii-type eigenvalue problems (Dusson and Maday, 2017) and the Feshbach–Schur analysis of Fourier discretisations (Dusson et al., 2023), which treats the same Schur-complement structure as our truncation bound (Appendix K.2), are further foundations; the error of Brillouin-zone quadrature (Cancès et al., 2020), the k-mesh part of the discretisation, is not bounded by our three-dimensional certificates, which certify the problem on a fixed mesh.

Verification and Validation. The distinction between verification (solving the equations right) and validation (solving the right equations) is standard in computational science and engineering (Oberkampf and Roy, 2010). In that practice, code verification relies on tests such as manufactured solutions, solution verification on estimated discretisation errors, and validation compares with experiment after budgeting the numerical error as an estimate. FORMAL IN SILICO keeps the distinction but replaces the estimates on the verification side by machine-checked enclosures, which is what makes the model-error interval (3) rigorous up to the measurement uncertainty. In electronic structure, the band-gap problem (Perdew and Levy, 1983; Sham and Schlüter, 1983) is the classic case of a model error that persists however accurate the numerics are; the reproducibility study of Lejaeghere et al. (2016) is the classic case of numerical settings confounding comparisons between codes, and Bosoni et al. (2024) give reproducible workflows that verify the precision of DFT implementations against each other and against all-electron references at converged settings, the natural yardstick for our code audits, which instead compare each code with the certified answer of one fixed discrete problem.

Computational Thermodynamics. Rigorous global phase-stability analysis, by deterministic global optimisation (McDonald and Floudas, 1995) and by interval Newton and branch-and-bound methods (Tessier et al., 2000), certified tangent-plane stability for chemical-engineering models decades ago, without machine-checked proofs; our CALPHAD checker applies a branch and bound with a Lean soundness theorem to binary CALPHAD models built from the TDB file. Uncertainty quantification of CALPHAD assessments (Otis and Liu, 2017b; Paulson et al., 2019) treats the parameters as uncertain, and measured invariant temperatures carry uncertainties of a kelvin or more; against these, sub-kelvin disagreements between a file and its printed table say nothing about accuracy, so our metric 1 is a finding about reproducibility, whether the published file regenerates the published table. Assessment guidelines (Schmid-Fetzer et al., 2007) already ask for checks of a description

outside its fitted range, such as liquid miscibility gaps appearing at high temperature; the certificates make such checks rigorous for the model the file defines. The frame of our systematic audit is the TDBDB index (van de Walle et al., 2018).

Formal Methods for Concurrent Systems. TLA$^+$ (Lamport, 2002) is used industrially to verify distributed protocols (Newcombe et al., 2015); trace validation (Cirstea et al., 2024) connects specifications with executions. We apply both to the SCF controller and to k-point parallelism, where errors are protocol errors rather than physics errors.

S Open Questions and Limitations

The Weakest Correctness. By the programme's own thesis, a result is only as trustworthy as its weakest correctness, and for every one-dimensional result in this paper, and for the two- and three-dimensional FEM and MD results, that is physical correctness, at V0 (Appendix Q.1). The validation infrastructure now exists (a ledger section, frozen and hashed experimental data sets, pre-registration and a `make verify-all` step), and it has been exercised on an empirical model with compiled data and on the self-consistent LDA band gaps of silicon, aluminium phosphide and gallium phosphide, where a proved model side turns the known underestimate of LDA into a deviation of the discretised model free of algebraic, rounding and software error (grade V1, Appendices K and Q.1). The S1 protocol for the Kohn–Sham LDA band gaps of Si, C and Ge (Slater exchange plus PW92 correlation, GTH-LDA q4, a static lattice with zero-point expansion removed, minimum cutoffs and k meshes) was frozen before any certified LDA gap of a real material existed; its three claims are registered and machine-checked to be unchanged until the comparison (O04–O06). Of the three materials only Si is pursued. For C the protocol's minimum cutoff (50 Ha, about 1300 plane waves per k point) puts a certificate check at an estimated 3700–4500 CPU hours with about 1 TB of memory, more than the machine allows until the checker streams its column results; for Ge the KS-LDA gap at the protocol's lattice constant is only about 31 meV, at Γ, so the zero-temperature constants grow like $1/T$ and $1/T^2$ with T half that gap, and a certificate would first need the d channel of the GTH pseudopotential in Lean, extended-precision eigenvectors and about $2 \cdot 10^4$ CPU hours (floating-point preflight estimates). Both claims (O05, O06) therefore stay registered and not compared, a limit of the present checker rather than of the protocol. Until then, most results here should be read as statements about models, proved to be solved correctly, and not about materials. The certified three-dimensional results are for LDA: a floating-point study found the GTH-PBE pseudopotential compatible with the certified kernel but the Newton–Kantorovich margin about 300-fold short at S1 scale with the present proof route, and PBE certification is not pursued; the GTH-PBE pseudopotential is certified only at its own layer, in the audit of Appendix K.1.

Claims below Target. Twenty-three ledger claims have not reached their target level (Table 7). Fourteen belong to the three-dimensional DFT track of Appendix H.1: statements with paper proofs at T1 (the continuum bounds DFT-C24–C26, C34, C39, the potential-space LDA continuum theorems C40 and C45, and the band-gap corollary W41), proved statements still without a checker (M35, W22, W23, T3), the planned continuous-band output of `certify-pw` (W15, not started) and two soundness theorems of the checker mode that binds S1 to its problem written in Lean (K33, K35, T3), whose checks `checkS1` and `runS1B` have not run on production certificates; the binding of the archived certificates itself reached T4 through the one-process check `runS1U` and a composition theorem (K38). Two are the registrations of the DFT code audits, recorded at T0 with target T1 (DFT-W200, W208). Three are CALPHAD claims: the Pb–Sn eutectic certificate, whose model is matched to a Lean-stated system by a run-time comparison rather than in Lean (CAL-B33, T3, target T4), and the verdicts of the NIST solder and published-assessment audits, combined from certified checker outputs by scripts outside Lean (CAL-B35, CAL-B37, T1, target T4). The remaining four are listed below; the equality of E_∞ with the continuous ground-state energy (DFT-C02) and the force-interval checker (DFT-N04) reached theirs on 8 October 2026 (T3,

with assumption A-DFT-08 for the equivalence of the Fourier-space model with operators on H^1; and T4).

- The Frama-C-proved particle-swarm kernel of `mini-CSP` (CSP-P01, T0, target T4), not yet written; the searcher is untrusted and its outputs are certified by the checker.
- The consistency of the thread-parallel k-point implementation with its model-checked protocol (DFT-P02, T0, target T2), evidenced by ThreadSanitizer, bitwise identity across thread counts and trace validation of both the controller state and the thread events.
- The consistency of the MPI-parallel three-dimensional driver of showcase S3 with `PWParallel` and its bitwise identity across process counts (DFT-R14, R15, T0, target T2), evidenced by trace validation and hash comparison in the 25 tested layouts; the FFT transpose, the Jacobian and the refinement phases are not in the trace.

Three further claims of Phase 0 reached their target T1 with paper proofs, since the literature (Solovej, 1991; Cancès et al., 2008; 2012) treats the 3D Coulomb interaction or polynomially decaying potentials and does not cover the 1D periodic Hartree kernel directly: well-posedness of the continuous 1D rHF model at zero and finite temperature (DFT-C01), the exponential a priori plane-wave estimate for analytic potentials (D06), and local convergence of linear-mixing SCF (A01). Writing the last proof corrected the stability condition the ledger had recorded: the iteration is stable for $0 < \alpha < 2/\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of the preconditioned dielectric operator, which a numerical check confirms on a system that is stable at $\alpha = 0.28$ and diverges at 0.3. Existence of solutions of Newton's equations for the Lennard-Jones chain (MD-C03) is now machine-proved (Appendix M). The Lean reading of TDB files (CAL-I01) is at T0 by nature: it is a definition, and only cross-checks can show that it reads a database as intended.

Scope of the First Realisation. The four original domains are mostly one-dimensional: FEM reaches two-dimensional Poisson problems and membrane eigenvalues, and MD a three-dimensional periodic box, both only in the first cases described in Appendices L and M; `mini-CSP` treats three-dimensional clusters, but only with the Lennard-Jones potential and only certifies local minima. `mini-DFT` certifies LDA (Slater exchange, PW92 correlation) at finite temperature and, under a gap, at zero temperature, and band structures at the Γ point and with k-points, except for zero-temperature LDA with k-points and for k-point bands of the named Gaussian potential. Its spin polarisation is collinear LSDA with Slater exchange or PW92 spin correlation, at the Γ point and with k-points, at finite and zero temperature, with energies and (at finite temperature) bands, but without zero-temperature bands or the Gaussian potential with k-points; exact exchange is closed-shell Hartree–Fock at finite temperature at the Γ point, without an energy interval; forces are certified only on Gaussian sites for rHF at the Γ point; it has no pseudopotentials or MPI parallelism. The three-dimensional checkers (Appendix K) bind fcc cells only, cover rHF and LDA with local and GTH pseudopotentials, bound Ewald sums and total energies, and have no forces, GGA or MPI. `mini-FEM` lacks variable coefficients in boundary-value problems, higher-order Lagrange elements, continuous eigenvalues with Neumann or Robin ends, certified two-dimensional bounds for non-polynomial loads, eigenvalues beyond the first in two dimensions, elasticity and three-dimensional meshes. `mini-MD` lacks neighbour lists, non-cubic cells, other potentials, thermostats and constraints, and its three-dimensional tubes cover about half a vibration period. `mini-CALPHAD` lacks order–disorder and multicomponent models, sublattice phases with three or more mixing sublattices or reciprocal parameters, magnetic phases with two mixing sublattices, pressure-dependent databases, and phases whose endmembers a binary database does not give (93 phase–system pairs in the audit, mostly Laves and σ phases whose endmembers appear only in ternaries); its convexity certificates cover non-magnetic substitutional solutions at single temperatures, its phase diagrams are certified complete at each grid temperature but not between temperatures, where only fixed compositions, the absence of a phase and binary eutectics and eutectoids are certified, and its thermodynamic properties are those of single phases at fixed composition (fixed internal state), not of equilibria and not at the Curie temperature. The CALPHAD solver audits compare pycalphad and a development build of OpenCalphad only: no solver

with global minimisation (Thermo-Calc, Pandat, FactSage) was compared. The largest non-global pycalphad equilibria, single-phase HCP in Gd–Ti and single-phase DHCP in Nd–Ti, at least 1.02 and 1.23×10^4 J/mol above certified two-phase states (Appendix N), remain undiagnosed; whether they come from composition sets rather than from coarse sampling is not established. The domain ledgers list these gaps item by item, and the parity tables order them. Of the TLA$^+$ modules planned for Phase 2 (halo exchange, atom migration, parallel assembly, checkpointing), none exists yet; Appendix V shows the intended style.

Trusted Base and Cost. The Lean compiler and runtime and the certificate parsers are trusted, the C compiler only for the C interval kernel, on which no checker output depends (Appendix T). The problem data in a certificate are written by an untrusted driver. For the two largest pieces the checkers now rebuild the problem themselves: the DFT external potential is compared with its Lean definition, and the CALPHAD model is built from the TDB file (Appendix I.2). What is trusted instead is that these Lean definitions, `ExtPot` and the roughly 930 lines of `modelOfTdb`, say what is intended (A-DFT-05, A-CAL-01). For S1 the registered problem is itself a Lean definition, stated with Mathlib alone, which a checker mode compares with the certificate exactly; what is trusted there is that this specification says what the protocol registered; the archived silicon certificates are bound to it, the SCF and one-sided zone certificates in one process and the two-sided result by a composition theorem whose glue (same file bytes, same executable, comparison of printed numbers) lies outside Lean (Appendix K). The remaining data are still taken from the certificate: the other DFT parameters are echoed exactly but can only be compared with the command line, Gaussian-potential zero-temperature LDA certificates and energy components carry only a deviation bound and Gaussian-potential k-point bands have no certificate, CALPHAD certificates checked without the TDB file (for instance when it is piped in) certify the model written in them, and FEM boundary data and right-hand sides and MD parameters are echoed exactly, in the two- and three-dimensional checkers too, but not recomputed. They are backstopped by cross-checks, by the drivers' input validation (Appendix I.2) and by known-answer tests whose positive controls show exactly what goes undetected otherwise (Appendix P). Exact rational checking becomes slow for large certificates. Supercell certificates scale roughly as the fourth power of the supercell dimension (about 8 minutes at dimension 85; under a minute for the k-point LDA certificates at $K = 4$), and MD checking takes about 50 ms per step for $N = 8$. Two-dimensional branch and bound for CALPHAD phases with two mixing sublattices took tens of minutes per certificate with a fixed splitting target; a running upper bound and the ellipse windows bring it to 1–4 s, and a box budget cuts degenerate certificates off with wider intervals. A third mixing sublattice or a ternary system adds a dimension. In three dimensions two remedies are in place, the exploitation of Bloch structure (time reversal and point groups) and exact arithmetic modulo machine-word primes, proved to compute the same numbers as the big-integer route: the S1 certificate for silicon, 253 MB, checks in 196 CPU hours and needs 93 GB of memory, so memory, not time, is the next limit. The two-sided zone checks took 939 (Si), 705 (AlP) and 1 842 CPU hours (GaP), the last two after integer and memoised versions of the lower-bound leaf test, proved equal to the original functions, made that step 14–69 times faster (W260–W267).

Open Research Questions. Several questions remain open at the level of the programme itself.

- *Comparing with experiment.* Measured quantities and model quantities rarely mean the same thing: band gaps need zero-point and temperature corrections, lattice constants zero-point expansion, and the corrections themselves come from other models or other measurements. How such corrections enter a V grade without contaminating the rigorous part of (3), and how to combine a rigorous interval with a statistical uncertainty in one statement, are open; so is how to keep fitting and validation sets apart when published functionals and databases do not always say which data they were fitted to.
- *Validation at scale.* The observation decomposition needs the computed interval to be much narrower than the deviation it bounds. For silicon at the S1 settings it is (the compared whole-zone interval is 0.071 eV wide against a deviation of

about 0.8 eV), but it is wider than the measurement uncertainty and than the meV-level uncertainty of a converged floating-point gap, so a deviation of the size of a discretisation error could not be resolved; narrower intervals need tighter margins in the driver and more octree leaves. The limit is thus cost: carbon at its protocol cutoff and germanium with its nearly closed LDA gap are out of reach of the present checker; lattice constants have a first certified bracket for the discrete model at a fixed basis, which still assumes that the end-point solutions are ground states (Appendix K), and magnetic moments need spin in three dimensions.

- *Sobolev spaces.* How deep should Sobolev theory be formalised, and which theorems come first? The first 2D results still avoid Sobolev spaces, by dual-norm statements against smooth test functions and Rayleigh quotients of Lipschitz functions (Appendix L), but the step from a dual bound to $\|\nabla(u - u_h)\|$, local constants, elasticity and mixed problems will need them, and the choice of what to assume and what to prove will shape G2 and much of G1's d-dimensional step.
- *Chaos and long times.* Pointwise trajectory bounds for chaotic many-body dynamics necessarily grow exponentially. For a near-integrable chain, a one-sided Lipschitz bound in the energy norm replaces the Lipschitz constant and carries certified mode energies of the FPUT chain to about 12% of its recurrence time (Appendix M), but the growth is still exponential. Long-time MD guarantees must take another form: shadowing, modified-Hamiltonian backward error analysis (Hairer et al., 2006), of which the exactly conserved modified energy of Verlet for the harmonic oscillator is the simplest case, or statements about statistics.
- *Randomness.* Langevin dynamics and random initial guesses involve probabilistic properties. TLA$^+$ describes nondeterminism but not probability, so another formalism is needed.
- *GPUs.* How asynchronous streams and weak memory models enter a TLA$^+$ specification is open.
- *Nonconvex DFT.* With exchange–correlation, the ground state is no longer characterised by convexity. Which certificates remain possible, e.g. local uniqueness plus a posteriori bounds, and what "ground state" should mean in a certificate are open questions. The non-variational smearing and LDA certificates of Appendix J are first answers, both giving uniqueness only near the computed solution; local uniqueness nevertheless proved strong enough to transfer symmetry, since it forces the k-point LDA solution to be cell-periodic, and disjoint balls of local uniqueness give certified lower bounds on the number of solutions (Appendix Q). Certifying which solution has the lowest free energy among *all* solutions, rather than among those found, remains open.

T Trusted Computing Base

U An Excerpt of a Trust Report

The following excerpt (numbers elided) is printed by `mini-dft --certify run.cert --verify`. Each line cites the ledger claim that justifies it, and lines prefixed with `Lean` come from the compiled checker; the `PROBLEM` and `POT_OK` lines bind the certificate to the potential named on the command line (Appendix J).

```
params  L = 4  N = 2  K = 16  M = 65  tol = 1e-11  pot = gauss  V0 = 2  replicas = 1
status  converged (iteration <n>)                          DFT-A02
grid    M = 65 >= 4K+1 = 65, no aliasing in exact arithmetic DFT-D04
charge  int rho in [<lo>, <hi>], N = 2 (interval arithmetic) DFT-D07, DFT-F01
eigen   2 eigenpairs per k-point, residual <= <r>           DFT-A03, DFT-F02
gap     no eigenvalue in [<a>, <b>]                         DFT-A04, DFT-F02
assume  A-DFT-01 1D rHF well-posed; A-DFT-02 LAPACK output is a candidate;
        A-DFT-03 analytic potential; A-DFT-04 gap
cert    written to run.cert                                 DFT-B01..B08
Lean    PROBLEM POT gauss V0 2 SIGMA 0.5 R 1   ... and L, K, NPAIR, KPOINTS
```

```
Lean    POT_OK gauss <l1 distance of vhat from the named potential>
Lean    CERT_OK  <E_K lo> <E_K hi>
Lean    GAUSS_OK <E_K lo> <E_K hi>   for the named Gaussian potential
Lean    LIMIT_OK <E_inf lo> <E_inf hi> <E_K - E_inf upper bound>
Lean    RHO_OK   <Coulomb-norm density error>
Lean    EIG_OK 0 <lo> <hi>   ... one line per eigenvalue, plus GAP_OK
```

V A Planned Software-Layer Specification: Atom Migration

The following TLA$^+$ module illustrates the software-layer obligations planned for Phase 2 (Table 4, MD, parallel row). Under spatial domain decomposition, an atom that crosses a subdomain boundary migrates from one rank to another. The safety property states that every atom is either owned by exactly one rank or is in exactly one in-flight message. The liveness property states that every sent atom is eventually received. The module is a starting point for TLC and is not yet in the repository.

```
---------------------------- MODULE AtomMigration ----------------------------
EXTENDS Naturals, FiniteSets
CONSTANTS Atoms, Ranks, InFlight       \* InFlight is a model value, not a rank
VARIABLES owner, msgs                  \* owner \in [Atoms -> Ranks \cup {InFlight}]
vars == <<owner, msgs>>
Init == /\ owner \in [Atoms -> Ranks]
        /\ msgs = {}
Send(a, src, dst) ==                   \* atom a leaves the subdomain of src
  /\ owner[a] = src /\ dst # src
  /\ owner' = [owner EXCEPT ![a] = InFlight]
  /\ msgs'  = msgs \cup {<<a, dst>>}
Recv(a, dst) ==                        \* dst receives atom a
  /\ <<a, dst>> \in msgs
  /\ owner' = [owner EXCEPT ![a] = dst]
  /\ msgs'  = msgs \ {<<a, dst>>}
Next == \/ \E a \in Atoms, s, d \in Ranks : Send(a, s, d)
        \/ \E a \in Atoms, d \in Ranks    : Recv(a, d)
Spec == Init /\ [][Next]_vars
             /\ \A a \in Atoms, d \in Ranks : WF_vars(Recv(a, d))
AtomsConserved ==                      \* safety
  \A a \in Atoms :
    LET inflight == {m \in msgs : m[1] = a} IN
      \/ owner[a] \in Ranks  /\ inflight = {}
      \/ owner[a] = InFlight /\ Cardinality(inflight) = 1
Delivered ==                           \* liveness
  \A a \in Atoms, d \in Ranks : (<<a, d>> \in msgs) ~> (owner[a] = d)
THEOREM Spec => []AtomsConserved /\ Delivered
=============================================================================
```

W Parity Milestones in Detail

G1 (DFTK). *M0, existing:* kinetic, external (Fourier coefficients) and Hartree terms; Fermi–Dirac smearing and entropy; Γ-centred and Monkhorst–Pack k-grids; dense diagonalisation; linear, Kerker and Anderson mixing; non-self-consistent bands and DOS (certified at Γ and, on the primitive cell, with k-points; the zone-boundary point not compared); supercells; thread-parallel k-points. *M1, 1D solvers, occupations and k-points:* Gaussian, Methfessel–Paxton and Marzari–Vanderbilt smearing; Monkhorst–Pack shifts and time-reversal reduction; grid-based $H\psi$ with LOBPCG and preconditioning; adaptive bands and diagonalisation tolerance; LDOS, dielectric and χ_0 mixing, potential mixing; Newton's method and direct minimisation. *M2, 1D physics terms:* LDA (aligned with DFTK; Slater exchange and PW92 correlation certified at Γ and with k-points for all four smearings and, under a gap, at zero temperature; done); collinear spin (done: spin-polarised Slater exchange or PW92 correlation in a uniform Zeeman field, certified at Γ and with k-points, at finite and zero temperature, with magnetisation, energies, bands and the Gaussian potential) and

magnetic terms (DFTK's vector potential, not done); local nonlinearity (Gross–Pitaevskii); exact exchange (partial: closed-shell Hartree–Fock SCF solutions at finite temperature at Γ, DFTK fixed-point check; no energy interval yet); atoms with local and nonlocal pseudopotential parts, Ewald and pseudopotential corrections; pair potentials; DFT+U. *M3, 1D post-processing and response:* forces (partial: certified on Gaussian sites for rHF at Γ, zero and finite temperature), geometry optimisation, stresses; χ_0, Hessian and DFPT; polarisability, phonons, elastic constants; a posteriori error estimation and refinement; currents. *M4, d dimensions:* a d-dimensional Lean lattice model (done); 2D/3D plane waves with FFTW (3D done without FFTW, 2D open); arbitrary lattices and symmetry (fcc with time reversal and point groups); HGH and UPF pseudopotentials (GTH s and p channels); DFTK's examples (silicon and Cohen–Bergstresser done, graphene and the GaAs surface open); Ewald sums and anyons (open). *M5, full 3D:* GGA and meta-GGA; 3D forces, stresses, phonons and elastic constants; ACE exact exchange; MPI k-point parallelism with bitwise identity on 1–64 processes. *M6, infrastructure:* GPU; arbitrary floating-point types and automatic differentiation; Wannier90, JLD2, VTK, JSON and plotting.

G2 (scikit-fem). *F0, 1D Poisson with P1 (complete):* meshes with arbitrary rational nodes; assembly; nonhomogeneous Dirichlet conditions; polynomial loads; untrusted solves; true nodal values and pointwise error bounds; Céa's lemma. *F1, 1D extensions:* Neumann and Robin conditions (also Neumann at both ends), non-polynomial loads, cubic Hermite beams, adaptive refinement and energy-norm errors (done); the eigenvalue problem $-u'' = \lambda u$ and the Dirichlet Sturm–Liouville problem (done: discrete eigenvalues certified by Sylvester inertia, continuous eigenvalues bounded from both sides; Neumann and Robin ends open); variable coefficients in boundary-value problems, P2 and higher-order elements, and a one-to-one comparison with scikit-fem's 1D examples (open). *F2, 2D triangles:* meshes with checker-verified validity and guaranteed upper bounds by equilibrated fluxes (done, polynomial loads); P1 triangles (aligned); the first Dirichlet eigenvalue of rectangles and L-shaped domains (done); the H^1_0 model (avoided so far by dual-norm statements); P2 triangles and quadrilaterals (open). *F3, more physics and 3D:* linear elasticity (Korn's inequality), Stokes and mixed elements (inf–sup), adaptive refinement in 2D, tetrahedral and hexahedral meshes, parallel assembly (TLA$^+$).

G3 (pycalphad). *C0, binary point equilibrium (complete):* TDB parsing (also in Lean, so that the checker builds the model from the file); substitutional solutions with Redlich–Kister excess; stoichiometric compounds; equilibrium phases, compositions and amounts; equilibrium Gibbs energy; chemical potentials; miscibility gaps; existence of equilibria. *C1, binary extensions:* phase diagrams on T–x grids, completeness of isothermal sections, Inden–Hillert–Jarl magnetism, sharper single-phase chemical potentials and driving forces, G, S, H, C_p of each phase at fixed composition, equilibria over temperature intervals, interstitial solutions and magnetic sublattice phases and compounds (done); invariant reaction temperatures (aligned: binary eutectics and eutectoids certified, peritectics open); H, S, C_p at equilibrium level and pressure-dependent databases (open). *C2, multicomponent and sublattices:* sublattice phases of the compound energy formalism with one or two mixing sublattices, conditions on chemical potentials and on activities (done); ternary and higher solutions (Muggianu extrapolation); order–disorder. *C3, further computations:* Scheil solidification with bounded error propagation; parameter fitting, in which the fitting itself stays untrusted and only the final database is certified; ionic-liquid and quasichemical models.

G4 (ASE). *M0, one-dimensional Lennard-Jones chain (complete):* all pairs without cutoff, velocity Verlet, local and global bounds for the true solution, energy and momentum. *M1, three-dimensional periodic box:* cubic cell with unwrapped coordinates, minimum image, velocity Verlet, local and global bounds for the true solution, existence and uniqueness, exact energy and total momentum (done); the C^4 switched cutoff (aligned, since ASE's own switch is only C^1). *M2:* neighbour lists, non-cubic and partly periodic cells, other pair and many-body potentials, thermostats and barostats, constraints, statistics such as temperature and radial distribution functions, and long-time statements (open).

Showcase	Foremost	Content and status
S1 Certified deviation from experiment	Physical	LDA band gaps and lattice constants of a set of semiconductors and insulators, compared systematically with experiment. The computed values are rigorous intervals (discrete and, where possible, continuum), so the deviation contains no algebraic, rounding or software error: with continuum intervals it would be the error of the model (functional, pseudopotential, Kohn–Sham gap for the quasiparticle gap), with discrete ones, as reached, that of the discretised model; a quantitative version of the band-gap problem (Perdew and Levy, 1983; Sham and Schlüter, 1983). Magnetic moments of magnetic metals likewise. *Status:* protocol for Si, C and Ge frozen; Si certified at the protocol's settings, with the one-sided zone-wide gap bound (at most 0.530902 eV) and floating-point convergence of the gap at these settings to 2.4 meV; compared with experiment (one-sided, then a two-sided supplement: gap in $[0.4146, 0.4856]$ eV, deviation $[0.741, 0.847]$ eV), total energy of the certified solution bounded to $8 \cdot 10^{-8}$ Ha, and every energy minimum of the fixed-basis discrete model bracketed in lattice constant to 0.004 Å if the end-point solutions are ground states (not compared with experiment); AlP and GaP (pre-registered) certified at their protocols' settings, gaps at most 1.478051 and 1.585748 eV, compared one-sided; C and Ge not pursued (Appendix S), GaAs and AlAs registered but not computed (cost).
S2 Not the ground state	Mathematical	A machine-checked proof that a converged self-consistent solution is not the ground state of the discrete model: the higher solution is certified (existence, uniqueness, a free-energy lower bound) and an admissible trial state has a strictly lower free-energy upper bound; floating-point computation cannot give such a proof. A converged census found the orderings in all candidates already known (the magnetic orders of γ-Fe, AFM-II NiO, symmetry breaking in stretched H_2), so S2 is a first in rigour, not a physical discovery. First example, done: simple-cubic hydrogen with a local pseudopotential (certified margin 330 meV); real-material example: NiO. The one-dimensional precedent is the certified multiplicity of LDA solutions in Appendix Q.
S3 Parallel correctness and bitwise reproducibility	Software	k-point, band and plane-wave parallelism and the distributed FFT transpose in 3D written as TLA$^+$ specifications and model-checked, run logs passed through trace validation; goal: bitwise-identical results independent of the process count; planted parallel bugs caught at the specification level as positive controls. *Status:* the protocol layer is model-checked and catches four planted bugs; an MPI driver of the S1 problem class implements it: on a small Si case its output is bitwise identical and its logs pass trace validation in all 25 tested layouts with between 1 and 32 processes, and at the S1 setting in six layouts; four bugs planted in the code are caught, and its S1 certificate is accepted by the Lean checker (Appendix I.2). The parallel evidence is tests only (T0): the FFT transpose, Jacobian and refinement phases are outside the trace, and bitwise identity is tested on one machine and library build.

Table 8: The three-dimensional showcases and their status. Of two optional ones, the validation of the empirical Cohen–Bergstresser pseudopotentials outside the optical data they were fitted to (physical) is done for twelve materials (Appendix Q.1), and short certified Born–Oppenheimer molecular dynamics (mathematical) has been set aside.

Phase	Scope	Exit criterion	Status
0. Minimal DFT	1D periodic rHF, plane waves, serial; build ledger, checkers, reports, trace validation	Every obligation reaches its target level; four benchmark groups pass; every run prints a trust report	Exit criteria met (literature items at T1; interval kernel proved on the C code)
1. DFT extensions	1D LDA; k-point sampling and its MPI parallelisation (first real parallel protocol); 3D plane waves with local pseudopotentials vs. DFTK	k-point parallel runs bitwise identical on 1 to 64 processes; every run log passes trace validation	Partly ahead of plan; LDA (Slater, PW92) certified (Γ and k-points, all four smearings); 3D plane waves with GTH pseudopotentials certified for Si (Appendix K); an MPI driver of the 3D point-group LDA problem is bitwise identical in 25 tested layouts with between 1 and 32 processes and passes trace validation (showcase S3, Appendix I.2; exit criterion 1–64 not yet met)
2. MD and FEM	1D Lennard-Jones chain with velocity Verlet; 1D Poisson with P1; TLA$^+$ modules for halo exchange, atom migration, parallel assembly, checkpointing	Both minimal versions print trust reports; Céa's lemma and Verlet symplecticity reach T3	Exit criteria met; 2D Poisson and 3D periodic MD certified; TLA$^+$ modules not started
3. Trusted kernels	Bring a few kernels to T4: plane-wave kinetic operator, FEM element-stiffness assembly, one Verlet step	Their floating-point error bounds are machine-checked against the code and enter the global error budget	Not started; starting point: the interval kernel
4. Benchmarks	Δ-gauge for DFT; NVE energy drift and radial distribution functions for MD; manufactured solutions for FEM	Published benchmark results, each with its trust ledger	Not started

Table 9: The roadmap phases, their exit criteria and their status at the time of writing.

Goal	Milestones	Current milestone	Done
G1: DFTK.jl 0.8.0	M0 existing; M1 1D solvers, occupations and k-points; M2 1D physics terms (LDA, spin, pseudopotentials); M3 1D post-processing and response; M4 d-dimensional Lean model, 2D/3D plane waves; M5 full 3D (GGA, forces, MPI); M6 infrastructure (GPU, AD)	M1: Gaussian, MV and MP smearing of any order done; MP grids, shifts, time reversal done. M2: LDA done, Slater exchange and PW92 correlation certified at Γ and with k-points for FD, Gaussian, MV and MP and, under a gap, at zero temperature; collinear spin done (Slater exchange and PW92, Zeeman field, Γ and k-points, finite and zero temperature); exact exchange partial (closed-shell HF, finite temperature, Γ). M3: forces partial (Gaussian sites, rHF, Γ). M4: d-dimensional Lean model done; 3D plane waves, fcc symmetry (time reversal, point groups), GTH pseudopotentials and the silicon and Cohen–Bergstresser examples partial (Appendix K)	10/46 (+2 aligned)
G2: scikit-fem 12.0.2	F0 1D Poisson P1; F1 1D extensions (Robin/Neumann, non-polynomial loads, Hermite beams, adaptivity, P2, eigenproblems); F2 2D triangles with an H^1 model and guaranteed a posteriori bounds; F3 elasticity, Stokes, 3D, parallel assembly	F0 complete; F1: pure Neumann, energy-norm errors and two-sided bounds for continuous Dirichlet Sturm–Liouville eigenvalues done; F2: P1 triangles with guaranteed Prager–Synge bounds (polynomial loads) and the first Dirichlet eigenvalue of rectangles and L-shaped domains done	17/27 (+1 aligned)
G3: pycalphad 0.11.1	C0 binary point equilibrium; C1 binary phase diagrams, invariant reactions, magnetism, thermodynamic properties; C2 multicomponent and sublattice models; C3 Scheil solidification, parameter fitting	C0 complete; C1: phase diagrams with certified complete isothermal sections, magnetism, driving forces, single-phase chemical potentials, the thermodynamic properties G, S, H, C_p of each phase, equilibria over temperature intervals, interstitial solutions and magnetic sublattice phases done, binary eutectics aligned; C2: sublattice phases with one or two mixing sublattices, chemical-potential and activity conditions done	21/28 (+1 aligned)
G4: ASE 3.29.0	M0 one-dimensional Lennard-Jones chain; M1 three-dimensional periodic box; M2 neighbour lists, general cells, other potentials, thermostats, constraints, statistics	M0 complete; M1: cubic periodic box, minimum image, velocity Verlet, true-trajectory bounds, existence and uniqueness, energy and momentum done, C^4 switched cutoff aligned	10/18 (+1 aligned)

Table 10: The four parity goals. "Done" counts features that are implemented, agree with the reference and are fully certified. Appendix W gives the milestone contents in more detail.

Reference	Cases	Agreement and containment
DFTK 0.8.0	Γ, k-points, FD/Gaussian/MP/MV, bands, DOS; LDA (six systems, plus two certified k-point systems); collinear LSDA (four systems, Slater exchange and PW92); forces (six); Hartree–Fock (fixed-point check, two systems)	Free energies $\leq 2.8 \cdot 10^{-14}$ (zero-temperature and FD energies $\leq 5 \cdot 10^{-15}$), eigenvalues $\leq 1.0 \cdot 10^{-12}$, bands $\leq 2.8 \cdot 10^{-12}$; LDA energies, components and eigenvalues agree; DFTK's ε_F, E and F lie inside the Lean intervals, LDA with k-points included, and so do its bands and DOS for finite-temperature SCF solutions (rHF and LDA) and, on the primitive cell, for k-point SCF solutions (rHF, LDA, zero-temperature rHF); LSDA ν, ε_F, E, F, bands and DOS inside the Lean intervals (ν and $\varepsilon_F \leq 8.2 \cdot 10^{-14}$ from the driver), zero-temperature LSDA ν equal and ε_F in the Lean gap, with k-points included, also with PW92; forces inside the `FORCE_OK` intervals; HF Fock matrix $\leq 1.7 \cdot 10^{-14}$ from DFTK's Hamiltonian at the certified centre
DFTK 0.8.0 (3D)	Cohen–Bergstresser bands (Si, Ge, GaAs); rHF Si model (Γ, up to 411 plane waves; k meshes up to $8 \times 8 \times 8$); real Si with GTH-LDA q4, rHF and LDA, Γ and meshes up to $8 \times 8 \times 8$, full mesh with explicit k points	DFTK's SCF solutions $2 \cdot 10^{-14}$–$5 \cdot 10^{-12}$ from the certificate centres, inside the radii; every DFTK eigenvalue compared lies in its certified interval (8192 of 8192 on the $8 \times 8 \times 8$ mesh, 7728 of them at k points the checker reaches only through symmetry); empty-lattice levels inside the Cohen–Bergstresser intervals
scikit-fem 12.0.2	11 P1 cases, 9 pure Neumann cases, 6 beam cases; energy-norm certificates; 2D P1 Poisson; Q1 first eigenvalues (unit square, L-shaped domain)	Nodal values $\leq 3 \cdot 10^{-13}$ (P1), $< 10^{-12}$ after removing the constant (pure Neumann) and $\leq 4 \cdot 10^{-12}$ (beam), within the certified `NODE_ERR`; scikit-fem's nodal values pass the energy-norm check with the same η; 2D nodal values $\leq 3 \cdot 10^{-16}$; discrete Q1 eigenvalues within $2.1 \cdot 10^{-14}$ relative of the checker's exact Rayleigh quotients
ASE 3.29.0	3 chains, up to 2000 steps; 3D periodic box (our potential as an ASE calculator; ASE's unswitched Lennard-Jones where no pair is in the switching region)	Positions $\leq 3.6 \cdot 10^{-15}$, momenta $\leq 5.0 \cdot 10^{-14}$, energies $\leq 3.2 \cdot 10^{-14}$; 3D Verlet positions $\leq 4.4 \cdot 10^{-16}$, forces $\leq 2.8 \cdot 10^{-14}$
pycalphad 0.11.1	29 equilibria in Al–Zn, Pb–Sn, Cu–Mg, Cr–Fe, Al–Fe with 141 driving forces; 15 with extreme dilution, a hull compound and sublattice phases (COST 507 Al–Y, Al–Cu, Cu–Mg; Al–Mg); 315 phase-diagram certificates in seven diagrams, with partition certificates at 67 temperatures; 16 chemical-potential and 17 activity conditions; G, S, H, C_p of single phases at 713 points in 21 systems; with 0.11.2, the Pb–Sn eutectic and the Fe–C eutectoid (pressure-free limit) and six equilibria with interstitial and magnetic sublattice phases	Compositions $< 10^{-6}$, $G_M < 10^{-5}$ J/mol; pycalphad's G_M, μ, phase compositions and driving forces lie inside the Lean intervals and windows (local driving-force maxima excepted, see text); internal-state energies of phases with two mixing sublattices agree to $1.4 \cdot 10^{-13}$ relative; `GM`, `SM`, `HM`, `CPM` lie within $3 \cdot 10^{-14}$ relative of the Lean intervals ($3.3 \cdot 10^{-12}$ for magnetic C_p); pycalphad's eutectic and eutectoid temperatures inside the certified intervals; G_M, μ and phase compositions of the six equilibria inside

Table 13: Cross-checks on identical discrete problems or databases.

Domain	Certified quantity	Width / bound	Check time
DFT	Ground-state energy E_K	$6 \cdot 10^{-10}$–$9 \cdot 10^{-9}$	0.8 s ($K = 16$)
DFT	Truncation error $E_K - E_\infty$	$5.9 \cdot 10^{-10}$ ($K = 8$)	–
DFT	Free energy $F_{T,K}$ (FD)	$4 \cdot 10^{-11}$–$2 \cdot 10^{-10}$	2 s ($K = 16$)
DFT	MV/MP SCF solution radius; free energy	$r \approx 10^{-12}$; $4 \cdot 10^{-14}$–$5 \cdot 10^{-13}$	4–8 s ($K = 8$)
DFT	All 33 eigenvalues of H^* (half-width)	1.9–$2.8 \cdot 10^{-4}$	5 s ($K = 16$)
DFT	LDA SCF solution radius	$r \approx 10^{-12}$–$8 \cdot 10^{-12}$	6–14 s ($K = 8$)
DFT	LDA energy; free energy	$3 \cdot 10^{-9}$–$2 \cdot 10^{-8}$; 10^{-15}–10^{-12}	–
DFT	k-point LDA radius; per-cell E; F	2–$9 \cdot 10^{-12}$; 10^{-9}–$2 \cdot 10^{-8}$; $3 \cdot 10^{-15}$–$8 \cdot 10^{-14}$	9–53 s ($K = 4$, $n_k = 2$–4)
DFT	Finite-T bands (half-width), rHF MV; LDA FD	$\sim 10^{-11}$; $\sim 10^{-9}$	–
DFT	k-point bands (half-width), rHF MV; LDA FD; rHF at $T = 0$	$\leq 5.6 \cdot 10^{-12}$; $\leq 3.9 \cdot 10^{-11}$; $\leq 3.5 \cdot 10^{-4}$	9 s; 50 s; 2 s ($n_k = 3$)
DFT	Collinear LSDA radius (intervals for ν, μ)	$3.3 \cdot 10^{-14}$–$9.9 \cdot 10^{-13}$	2.7–6.1 s ($K = 6$–8)
DFT	k-point LSDA, per-cell ν width	$1.6 \cdot 10^{-13}$; $5.3 \cdot 10^{-13}$	20 s; 49 s ($n_k = 3$; 4)
DFT	Zero-T k-point LSDA radius	$5.5 \cdot 10^{-15}$–$6.4 \cdot 10^{-14}$	6–19 s ($n_k = 2$, 3)
DFT	LSDA bands; DOS (half-width)	$1.6 \cdot 10^{-11}$–$3.0 \cdot 10^{-10}$; $\leq 4 \cdot 10^{-8}$	13 s (Γ); 65 s ($n_k = 3$)
DFT	LSDA energy; free energy	$1.5 \cdot 10^{-10}$–$1.0 \cdot 10^{-8}$; $4 \cdot 10^{-15}$–$6.6 \cdot 10^{-14}$	8–55 s
DFT	LSDA with PW92 correlation, radius	$3.5 \cdot 10^{-14}$–$4.3 \cdot 10^{-13}$	10–22 s; 61 s ($n_k = 3$)
DFT	PW92 LSDA energy; free energy	$1.3 \cdot 10^{-8}$–$5.9 \cdot 10^{-7}$; $3.6 \cdot 10^{-14}$–$1.6 \cdot 10^{-9}$	24–116 s
DFT	Site forces (half-width), $T = 0$; $T = 0.05$	$1.8 \cdot 10^{-4}$; 4.5–$4.7 \cdot 10^{-5}$	2.4–3.8 s ($K = 16$)
DFT	Hartree–Fock radius	$4.5 \cdot 10^{-15}$–$9.3 \cdot 10^{-14}$	0.4–3.1 s ($K = 4$–6)
DFT	Quartic oscillator: $\lambda_{\min}(H_{40})$; continuum E_0	$8.7 \cdot 10^{-19}$; $1.1 \cdot 10^{-16}$	about 1 s each
3D DFT	Cohen–Bergstresser band energies (Si, Ge, GaAs)	about $4 \cdot 10^{-9}$ (10^{-7} eV)	6–14 s per k point
3D DFT	real Si (GTH), rHF, Γ, $n = 411$: radius; band widths	$6.2 \cdot 10^{-9}$; $1.3 \cdot 10^{-7}$	1.7 h
3D DFT	S2, stretched hydrogen: certified energy margin	330 meV	–
3D DFT	S1, real Si KS-LDA, $8 \times 8 \times 8$, $n \approx 410$: radius; band widths; mesh gap	$2.1 \cdot 10^{-10}$; $\leq 8.6 \cdot 10^{-9}$; $1.7 \cdot 10^{-8}$	7.9 h (32 cores, 196 CPU h)

Table 14: Certificate tightness and cost in DFT. Widths are in Hartree (the oscillator is dimensionless); the first block is one-dimensional, the second three-dimensional. FEM, MD and CALPHAD follow in Table 15.

Domain	Certified quantity	Width / bound	Check time
FEM	True solution at nodes, P1	$\leq 2 \cdot 10^{-15}$	1.4 s ($N = 2000$, polynomial f)
FEM	sup $\lvert u - V_h \rvert$ vs. sampled error	ratio 0.87–1.00	–
FEM	Energy-norm bound minus true error ($-u'' = e^x$, 16 elements)	$6 \cdot 10^{-19}$	–
FEM	Sturm–Liouville λ_1 ($-u''$, 512 elements), relative width	$3.8 \cdot 10^{-6}$	–
FEM	2D Prager–Synge η vs. true error (unit square)	ratio about 1.2	4.5 s (8192 triangles)
FEM	2D λ_1, unit square; L-shaped (128^2), relative width	$5.0 \cdot 10^{-5}$; $1.3 \cdot 10^{-3}$	–
MD	Local trajectory error per step	$2.8 \cdot 10^{-9}$, ratio ≤ 0.99	50 s (1000 steps, $N = 8$)
MD	3D periodic box ($N = 32$): local bound; tube length	$1.5 \cdot 10^{-7}$; $t = 0.148$ (74 of 100 steps)	46 s (16 cores)
MD	Pendulum period ($k = 1$, $\theta_0 = 1$)	$1.2 \cdot 10^{-8}$	about 4 s
MD	FPUT mode-energy envelopes (31 masses)	5 modes to $t = 200$; 3 modes to $t = 1272.875$	about 13 min ($t = 200$, one core)
CALPHAD	Chemical potential, two-phase / one-phase	$\sim 10^{-9}$ / 10^{-8}–10^{-7} J/mol	0.1–4 s per point
CALPHAD	Phase boundary (window width)	$\leq 2 \cdot 10^{-6}$	315 certificates, 7 diagrams
CALPHAD	Unresolved interval at a phase boundary (partition)	$\leq 5.4 \cdot 10^{-8}$	67 temperatures, 7 systems
CALPHAD	μ_A at prescribed μ_B	$\sim 2 \cdot 10^{-9}$ J/mol	16 points
CALPHAD	G, S, H, C_p of a phase at fixed composition	$\sim 10^{-25}$ (exact rational)	0.14–0.74 s per certificate
CALPHAD	Onset of a miscibility gap (convexity and point certificates)	1 K bracket	0.1–0.2 s per convexity certificate
CALPHAD	Eutectic (Pb–Sn), eutectoid (Fe–C) temperature	0.02 K	about 70 s (Pb–Sn, driver and check)

Table 15: Certificate tightness and cost in FEM, MD and CALPHAD. Widths are in solution units (FEM, MD) or as stated. Ratios compare a high-accuracy reference error to the certified bound (1 = sharp). FEM nodal widths are limited by decimal output.

Database	System	No gap	Gap	Liquid at top grid T
COST 507	Fe–Si	2177 K	2178 K	3500 K: 0.548 / 0.994
COST 507	Mg–Si	2554 K	2555 K	2900 K: 0.568 / 0.866
COST 507	Cu–Si	2603 K	2604 K	3100 K: 0.029 / 0.388
COST 507	Al–Ce	2758 K	2759 K	2800 K: 0.708 / 0.836
Dupin (2001)	Al–Ni	5053 K	5054 K	5900 K: 0.286 / 0.664

Table 17: Anomalous liquid miscibility gaps that open on heating, from the certified audit. At the "No gap" temperature a convexity certificate proves the liquid convex; at the "Gap" temperature a point certificate of the liquid alone has windows that exclude x_0; whether the liquid changes only once in between is not certified. The last column gives the two certified liquid compositions at the highest grid temperature. Dupin (2001) is the Al–Ni database `alni_dupin_2001` shipped with pycalphad.

Component	Version	Role
Lean 4 kernel	v4.33.1	Checks all Lean proofs
Mathlib	v4.33.1	Checked by the kernel; only `propext`, `Classical.choice`, `Quot.sound` are trusted additionally
Lean compiler	v4.33.1	Compiles the eleven checker executables; their parsers (including reading a TDB file byte by byte; some, such as those of the zone checkers, are `partial` definitions, opaque to the proofs) and decimal printers are also trusted
Lean runtime	v4.33.1	Tasks, hash tables and machine-word integers used by the checkers; the parallel and hashed computations are proved equal to the sequential definitions, only the runtime's implementation is trusted
Trusted only for the software-layer claims (SCF controller, k-point protocol, C interval kernel); no checker output line depends on the components below		
TLC	2.19	Model checker for TLA$^+$ specifications
Frama-C (WP, RTE)	33.0	Generates the run-time safety and floating-point goals for the interval kernel; its library contracts for `sqrt`, `fabs`, `isfinite` are trusted
Why3, Alt-Ergo, CVC5, Z3	1.8.2, 2.6.4, 1.2.1, 5.1.0	Discharge WP goals; Why3's `ieee_float` theory is the floating-point semantics
C compiler	system clang	Unverified; CompCert is an option for the kernel
IEEE 754	hardware	Round-to-nearest binary64 as modelled by WP, C's `nextafter` as specified in the kernel, and `INFINITY` $= +\infty$ (the proof's one `admit`); compiled with `-ffp-contract=off`, no `-ffast-math`

Table 18: Trusted base. As for any computation, the operating system and the hardware that run the checker executables are trusted as well. An axiom audit in `make verify-all` checks that every constant in the project depends only on the three standard axioms; any `sorry` or project-declared axiom is reported. Each domain additionally names its modelling conventions, e.g. that the certified problem uses the exact rationals written in the certificate, or that the Lean definitions of the DFT external potentials and of how a TDB file is read say what is intended (backstopped by cross-checks).