

AI-Augmented Peer Review by Orthogonal Constraint: A Retrospective Case Study of Mathematical, Concordance, Chemical, and Meta-Review

Charles D. Schaper, Ph.D.

Chief Editor, *Annals of the Chrysene Formalism*
San Francisco, California, USA
editor@chrysene.com

Abstract

The increasing use of artificial intelligence in scientific writing and mathematical formalization creates a corresponding need for review methods capable of validating rapidly generated, multidisciplinary technical work. We present and retrospectively evaluate an AI-augmented peer-review architecture in which distinct review stages are assigned orthogonal objectives: mathematical correctness, concordance with an established formal framework, chemical admissibility of molecular claims, targeted verification of revisions, and final meta-review. The method is examined through the development of *The Chrysene Formalism Revisited: A Twelve-Axiom Foundation for Molecular Characterization, Tensorial Noncommutative Geometry, and Operator Response*, a manuscript whose mathematical development is externally constrained by a molecular reference structure.

The complete review trajectory was reconstructed as a coded revision ledger. Sixty-four review/process actions were identified, of which 63 were manuscript-directed. Twenty-two revisions (34.9%) were classified as major, 20 (31.7%) as structural, and 59 (93.7%) were fully adopted. Forty-eight actions (76.2%) were explicitly verified in a subsequent review cycle, increasing to 55 (87.3%) when indirect later verification was included. The review disciplines generated markedly different correction profiles. Mathematics produced 35 of the 63 manuscript-directed actions and accounted for all coded proof/analytic repairs and theorem-strength corrections. Concordance generated nine actions, all of which were major and structural, including every action classified as structural restoration. Chemistry generated eight actions concentrated primarily in molecular qualification and nomenclature.

The pivotal event was the failure of a mathematically improved manuscript to pass concordance review, requiring substantial reconstruction before renewed chemistry and mathematics review. The case therefore suggests that the value of AI augmentation may lie not simply in increasing review volume, but in separating validation objectives, preserving reviewer disagreement as diagnostic information, distinguishing criticism from revision and verification, and maintaining an auditable record of manuscript development. Because this is a single retrospective case study without a conventional peer-review control group, the results do not establish the superiority of AI-augmented review. They instead provide a reproducible framework for prospective evaluation across future manuscripts.

Keywords: AI-augmented peer review; artificial intelligence; peer-review methodology; multidisciplinary review; mathematical verification; concordance review; chemistry review; meta-review; revision auditing; human–AI collaboration; scientific reproducibility; Chrysene Formalism

1 Introduction

Artificial-intelligence-assisted scientific writing creates an unusual asymmetry between the rate at which technical material can be generated and the rate at which it can be validated. A modern generative system can assist in constructing definitions, axioms, proofs, operator formulations, literature syntheses, and multidisciplinary arguments within a single drafting environment. That increase in generative capacity does not remove the traditional sources of scientific error. A mathematically sophisticated manuscript may still contain an ill-typed object,

an unstated hypothesis, a circular argument, an incorrect operator domain, or a theorem stronger than its proof. A formally consistent manuscript may also drift away from the scientific framework that it claims to develop. Conversely, a conceptually faithful mathematical construction may assign physical or chemical meaning to structures that the underlying system cannot realize.

These failure modes become especially important when a research program develops horizontally across several disciplines rather than vertically within one established specialty. In such work, graph theory, functional analysis, operator alge-

bras, noncommutative geometry, systems concepts, molecular chemistry, physical response, and application-specific interpretation may all occur within the same formal architecture. A favorable evaluation of one layer does not validate the others. Mathematical elegance cannot establish chemical realizability; chemical plausibility cannot repair an invalid theorem; and internal mathematical consistency does not by itself establish that a revised construction still belongs to the formal system it claims to extend.

The present study examines a review architecture designed around that distinction. We refer to the procedure as *AI-augmented peer review*. Its defining feature is not simply the use of artificial intelligence to generate additional reviewer reports. Rather, separate review stages are assigned deliberately different validity questions, revisions are recorded as distinct manuscript actions, and important repairs may be returned to later review for verification. The procedure was developed during the revision of a single multidisciplinary foundational manuscript and is evaluated here retrospectively from the resulting manuscript, review, revision, and meta-review record.

1.1 Validation by Orthogonal Review

The central methodological principle is *orthogonal review*. The term is used operationally rather than geometrically: different reviewers are instructed to interrogate different dimensions of manuscript validity instead of attempting to collapse all judgments into one general assessment.

For the case studied here, the three principal review questions were

M : Is the mathematics valid at the strength claimed?
 C : Is the construction concordant with the formal framework?
 H : Is the molecular interpretation chemically admissible?

(1)

A foundational statement intended simultaneously as mathematics, as a statement within the Chrysene Formalism, and as a molecularly grounded construction should therefore satisfy the intersection

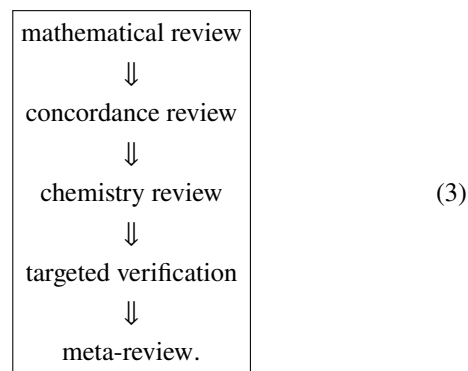
$$\mathcal{F}_{\text{adm}} = M \cap C \cap H. \quad (2)$$

The three conditions are not intended as substitutable components of a score. Failure of one cannot be compensated by exceptional performance in another.

Mathematical review asks whether objects are correctly typed, domains and codomains are defined, axioms are coherent, dependencies are non-circular, proofs are valid, and theorem statements do not exceed their hypotheses. Concordance review asks a different question: whether mathematical revision has preserved the defining architecture and purpose of the formal system being developed. A mathematically valid reformulation can therefore fail concordance if, in gaining generality or elegance, it replaces the structures that make the work part of the intended program. Chemistry review supplies an additional external constraint by asking whether claims assigned molecular meaning remain compatible with the reference molecule and whether formal existence

constructions are being distinguished from chemically realizable structures.

The complete review architecture used in the case study consequently took the form



The arrows indicate review progression rather than irreversible closure. A chemistry or concordance revision that changes mathematical content can return the manuscript to mathematical review. Likewise, the purpose of targeted verification is not to generate another independent global opinion, but to determine whether a previously identified defect has actually been repaired.

This distinction separates three states that are frequently combined in ordinary revision histories:

criticism $\neq$ revision $\neq$ verification.

(4)

A reviewer may correctly identify a problem that the author does not adopt; an apparently suitable revision may fail to correct the original defect; and a correction that was appropriate for one manuscript version may later be superseded during a larger reconstruction. Preserving these distinctions makes the review history itself an analyzable object rather than merely a collection of scores and comments.

1.2 The Chrysene Formalism as an Externally Anchored Test Case

The Chrysene Formalism provides an unusually stringent environment in which to examine this type of review. The broader program seeks to develop a foundational mathematical language capable of supporting application-specific mathematics across widely separated domains, including molecular therapeutics, clinical and biological systems, quantum information, energy conversion, optical response, finance, origins research, and cosmological modeling. The validity of any particular application remains an independent scientific question. The relevant methodological feature is that the formalism attempts to permit broad mathematical development while retaining a common foundational reference structure.

That reference is molecular. The formalism is grounded in selected features of the topology, symmetry, directional organization, locant architecture, and chemically admissible realization of chrysene. The molecule is not an “invariant” in the sense that every property is preserved under every mathematical operation.

Rather, it functions as an *externally anchored reference structure*: selected molecular properties remain fixed constraints against which proposed foundational extensions can be evaluated.

This distinction is important for AI-assisted mathematical development. An internally consistent formal system can be enlarged, generalized, or reorganized almost indefinitely. If the same generative process is also free to redefine the scientific object to which the mathematics refers, mathematical development can gradually become detached from the problem that originally constrained it. An external molecular reference limits that freedom. A new axiom, operator, or theorem may be mathematically abstract, but when it is presented as part of a molecularly grounded formalism its interpretation must remain compatible with the reference architecture or be stated explicitly as a realization hypothesis whose physical implementation has not yet been established.

The molecular anchor therefore contributes to falsifiability without serving as a substitute for proof. Chrysene does not determine that a particular crossed product, spectral triple, finite-field multiplication, transfer operator, or response tensor must exist. Instead, it restricts which foundational assertions may legitimately be attributed to the molecular system. Downstream claims can then fail at more than one level: a theorem can fail mathematically, a construction can fail concordance, or a molecular interpretation can fail chemically.

This feature proved important during the review process studied here. The governing principle that ultimately emerged was

$$\boxed{\begin{array}{l} \text{restore foundational structure} \\ \text{without restoring unsupported implication.} \end{array}} \quad (5)$$

For example, a native four-state structural alphabet could be restored without assuming a $GF(4)$ field multiplication; directional inequivalence could be foundational without implying transfer or optical anisotropy; and an admissibility boundary could be defined without treating the corresponding projection as a physical scission process. The molecular reference thus constrained what the formalism was about, while mathematical review constrained what could actually be proved from it.

The present paper does not claim that chrysene is uniquely capable of serving as such a reference structure, nor does it attempt to establish that the Chrysene Formalism should constitute a distinct new mathematical discipline. Those questions form a broader research problem. Here, the Chrysene Formalism is used more narrowly as a particularly informative case for studying whether external scientific anchoring can make AI-augmented multidisciplinary review more explicit, auditable, and falsifiable.

1.3 The Review of *The Chrysene Formalism Revisited*

The case study concerns the manuscript *The Chrysene Formalism Revisited: A Twelve-Axiom Foundation for Molecular Characterization, Tensorial Noncommutative Geometry, and Operator Response*. The manuscript was prepared with AI assistance from a pre-existing source corpus consisting of two monographs and

two earlier articles. It then underwent successive mathematical, concordance, chemical, targeted-verification, and meta-review cycles.

The trajectory was not a monotonic sequence of increasingly favorable scores. The first mathematically improved formulation subsequently failed concordance. Examination of that failure showed that, while the mathematical presentation had become increasingly rigorous, parts of the formalism had drifted away from foundational concepts in the source corpus. The manuscript therefore required substantial reconstruction before it could proceed. After concordance was restored, chemistry review generated additional corrections, the reconstructed manuscript returned to standalone mathematics review, specific mathematical repairs underwent targeted verification, and the complete subject-specific record was finally evaluated by independent meta-review.

This sequence makes the manuscript useful as more than an anecdotal example of repeated AI editing. It contains a documented instance in which success under one review criterion did not imply success under another, followed by a return to the first criterion after the second had materially changed the manuscript. The resulting record therefore permits the behavior of the review architecture to be examined retrospectively.

To do so, the complete review history was reconstructed into a coded revision ledger. The unit of analysis was a distinct review-driven manuscript action rather than a reviewer comment or a LaTeX code block. Each action was coded by review phase, disciplinary source, revision type, severity, structural versus local character, acceptance, supersession, and subsequent verification.

The resulting dataset contained 64 coded review/process actions, of which 63 were manuscript-directed. Twenty-two of the 63 manuscript-directed actions were classified as major and 20 as structural. Fifty-nine were fully adopted. Forty-eight were explicitly checked in a later review cycle, increasing to 55 when indirect later verification was included. More importantly, the revision types were not distributed uniformly across review disciplines. Mathematics generated 35 of the 63 manuscript-directed actions and accounted for all coded proof/analytic repairs and all theorem-strength corrections. Concordance generated nine actions, all of which were major and structural, including every action coded as structural restoration. Chemistry generated eight actions, concentrated primarily in molecular qualification and nomenclature.

These distributions provide the principal empirical motivation for the present paper. The apparent value of the layered process was not merely that additional reviewers produced additional comments. The different review stages detected measurably different classes of defect.

1.4 Objectives and Scope

The study therefore has five principal objectives:

1. to define a reproducible AI-augmented review architecture that separates mathematical validity, framework concordance, molecular admissibility, targeted verification, and

final meta-review;

2. to reconstruct the complete review trajectory of a complex multidisciplinary foundational manuscript as an auditable sequence of manuscript states, reviewer criticisms, revision actions, and verification outcomes;
3. to quantify the severity, structural character, adoption, and verification of the resulting review-driven revisions;
4. to determine whether mathematics, concordance, and chemistry generated redundant criticism or instead identified distinct classes of manuscript defect; and
5. to use the retrospective case study to formulate a prospective protocol in which future papers can record review, revision, and verification contemporaneously.

The study is intentionally limited in what it claims. It is not a controlled comparison between AI-augmented and conventional human peer review. It does not establish that reviewer scores are calibrated measures of manuscript quality, and it does not infer general performance characteristics from a single case. The coding itself is retrospective rather than prospectively randomized or independently blinded.

The narrower question is whether a layered review system can produce a traceable record in which distinct validity criteria generate distinct, actionable, and subsequently testable corrections. The Methods section defines the source corpus, review stages, coding rules, and descriptive analysis. The Results quantify the resulting revision record and compare correction profiles across review disciplines. The Discussion examines the pivotal concordance failure, reviewer disagreement, the distinction between revision and verification, the complementary roles of human and AI judgment, the function of an external reference structure, and the prospective review protocol that follows from the case.

2 Materials and Methods

2.1 Study Design and Source Corpus

The present study is a retrospective single-manuscript process study of an AI-augmented peer-review procedure. The object of analysis is the development and review of *The Chrysene Formalism Revisited: A Twelve-Axiom Foundation for Molecular Characterization, Tensorial Noncommutative Geometry, and Operator Response*. The purpose of the study is not to estimate the general performance of artificial-intelligence systems or to compare AI review statistically with conventional human peer review. Rather, the manuscript and its complete review trajectory are treated as a documented case through which the internal behavior of a layered review architecture can be reconstructed.

The manuscript was generated from a pre-existing Chrysene Formalism source corpus consisting of two monographs and two earlier articles. We denote this corpus by

$$\mathcal{S}_{\text{source}} = \{S_1^{\text{mon}}, S_2^{\text{mon}}, S_1^{\text{art}}, S_2^{\text{art}}\}. \quad (6)$$

These documents supplied the previously stated molecular, mathematical, operator-theoretic, and structural concepts against which later concordance could be assessed. They were not treated as proof that the new manuscript was mathematically correct. Their role was instead to define the historical and conceptual framework that the new manuscript claimed to reconstruct and extend.

AI assistance was used in preparing the initial manuscript from this source material. That manuscript was then subjected to successive review and revision cycles. For purposes of retrospective analysis, only manuscript states corresponding to meaningful review transitions were retained. Minor compilations involving page layout, line breaking, column fitting, equation wrapping, or other purely typesetting changes were not treated as distinct scientific manuscript states.

A manuscript state is denoted by

$$M_0, M_1, \dots, M_f, \quad (7)$$

where M_0 denotes the initial review manuscript and M_f denotes the final controlling manuscript presented to meta-review. Intermediate states correspond to revisions following substantive mathematical, concordance, chemistry, or verification review.

2.2 Manuscript-State and Review Provenance

The review process was reconstructed as a sequence of manuscript, review, and revision events rather than as a single final editorial decision. Schematically, a review transition has the form

$$M_i \longrightarrow R_{i,j} \longrightarrow \Delta_{i,j} \longrightarrow M_{i+1}, \quad (8)$$

where $R_{i,j}$ denotes the j -th review applied to manuscript state M_i , and $\Delta_{i,j}$ denotes the corresponding set of review-driven revision actions. More than one independent review could be applied to the same manuscript state, and a later review could require return to a previously completed disciplinary layer.

The review record was preserved chronologically. Earlier review reports were not rewritten after later corrections had been made. This distinction was particularly important for meta-review: a criticism appearing in an earlier report was treated as evidence that a defect had occurred, but it was not counted as a defect of the final manuscript if the later text had repaired it.

Whenever possible, independent first-pass reviewers were instructed not to rely upon scores, comments, or conclusions produced by other reviewers. Numeric scores were therefore treated as reviewer-specific diagnostic summaries rather than measurements to be averaged into a single estimate of manuscript quality. Where a 100-point rubric was used, ten review categories were each scored on a 0–10 scale. The resulting total was retained as part of the review record, but editorial decisions were based upon the substantive criticisms and their disposition rather than upon an arithmetic mean of scores.

The exact review prompts, requested outputs, and scoring rubrics were retained as part of the methodological record. This permits the review criteria themselves to be inspected independently of the resulting reviewer reports.

2.3 Standalone Mathematical Review

Mathematical review was designed to answer a deliberately narrow question:

Does the manuscript establish mathematically what it claims to establish? (9)

The mathematics reviewers were instructed to treat the manuscript as a standalone mathematical submission. Successful chemistry or concordance review could not be used as evidence that a mathematical argument was correct, and authorial intent could not be used to repair a missing definition, domain condition, hypothesis, or proof step.

The mathematical audit included, as applicable,

1. definitions, object types, domains, codomains, and quantifiers;
2. logical dependency among definitions, axioms, propositions, and theorems;
3. circularity and hidden use of theorem conclusions as assumptions;
4. finite graph and polyhex classification;
5. molecular-characterization theorem strength;
6. tensor-network and $\mathbb{Z}^3$ -fibration mathematics;
7. quasi-local C^* -algebras and crossed products;
8. GNS, von Neumann, commutant, and opposite-algebra constructions;
9. spectral triples, compact resolvent, domain invariance, and metric constructions;
10. Schur-complement, transfer, linear-response, and admissibility mathematics; and
11. existence and non-vacuity of the complete axiomatic system.

Particular attention was given to distinguishing statements of different logical status. A definition was not treated as a theorem, an axiom was not credited as a derived consequence, a realization hypothesis was not treated as universally proved, and a deferred theorem target was not penalized merely because it remained unproved. Conversely, any statement represented as an established theorem was required to follow from the stated axioms and hypotheses.

The study contained an early developmental mathematics-review cycle and a later formal standalone mathematics-review cycle. The latter was conducted after the concordance reconstruction and chemistry review so that the final mathematical architecture could be examined again independently from first principles. When mathematical reviewers identified a repairable problem, explicit replacement text, definitions, equations, proof steps, or LaTeX insertions were generated where appropriate. Adoption of a suggested replacement was recorded separately from the fact that the criticism had originally been made.

2.4 Concordance Review

Concordance review was introduced to test a question not answered by mathematical correctness:

Does the revised construction remain the formal system that it claims to express? (10)

The concordance reviewer compared the manuscript with S_{source} and examined whether structures central to the earlier Chrysene Formalism had been preserved, rigorously reconstructed, inadvertently removed, or replaced by mathematically convenient substitutes.

The concordance audit therefore examined the status of, among other features, the chrysene molecular architecture, the literal Chrysene Tensor Space, the $\mathbb{Z}^3$ organization, sequence-dependent noncommutativity, the $3, 6, 9, 12 \leftrightarrow 2, 5, 8, 11$ dual cipher, the native four-state structure, and noncommutative spectral geometry. Concordance did not require restoration of every strong historical implication. Instead, the governing rule was

restore foundational structure
without restoring unsupported implication. (11)

Thus, a foundational object could be restored while a stronger claim formerly associated with that object could be demoted to a theorem target requiring additional hypotheses. Examples include distinguishing a native four-state alphabet from a proven $GF(4)$ field multiplication, geometric anisotropy from transfer anisotropy, and an admissibility projection from a physical scission process.

Concordance review was also the stage in which human editorial judgment was assigned its strongest interpretive role. A reviewer could identify apparent drift, but determining whether a mathematically improved construction had ceased to serve the intended conceptual program required examination of both the source corpus and the stated direction of the journal. Human intervention was therefore permitted to accept, reject, refine, or broaden an AI-generated concordance criticism before a major reconstruction was undertaken.

2.5 Chemistry Review

Chemistry review was conducted independently from the proof audit. Its primary question was

Are statements given molecular meaning chemically correct and appropriate? (12)

The chemistry reviewers were not asked to determine whether an operator-algebraic proof was valid. Their task was instead to examine the molecular, structural, symmetry, aromaticity, and physical-chemistry statements on which the mathematical interpretation depended.

The audit included:

1. chrysene molecular identity and connectivity;
2. standard locant numbering and primary/dual perimeter assignments;

3. point-group symmetry, achirality, and substitution-dependent symmetry reduction;
4. polyhex and benzenoid interpretation;
5. bipartite and balanced-bipartite statements;
6. bay-region and K-region nomenclature;
7. aromatic and conjugated- π language;
8. distinction between abstract bridges and chemically realizable linkers;
9. separation of molecular geometry from electronic-transfer conclusions; and
10. polarization and response statements when they were assigned physical interpretation.

A central methodological rule was that an abstract mathematical existence witness need not itself constitute a chemically synthesizable molecule. It was therefore acceptable for an existence model to use minimal or degenerate mathematical data, provided that the manuscript explicitly labeled those objects as formal witnesses and did not represent them as specific chemical linkers, reactions, or experimentally realized structures.

Likewise, chemically plausible intuition was not allowed to convert a realization hypothesis into an axiom. Molecular statements concerning conjugation, stacking, transfer, or response were required to be assigned a logical status consistent with the available evidence.

2.6 Targeted Post-Revision Verification

A full peer review was not automatically repeated after every revision. When a reviewer had identified a limited set of substantive defects and those defects had been specifically repaired, the manuscript could instead undergo targeted verification.

The targeted verifier received the revised manuscript together with the defined list of repairs requiring confirmation. The task was not to generate a new global score or search for an entirely new theory of the paper. It was to determine, item by item,

criticism $\longrightarrow$ implemented repair $\longrightarrow$ verified resolution. (13)

A repair was accepted only if the revised text corrected the original problem without introducing a new inconsistency elsewhere in the axiomatic system.

For the case examined here, targeted mathematical verification included the molecular point-group criterion, canonical versus realization-specific geometry, graph/node typing, theorem strength, bridge compatibility, lattice registry transport, unbounded-operator domain conditions, self-adjointness, crossed-product conventions, Schur-complement adjoint structure, and the explicit existence model. The verification stage also distinguished substantive defects from residual wording or copyediting issues.

2.7 Final Meta-Review

The final review stage was a meta-review of the complete subject-specific record. The meta-reviewer received:

1. the final manuscript;
2. the mathematics reviews;
3. the concordance-review record;
4. the chemistry reviews; and
5. the targeted post-revision verification.

The principal meta-review rule was that the final manuscript was the controlling document. Earlier reviews were evidence concerning problems previously identified, but they were not substitutes for examining the revised text. Thus, a criticism was not carried forward merely because it had once appeared in a review report.

The meta-review was instructed not to average previous numerical scores. Instead, it performed a resolution audit. Material criticisms were classified according to whether they were

$$\left\{ \begin{array}{l} \text{resolved,} \\ \text{partially resolved,} \\ \text{unresolved,} \\ \text{no longer applicable.} \end{array} \right. \quad (14)$$

The meta-review also examined whether the different disciplinary reviews contradicted one another, whether any central theorem remained unsupported, whether the chemistry and mathematics were being used at their appropriate logical levels, and whether the Abstract, axioms, principal theorems, appendices, Discussion, and Conclusion stated mutually compatible claims.

Remaining issues were further classified as blocking, minor, or editorial. A recommendation for publication therefore required not that the manuscript be incapable of further improvement, but that no unresolved issue remained that invalidated its principal claims within the scope stated by the paper.

2.8 Construction of the Retrospective Revision Ledger

After completion of the review process, the manuscript history was reconstructed into a coded revision ledger. The unit of analysis was a *review-driven revision action*. A revision action was defined as one distinct insertion, replacement, deletion, redefinition, equation correction, proof repair, theorem-scope correction, logical-status correction, or related manuscript change proposed in response to review.

The unit was intentionally not defined as a LaTeX code block. A single conceptual repair could require several insertions at different locations, while one replacement block could simultaneously resolve more than one local wording defect. The coding therefore attempted to identify the underlying review action rather than the mechanics of text insertion.

Purely typographical interventions were excluded from the scientific revision count. Examples excluded from the ledger included:

- resizing or line-breaking equations for the final two-column layout;
- moving tables to satisfy column constraints;
- filling white space produced by journal pagination;
- changing display formatting without changing mathematical content; and
- other post-review page-composition adjustments.

Each coded action contained the following fields:

$$\mathcal{R}_i = (\text{phase, discipline, revision type, severity, structural/local, accepted, superseded, verified}), \quad (15)$$

together with free-text fields identifying the original criticism, recommended change, whether a LaTeX patch was generated, the resulting manuscript stage, and the later verification outcome.

Phase. Actions were assigned to the review stage in which the operative criticism or revision occurred. The coded stages were:

1. preliminary mathematics development;
2. concordance failure and reconstruction;
3. concordance and molecular qualification;
4. independent chemistry review;
5. formal standalone mathematics review;
6. targeted mathematical verification; and
7. meta-review and final editorial cleanup.

Discipline. The primary source of the correction was coded as mathematics, concordance, chemistry, mixed, editorial, or publication-process. “Mixed” was used where a revision simultaneously served two substantive review domains and assigning it exclusively to either would misrepresent its purpose.

Revision type. Revision types included axiom/definition, proof/analytic, scope/qualification, typing/notation, editorial/reference, structural restoration, theorem/claim, dependency/architecture, chemistry/nomenclature, existence model, and publication process.

Severity. A revision was coded as *major* when it altered a foundational definition, principal theorem, logical architecture, existence argument, or substantial part of the formal system. A *minor* revision corrected a substantive but bounded defect without requiring architectural reconstruction. *Editorial* actions affected wording, notation, references, or presentation without altering a substantive result. A separate *process* category was retained for non-manuscript publication safeguards.

Structural versus local. A structural action changed the architecture, scope, foundational objects, theorem hierarchy, or organization of the formal system. A local action modified a bounded definition, equation, proof step, qualification, terminology choice, or editorial passage.

Acceptance. Recommendation disposition was coded as fully accepted, partially accepted, not accepted, optional, conditional, or not applicable. An optional suggestion was not counted as a rejected recommendation when the author chose not to implement it.

Supersession. A revision was marked superseded when a later reconstruction replaced the earlier repair rather than merely refining or extending it. This allowed early improvements to remain visible in the historical record without implying that their original wording survived into the final manuscript.

Verification. A repair was coded as explicitly verified when a later reviewer directly examined the implemented correction. It was coded as indirectly verified when subsequent review accepted the relevant section or construction without an item-specific recheck. Final copyediting performed after meta-review was recorded separately as having no additional review cycle.

2.9 Descriptive Statistical Analysis

The coded ledger was analyzed descriptively. Because the study contains a single manuscript and the revision events are neither independent samples nor observations from a defined population, inferential hypothesis testing was not used. Counts and percentages were instead calculated to characterize the behavior of the review process.

The primary denominator for manuscript-revision statistics was

$$N_{\text{man}} = \# \{ \mathcal{R}_i : \mathcal{R}_i \text{ is manuscript-directed} \}. \quad (16)$$

Publication-process recommendations that did not alter the manuscript were excluded from this denominator.

The structural-revision fraction was calculated as

$$f_{\text{str}} = \frac{N_{\text{structural}}}{N_{\text{man}}}, \quad (17)$$

with

$$f_{\text{local}} = 1 - f_{\text{str}}. \quad (18)$$

Full adoption was calculated as

$$f_{\text{adopt}} = \frac{N_{\text{accepted=yes}}}{N_{\text{man}}}. \quad (19)$$

Because optional and conditional suggestions are not equivalent to definite revision requests, a second adoption statistic was calculated over definite recommendations:

$$f_{\text{adopt}}^{\text{def}} = \frac{N_{\text{yes}}}{N_{\text{yes}} + N_{\text{partial}} + N_{\text{no}}}. \quad (20)$$

Table 1: Overall characteristics of the retrospectively coded manuscript-revision record. Percentages use the 63 manuscript-directed actions as denominator except where the 61 definite recommendations are explicitly indicated.

Measure	n/N	%
Major revisions	22/63	34.9
Minor revisions	34/63	54.0
Editorial revisions	7/63	11.1
Structural revisions	20/63	31.7
Local revisions	43/63	68.3
Fully adopted	59/63	93.7
Explicitly verified	48/63	76.2
Explicit or indirect verification	55/63	87.3
Full adoption, definite recommendations	59/61	96.7
Full or partial adoption, definite recommendations	60/61	98.4

The complete coded dataset contains 64 review/process actions. One action concerned an external publication-process safeguard rather than a manuscript revision and is excluded from the $N_{\text{man}} = 63$ denominator. “Definite recommendations” exclude one optional and one conditional recommendation.

The corresponding full-or-partial adoption rate was

$$f_{\text{adopt+partial}}^{\text{def}} = \frac{N_{\text{yes}} + N_{\text{partial}}}{N_{\text{yes}} + N_{\text{partial}} + N_{\text{no}}}. \quad (21)$$

The explicit verification rate was defined by

$$f_{\text{ver}} = \frac{N_{\text{verified=yes}}}{N_{\text{man}}}, \quad (22)$$

while broader verification coverage included later indirect support:

$$f_{\text{cov}} = \frac{N_{\text{verified=yes}} + N_{\text{verified=indirect}}}{N_{\text{man}}}. \quad (23)$$

Revision counts were then cross-tabulated by review phase, discipline, severity, structural/local status, and revision type. Particular attention was given to whether the disciplinary reviews generated similar or different correction profiles. For a review discipline d and revision class t , the cross-tabulated count was

$$N_{d,t} = \# \{ \mathcal{R}_i : \text{discipline}(\mathcal{R}_i) = d, \text{type}(\mathcal{R}_i) = t \}. \quad (24)$$

This permitted examination of whether proof repairs, theorem-strength corrections, structural restorations, chemical qualifications, and editorial changes were distributed uniformly across the review system or concentrated in particular review layers.

The resulting statistics are intended as descriptive properties of this case study. They are not estimates of the expected performance of AI-augmented peer review in other journals or disciplines. The retrospective analysis is instead used to determine whether the review architecture generated an auditable sequence of distinct, adopted, and subsequently verified corrections, and to identify which parts of that architecture were responsible for different classes of manuscript change.

2.10 Prespecification of the Descriptive Results

Before drafting the narrative Results section, the coded ledger was used to fix the principal quantitative objects to be reported.

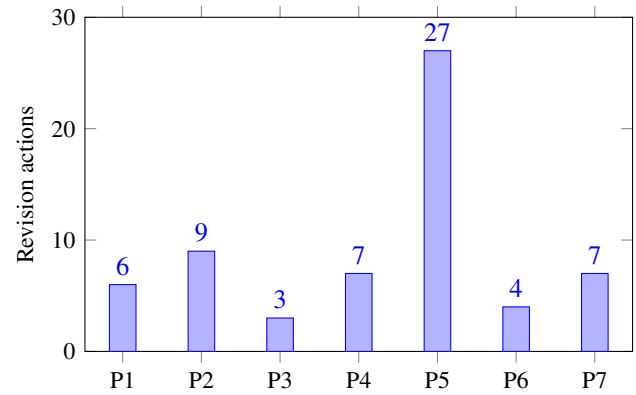


Figure 1: Distribution of the 63 manuscript-directed revision actions across review stages. P1, preliminary mathematics development; P2, concordance failure and reconstruction; P3, concordance and molecular qualification; P4, independent chemistry review; P5, formal standalone mathematics review; P6, targeted mathematical verification; and P7, meta-review and final editorial cleanup.

Table 1 records the overall revision burden, adoption, and verification rates; Table 2 records the distribution of revision actions across review stages; and Table 3 cross-tabulates correction type against the primary disciplinary source of review. Figures 1 and 2 provide graphical summaries of review-stage activity and structural revision intensity, respectively.

These tables and figures were specified from the completed coded ledger before the narrative interpretation of the Results was drafted. This ordering was used to reduce the possibility that categories, denominators, or graphical emphasis would be altered retrospectively to support a preferred narrative. Subsequent Results text therefore describes these fixed descriptive objects rather than determining their numerical content.

3 Results

The retrospective ledger contained 64 coded review/process actions. One action concerned a publication-process safeguard rather than a change to the manuscript and was therefore excluded from the primary revision denominator. The quantitative analysis consequently used

$$N_{\text{man}} = 63 \quad (25)$$

manuscript-directed revision actions. The following results describe the distribution, severity, structural character, adoption, and verification of those actions. Because the study concerns one manuscript and one reconstructed review trajectory, the results are descriptive rather than inferential.

3.1 Overall Revision Burden, Adoption, and Verification

The review process generated a substantial volume of substantive revision. Of the 63 manuscript-directed actions, 22 were classified as major, 34 as minor, and 7 as editorial, corresponding

Table 2: Distribution and character of manuscript-directed revision actions across review stages. The structural column identifies changes to foundational architecture, theorem scope, or organization rather than bounded local corrections.

Review stage	Actions	% of 63	Major	Structural	Explicitly verified
P1. Preliminary mathematics development	6	9.5	3	3	0
P2. Concordance failure and reconstruction	9	14.3	9	9	9
P3. Concordance and molecular qualification	3	4.8	0	0	3
P4. Independent chemistry review	7	11.1	0	0	7
P5. Formal standalone mathematics review	27	42.9	10	8	25
P6. Targeted mathematical verification	4	6.3	0	0	4
P7. Meta-review and final editorial cleanup	7	11.1	0	0	0
Total	63	100.0	22	20	48

The absence of explicit verification in P1 does not imply that the early revisions were unsupported; several received indirect support from later review but were subsequently superseded during the major concordance reconstruction. P7 consists principally of corrections generated by the final meta-review and therefore had no subsequent review cycle within the present study.

Table 3: Cross-tabulation of revision type by primary review discipline. This table provides the principal quantitative test of whether the review layers generated redundant or differentiated correction profiles.

Revision type	Math.	Concord.	Chem.	Mixed	Editorial	Total
Axiom / definition	8	0	1	0	0	9
Proof / analytic	9	0	0	0	0	9
Theorem / claim	6	0	0	0	0	6
Typing / notation	6	0	0	0	2	8
Existence model	2	0	0	0	0	2
Dependency / architecture	2	1	0	0	0	3
Structural restoration	0	7	0	0	0	7
Scope / qualification	1	1	4	3	0	9
Chemistry / nomenclature	0	0	3	0	0	3
Editorial / reference	1	0	0	1	5	7
Total	35	9	8	4	7	63

to 34.9%, 54.0%, and 11.1%, respectively (Table 1). Thus, nearly nine tenths of the recorded actions altered mathematical, scientific, or logical content rather than serving only an editorial function.

The distinction between structural and local revision produced a different view of the same record. Twenty actions (31.7%) were classified as structural and 43 (68.3%) as local. Structural actions included changes to foundational objects, theorem scope, dependency architecture, or the organization of the formal system. Local actions corrected bounded definitions, equations, proof steps, terminology, qualifications, or editorial passages. The majority of corrections were therefore local, but approximately one third of the complete revision record altered the structure or scope of the formalism itself.

Reviewer recommendations were adopted at a high rate. Fifty-nine of the 63 manuscript-directed actions were fully incorporated, giving an overall full-adoption rate of

$$\frac{59}{63} = 93.7\%. \quad (26)$$

Because the ledger contained one optional and one conditional recommendation, a second calculation was performed over the 61 definite recommendations. Fifty-nine of these were fully adopted (96.7%), and 60 were fully or partially adopted (98.4%). Thus, almost every definite review recommendation produced a corresponding manuscript change.

Revision and verification were treated separately. Forty-eight of the 63 manuscript-directed actions (76.2%) were explicitly examined in a later review cycle. When later indirect support was also counted, 55 actions (87.3%) had subsequent verification coverage. The review record therefore contained not only a large number of corrections but, for most technical corrections, a later stage capable of determining whether the requested repair had actually been implemented.

3.2 Distribution of Revision Actions Across Review Stages

Revision activity was unevenly distributed across the seven coded review stages (Figure 1; Table 2). The formal standalone mathematics review generated the largest number of manuscript-directed actions:

$$27/63 = 42.9\%. \quad (27)$$

The concordance failure and reconstruction generated 9 actions (14.3%), while the independent chemistry review and final meta-review/editorial stage each generated 7 (11.1%). Preliminary mathematics development generated 6 actions (9.5%), targeted mathematical verification generated 4 (6.3%), and concordance-associated molecular qualification generated 3 (4.8%).

The stage producing the largest number of revisions was not, however, the stage producing the greatest structural concentration.

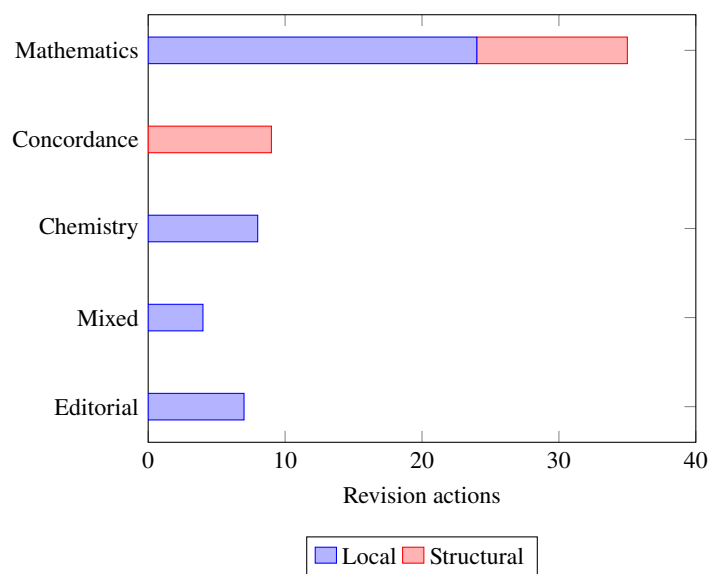


Figure 2: Structural versus local manuscript revisions by primary review discipline. Concordance had the highest structural intensity: all nine concordance actions were structural. Mathematics generated the largest absolute number of revisions but contained both structural and local corrections. Chemistry, mixed qualification, and final editorial review generated only local changes in the coded case record.

The concordance-reconstruction stage contained nine revision actions, and all nine were classified as both major and structural:

$$\frac{9}{9} = 100\% \quad \text{major and structural.} \quad (28)$$

These revisions restored or reconstructed foundational components of the formalism rather than correcting isolated passages. By contrast, the formal standalone mathematics stage generated 27 actions, of which 10 were major and 8 structural. Its larger absolute revision burden therefore reflected a mixture of architectural corrections and a greater number of local mathematical repairs.

The verification pattern also differed by stage. All nine concordance-reconstruction actions, all three molecular-qualification actions, all seven chemistry-review actions, and all four targeted-verification actions were subsequently explicitly checked. Within the formal standalone mathematics stage, 25 of 27 actions were explicitly verified. The preliminary mathematics-development cycle had no action-specific explicit verification in the final coded scheme, although several of its corrections received later indirect support or were replaced during the subsequent concordance reconstruction. The seven final meta-review/editorial actions had no additional review cycle because they occurred at the end of the recorded process.

3.3 Discipline-Specific Correction Profiles

When the 63 manuscript-directed actions were classified by the primary disciplinary source of the criticism, mathematics

accounted for the largest fraction:

$$\frac{35}{63} = 55.6\%. \quad (29)$$

Concordance accounted for 9 actions (14.3%), chemistry for 8 (12.7%), mixed review for 4 (6.3%), and final editorial review for 7 (11.1%).

The more important result was not the difference in absolute counts but the distribution of *types* of correction across the review disciplines (Table 3). Mathematical review generated all nine coded proof/analytic corrections and all six theorem/claim corrections. It also generated eight of the nine axiom/definition changes, six of the eight typing/notation changes, both explicit-existence-model revisions, and two of the three dependency/architecture revisions. Mathematical review therefore concentrated strongly on proof validity, formal typing, theorem strength, and analytic realization.

Concordance displayed a markedly different correction profile. Seven of its nine actions were classified as *structural restoration*, and no other discipline generated an action in that category. Its remaining two actions concerned dependency/architecture and scope/qualification. The principal concordance contribution was therefore not incremental proof repair but restoration of foundational structures that had been weakened or lost during prior mathematical reformulation.

Chemistry produced a third correction profile. Four of the eight chemistry-coded actions concerned scope or qualification of molecular claims, three concerned chemistry or nomenclature directly, and one concerned an axiom/definition distinction associated with molecular symmetry. Chemistry produced no proof/analytic, theorem-strength, or structural-restoration actions. Its role in the observed review record was primarily to constrain molecular interpretation and ensure that mathematical abstractions were not assigned chemical meanings stronger than the molecular evidence supported.

The mixed-review actions were similarly concentrated in scope/qualification, whereas the final editorial layer was dominated by wording, references, and notation. The resulting cross-tabulation therefore shows that the review layers did not simply reproduce the same class of criticism with different disciplinary terminology. Instead, distinct review objectives were associated with distinct patterns of manuscript correction.

3.4 Structural Intensity of the Review Disciplines

The structural/local coding further distinguished the review functions (Figure 2). Of the 35 mathematics-generated actions, 11 were structural and 24 were local, so the structural fraction within mathematical review was

$$\frac{11}{35} = 31.4\%. \quad (30)$$

Mathematics therefore contributed substantially to architectural change while still producing predominantly local corrections.

Concordance differed qualitatively. All nine concordance actions were structural:

$$\frac{9}{9} = 100\%. \quad (31)$$

No chemistry, mixed, or editorial action was classified as structural. All eight chemistry actions, all four mixed actions, and all seven editorial actions were local. Consequently, the entire set of 20 structural revisions in the manuscript was generated by only two disciplinary sources:

$$20 = 11_{\text{mathematics}} + 9_{\text{concordance}}. \quad (32)$$

Severity showed a comparable concentration. Mathematics generated 13 major actions and concordance generated 9; together they accounted for all 22 major manuscript revisions. Every concordance action was major, whereas 13 of 35 mathematics actions were major. Chemistry, mixed, and editorial review generated no action coded as major.

These results indicate that revision volume and structural impact were not interchangeable measures. Mathematics produced more than three times as many actions as concordance, yet the concordance stage produced a higher concentration of foundational change. The retrospective record therefore distinguishes at least two forms of review intensity: the number of defects identified and the degree to which correcting those defects altered the architecture of the manuscript.

3.5 Empirical Differentiation of Review Functions

The combined results provide a descriptive test of the orthogonal review design introduced in Section 1.1. If the review layers were functionally redundant, similar classes of corrections would be expected to recur across mathematics, concordance, and chemistry. The observed record instead showed marked concentration of different revision classes within different review domains.

In particular,

$$\begin{aligned} \text{mathematics} &\longrightarrow \text{proof, typing, theorem strength, analytic validity,} \\ \text{concordance} &\longrightarrow \text{foundational structural restoration,} \\ \text{chemistry} &\longrightarrow \text{molecular qualification and nomenclature.} \end{aligned} \quad (33)$$

This pattern was not absolute: mathematics also produced structural changes, concordance produced one scope correction, and chemistry produced one axiom/definition correction. Nevertheless, the dominant correction profiles were sufficiently differentiated that each review layer contributed information not reproduced by the others.

The concordance stage provides the clearest example. The manuscript had already undergone mathematical development before concordance was applied, yet concordance generated nine additional actions, all of which were major and structural. Subsequent review did not reverse those restorations; rather, the reconstructed structures were later subjected to chemistry and

renewed mathematics review. Thus, success under mathematical review did not eliminate the possibility of a different class of framework-level failure.

Conversely, successful concordance did not eliminate the need for later mathematical scrutiny. The final standalone mathematics cycle generated 27 additional corrections after the conceptual framework had been restored. Twenty-five of these were subsequently explicitly verified. The observed process therefore behaved as a layered system in which different review objectives remained capable of detecting different defects at different stages of the same manuscript.

Taken together, the quantitative record supports three descriptive conclusions. First, AI-augmented review generated a large and traceable body of manuscript correction rather than merely a set of final reviewer scores. Second, the disciplinary stages displayed different correction profiles, with mathematics dominating proof and formal validity, concordance dominating structural restoration, and chemistry dominating molecular qualification. Third, most technical repairs were either explicitly or indirectly examined in a later review cycle. The implications of these findings for reviewer disagreement, human editorial intervention, external reference constraints, and prospective AI-augmented review are considered in the following section.

4 Discussion

The present case study was designed to examine whether a multidisciplinary manuscript could be reviewed more effectively by separating different dimensions of validity rather than asking a single review process to adjudicate them simultaneously. The retrospective record supports that distinction. Mathematical review, concordance review, and chemistry review generated markedly different correction profiles. Mathematics produced the largest number of revision actions and was responsible for essentially all proof, analytic, typing, and theorem-strength repairs. Concordance produced fewer actions but the highest concentration of architectural change: all nine actions in the principal concordance-reconstruction cycle were classified as both major and structural. Chemistry, by contrast, primarily constrained molecular terminology, interpretation, and claim scope.

The importance of this result lies less in the absolute number of corrections than in their distribution. If the three review layers had merely reproduced one another's judgments, a staged architecture would add procedural complexity without adding substantial information. Instead, the observed correction profiles were differentiated in a manner consistent with the objectives assigned to the reviewers. The case therefore provides preliminary evidence that explicitly separating review objectives can expose failure modes that might otherwise remain hidden within an undifferentiated assessment of overall manuscript quality.

At the same time, the present study does not establish that AI-augmented review is generally superior to conventional peer review. It concerns one manuscript, one evolving formal system, and one retrospectively reconstructed review trajectory. Its

value is therefore methodological rather than comparative: it demonstrates a review architecture that produced an auditable sequence of criticisms, revisions, and later verification, and it identifies several features that can be tested prospectively in future manuscripts.

4.1 Concordance Failure as the Pivotal Review Event

The most consequential event in the review history was not a mathematical rejection or a chemical error. It was the failure of a mathematically improved manuscript to satisfy concordance with the formal system that it claimed to reconstruct. This event provides the clearest illustration of why mathematical correctness and framework identity cannot be treated as interchangeable review criteria.

Prior to concordance review, substantial effort had been directed toward making the manuscript mathematically precise. Definitions were formalized, logical distinctions were sharpened, and stronger mathematical structures were introduced. Yet the concordance review showed that this process had also weakened or removed concepts that were central to the earlier Chrysene Formalism. The resulting manuscript was becoming a more self-contained mathematical theory while simultaneously becoming less faithful to the system that had motivated it.

Human editorial examination of the concordance result supported that diagnosis. The problem was not that the mathematics had become too advanced. Rather, mathematical optimization had begun to determine the direction of the framework instead of serving it. The manuscript was therefore substantially reconstructed so that the improved mathematics was redirected toward the foundational architecture of the Chrysene Formalism.

The reconstruction restored, among other elements, a literal tensor space, the $\mathbb{Z}^3$ organization, the primary/dual cipher, sequence-dependent noncommutativity, and spectral noncommutative-geometric structure. Importantly, this restoration did not reinstate every strong implication appearing in earlier formulations. The revised principle was instead

$$\boxed{\begin{array}{l} \text{restore foundational structure} \\ \text{without restoring unsupported implication.} \end{array}} \quad (34)$$

Thus, the four-state alphabet could be foundational while a $GF(4)$ field multiplication remained deferred; directional inequivalence could be foundational while transfer anisotropy remained a theorem target; and an admissibility boundary could be defined without treating it as autonomous physical scission.

The concordance episode can therefore be viewed as a natural within-case test of the layered review hypothesis. A manuscript that had already benefited from mathematical review nevertheless contained a different class of defect that mathematical review had not been designed to identify. The later return to mathematics then showed the converse: successful concordance did not imply that all remaining mathematical defects had been removed. The two review functions were therefore complementary rather than hierarchical.

4.2 Reviewer Disagreement as Diagnostic Information

The formal standalone mathematics cycle also demonstrated the value of retaining disagreement among independent reviewers. The three mathematics assessments differed substantially in severity, producing overall scores of

$$81, \quad 88, \quad 100. \quad (35)$$

An arithmetic average of these values would have obscured the most important information in the reviews. The strictest reviewer identified problems involving graph/node typing, $O(3)$ -conjugacy, theorem strength, compact-resolvent realization, Schur-complement adjoints, and stale finite-classification and analytic-existence language. Several of these criticisms produced material changes to the final manuscript.

The favorable review was therefore not treated as evidence that the stricter review must have been incorrect, and the stricter review was not treated as a permanent indictment after the manuscript had been repaired. Instead, disagreement triggered examination of the underlying claims. Once the identified defects had been corrected, the strict reviewer was asked to perform a targeted verification of those repairs rather than to repeat the entire review from the beginning.

This suggests a different interpretation of inter-reviewer variance in AI-augmented review. Reviewer disagreement need not be reduced immediately to consensus. It can function as a search mechanism for regions of the manuscript in which reviewer assumptions, standards, or levels of scrutiny differ. In such a system, the relevant question is not

Which reviewer supplied the “correct” score?

but rather

Which criticisms survive direct examination of the text?

The final manuscript, rather than the prestige or severity of any particular reviewer, becomes the object that resolves the disagreement.

This approach also explains why numeric scores were not averaged in the meta-review. Scores summarized reviewer judgments at particular manuscript states; they did not constitute calibrated measurements of an invariant quantity. A score assigned before a defect was repaired should not be interpreted as a score of the final manuscript, while a high score that overlooked a genuine defect should not prevent that defect from being corrected.

4.3 Criticism, Revision, and Verification as Distinct Stages

The review history also supports the distinction introduced in Equation (36):

$$\text{criticism} \neq \text{revision} \neq \text{verification}. \quad (36)$$

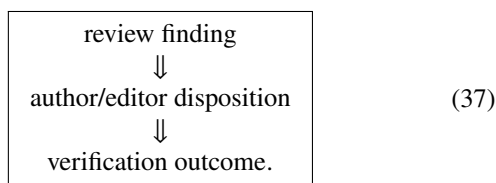
A reviewer comment has limited methodological value if it is not translated into an identifiable manuscript action. Likewise, the existence of a revised passage does not demonstrate that the revision actually resolved the original problem.

The retrospective ledger therefore recorded recommendation disposition independently from later verification. Fifty-nine of the 63 manuscript-directed actions were fully adopted. Among the 61 definite recommendations, 59 were fully adopted and 60 were fully or partially adopted. These high adoption rates indicate that the review process materially altered the manuscript rather than merely generating commentary.

More importantly, 48 of the 63 manuscript-directed actions were later verified explicitly, and 55 had either explicit or indirect verification coverage. This additional layer is especially relevant for AI-assisted revision because an AI system can generate a technically plausible replacement that appears locally correct while introducing a new inconsistency elsewhere. Returning the revised material to a later reviewer therefore serves a function analogous to regression testing in software or independent checking in formal mathematics.

The targeted-verification stage was particularly efficient in this respect. Once the substantive issues had been identified and repaired, the strict reviewer was not asked to regenerate another complete 100-point assessment. Instead, it was instructed to examine the specific repairs and determine whether they had been implemented correctly and without introducing new defects. This reduced redundant review while preserving accountability for the original criticisms.

A prospective implementation of the review system should retain this separation explicitly. Each recommendation should have at least three possible records:



Such a record permits later investigators to determine not merely what reviewers said, but which criticisms actually changed the manuscript and whether those changes survived subsequent examination.

4.4 Complementary Roles of Human and AI Review

The case study does not support a simple replacement model in which AI review substitutes for human editorial judgment. Instead, the review history suggests a division of labor.

AI review was particularly useful for tasks that benefited from repetition, formal decomposition, and narrow technical instructions. These included checking object types, identifying stale notation, testing theorem strength, examining proof domains, comparing multiple axiom layers, generating replacement LaTeX, conducting independent chemistry audits, and returning to previously identified defects after revision. The ability to

conduct several independent reviews also made it practical to expose disagreement rather than forcing a single reviewer to function as the sole gatekeeper.

Human judgment became most important when the question was not whether a formula was valid, but what role that formula should play in the research program. The concordance failure illustrates this point. An AI reviewer could identify that foundational concepts from the source corpus had been weakened or removed. Determining whether those differences represented desirable mathematical abstraction or unacceptable departure from the intended formalism required editorial interpretation of the program itself.

Human review was also required to decide which AI recommendations to adopt, which to reject, which to treat as optional, and when a localized repair was no longer sufficient. The review architecture should therefore be described as *AI-augmented peer review*, not autonomous AI peer review. AI systems supplied additional review capacity and specialized verification, while human editorial authority remained responsible for the identity, scope, and publication decision of the research program.

This division may become increasingly important as AI-generated mathematics becomes more sophisticated. Improved generative ability can make a manuscript more elegant, more formal, and more internally coherent without guaranteeing that it is solving the intended problem. A review architecture should therefore increase validation capacity alongside generation capacity rather than assume that improvements in one imply improvements in the other.

4.5 External Reference Structure, Constraint, and Falsifiability

The Chrysene Formalism may be unusually well suited to this type of review because its abstract mathematics is anchored to an external molecular reference structure. That feature should not be confused with a claim that the molecule itself proves the mathematics. Rather, the molecular architecture restricts which foundational statements may legitimately be assigned molecular meaning.

One can represent the admissible foundational region schematically as

$$\mathcal{F}_{\text{adm}} = M \cap C \cap H, \quad (38)$$

where M denotes mathematical validity, C denotes concordance with the defined formal program, and H denotes chemical admissibility when molecular interpretation is asserted.

This construction provides an external constraint on otherwise broad mathematical development. The formalism can introduce operator algebras, spectral triples, tensor networks, finite-state encodings, metric structures, transfer operators, or response models, but those objects cannot automatically be projected back onto the molecule without satisfying the relevant molecular constraints. Conversely, a chemical feature of chrysene does not automatically become a theorem of the abstract system unless the axioms and realization hypotheses support that implication.

The value of the molecular reference is therefore partly negative: it permits proposed extensions to fail. A candidate axiom can be rejected because it is mathematically inconsistent, because it destroys concordance with the established formal system, or because its claimed molecular realization is incompatible with the reference architecture. Future theorem programs can likewise be formulated so that stronger claims remain falsifiable rather than being inserted into the axioms.

This feature may be particularly relevant to AI-assisted development. A generative system operating only under internal mathematical coherence can continuously enlarge or reformulate a theory. An externally anchored reference adds a constraint that is not generated solely by the same mathematical process. In principle, this may help limit drift toward increasingly generic mathematical constructions that are elegant but no longer tied to the original scientific object.

The present case does not establish that chrysene is uniquely suited to serve such a role, nor does it establish that the Chrysene Formalism should constitute a new mathematical discipline. It does, however, motivate a separate hypothesis: that an externally anchored reference structure may permit rapid AI-assisted mathematical development while retaining a stable basis for admissibility, verification, and falsifiability. That hypothesis is broader than the peer-review question examined here and should be investigated independently.

4.6 Limitations

Several limitations constrain the interpretation of the present results.

First, this is a single-manuscript case study. The 63 manuscript-directed revision actions describe one review trajectory and cannot be interpreted as estimates of expected revision frequency for other papers, journals, disciplines, or AI systems.

Second, the analysis was retrospective. The revision ledger was reconstructed after the principal review cycles had been completed. Although the underlying manuscript versions, review reports, prompts, and revision history were retained, retrospective coding can introduce classification judgment. Future studies should code recommendations prospectively at the time they are generated.

Third, no independent blinded coding team was used for the present revision ledger. Inter-rater agreement for classifications such as “major,” “structural,” or “scope/qualification” was therefore not estimated. Prospective studies could improve this methodology by using two independent coders and reporting agreement statistics before disagreements are adjudicated.

Fourth, the AI systems used for drafting and review were not assigned through a randomized experimental design. Model capabilities, instructions, context, and review severity differed across review instances. The resulting reviewer scores should therefore be treated as process descriptors rather than psychometrically calibrated measurements.

Fifth, parts of the drafting and review process occurred within the same broader AI-assisted research environment. Although independent review prompts were designed to prevent reviewers

from relying upon earlier scores or disciplinary conclusions, the study does not provide the independence of a blinded external human review panel.

Sixth, human editorial intervention was integral to the process, especially during concordance reconstruction. The results therefore cannot be attributed solely to artificial intelligence. This is a feature of the proposed AI-augmented methodology, but it prevents interpretation of the study as an evaluation of autonomous AI review.

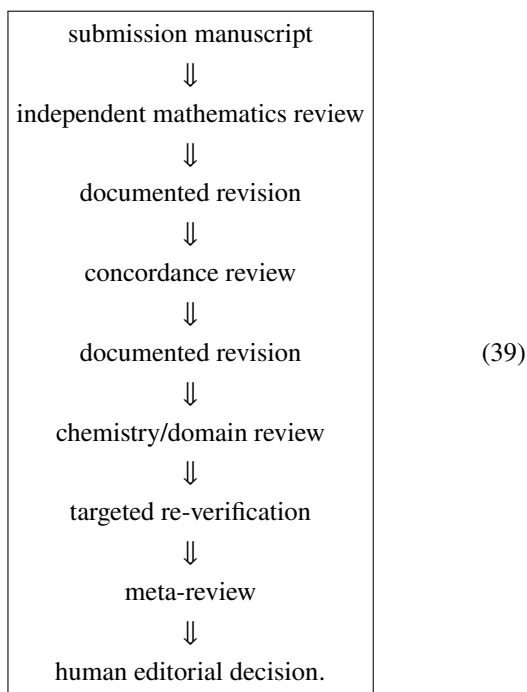
Seventh, the study contains no conventional peer-review control group. It therefore cannot determine whether the same defects would or would not have been detected by a particular panel of human reviewers, nor whether the observed process is faster, cheaper, or more accurate than conventional review under controlled conditions.

Finally, the external-reference property of the Chrysene Formalism may make this case unusually favorable for constraint-based review. Other fields may lack a comparable reference object against which concordance and physical admissibility can be evaluated. The general method should therefore be tested in additional settings before the present architecture is generalized beyond similarly constrained multidisciplinary systems.

4.7 Prospective Protocol for Future Manuscripts

The principal methodological improvement suggested by the retrospective analysis is to convert the present reconstruction into a prospective review protocol. Future manuscripts in the journal can record review events contemporaneously rather than reconstructing them after publication.

A proposed workflow is



For every recommendation, the journal can prospectively record:

1. manuscript version;
2. reviewer identifier;
3. review discipline;
4. exact criticism;
5. severity;
6. proposed revision;
7. author/editor disposition;
8. resulting manuscript version;
9. verification status; and
10. final meta-review disposition.

The review record should preserve earlier reports rather than rewrite them after corrections. A reviewer criticism that was valid for an earlier manuscript should remain visible historically even when it is no longer applicable to the final version. This makes the revision history an auditable record of manuscript development rather than a retrospectively sanitized account.

Independent first-pass reviewers should, when practical, remain unaware of other reviewer scores and conclusions until they have completed their own assessments. Reviewer disagreement should then be resolved by examining the manuscript rather than by majority vote or score averaging.

The final meta-review should continue to treat the final manuscript as controlling. Its task is to determine whether earlier criticisms have been resolved and whether the different disciplinary layers remain mutually consistent. It should not penalize the manuscript merely because an earlier draft contained a defect that has since been corrected.

Prospective data collection will also permit stronger methodological analysis. Across multiple manuscripts, future studies could examine reviewer agreement, correction rates, revision latency, error recurrence, proportion of AI recommendations rejected by human editors, discipline-specific defect profiles, and the probability that a repair survives independent verification. Such analyses would move AI-augmented peer review from a single retrospective case study toward a reproducible empirical methodology.

5 Conclusion

This retrospective case study examined the development of a multidisciplinary mathematical manuscript through a layered AI-augmented peer-review process. The reconstructed record contained 63 manuscript-directed revision actions. The principal finding was not simply that a large number of revisions occurred, but that the different review layers generated different classes of correction.

Mathematical review produced the largest number of actions and dominated proof repair, analytic validity, object typing,

and theorem-strength correction. Concordance review produced fewer actions but the greatest concentration of architectural change: all nine actions in the principal concordance-reconstruction cycle were classified as major and structural. Chemistry review primarily constrained molecular terminology, physical interpretation, and claim scope. These patterns are consistent with the intended separation of review objectives.

The review process also distinguished criticism from revision and revision from verification. Most definite recommendations were implemented, and most technical repairs received subsequent explicit or indirect review. Reviewer disagreement was not resolved by averaging scores; it was used to identify claims requiring further examination. The strictest mathematical criticisms could therefore improve the manuscript without remaining permanent objections after their underlying defects had been repaired.

The experience further suggests that AI review is most useful when it augments rather than replaces human editorial judgment. AI systems provided scalable mathematical, chemical, and consistency audits and made repeated targeted verification practical. Human judgment was particularly important in determining conceptual concordance and in deciding when mathematical refinement had begun to alter the identity of the research program itself.

The Chrysene Formalism provided an especially useful test environment because its abstract mathematical development remains linked to an external molecular reference structure. This anchoring supplied an additional admissibility constraint beyond internal mathematical consistency and allowed proposed extensions to remain falsifiable at multiple levels. Whether such externally anchored formalisms provide a more general strategy for rapid AI-assisted mathematical development is a separate question that warrants independent study.

The present results therefore do not establish the superiority of AI-augmented peer review over conventional peer review. They demonstrate something narrower: a staged system of mathematical, concordance, chemical, targeted-verification, and meta-review can produce an auditable record in which different review objectives detect measurably different classes of defect. Converting this retrospective process into a prospective protocol for future manuscripts will permit the method itself to be tested, refined, and, where necessary, falsified.

A Modern Peer Review: Evidence, Reform, and AI-Augmented Evaluation

A.1 Scope of the Review

Peer review occupies a central position in scientific publishing, but its institutional importance has historically exceeded the strength of the empirical evidence available to characterize its performance. Critical examinations have emphasized that peer review serves several functions simultaneously: it informs editorial selection, identifies defects, improves presentation, contributes to scientific legitimation, and distributes professional

recognition [Smi06, TRH20]. These functions need not produce the same evaluative outcome, and the term “peer review” itself encompasses a wide range of practices involving different degrees of anonymity, openness, editorial mediation, iteration, and post-publication evaluation [RH17, HH18].

This appendix provides a narrative review of literature most relevant to the methodology developed in the present article. It is not intended as a systematic review of all peer-review research. The literature is organized around six issues that bear directly on the case study: reviewer reliability and disagreement; attempts to improve review quality; transparency and evolving publication models; separation of review stages; meta-review and evaluation of reviews; and the emerging use of artificial intelligence in manuscript assessment.

A.2 Reviewer Reliability, Disagreement, and Editorial Judgment

A persistent finding in the peer-review literature is that independent reviewers need not reach closely concordant judgments about the same manuscript. A multilevel meta-analysis by Bornmann, Mutz, and Daniel found limited inter-rater reliability across a large collection of journal peer-review studies [BMD10]. The implication is not necessarily that one reviewer is correct and another is incorrect. Reviewers may attend to different dimensions of the same manuscript, apply different thresholds, possess different areas of expertise, or assign different weights to novelty, rigor, presentation, and significance.

The distinction between disagreement and error is important. Literature examining bias in peer review likewise shows that judgments are embedded within social and epistemic processes and that simple accounts of reviewer bias can themselves be difficult to operationalize [LSZC13]. Consequently, variance among reviews should not automatically be interpreted either as evidence of arbitrariness or as a quantity to be eliminated through averaging.

Recent longitudinal evidence strengthens this interpretation. De las Cuevas analyzed reviewer recommendations across successive review rounds and found that reviewer discordance was common, particularly at the beginning of the review process, while final editorial decisions reflected but did not mechanically reproduce reviewer recommendations [De 26]. This supports an adjudicative conception of peer review in which reviewers supply heterogeneous evidence and editorial judgment integrates that evidence rather than treating review recommendations as votes.

This literature is directly relevant to the treatment of reviewer scores in the present study. The substantially different mathematical review scores were not averaged to determine whether a manuscript version was correct. Instead, the underlying criticisms were examined individually, and later manuscript states were used to determine whether each criticism remained applicable. Such an approach is consistent with the broader observation that inter-reviewer agreement is an imperfect proxy for review validity [BMD10, TRH20, De 26].

A.3 Review Quality and Attempts to Improve Peer Review

Empirical study of peer review requires some method for distinguishing a useful review from an unhelpful one. Van Rooyen, Black, and Godlee developed the Review Quality Instrument to evaluate such features as attention to methodological strengths and weaknesses, interpretation, presentation, constructiveness, and substantiation of reviewer comments [vRBG99]. The development of instruments of this kind established that peer-review reports themselves can become objects of structured empirical analysis.

Randomized studies have examined whether conventional modifications to peer review materially change its quality. Experiments on blinding and reviewer identification found that changes in anonymity can alter features of the process without necessarily producing large improvements in review quality [vRGE+98, vRGE+99]. A randomized trial of reviewer training likewise found limited evidence that short training interventions generated substantial durable improvement in review quality [SBE+04].

Systematic assessments consequently provide a more cautious picture than the widespread institutional reliance on peer review might suggest. Jefferson and colleagues found that the empirical evidence supporting editorial peer review as a quality-control mechanism was limited and methodologically difficult to evaluate [JRBFD07]. A later systematic review and meta-analysis by Bruce and colleagues examined interventions including training, checklists, statistical review, blinding, and open review, again showing that the effects of interventions vary and that no simple procedural modification solves the full review problem [BCT+16].

Studies comparing submitted and revised manuscripts also indicate that peer review can improve reporting while still failing to detect important deficiencies. Hopewell and colleagues found that many reviewer-requested changes to reports of randomized trials were beneficial, but that reviewers did not identify all important reporting deficiencies and occasionally requested changes that could have adverse effects on reporting [HCB+14]. Peer review should therefore not be modeled as an error-free filter.

These findings motivate two features of the present methodology. First, a reviewer recommendation is not treated as correct merely because it originated in peer review. Second, implementation of a reviewer recommendation is not treated as proof that the underlying problem has been solved. The distinction among criticism, revision, and subsequent verification provides an additional control that is not captured by a conventional single review round.

A.4 Open Peer Review, Transparency, and the Evolving Scholarly Record

Modern peer review increasingly includes practices that separate evaluation from the traditional closed sequence of submission, anonymous review, acceptance, and final publication.

Ross-Hellauer’s systematic review demonstrated that “open peer review” is itself an umbrella concept encompassing multiple practices, including disclosure of identities, publication of reviewer reports, broader participation, and interaction among authors and reviewers [RH17]. Horbach and Halffman similarly describe peer review as an evolving family of institutional forms rather than a single stable mechanism [HH18].

Transparency has particular methodological value when review is itself being studied. Publication of reviews and author responses makes it possible to reconstruct how a manuscript changed, which deficiencies reviewers identified, and which recommendations authors followed. The retrospective study by Hopewell and colleagues, for example, depended upon access to successive manuscript versions and associated open reviewer reports [HCB⁺14]. The ability to observe the review process therefore transforms peer review from a largely hidden editorial event into an analyzable sequence of interventions.

Alternative publication models extend this principle further. Registered Reports move a substantial portion of peer review to the study-design stage, before results are known [Cha13]. The significance of this model for the present methodology is broader than its original application to experimental research: it demonstrates that peer review can be intentionally partitioned so that different questions are evaluated at different points in the research process.

Other reforms separate dissemination from evaluation. The publish–review–curate family of models makes manuscripts available before or during formal assessment and permits public review and editorial evaluation to remain visible components of the scholarly record. The eLife model is one prominent implementation of this approach [EAB⁺22]. These developments reinforce the principle that a scientific article need not be understood as an immutable object whose evaluation disappears once publication occurs. Manuscript versions, review reports, responses, and later assessments can instead form a traceable intellectual history.

The present study adopts a related but distinct principle. Earlier reviews are preserved as historically valid records of the manuscript state that each reviewer examined, while the final manuscript remains the controlling object for publication judgment. A defect in an earlier version is therefore neither erased from the review history nor automatically attributed to the final version after correction.

A.5 Reviewing Reviews and the Role of Meta-Review

Large conferences and other high-volume reviewing environments have made the synthesis of multiple reviews an explicit editorial task. Computational research on peer review has consequently addressed not only reviewer assignment and scoring but also reviewer disagreement, incentives, calibration, and process design [Sha22]. This literature treats peer review as a system whose architecture can itself be experimentally studied and improved.

The quality of a peer-review report is not necessarily easier to judge than the quality of the underlying manuscript. Goldberg and colleagues studied evaluations of reviews at NeurIPS and showed that the assessment of review quality can itself be influenced by features such as review length [GSC⁺25]. This result is especially relevant to AI-assisted evaluation because fluent, detailed, or lengthy reports should not automatically be interpreted as more rigorous.

Meta-review therefore serves a function different from generating an additional ordinary review. Its task is to synthesize heterogeneous reviewer evidence, identify conflicts among reviews, and determine which issues remain applicable to the controlling manuscript. Recent evidence that reviewer disagreement may persist across rounds and that editors integrate rather than mechanically follow reviewer recommendations further supports this role [De 26].

The meta-review architecture of the present study was designed around this distinction. It did not average prior numerical scores. Instead, it required a resolution audit of publication-relevant criticisms and classified them as resolved, partially resolved, unresolved, or no longer applicable. This approach treats the review record as evidence about the manuscript’s developmental history while preserving the final manuscript as the object on which the publication decision must ultimately rest.

A.6 From Computational Peer Review to Generative AI Assistance

The application of computation to peer review predates modern generative artificial intelligence. Computational research has examined reviewer assignment, conflicts of interest, incentives, aggregation, manipulation, and other aspects of review-system design [Sha22]. Large language models extend this computational role from workflow support into the production and synthesis of substantive natural-language criticism.

Early discussion of generative AI in scholarly review emphasized both potential efficiency gains and risks. Hosseini and Horbach identified possible uses of large language models for reducing reviewer workload while warning that such systems can reproduce bias, generate erroneous statements, and create accountability and confidentiality problems [HH23]. Editorial guidance has similarly stressed that AI tools do not assume responsibility for reviewer judgments and that their use must remain compatible with confidentiality, accountability, and disclosure obligations [FKTBD23].

Journal policies have not been uniform. An analysis of leading medical journals found substantial variation in whether and how artificial intelligence was addressed in peer-review guidance [LXC⁺24]. The policy environment has therefore evolved alongside the underlying technical capabilities, reinforcing the need to distinguish what an AI system *can* do from what a particular journal permits it to do.

Empirical studies have begun to evaluate the substantive usefulness of large-language-model-generated scientific feedback. Liang and colleagues compared GPT-4-generated manuscript

feedback with human peer-review feedback across thousands of papers and also conducted a prospective researcher study [LZC⁺24]. Their results indicate that LLM-generated feedback can identify issues overlapping with those raised by human reviewers and can be perceived as useful by researchers, while also showing that AI feedback is not equivalent to human expert judgment.

At the same time, generative AI is already entering real peer-review workflows. Liang and colleagues developed a corpus-level method for estimating LLM-modified text and applied it to reviews from major AI conferences, finding evidence consistent with nontrivial LLM use in submitted peer reviews [LIZ⁺24]. The question of AI in review is therefore no longer exclusively prospective.

AI assistance may also operate at the meta-review level. Sun and colleagues developed MetaWriter, a system intended to help synthesize multiple peer reviews into a meta-review. Their study found efficiency and coverage benefits while also identifying concerns about trust, over-reliance, and human agency [STHD24]. These observations align closely with the architecture used in the present case study: AI can be useful for search, decomposition, comparison, and synthesis, but final judgment should remain distinguishable from automated text production.

A.7 Relation of the Literature to Orthogonal AI-Augmented Review

The literature reviewed above suggests several principles relevant to the present methodology.

First, reviewer disagreement is common enough that agreement should not be treated as the sole indicator of validity [BMD10, De 26]. Independent reviews can instead be useful precisely because they expose different weaknesses.

Second, peer-review reports are themselves imperfect evaluative objects. Reviewer training, blinding, openness, and other conventional interventions do not eliminate all weaknesses in the process [vRGE⁺98, vRGE⁺99, SBE⁺04, BCT⁺16]. A methodology that permits reviewer recommendations to be rejected, revised, or subsequently verified is therefore preferable to assuming reviewer infallibility.

Third, transparency and version preservation make review empirically auditable [RH17, TRH20]. The preservation of prompts, review reports, manuscript states, revision actions, and verification outcomes in the present study follows directly from this principle.

Fourth, modern review systems increasingly separate evaluative tasks that were historically compressed into a single pre-publication decision. Registered Reports separate design review from result review, open-review systems expose reviewer reasoning, and publish-review-curate models separate dissemination from later evaluation [Cha13, RH17, EAB⁺22]. The present mathematics-concordance-chemistry-verification-meta-review architecture extends this general logic by separating review according to the *type of validity claim* being assessed.

Fifth, the emerging AI literature suggests that large language models can contribute useful scientific criticism and review

synthesis, but also that fluent output must not be confused with reliable scientific judgment [HH23, LZC⁺24, STHD24, GSC⁺25]. Human editorial responsibility, explicit reviewer instructions, and later verification therefore remain important components of an AI-augmented system.

Within the literature considered here, these precedents support the individual components of the present architecture, but they do not establish the effectiveness of the exact orthogonal-review sequence used in this case study. The contribution of the present work is therefore empirical and procedural rather than a claim that prior peer-review research already validates this particular system. The retrospective Chrysene case provides an initial dataset from which the architecture can be specified prospectively and tested across future manuscripts.

B Supplementary Materials and Review-Provenance Record

The supplementary materials accompanying this article are intended to make the AI-augmented peer-review case study independently inspectable at a level substantially deeper than is possible from the narrative description in the main text. Because the object of study is the *review process itself*, the relevant evidence consists not only of the final manuscript and reviewer conclusions, but also of the source documents from which the manuscript was constructed, meaningful intermediate manuscript states, exact review instructions, reviewer reports, revision records, targeted verification, meta-review, and the coded dataset used for the retrospective statistical analysis.

The supplements therefore preserve the distinction

$$\begin{array}{l} \text{source material} \longrightarrow \text{manuscript state} \longrightarrow \text{review} \\ \longrightarrow \text{revision} \longrightarrow \text{verification} \longrightarrow \text{final disposition.} \end{array} \quad (40)$$

This organization is important because the review reports describe different states of the manuscript. A criticism that was valid when first raised may no longer apply after revision, while the existence of a later correction does not erase the historical fact that the defect was identified. The supplementary archive consequently preserves earlier reviews as records of the manuscript versions actually examined rather than rewriting those reviews retrospectively to match the final paper.

The final manuscript remains the controlling document for the ultimate publication judgment. Earlier manuscript versions and reviews are provided to reconstruct how that final state was reached and to permit independent examination of the review methodology. Numerical reviewer scores should likewise be interpreted in the context of the manuscript state and review rubric under which they were produced; they are not presented as interchangeable measurements of a single invariant quantity.

B.1 Structure of the Supplementary Archive

The supplementary record is organized into ten components.

- S1. Foundational Source Corpus.** The two original monographs and two earlier articles from which the reviewed manuscript was constructed, together with a source-corpus index identifying each document and its role in the prior Chrysene Formalism.
- S2. Manuscript Evolution Corpus.** Scientifically meaningful manuscript states spanning the review trajectory, including the original review manuscript, the mathematically revised manuscript, the version evaluated for concordance, the reconstructed post-concordance manuscript, subsequent chemistry- and mathematics-revised versions, the targeted-verification manuscript, and the final controlling manuscript. Purely typographical or page-layout compilations are omitted.
- S3. Mathematics Review Record.** The independent mathematical reviews used during manuscript development, including the preliminary developmental cycle where available and the later formal standalone mathematics reviews. Reassessments performed after mathematical correction are retained separately so that changes in reviewer judgment can be linked to changes in manuscript state.
- S4. Concordance Review Record.** The review materials used to determine whether the mathematically revised manuscript remained consistent with the foundational Chrysene Formalism, including the failed concordance assessment, the resulting reconstruction criteria, relevant human editorial assessment, and the subsequent successful concordance review.
- S5. Chemistry Review Record.** The independent chemistry reviews and the associated resolution record for issues involving molecular identity, locant numbering, point-group symmetry, benzenoid interpretation, bay/K-region nomenclature, conjugation language, bridge interpretation, and physical-chemistry scope.
- S6. Targeted Verification and Meta-Review.** The post-revision mathematical verification and the final independent meta-reviews. These materials document whether previously identified defects were repaired and whether any publication-relevant issue remained in the final controlling manuscript.
- S7. Review Prompts and Scoring Rubrics.** The exact substantive instructions supplied to the mathematics, concordance, chemistry, targeted-verification, and meta-review systems, including the scoring rubrics and instructions concerning reviewer independence, theorem status, disciplinary scope, and treatment of the final manuscript as controlling.
- S8. Coded Revision Dataset and Statistical Analysis.** The retrospective revision ledger, statistical workbook, machine-readable tabular data where provided, coding definitions, and formulas used to derive the descriptive statistics reported in the main article.
- S9. Review-Provenance Map.** A structured mapping among manuscript states, review events, revision actions, and verification outcomes. This component is intended to make the developmental sequence reproducible without requiring the reader to infer chronology from filenames or review dates alone.
- S10. Human Editorial Intervention Record.** A record of substantive points at which human editorial judgment accepted, rejected, modified, or prioritized AI-generated review findings. Particular attention is given to the concordance cycle, where human interpretation of the intended research program was most consequential.

B.2 Purpose of the Supplemental Record

The supplements serve three related purposes. First, they provide *provenance*. Readers can identify which manuscript version was examined by each reviewer and which subsequent manuscript state implemented the resulting corrections. Second, they provide *methodological transparency*. The exact reviewer instructions and scoring criteria can be inspected rather than inferred from the published review reports. Third, they provide *reproducibility of the retrospective analysis*. The coded revision ledger exposes the observations from which the reported counts, percentages, and disciplinary correction profiles were calculated.

The supplemental materials are therefore not intended merely as an archive of ancillary files. They constitute part of the evidentiary basis for the methodological claims made in this article. In particular, the assertion that mathematical, concordance, and chemistry review produced different classes of correction can be examined directly against the underlying reviews and revision ledger.

This record also distinguishes transparency from retrospective revision of history. Earlier reviewer criticisms are preserved even when later corrections rendered them no longer applicable. Likewise, an early favorable assessment is preserved even if another reviewer subsequently identified a defect that required correction. The archive is therefore designed to document the actual developmental trajectory rather than present an artificially linear progression toward publication.

B.3 Versioning and Prospective Extension

The present supplemental record was reconstructed retrospectively from the manuscript and review history. Future articles using the proposed AI-augmented review architecture should create the same record prospectively. Each substantive manuscript state can be assigned a stable identifier

$$M_0, M_1, \dots, M_f, \quad (41)$$

and each review event a corresponding identifier

$$R_1, R_2, \dots, R_q. \quad (42)$$

Individual review-driven revision actions may then be linked directly to both the source review and the manuscript state in which the repair first appears.

A prospective record of this form would permit the review history to be represented as a directed provenance structure,

$$M_i \longrightarrow R_j \longrightarrow \Delta_{ij} \longrightarrow M_{i+1}, \quad (43)$$

with later verification linked back to the corresponding Δ_{ij} . Across multiple articles, such records would make it possible to compare reviewer agreement, correction type, adoption rates, verification rates, recurrence of defects, revision latency, and the relative contributions of human and AI review.

The supplementary package accompanying the present article should therefore be regarded both as the evidentiary archive for the retrospective case study and as a prototype for the prospective review record proposed for future papers in the *Annals of the Chrysene Formalism*.

References

- [BCT⁺16] Rachel Bruce, Anthony Chauvin, Ludovic Trinquart, Philippe Ravaut, and Isabelle Boutron. Impact of interventions to improve the quality of peer review of biomedical journals: A systematic review and meta-analysis. *BMC Medicine*, 14:85, 2016.
- [BMD10] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, 5(12):e14331, 2010.
- [Cha13] Christopher D. Chambers. Registered reports: A new publishing initiative at Cortex. *Cortex*, 49(3):609–610, 2013.
- [De 26] Carlos De las Cuevas. Reviewer recommendations and editorial outcomes in peer review: A longitudinal analysis of agreement and disagreement across review rounds. *Research Integrity and Peer Review*, 11(1):15, 2026.
- [EAB⁺22] Michael B. Eisen, Anna Akhmanova, Timothy E. Behrens, Jörn Diedrichsen, Diane M. Harper, Michaela D. Iordanova, Detlef Weigel, and Mone Zaidi. Peer review without gatekeeping. *eLife*, 11:e83889, 2022.
- [FKTBD23] Annette Flanagan, Jacob Kendall-Taylor, and Kirsten Bibbins-Domingo. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA*, 330(8):702–703, 2023.
- [GSC⁺25] Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PLOS ONE*, 20(4):e0320444, 2025.
- [HCB⁺14] Sally Hopewell, Gary S. Collins, Isabelle Boutron, Ly-Mee Yu, Jonathan Cook, Milensu Shanyinde, Rose Wharton, Larissa Shamseer, and Douglas G. Altman. Impact of peer review on reports of randomised trials published in open peer review journals: Retrospective before and after study. *BMJ*, 349:g4145, 2014.
- [HH18] Serge P. J. M. Horbach and Willem Halffman. The changing forms and expectations of peer review. *Research Integrity and Peer Review*, 3:8, 2018.
- [HH23] Mohammad Hosseini and Serge P. J. M. Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8:4, 2023.
- [JRBFD07] Tom Jefferson, Melanie Rudin, Suzanne Brodneyn Folse, and Frank Davidoff. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews*, (2):MR000016, 2007.
- [LIZ⁺24] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29575–29620. PMLR, 2024.
- [LSZC13] Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.
- [LXC⁺24] Zhi-Qiang Li, Hui-Lin Xu, Hui-Juan Cao, Zhao-Lan Liu, Yu-Tong Fei, and Jian-Ping Liu. Use of artificial intelligence in peer review among top 100 medical journals. *JAMA Network Open*, 7(12):e2448609, 2024.
- [LZC⁺24] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8), 2024.
- [RH17] Tony Ross-Hellauer. What is open peer review? a systematic review. *F1000Research*, 6:588, 2017.

- [SBE⁺04] Sara Schroter, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee, and Richard Smith. Effects of training on quality of peer review: Randomised controlled trial. *BMJ*, 328(7441):673, 2004.
- [Sha22] Nihar B. Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.
- [Smi06] Richard Smith. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006.
- [STHD24] Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. MetaWriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32, 2024.
- [TRH20] Jonathan P. Tennant and Tony Ross-Hellauer. The limitations to our understanding of peer review. *Research Integrity and Peer Review*, 5:6, 2020.
- [vRBG99] Susan van Rooyen, Nick Black, and Fiona Godlee. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology*, 52(7):625–629, 1999.
- [vRGE⁺98] Susan van Rooyen, Fiona Godlee, Stephen Evans, Richard Smith, and Nick Black. Effect of blinding and unmasking on the quality of peer review: A randomized trial. *JAMA*, 280(3):234–237, 1998.
- [vRGE⁺99] Susan van Rooyen, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. Effect of open peer review on quality of reviews and on reviewers’ recommendations: A randomised trial. *BMJ*, 318(7175):23–27, 1999.