ANNALS OF THE CHRYSENE FORMALISM
EDITORIAL | AI-AUGMENTED PEER-REVIEWED | VOL. 1, NO. 5 | NOVEMBER 2026

Executable Certification: Verification Code and Machine-Checked Proof as Accretive Layers in Orthogonal Review

Charles D. Schaper, Ph.D.
Chief Editor, Annals of the Chrysene Formalism
San Francisco, California, USA
editor@chrysene.com

Abstract

A previous paper in this journal established *orthogonal review*: a procedure in which review stages are assigned deliberately different validity questions rather than collapsed into a single judgment. This editorial announces an accretive extension of that procedure. Two artifacts are now published beside qualifying papers — a verification-code dossier, in which declared quantitative claims within scope are recomputed by a program that asserts rather than prints, and a formal-proof dossier, in which declared discrete content of the paper is machine-checked by the Lean 4 kernel. The methodological claim is that these layers supply a second kind of orthogonality. The three review disciplines of the earlier architecture are independent in the *question* they ask but identical in how they fail, namely by forming a wrong opinion; a verification harness and a proof kernel have no opinion to be wrong, and signal detected disagreement by a failed check. Their reliability remains conditional on correct encodings, adequate checks, and the trusted implementation. The two new layers are partial rather than total predicates on a manuscript, and their domains are fixed by status labels together with the declared checking scope; we state those domains precisely and show that neither contains the other. We report the defects the layers have found in two published papers and in one unpublished dossier — the last caught in pre-production by an independent checking system — and we grade those findings, since several are conservative and one was not. We also report one false positive generated by the review layer itself. For a journal that applies mathematical rigour to questions such as biogenesis, where the experimental loop is slow and in places not yet runnable, machine-checked internal consistency is the external constraint available before experiment, and its principal value is that it converts a falsification register into quantities an experiment can measure. The layers do not establish empirical chemical claims; numerical checks reach declared real-valued quantities, while the present Core Lean tier is limited to its stated discrete domain. They do not substitute for adversarial review; what they replace is testimony, by an artifact the reader can execute.

Keywords: formal verification; Lean 4; proof assistant; AI-augmented peer review; orthogonal review; reproducibility; falsifiability; biogenesis; machine-checked proof; Chrysene Formalism

1 Accretion, Not Replacement

The review architecture described in Article 41 of this journal [Sch26a] is unchanged by what follows. Its five stages — standalone mathematical review, concordance review, chemistry review, targeted post-revision verification, and final meta-review — continue to govern, as does its admissibility condition

$$\mathcal{F}_{\text{adm}} = \mathsf{M} \cap \mathsf{C} \cap \mathsf{H}, \tag{1}$$

in which M asks whether the mathematics is valid at the strength claimed, C whether the construction remains concordant with the formal framework it extends, and H whether the molecular interpretation is chemically admissible. Nothing here relaxes those conditions or substitutes for them.

What is added is a certification tier. Two artifacts are now prepared alongside qualifying papers and published as separate documents reachable from the article record: a *verification-code dossier* and a *formal-proof dossier*. Both are executable. Both are intended to be run by the reader rather than believed on the editor's word.

The tier reaches only part of a manuscript, and saying which part is this paper's technical content rather than a caveat appended to it. Section 3 states the domains formally; Section 4 observes that they are predicted by the status labels the journal already

Citation: Schaper, C. D. (2026). Executable Certification: Verification Code and Machine-Checked Proof as Accretive Layers in Orthogonal Review. *Annals of the Chrysene Formalism*, 1(5):1955–1965.
https://chrysene.com/index.php/acf/article/view/101
© 2026 Charles D. Schaper. Distributed under the CC Attribution-NonCommercial-NoDerivatives 4.0. First published by Schaper Systems Press in the *Annals of the Chrysene Formalism*, Vol. 1, No. 5 (2026). ISSN 3144-2021.

uses; and Section 10 states plainly what lies outside them, which for the research articles of this volume is most of what they assert.

1.1 What Article 41 Left Open

Article 41 reported that 48 of 63 manuscript-directed revision actions were explicitly verified, rising to 55 when indirect verification was counted — a coverage of 87.3%. That figure is a real one, but the word *verified* carries a specific and limited meaning there: a later reviewer examined the implemented repair and judged it sound. The verification of a judgment is therefore itself a judgment. It is testimony about testimony, and it inherits the failure mode of the thing it reports on.

This leaves a residual class of defect: those that survive every judgment layer because judgment is the wrong instrument for them. Three kinds recur, and each is witnessed by a defect reported in Section 5.

The first is arithmetic that is wrong by a power. A reviewer reading a table of dissipated energies against fork speeds will check that the magnitudes are plausible and that the reading beneath the table follows from it. A reviewer will not, in the ordinary case, recompute a four-row table from its stated formula, and if the table is internally smooth — each row a fixed factor from the last — nothing in its appearance signals that the factor is wrong.

The second is a figure that appears in a paper but was never produced by the computation the paper attributes it to. This is a transcription defect, and it is invisible to every reviewer who does not hold both the document and the emitting program at once.

The third is the hardest: a derivation whose conclusion is correct and whose reasoning is not. Here every printed number is right, so no numerical check can reach the defect, and the reviewer who verifies the numbers will find nothing wrong. Section 5.2 reports such a case.

1.2 Three Axes of Independence

Article 41 obtained independence between reviewers by varying the *question*: M, C and H interrogate different dimensions of validity, and a favourable answer to one is not evidence for another. That remains the architecture's central idea. But it leaves one thing constant across all three reviewers, namely the mechanism by which each of them can be wrong. All three are judgments, and all three fail in the same way: by forming a wrong opinion.

Two further axes are available, and the certification tier uses both. Table 1 sets them beside the first.

The second axis varies the implementation forming the opinion. Where a manuscript or an artifact has been prepared with the assistance of one generative system, an independent system may be asked to check it. This does not change the *kind* of thing a review is — the second system also forms an opinion, and can also be wrong — but it can reduce shared-context error when the two systems do not share a drafting context or accumulated commitment to the construction. Shared assumptions, training data, and specifications can still correlate their errors. Section 5.3 reports a defect caught on this axis.

Table 1: Three axes of reviewer independence. The first is the contribution of Article 41; the second and third are added here. Only the third changes how a layer can fail.

Axis	Independence of	Failure
$\mathsf{M} \cap \mathsf{C} \cap \mathsf{H}$	the question asked	wrong opinion
Cross-system	the implementation asking	wrong opinion, potentially decorrelated
V, L	the checking mechanism	failed check

The third axis removes opinion. A verification harness does not judge a manuscript; it recomputes the manuscript's quantities and halts with a nonzero status if they disagree. A proof kernel does not judge a theorem; it checks a term against a type and refuses the term if the check fails. Neither can be persuaded, and neither can be tired, and — more to the point — neither has a view that could be mistaken. Their explicit rejection mechanisms complement the first two axes. This mechanical distinction does not eliminate shared specification, translation, implementation, or coverage errors.

> **Remark 1** (Not a claim of statistical independence)**.** *The three axes use different checking mechanisms; they do not have mutually exclusive failure modes. A manuscript, harness, and formalization can share a specification error or omit the same hypothesis. No claim of statistical independence or residual defect rate follows. Conditional on faithful encoding, an adequate check, and correct execution, a failing encoded condition is rejected by that check. Detecting every defect in the prose requires additional coverage and translation arguments, not mechanical diversity alone.*

2 What the Two Artifacts Are

This section is written for a reader who has not used a proof assistant. Both artifacts are simple objects; what matters about each is a single property, and in both cases that property is the ability to fail.

2.1 Verification Code

A verification-code dossier publishes a program that rebuilds the declared quantitative claims within its scope from the paper's

own definitions and compares the result against the figure the paper prints. The program is written independently of whatever computation produced the paper's numbers in the first place, so that agreement between them is evidence rather than a restatement.

The property that matters is that the program *asserts* rather than *reports*. A reporting program prints a table and leaves a human to compare it against the manuscript; an asserting program carries, for each claim, a statement of what the claim is, a tolerance where the quantity is physical, and a verdict. If any verdict is negative the program halts with a nonzero exit status. Definition 1 makes the distinction precise.

Definition 1 (Reporting and asserting programs)**.** *Let D range over well-formed tuples of defining manuscript data and claimed quantities, and let $Q(D)$ be the declared agreement predicate, including any numerical tolerances. Write $\chi(P(D))$ for the exit status of a terminating checking run. A program is* reporting-only *when it prints results without using Q to determine success or failure. It is* asserting for Q *when $\chi(P(D)) = 0$ if and only if $Q(D)$ holds; a false declared condition must produce a nonzero exit status. An unrelated crash or merely nonconstant status is not a certificate. A successful run certifies only Q on the supplied data, subject to the correctness of the implementation and its correspondence with the manuscript. Mutation tests probe this correspondence through specified changes $\mu(D)$ expected to make Q false.*

The distinction is not pedantic, and the journal has been on the wrong side of it. The block scripts published in the appendices of Articles 62 and 64 [Sch26b, Sch26c] printed tables of cut sets, carbon ledgers, return-map multipliers and pool ratios, each followed by a reading. They asserted nothing. They did not condition success on agreement with the printed claims, so by Definition 1 they carried no such certificate; a reader wishing to know whether the paper agreed with its own appendix had to perform the comparison by eye, which reinstates precisely the judgment layer the program appeared to relieve.

Rewriting them as harnesses — one assertion per printed quantity, a stated tolerance, a nonzero exit on any disagreement — is what turned them from exposition into evidence. The two harnesses now published run 108 checks over Article 62 and 80 over Article 64. Every defect reported in Section 5.1 was found in the course of that rewriting, and none had been found in the judgment review that preceded it.

Two limits should be stated at once. A verification harness checks transcription and arithmetic, not truth: it establishes that a paper agrees with a computation from its own definitions, and both can be wrong together if the definitions are wrong. And it is silent wherever a claim is not arithmetic, which includes every claim of chemistry.

2.2 Machine-Checked Proof in Lean 4

A formal-proof dossier is a different kind of object, and the difference is worth stating carefully because the two are easily conflated.

Lean 4 [dMU21] is an interactive theorem prover built on dependent type theory. In it, a mathematical statement is a *type* and a proof of that statement is a *term* of that type. Checking a proof is therefore type-checking: a small trusted program, the kernel, verifies that the term the author supplied really does have the type the author claimed. A term that fails the check is not a bad proof; it is not a proof of that statement at all, and the system will not accept it as one.

Three mechanics suffice to read the dossiers this journal publishes, and each appears in the Article 62 artifact.

`rfl` closes a goal when both sides of an equation reduce to the same normal form by computation. The statement that the eighteen single-edge cut sets of the chrysene perimeter all leave Hamiltonian paths is proved this way: the kernel constructs each path and checks each membership, and the two sides of the claim become literally the same object.

`decide` closes a goal when the proposition is decidable and its decision procedure evaluates to *true*. It is the natural tactic for a finite claim, but it is not free: evaluating a decision procedure inside the kernel can exceed the elaborator's recursion limit, and in the present artifact it did so on the 153-element two-edge enumeration. That failure is instructive and is taken up below.

`#print axioms` is not a tactic but an audit. It reports, transitively, every axiom a proof depends on. Lean's standard axioms are `propext`, `Classical.choice` and `Quot.sound`, and a proof resting on them is in no way suspect; many developments use one or more of these axioms. But the audit also reports a fourth possibility, and that one is decisive.

Remark 2 (Why the axiom audit is the editorially important part)**.** *When a tactic fails, Lean reports an error and may retain a placeholder proof during error recovery, marked as depending on* `sorryAx`*. Such a declaration can appear in later diagnostic output, but a failed batch compilation is not a certified build. An explicit* `sorry` *can also leave an admitted declaration with a warning. Certification therefore requires both a successful compilation and an axiom audit excluding* `sorryAx` *and undeclared assumptions; the audit does not replace the exit-status check.*

The consequence for a journal is immediate. A formal-proof dossier that prints a table of theorem identifiers, and a listing of the file, and no axiom audit, can display a page of green theorems that the kernel considers unproved. The reader has no way to tell, because the error appeared on a console that is not part of the published document.

This is not hypothetical. It is exactly the defect described in Section 5.3, and it is why Convention 1 makes the verbatim

audit log — not a summary of it — a requirement of the dossier rather than an ornament.

The recursion-limit failure mentioned above makes the same point from the other direction. When `decide` exceeded the limit during preparation of the Article 62 artifact, the tactic failed, the declaration was admitted, and the audit reported `sorryAx`. Nothing else in the file changed appearance. The defect was visible only in the audit, and the repair — raising the recursion limit and proving the enumeration by `rfl` instead — was found because the audit was read.

3 The New Layers Are Partial Predicates

Article 41's M, C and H are total on a manuscript: every statement in a paper is subject to all three, and a paper that fails any one of them fails. The new layers are not total, and the difference is structural rather than incidental. A verification harness has nothing to say about a claim that is not arithmetic; a proof kernel has nothing to say about a claim of chemistry. Treating them as total would make the certification tier either vacuous or dishonest.

We therefore give them domains, indexed by the status labels the journal's research articles already carry: [A] for statements that are finite or axiom-free, [D] for those derived under named hypotheses, [C] for imported results, and [E] for molecular readings that enter no proof.

Definition 2 (Status-indexed domains)**.** *Let* Σ *be a manuscript whose statements* σ *carry status labels* $\lambda(\sigma) \in \{[\mathrm{A}], [\mathrm{D}], [\mathrm{C}], [\mathrm{E}]\}$. *Define*

$$\mathrm{dom}(\mathsf{L}) = \{\sigma : \lambda(\sigma) = [\mathrm{A}], \ \sigma \text{ enumerable or closed over } \mathbb{N}\}, \tag{2}$$

$$\mathrm{dom}(\mathsf{V}) = \{\sigma : \lambda(\sigma) = [\mathrm{A}], \ \sigma \text{ finite}\} \cup \{\sigma : \lambda(\sigma) = [\mathrm{D}], \ \sigma \text{ reduces to arithmetic over stated constants}\}. \tag{3}$$

Here finite *means assigned to a declared, finitely bounded exhaustive evaluation under this editorial policy; it does not mean "has some finite proof." The* [A] *content is classified as finite-enumeration content or universally quantified discrete content requiring a symbolic proof, with the classification and scope stated in the dossier. Finite-enumeration claims also fall within the enumerable clause of* (2)*. Thus the domains describe the present certification policy, not the intrinsic limits of programs or Lean. In particular, unbounded discrete identities can have finite symbolic proofs, and Lean with additional libraries can formalize real analysis. The finite restriction in* (3) *prevents its first clause from containing all* [A] *statements by definition.*

Definition 3 (Extended admissibility)**.**

$$\mathcal{F}'_{\mathrm{adm}} = \mathsf{M} \cap \mathsf{C} \cap \mathsf{H} \cap \mathsf{V}|_{\mathrm{dom}(\mathsf{V})} \cap \mathsf{L}|_{\mathrm{dom}(\mathsf{L})}. \tag{4}$$

Here $\mathsf{V}|_{\mathrm{dom}(\mathsf{V})}$ *denotes the set of manuscripts* Σ *such that* $\mathsf{V}(\sigma)$ *holds for every* $\sigma \in \Sigma \cap \mathrm{dom}(\mathsf{V})$*, and likewise for* $\mathsf{L}|_{\mathrm{dom}(\mathsf{L})}$*; the restricted layers are thereby lifted from predicates on statements to subsets of the manuscript space, which is what* (1) *intersects. The universal quantification is over a possibly empty set, so a manuscript with* $\Sigma \cap \mathrm{dom}(\mathsf{V}) = \varnothing$ *satisfies the condition vacuously — which is the formal content of the third rollout rule of Section 9.*
As in Article 41 the conditions are not substitutable: a restriction cannot be discharged by performance elsewhere, and a statement lying outside both new domains is governed by $\mathsf{M} \cap \mathsf{C} \cap \mathsf{H}$ *exactly as before.*

It would be convenient if the two new domains were nested — if, say, everything a kernel could check were also checkable by a harness, so that the formal-proof dossier were simply a stronger version of the verification dossier. They are not, and the failure of nesting is in both directions.

Proposition 1 (The two domains are not nested)**.** $\mathrm{dom}(\mathsf{L}) \not\subseteq \mathrm{dom}(\mathsf{V})$ *and* $\mathrm{dom}(\mathsf{V}) \not\subseteq \mathrm{dom}(\mathsf{L})$.

Proof. Under the declared policy, the two witnesses occupy different classes. The capacity identity $4^m = 2^m \cdot 2^m$ for every $m \in \mathbb{N}$ is assigned to universally quantified discrete [A] content. It has an inductive proof in $\mathrm{dom}(\mathsf{L})$; finitely many sampled values alone do not certify that universally quantified statement, so it is not assigned to $\mathrm{dom}(\mathsf{V})$. This is a scope decision, not an impossibility of symbolic computation by a program.
Conversely, the claimed numerical value of the pool ratio $\nu_2 = 8\pi L_p h/(aE^*)$, evaluated from declared constants within a stated tolerance, is arithmetic [D] content. It belongs to $\mathrm{dom}(\mathsf{V})$ and not to the present [A]-only $\mathrm{dom}(\mathsf{L})$. Both witnesses are present in the described policy, so neither domain contains the other. No limitation of Lean's expressive power is needed for this conclusion. □

Remark 3 (Where the two overlap, and why that matters)**.** *On the intersection* $\mathrm{dom}(\mathsf{V}) \cap \mathrm{dom}(\mathsf{L})$ *— the finite* [A] *content — both layers apply, and their agreement is informative rather than redundant precisely because they fail differently.*

> *The cut-set enumeration of Article 62 is the worked case. All* $\binom{18}{2} = 153$ *two-edge cut sets are recomputed numerically by the harness and proved exhaustively by the kernel, and the two computations share no code, no language and no author.*

4 The Gradient Follows the Journal's Own Status Labels

The status convention was adopted for a different purpose. Its function in the research articles is to mark what a statement rests on, so that a reader can see at a glance whether a result is finite, derived under hypotheses, imported, or a molecular reading entering no proof. It was not designed to predict which instrument could verify a statement, and no such use was contemplated when it was introduced.

It nevertheless does so. Table 2 gives the correspondence observed across the corpus audited to date, with a witness from Article 62 in each row.

That a convention introduced to record provenance should also predict verifiability is, we think, evidence that the labels track something real about the statements they mark — roughly, how much of a statement's content is carried by its own definitions as against by the world. A statement whose truth is settled by its definitions is one a kernel can settle; a statement whose truth depends on a measured permeability is not, however carefully it is written.

The observation should not be overstated. It is a correspondence on a corpus of a few papers, checked by one editorial group, and it is recorded in Section 12 as a claim that a later paper could refute.

5 The Catch Record

Every defect reported below had passed drafting, referee report, revision and finalization under the architecture of Article 41. They are grouped by the layer that found them.

5.1 Caught by Executable Arithmetic

Four defects were found in the course of converting the block scripts of Articles 62 and 64 into asserting harnesses.

An energy column scaling as the square of a quantity that enters linearly. Article 64's appendix tabulates the energy dissipated in unwinding a strand against the speed of the replication fork. The dissipated energy is $N\zeta(\omega r)^2(L/v)$ with $\omega \propto v$, hence linear in v; the published column advanced by two decades per decade of speed. Three of its four rows were wrong, and — the detail that makes this a transcription defect rather than an arithmetic one — the script to which the appendix attributes the column never produced those rows. It emits the linear ladder.

An active phase labelled as a period. The same paper fixes a twelve-hour illuminated phase of 4.3×10^4 s. Two blocks of its appendix use that figure for the forcing period T itself. At the stated duty of one half the two cannot coincide: if the light phase is twelve hours the diurnal period is twenty-four. The error is a factor of two in a quantity that an experimentalist would read directly off the paper.

A constancy claim with a counterexample in its own first row. Article 62's pool-ratio block closes by stating that the ratio ν is independent of strand length. For ν_2 this holds at every tabulated length. For ν_1 it holds only above the coil threshold $n > 147$ bp, and the table's first row is $n = 100$, below it; the block's own output reads $\nu_1 = 251$ in that row against 438 in every other. The exception was printed and was not noted.

Six drafting markers inside listing environments. Article 62 shipped with six `<TODO>` markers, five of them in the block describing the record on which the later blocks run. They survived a sweep for drafting macros because that sweep searched for coloured commands and these were plain text inside verbatim listings. Among the quantities left as placeholders were two tether lengths that the same appendix computes, nine lines further down.

These findings are not of equal weight, and reporting them as though they were would overstate the layer's value. Three of the four are conservative in the sense that the paper's conclusion survives unchanged, and in two cases survives with more margin than it claimed: the energy column *overstates* the cost of unwinding, so the reading beneath it — that the cost is negligible — holds *a fortiori*; the constancy exception affects a ratio the architecture does not rest on. The drafting markers are an embarrassment rather than an error, since the quantities they stand for are computed elsewhere in the same appendix. Only the mislabelled period is a defect of the kind that could propagate into work downstream, and it is treated as such in Section 8.1. An audit that does not grade its own findings is advertising.

5.2 Caught by Cross-Layer Recomputation

The most instructive single defect was found by neither layer acting alone.

Article 62 locates two substituted positions on the chrysene perimeter and asks how the molecule may be oriented within a membrane leaflet. A proposition confined the bilayer normal to a window of some 44°, between the short-axis direction and the direction joining the two substituted positions, which it gave as 46° to the long axis. A corollary then reported the depth and footprint the skeleton presents across that window.

The 46° is the angle between two *lines* and is correct as such. The vector joining the two positions lies at 133.9°. The computation that produced the figure had taken the magnitudes of the two components before forming the angle — an operation

Table 2: Reachable certification layer by status label, with a witness from Article 62 in each row. The correspondence is an observation on the corpus audited to date, not a theorem.

Status	Character	Reachable by	Witness
[A]	finite, axiom-free	kernel, exhaustively	all 153 two-edge cut sets
[A]	closed over $\mathbb{N}$	kernel, by induction	$4^m = 2^m \cdot 2^m$ for every m
[D]	arithmetic over stated constants	harness, with tolerance	pool ratio $\nu_2 \approx 66$
[D]	resting on physical constants	judgment, with sensitivity sweep	footprint over the tilt range
[C], [E]	imported; molecular reading	judgment only	patent figure attestations

that folds an obtuse bearing into its acute supplementary line angle — and the proposition then read the result as a direction. The consequence is sharp: the depth separation on which the proposition's own criterion depends *vanishes* at 43.9°, so the window had been opened at within three degrees of the one orientation the criterion forbids, where the separation is 0.18 Å against 3.64 Å at the other end.

What makes the case instructive is that *every printed figure was correct*. The depth and footprint intervals the corollary reported were the true intervals over the range the paper's hypotheses actually supply, which begins near 68° rather than at 46°. Checking those reported numbers alone could not have reached the defect, because the arithmetic of those values was not wrong. Only the derivation was.

It surfaced when the admissible range and the intervals computed over it were recomputed against each other and found to disagree — that is, when two statements of the manuscript were checked for coherence rather than each being checked against a computation. The moral is a limit on what Section 2.1 claims: a verification harness checks agreement between a manuscript and a computation; it checks agreement between two manuscript statements only if that relation is explicitly encoded as a check. A paper can therefore remain internally incoherent while all its individually checked numbers verify.

The correction preserved every figure of the corollary and replaced the derivation of the range, which now rests on the requirement that the unit occupy one leaflet. The superseded proposition is recorded in the affected paper's withdrawals register.

5.3 Caught in Pre-Production by an Independent System

The remaining defect was found in an unpublished dossier passing through the editorial process, and was corrected before publication.

The dossier presented six theorems with a table of identifiers, a listing of the formal file, and an audit log stating that each theorem depended on no axioms. The listing as prepared for the document did not compile. Five of the six theorems returned `sorryAx`; only the sixth was proved. The audit log as typeset asserted the opposite of what the kernel reports.

Two causes were present, both consistent with damage introduced in preparing the document rather than with an error in the checked source: a tactic that cannot close a conjunction had been applied to five conjunctive statements, and one definition carried a sign that inverted the relation its theorem asserted. Repairing both restored all six theorems and the axiom-free audit the dossier had claimed.

Two points follow, and the second is the general one.

First, this is the cross-system axis of Table 1 doing work. The artifact was prepared with the assistance of one generative system and checked by a second, independent one, and what the second returned was not an opinion that could be weighed against the first. It was a kernel transcript. The disagreement was not resolvable by discussion.

Second — and this is the part that generalizes past mathematics — the defect was introduced *between* the file that was checked and the document that was prepared. The source was sound; the artifact was not. Transcription between the two is an error channel, and it is one that no amount of care in the mathematics defends against. The requirement that follows is stated in Section 6: the artifact that ships must be the artifact that was checked, and the journal must be able to demonstrate it.

6 The Failure Mode the Machinery Introduces

A certification tier creates a new way for a journal to be wrong. A paper with no verification dossier makes no claim about machine checking, and a reader knows to treat its arithmetic as unaudited. A paper with a verification dossier makes a strong claim, and if the dossier is defective the claim is worse than absent: it is a false assurance, delivered in the register of certainty, and a reader who defers to it is worse off than one who had nothing.

Section 5.3 is the worked example and was not deliberate. No one typed a false audit log. A document was prepared from a file, the preparation damaged the file, and the log that was typeset came from a run that no longer corresponded to the listing beside

it. Every step was ordinary.

The journal therefore adopts the following as a condition of publication rather than as good practice.

> **Convention 1** (Dossier contract)**.** *A verification-code or formal-proof dossier published by this journal shall carry:*
>
> 1. *a* ***scope section*** *naming what the layer does not reach, in the terms of Definition 2;*
> 2. *a* ***tamper test*** *in the sense of Definition 4, reporting a set of perturbations and the number of checks each one breaks;*
> 3. *the* ***verbatim standard output*** *of the checking run, not a summary of it, together with its exit status and the version string of the toolchain that produced it;*
> 4. *a listing* ***generated programmatically*** *from the file that produced that output, and verified byte-identical to it; and*
> 5. ***no external dependencies****, so that a reader can reproduce the result without the journal's infrastructure. The Article 62 formal artifact is written in Core Lean 4 and runs in a browser.*

The fourth item is the one that addresses Section 5.3 directly, and it is the one most easily neglected, because pasting a listing into a document is the obvious way to prepare one and leaves no trace when it goes wrong. In the dossiers accompanying Articles 62 and 64 the embedded listing is extracted from the checked file by the build script and compared byte-for-byte against it; the comparison is part of the build and fails it.

7 Anti-Vacuity: a Certificate Must Be Able to Fail

An axiom audit identifies the axioms on which the checked declarations depend. It does not establish that the theorems say anything. A file of trivialities — $2 + 2 = 4$, restated twenty-three times with different names — audits perfectly, exits zero, and does not certify the substantive claims of the paper beside it.

The same hazard attends a verification harness. A harness whose assertions are insensitive to the manuscript's defining data will pass under those perturbations, and its passing conveys no evidence of sensitivity to the affected claims. This is not a remote possibility; it is the natural failure mode of a harness written by someone who already believes the manuscript, since the easiest assertion to write is one that cannot fail.

The journal therefore requires a demonstration of sensitivity.

> **Definition 4** (Mutation sensitivity)**.** *Let A be a certification artifact over defining data D, with a successful baseline run. Let μ be a specified perturbation expected to falsify at least one declared check. The artifact is* sensitive *to μ if the mutated run fails. A dossier reports the perturbations, expected affected checks, observed failures, and exit statuses. Failure to detect an expected-invalid mutation exposes a gap in the tested sensitivity. A mutation that preserves the certified property need not fail; passing a finite mutation suite does not establish completeness.*

For the Article 62 formal artifact the reported perturbations are of its defining chemical data. Moving one substituted locant from C12 to C11 leaves seven of the twenty-three theorems unproved; deleting one of the three interior chords leaves one; lengthening the cytosine extension of the longer tether from three atoms to four leaves one. For the corresponding harness, six perturbations break between one and ten of its 108 checks. Every perturbation tested was caught and every mutated run exited nonzero.

The spread is itself informative and is reported rather than averaged. A substituted locant is load-bearing across the file, entering the single-edge counts, the two-edge partition, the named cut sets and the chain numbering; a single interior chord enters only the cyclomatic count. A perturbation that breaks one check is not a weaker result than one that breaks seven, provided the check it breaks is the one that depends on the datum perturbed.

8 Why This Matters for Biogenesis

The preceding sections would apply to any journal publishing quantitative work. This one does not, and it is the reason the certification tier is being introduced now rather than noted as a possibility.

This journal applies mathematical rigour to questions including biogenesis: the chemical and physical conditions under which an encoded, self-maintaining system could have arisen. The distinguishing feature of that subject, for present purposes, is not that it is difficult but that its experimental loop is slow, expensive, and in places not currently runnable at all. A proposed architecture for a non-enzymatic replicating enclosure cannot be put in a flask next week and observed.

A paper in such a field is therefore, in substantial part, a *consistency claim*: an assertion that a particular architecture hangs together — that its material ledger balances, that its geometry admits the dispositions it requires, that its timescales separate in the directions it needs, and that its information-carrying capacity is what it says. Machine checking is the external constraint available to such a claim before experiment, and it does two things.

8.1 Removing Defects That Would Misdirect Experiment

An architecture proposed as a guide to experiment inherits every arithmetic defect of the paper proposing it. The mislabelled period of Section 5.1 is the clear case: an experimentalist designing an illumination schedule from that paper would have taken a twelve-hour phase for a twenty-four-hour period, and would have been wrong by a factor of two in the one parameter the architecture is entrained by. The defect was conservative for the paper's own conclusions. It would not have been conservative for the experiment.

This is the unglamorous half of the argument, and it is the half most easily dismissed as mere tidiness. It is not tidiness. In a field where a single experimental cycle may take months, a factor of two in a design parameter is a substantial fraction of a research programme.

8.2 Sharpening the Falsification Register

The stronger half is that verification converts a falsification register from a list of concerns into a set of measurable quantities.

Article 64 derives three independent caps on the length of a genome such an architecture could carry: an information-theoretic threshold, a geometric span, and the requirement that replication complete within the illuminated phase. The three admit a common band — and the band is *empty* above a per-base addition time of 143 s. A measured value above that figure does not strain the account; it refutes it outright. The figure, the emptiness of the band above it, and the point at which the clock cap begins to bind are all machine-verified.

That is a different kind of statement from a qualitative prediction. It tells an experimentalist which single measurement would end the architecture, and it does so with the arithmetic checked. A falsification register whose entries are verified is a research programme's most useful output, because it directs effort at the measurements that can settle something.

The limit must be stated as plainly as the claim. Internal consistency is not evidence that an architecture is realized in nature, and no quantity of machine checking makes it so. A perfectly verified account of a chemistry that never occurred is perfectly verified and false. What consistency establishes is that the account is worth the cost of testing — that an experiment addressed to it will be addressed to a coherent proposal rather than to an arithmetic slip. In a field where experiments are cheap, this matters little. In this one it matters a great deal.

8.3 Supporting the Theoretical Arc

The argument so far concerns experiment. There is a second use, internal to the mathematics, and over the long run it may be the more important of the two.

The stated objective of this programme is a continuous derivation: from a self-separating source, through the formulation of a genetic code, to the organization of modern biology. Such an arc is not one paper but a chain of them, and a chain has a failure mode that a single paper does not. Each link imports constants, structures and hypotheses from its predecessors. Article 62 derives a three-atom difference between two tethers; Article 63 imports that figure and builds a readout contrast upon it; what follows imports the contrast. An error at any link is not confined to the paper that makes it — it propagates, silently, into every result downstream, and the longer the arc the lower the probability that it is sound if the links are checked only by reading.

Verification changes the arithmetic of that risk in two ways.

The first is that it makes an import auditable. A constant that is machine-checked where it is derived, cited by label where it is used, and re-audited when it moves, is a traceable dependency rather than a remembered one. The re-audit clause of Remark 4 exists for this reason, and its cost is the price of a corpus whose later papers may rely on its earlier ones.

The second is less obvious and is the stronger effect: formalization forces a statement to be decided before it can be written. A proof assistant will not accept a theorem whose hypotheses are left to the reader's charity, and the discipline of supplying them exposes commitments that prose can leave implicit. The present corpus supplies a case. In preparing the Article 62 formal artifact, a theorem about the locant of a retained unsaturation failed to compile, because the statement presumed an orientation of the product chain that the definition it referred to did not implement. The prose had been consistent, because prose can carry an unstated convention; the formal statement could not. The convention is now explicit and is itself a theorem.

Over a long arc that discipline compounds. Each formalized link leaves behind a set of hypotheses stated precisely enough to be imported without renegotiation, and a successor paper that imports them inherits a definition rather than an impression.

One qualification is required, and it limits the objective rather than the method. An arc terminating in modern biology cannot be *proven* in the sense a kernel understands, because its later reaches contain empirical content — that a chemistry occurred, that a structure was selected — and no amount of formalization bears on whether a thing happened. What can be achieved is narrower and still substantial: that the *mathematical* skeleton of each link is machine-checked, and that at every step the empirical content is separated from the derived content and labelled as such. That is the function the status labels already perform, and Section 4 is the observation that they perform it well enough to predict which instrument applies.

It follows that verification coverage will *thin* as the arc extends, not thicken. The early links are combinatorial and substantially formalizable; the late ones are biological and largely are not. A programme that expected the opposite would be disappointed by

its own progress. The useful statement is the one the gradient of Section 4 already supports: at each stage the corpus can say which part of its own claim is settled by its definitions, which part rests on named hypotheses, and which part awaits the world.

9 Rollout, and Retroactive Audit

The editorial board's expectation is that the certification tier will be applied to all papers, irrespective of perspective or subject, and — in keeping with the treatment of this journal as a living document — retroactively to the existing corpus.

Which of the two dossiers a paper receives is determined by its content rather than by editorial judgment:

- a paper receives a **verification-code dossier** if it carries any finite statement labelled [A], or any [D] statement reducible to arithmetic over stated constants;
- it receives a **formal-proof dossier** if its [A] content is enumerable or closed over $\mathbb{N}$, in the sense of Definition 2;
- a paper qualifying for neither receives neither, and its article record says so explicitly rather than omitting the dossier silently.

The third clause matters more than it appears to. A reader encountering a paper without a dossier should be able to tell whether the tier was inapplicable or merely not yet applied, and an absent artifact that is not accounted for is indistinguishable from an artifact that failed.

> **Remark 4** (What a retroactive audit commits the journal to)**.** *The commitment should be stated before the programme begins rather than discovered during it.*
> *Auditing two papers produced the defects of Section 5. A corpus-wide audit should be expected to produce corrections at a comparable rate, and a programme that audits a corpus without publishing what it finds is not an audit. The journal therefore commits in advance: corrections will be published as they are found; each will be recorded in the withdrawals register of the affected paper; the superseded statement will remain visible rather than being silently replaced; and where a correction changes a figure that a later paper imports, the dependent paper will be re-audited.*
> *The last clause is the expensive one and is included because the corpus is tightly coupled. The tether difference verified in Article 62 is imported by Article 63 and underwrites its readout contrast; the unit footprint of Article 62 is imported by Article 64. A correction that moved either would not be contained.*
> *For the period convention discussed in Section 5.1, this editorial uses* $T_{\text{light}} = 43{,}200\,\text{s}$ *and duty fraction* $d = 1/2$*, hence* $T_{\text{cycle}} = T_{\text{light}}/d = 86{,}400\,\text{s}$*. A downstream import must identify which duration it means; an inherited* 43,200 s *symbol cannot silently serve as both. This convention specifies the re-audit target and does not assert that all dependent articles have already been corrected.*
> *A re-audit must also check the citation apparatus. Provisional volume, issue, and page data must be compared against the published record. These bibliographic fields are machine-checkable when represented in a manifest; a numerical harness omitting that manifest does not verify them merely by checking the paper's equations.*

10 Limitations

Six, stated without mitigation.

(i) The layers reach a minority of what the research articles assert. They do not establish empirical chemical claims or patent-figure attestations. The verification layer does reach specified real-valued calculations with tolerances; the present Core Lean layer covers only its declared discrete formal content. The realizability of a cut set by a named reagent, the direction an elimination takes, the attestation of a feature by a figure in a cited document — none of these is touched. Most of what Articles 62 to 65 assert lies outside both domains of Definition 2.

(ii) They did not find the most serious defect reported here. The derivation defect of Section 5.2 was found by recomputing two statements against each other, which is an act of judgment and not a layer of the tier.

(iii) The review layer generates false positives. At least one is recorded in Section 10.1. The rate at which a layer raises false alarms is part of its cost, and this editorial reports a single case because a single case is what the audited corpus contains — not because the rate is known to be low.

(iv) There is no control group and the corpus is one journal's. Nothing here estimates how such a tier would perform elsewhere, and the descriptive statistics of Article 41 are not extended by this paper.

(v) The journal is single-author. Author, editor and referee are the same person, so these layers substitute for one component of adversarial review — the detection of mechanical and formal error — and not for adversarial review itself. No verification harness will notice that a paper's framing is self-serving, that its literature is selectively cited, or that its central idea is uninteresting. The claim made here is narrower than "more reliable" and is stated exactly: what the tier replaces is *testimony*, by an artifact the reader can execute.

(vi) The checking system assisted in preparing this editorial. The independent system that produced the artifacts and found several of the defects of Section 5 also assisted in drafting this paper, which therefore reports in part on its own performance. The mitigation is structural rather than rhetorical: the artifacts are published, they are executable, and the claims of Section 5

can be checked by a reader without reference to this account of them.

10.1 A False Positive, Recorded

During the audit of Article 62 the review layer raised a criticism that proved to be wrong, and the case is recorded in full because a process report that records only its successes is a sales document.

The paper budgets the extended length of a tethered unit against the hydrophobic core of a bilayer and concludes that the capped arm must fold, losing a stated amount of its extension. The criticism held that the folding hypothesis failed at the shallow end of the admissible tilt range, where the required loss is largest.

It did not. The comparison had been made between the arm's freely jointed extension and the *excess* over the core, when the quantity that must be compared is the arm's *remaining* length: an arm of 11.4 Å that extends to 3.8 Å when freely jointed gives up 7.6 Å unforced, which exceeds the largest loss the range requires under either the corrected or the superseded reading. The hypothesis holds throughout, and the paper's original claim was right as published.

The criticism was withdrawn before it reached a revision, and the subsequent artifact asserts the comparison in both directions so that it cannot be made backwards again. We note that the error was one of judgment, made in the course of operating a layer whose purpose is to reduce reliance on judgment, and that it was found by re-derivation rather than by any tool.

11 Prospective Protocol

Article 41 closed by proposing that its retrospective reconstruction be converted into a prospective protocol. This editorial discharges that proposal for the certification tier.

The workflow of Article 41 is extended by two stages placed after targeted verification and before meta-review: construction of the applicable dossiers, and an independent check of each dossier by a system that did not prepare it. For every artifact the journal records the toolchain and version string, the exit status, the set of perturbations tested and the sensitivity of each, the declared domain in the sense of Definition 2, and the result of the byte-identity comparison required by Convention 1.

Accumulated across papers, that record permits questions the present corpus cannot answer: what proportion of a paper lies within each domain; whether defect profiles differ by layer as Article 41 found they differ by discipline; what rate of false positives the judgment layers sustain; and whether the status-label correspondence of Section 4 continues to hold as the corpus grows. These are the measurements that would move the procedure from a reported practice toward an evaluated one.

12 Falsification Register

The claims and implementation of this editorial can be challenged by the following findings; the last is an audit warning rather than a logical refutation.

1. **A defect in a certified statement.** A reader finds, in a statement a published dossier certifies, an error of the kind the dossier claims to exclude. This would refute the tier's central claim directly.

2. **An irreproducible log.** A published audit or console log cannot be reproduced on a clean installation of the stated toolchain at the stated version. This would indicate that Convention 1 is not being met in practice.

3. **An insensitive artifact.** A specified expected-invalid mutation in a published dossier breaks no check. This refutes its claimed sensitivity to that mutation, not every unaffected certificate in the artifact.

4. **A failure of the gradient.** A later paper carries a statement whose status label does not predict which layer can reach it, in either direction. Section 4 would then be recording a coincidence of the present corpus.

5. **A retroactive audit that finds nothing.** A defect-free audit is not a logical refutation and can be a legitimate result. It warrants checking coverage and mutation sensitivity rather than inferring reliability solely from zero detected defects. Two audited papers do not establish an expected corpus-wide defect rate.

Entry 5 is stated deliberately. A certification programme is most dangerous to a journal at the point where it begins to return only good news.

13 Companion Artifacts and Scope of the Present Check

The Article 101 verification and formal-proof dossiers check this editorial's domain policy, manuscript-level lifting, elementary arithmetic witnesses, and specified cycle combinatorics. Their source, tool versions, verbatim logs, and mutation results are supplied separately. Historical counts and case descriptions imported from Articles 41, 62, and 64 remain attributed reports: the present supplements do not re-execute those articles' original dossiers or certify chemical realizability. The finite model makes the policy explicit without presenting its classification as a theorem about every possible verification technology.

14 Conclusion

The procedure described in Article 41 separated the questions a review asks. What is added here separates the ways a review can fail. A verification harness and a proof kernel are not better reviewers than the mathematical, concordance and chemistry layers that precede them; over most of a manuscript they are not reviewers at all, being silent on everything outside the domains of Definition 2. Within those domains they are a different instrument, and the defects reported in Section 5 are the ones that instrument reached after the others had passed.

The limits are not incidental to the proposal and have been stated at every point where they bear: the layers are partial; they are silent on chemistry and on the physical claims that carry most of this journal's content; they did not find the most serious defect reported here; they introduce a failure mode of their own, which Convention 1 exists to contain; and in a single-author journal they substitute for one component of adversarial review and not for the thing itself.

What they change is the standing of a claim of correctness. Under the architecture of Article 41, a reader who wished to know whether a paper's arithmetic was sound had the editor's assurance that a reviewer had looked. Under the tier described here, that reader has a file, a toolchain version, an exit status, and a mutation test, and may reach a different conclusion from the editor's. For a journal whose subject admits few experiments and whose author is also its editor, the substitution of an executable artifact for an assurance is not a refinement of the review process. It is the part of the review process a reader can check.

References

[dMU21] L. de Moura and S. Ullrich, "The Lean 4 Theorem Prover and Programming Language," in *Automated Deduction — CADE 28*, Lecture Notes in Computer Science, vol. 12699, Springer, 2021, pp. 625–635. DOI: 10.1007/978-3-030-79876-5_37.

[Sch26a] C. D. Schaper, "AI-Augmented Peer Review by Orthogonal Constraint: A Retrospective Case Study of Mathematical, Concordance, Chemical, and Meta-Review," *Annals of the Chrysene Formalism*, vol. 1, no. 2, pp. 561–581, 2026. DOI: 10.68366/kjwc7142.

[Sch26b] C. D. Schaper, "The Enclosure via the Encoded Chrysene Cluster: Integrated Synthesis of Nucleic Acids and Phospholipids, and the Orientational Encoding of the Membrane," *Annals of the Chrysene Formalism*, vol. 1, no. 3, pp. 1279–1320, 2026. DOI: 10.68366/24m3wx88.

[Sch26c] C. D. Schaper, "The Photoentrained Cycle: Existence and Stability of a Periodic Orbit in a Forced Hybrid Protocell," *Annals of the Chrysene Formalism*, vol. 1, no. 3, pp. 1348–1370, 2026. DOI: 10.68366/h9rzkd75.