Preprints.org

Article

Not peer-reviewed version

Temporal Convergence Framework: Distinguishing Structure from Coincidence in High-Precision, Low-Dimensionality Parameter Spaces

Andrew Michael Brilliant *

Posted Date: 29 December 2025

doi: 10.20944/preprints202511.0138.v4

Keywords: temporal convergence; pattern evaluation; lattice QCD; discriminatory power; pre-registration; evaluation standards

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Temporal Convergence Framework: Distinguishing Structure from Coincidence in High-Precision, Low-Dimensionality Parameter Spaces

Andrew Michael Brilliant

Applied Dynamics Research, Sapporo, Japan; a.brilliant@ieee.org

Abstract

Peer review of empirical patterns in high-precision, low-dimensionality parameter spaces relies on implicit evaluation standards. When $N = 3$ parameters at 2% precision permit thousands of statistically significant formulas, reviewers must distinguish structure from coincidence, but the criteria for doing so remain unarticulated. We found no published record of community debate establishing explicit standards, despite decades of informal application. This paper proposes one such articulation: seven criteria emphasizing temporal convergence through timestamped predictions. We offer specific thresholds not because we believe them correct, but because explicit proposals can be calibrated while implicit standards cannot. The need for explicit standards is timely. Lattice QCD has only recently achieved the precision necessary for discriminatory tests of quark mass relations. Historical precedents from lepton phenomenology (Koide, Gell-Mann–Okubo) provide limited guidance: leptons offer $\sim$35,000$\times$ greater discriminatory power than light quarks, involve no RG running, and constitute a fundamentally different measurement regime. The historical record is further compromised by survivorship bias: patterns that diverged are largely unrecorded. Historical cases motivate the problem by illustrating why implicit evaluation proved adequate for leptons but may prove inadequate for quarks. They cannot validate the proposed solution. Validation is prospective by design: starting from this publication, patterns evaluated under this framework will be tracked publicly. The framework succeeds if it proves predictively useful; it fails if it requires constant post-hoc adjustment, judged by its own temporal convergence criterion. If this proposal provokes disagreement that leads to better criteria, it will have served its purpose. If it is ignored, the current system of implicit evaluation continues unchanged. We consider both engagement and refinement to be success.

Keywords: temporal convergence; pattern evaluation; lattice QCD; discriminatory power; pre-registration; evaluation standards

---

1. Introduction

1.1. Motivation: A Transitional Precision Regime

The ability to perform discriminatory tests of empirical patterns in quark masses is new. This paper addresses a methodological gap created by recent advances in lattice QCD precision.

Historical impossibility. In earlier precision regimes, researchers searched for Koide-like relations in quark masses and found nothing conclusive. This was not because patterns were absent but because measurement uncertainties could not distinguish signal from noise. Quark masses were known to perhaps 20–50% precision; any simple formula would agree within errors.

The lattice QCD revolution. The Flavour Lattice Averaging Group (FLAG) 2024 review reports light quark mass ratios at 2–3% precision [1]. For the first time, the parameter space is both finite ($N = 3$ light quarks) and precise enough for meaningful tests. This creates a regime that has never existed before.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

The new problem. Paradoxically, this precision creates difficulty: statistical significance becomes readily obtainable. When search space vastly exceeds parameter dimensionality, combinatorial exploration guarantees discovery of statistically significant relationships, whether meaningful or coincidental.

A finite window. This framework addresses a transitional regime. In the past, quark mass precision was insufficient for discriminatory tests; any simple formula agreed within errors. In the future, continued lattice QCD improvement will achieve precision where single measurements discriminate definitively, as already occurs for leptons. The present is the window where precision enables meaningful tests but coincidences still pass statistical filters. Temporal convergence fills this gap. The methodology proposed here is timely precisely because the window where it matters most is finite, perhaps a decade as systematic improvements continue.

1.2. Discriminatory Power: Quarks vs. Leptons

To understand why historical precedents provide limited guidance, we quantify the discriminatory power available in each regime.

Leptons (Koide's situation, 1982–present):

Particle	Mass	Precision
Electron	0.511 MeV	$\sim 10^{-10}$
Muon	105.7 MeV	$\sim 10^{-8}$
Tau	1776.9 MeV	$\sim 10^{-4}$

- Dynamic range: $\tau/e \approx 3{,}500$
- Worst precision (τ): ∼0.007%
- All masses are pole masses: static, directly measurable, no scale dependence
- Two anchors (e, μ) known to extraordinary precision
- **Distinguishable positions across range: ∼50,000,000**

When Koide found a formula relating these masses to sub-percent precision [3], this was genuinely surprising. The probability of a random simple formula achieving such agreement by chance is negligible.

Light quarks (current situation):

Particle	Mass (2 GeV)	Precision
Up	2.16 MeV	∼3%
Down	4.7 MeV	∼1.5%
Strange	93.5 MeV	∼1%

- Dynamic range: $s/u \approx 43$
- Worst precision (u): ∼3%
- All masses are $\overline{\text{MS}}$ running masses that evolve with energy scale
- No anchors: all three masses have comparable relative uncertainties
- **Distinguishable positions across range: ∼1,400**

Finding a formula that "works" within current uncertainties is expected by combinatorics. With only ∼1,400 distinguishable slots, simple formulas routinely achieve sub-sigma agreement by chance.

Ratio of discriminatory power: 50,000,000 / 1,400 ≈ 35,000×

Leptons offer 35,000× greater discriminatory power than light quarks. Historical intuitions calibrated on lepton phenomenology may systematically underestimate coincidence rates in quark phenomenology.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

1.3. The Full Quark Spectrum Problem

Extending beyond light quarks makes the situation worse, not better.

The six-quark spectrum:

Quark	Mass	Type
Up	∼2.2 MeV	$\overline{\text{MS}}$, runs
Down	∼4.7 MeV	$\overline{\text{MS}}$, runs
Strange	∼93 MeV	$\overline{\text{MS}}$, runs
Charm	∼1.27 GeV	$\overline{\text{MS}}$, runs
Bottom	∼4.18 GeV	$\overline{\text{MS}}$, runs
Top	∼173 GeV	Pole mass

- Total dynamic range: $t/u \approx 80{,}000$
- But this range spans *incompatible mass definitions*
- Light quarks evolve together under renormalization group evolution (ratios preserved)
- Heavy quarks (c, b) occupy intermediate regime
- Top is fundamentally different: a pole mass, not comparable to running masses

The geometric problem:

With leptons, Koide had three static points and sought a pattern among them.

With quarks, one faces:

- Three points (u, d, s) that evolve along correlated trajectories as the renormalization scale changes
- Two points (c, b) in an intermediate regime
- One fixed point (t) that is statistically incomparable, with a different mass definition entirely

Any formula relating all six quarks must either:

1. Be scale-invariant (using only ratios at the same scale), or
2. Bridge fundamentally different mass definitions (physically questionable)

The parameter space is not just larger than leptons; it is geometrically deformed in ways that make pattern-finding simultaneously easier (more parameters, more possible relations) and less meaningful (coincidences proliferate, mass definitions conflict).

1.4. The Methodological Gap

Given this new regime, one might expect published standards guiding evaluation of empirical quark mass patterns. We found none.

This methodological gap exists within broader discussions of research strategy in fundamental physics. Smolin [10] has argued that progress benefits from diversity of theoretical approaches, suggesting that productive exploration requires methodological infrastructure to evaluate a wider range of ideas efficiently. The framework proposed here offers one such tool: explicit criteria that enable systematic evaluation without requiring case-by-case implicit judgment. By making intake broader but filtering faster, such infrastructure could reduce the cost of exploration for individual researchers while maintaining rigorous standards for publication.

The community has developed implicit evaluation standards through decades of experience, but these standards remain unarticulated. Reviewers assess patterns based on intuitions calibrated largely on lepton phenomenology. Authors receive rejections without clear criteria for improvement. Researchers from adjacent fields cannot identify what standards to meet.

The historical record cannot resolve this gap:

- **Survivorship bias**: Koide and GMO persist in the literature because they converged. Failed patterns were never published, or were published and forgotten.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

- **Apples to oranges**: Comparing lepton phenomenology (1982) to quark phenomenology (2024) involves different particles, different precision regimes, different mass definitions, and different discriminatory power.
- **Post-hoc evaluation**: Any framework "validated" against historical cases is partly constructed by examining those cases. This is not validation but motivation.

1.5. Historical Barriers to Empirical Pattern Work

The cost-benefit considerations for empirical pattern work in quark masses have historically been unfavorable. Insufficient precision meant inability to distinguish structure from coincidence. Publishing a relationship that later fails carries reputational cost. Publishing a relationship that holds but lacks theoretical mechanism invites dismissal as "numerology." The rational response has been avoidance: safer to work on other problems.

Explicit criteria with temporal tracking change this calculus. Patterns are tracked provisionally rather than asserted definitively. Divergence is revealed by data, not by criticism. A pattern that self-falsifies through Criterion 4 represents legitimate scientific contribution: documentation of what does not work, with the same rigor as documentation of what does. The framework absorbs the risk that previously fell on individual researchers.

This matters for the field's ability to explore the newly-accessible quark mass parameter space. Without explicit criteria, the rational avoidance continues even as precision improves. With explicit criteria, researchers can contribute to a tracked corpus where provisional patterns accumulate evidence over time, and self-falsification is recognized as valuable rather than embarrassing.

1.6. Accessibility of Evaluation Standards

Explicit criteria enable researchers from adjacent fields to understand community standards and assess their own work before seeking reviewer input.

Currently, a computational researcher from a field with different precision norms cannot easily find documentation of what HEP phenomenology requires. Publication thresholds, acceptable methodologies, and evaluation criteria remain implicit, accessible to domain insiders but opaque to potential contributors. This limits participation to those with prior access to community norms.

Documented standards address this asymmetry. Researchers can consult explicit criteria, understand the precision regime, and self-assess before investing significant effort. Advisors provide concrete guidance tied to defined benchmarks. Reviewers focus on substantive scientific judgment rather than re-articulating expectations case by case.

The goal is inclusion through clarity. Patterns stratify by documented status: "theoretical mechanism unknown" becomes a category for continued tracking, not grounds for rejection. Cross-domain contribution becomes feasible when expectations are transparent.

1.7. Purpose of This Paper

We propose explicit evaluation criteria for one reason: **explicit proposals can be calibrated while implicit standards cannot.**

If the community finds criterion X too strict, that discussion is progress. If reviewers find criterion Y irrelevant, removing it clarifies what actually matters.

We do not claim these criteria are correct. We claim the discussion should exist and currently does not.

2. Framework: Seven Criteria

We propose seven criteria organized hierarchically. Criteria 1–6 address empirical validation accessible to phenomenologists. Criterion 7 addresses theoretical context.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

2.1. Criterion 1: Scale Invariance Under Renormalization Group Evolution

Mass ratios must be scale-invariant under QCD running. Scale-dependent relationships likely reflect artifacts of scale choice rather than physical structure.

Open questions for community calibration:

- What deviation threshold constitutes "scale-invariant"? The proposed 10^{-4} ensures negligible running effects, but may be unnecessarily stringent for practical purposes.
- Over what energy range should invariance be tested? The 1 GeV to TeV range spans typical phenomenological applications, but specific contexts may warrant different bounds.

Empirical context: Light quark mass ratios m_s/m_d and m_d/m_u are preserved to high precision under QCD renormalization group evolution: ratios run together, so relationships among ratios remain stable. This is a feature of QCD, not a constraint on patterns: relationships expressed as ratios automatically satisfy scale invariance to the extent that running is flavor-universal at leading order.

Proposed starting point: Deviations $< 10^{-4}$ across 1 GeV to TeV scales. Relationships not expressed as scale-invariant ratios require explicit justification.

2.2. Criterion 2: Compression of Degrees of Freedom

Patterns must reduce N parameters to fewer degrees of freedom through unified constraints. A pattern that does not compress information provides no predictive content.

Open questions for community calibration:

- Is $N \to N-1$ compression sufficient, or should stronger compression be required for patterns involving more parameters?
- How should approximate compression be handled? A relationship that predicts one mass to 1% given the others provides less compression than an exact relationship, but still provides predictive content.

Empirical context: Koide's formula compresses three lepton masses to two degrees of freedom: given any two masses, the third is predicted. The Gell-Mann–Okubo relation similarly constrains hadron multiplets, reducing independent parameters. Compression is what distinguishes a pattern from a tautology.

Proposed starting point: N parameters reduced to at most $N-1$ degrees of freedom. Patterns that achieve only approximate compression should be assessed by the precision of their predictions.

2.3. Criterion 3: Statistical Agreement

Patterns must agree with measurements within statistical bounds. This criterion is necessary but not sufficient: given the low discriminatory power in the quark sector, statistical agreement alone provides minimal evidence of structure over coincidence.

Open questions for community calibration:

- What deviation threshold is appropriate? The choice between 1σ, 2σ, or some other bound affects the filter's stringency.
- Should the threshold depend on discriminatory power? As shown in Section 1.2, a 1σ match in the lepton sector represents far stronger evidence than a 1σ match in the quark sector.
- How should correlated uncertainties be handled when patterns involve multiple mass ratios?

Empirical context: The Koide formula achieves $< 0.1\sigma$ agreement with lepton masses. In the quark sector, the limited discriminatory power means simple formulas routinely achieve comparable agreement by chance (Section 1.2).

Proposed starting point: $< 1\sigma$ deviation, with the recognition that this threshold alone admits many coincidences at current quark mass precision. Criterion 4 provides the primary filter; statistical agreement serves as a necessary precondition rather than sufficient evidence.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

2.4. Criterion 4: Temporal Convergence

Patterns must demonstrate directional convergence or stability across data releases. This is the core discriminator. When error bars shrink, genuine patterns show residuals shrinking or stable. Coincidences show residuals growing—precision reveals the offset that broader uncertainties obscured.

Requirements:

- Pre-registration via timestamped repository before new data releases
- Central values converging toward (or stable around) prediction as precision improves

Why temporal convergence provides robust protection:
Pre-registration creates an immutable record. Authors cannot:

- Select data vintages post-hoc
- Adjust formulas after seeing results
- Mine through hypothesis variations retroactively
- Claim prescience after observing convergence

The only way to pass Criterion 4 is genuine predictive success across independent experimental cycles.

Open questions for community calibration:

- How many data releases constitute sufficient temporal evidence? Two releases showing convergence could reflect luck; five releases provide stronger evidence but require patience.
- How should convergence be quantified? Residuals shrinking proportionally with uncertainties, remaining stable in absolute terms, or some weighted combination?
- What constitutes a "new" data release versus an update? FLAG reviews aggregate multiple collaboration results; should each collaboration's update count independently?

Empirical context: Koide's formula has survived four decades of improving lepton mass measurements—a temporal track record that coincidence cannot plausibly explain. Historical patterns that diverged (Section 3.2) demonstrate the converse: initial statistical agreement revealed as coincidental when precision improved.

Proposed starting point: Convergence or stability demonstrated across ≥ 3 independent data releases, with pre-registration established before each release. Patterns showing systematic divergence receive $T = -1$ classification and are filtered from further consideration.

2.5. Criterion 5: Mathematical Simplicity

Patterns should involve simple mathematical structure. Complex formulas with many free parameters can fit anything; simplicity constrains hypothesis space.

Open questions for community calibration:

- How should "simplicity" be operationalized? Counting operations, parameters, or information-theoretic measures each yield different orderings.
- Should simplicity be absolute or relative to predictive power? A slightly more complex formula achieving 0.01σ agreement may warrant consideration over a simpler formula at 0.5σ.
- How should transcendental constants (π, e, ϕ) be weighted relative to integers?

Empirical context: Historical patterns that persisted tend toward notable simplicity. Koide's formula involves one equation with square roots and the fraction $2/3$. The Gell-Mann–Okubo relation involves linear combinations with small integer coefficients. Patterns that diverged (Nambu, Lenz) were comparably simple—simplicity did not save them, but complexity would have made their initial consideration unwarranted.

Proposed starting point: Basic arithmetic operations, integer exponents ≤ 5, standard constants (π, small integers). We recognize this threshold requires sharper operationalization—formula complexity is partly in the eye of the beholder. The principle (simpler is better, all else equal) is sound even if the specific boundary needs calibration.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

2.6. Criterion 6: Independent Validation

Patterns must show consistency across independent determinations. A pattern that appears in one collaboration's results but not others may reflect systematic artifacts rather than physical structure.

Open questions for community calibration:

- How many independent determinations constitute sufficient validation? Three collaborations with genuinely independent systematics may suffice; three analyses of the same underlying data do not.
- How should "independence" be assessed when collaborations share methodological choices, gauge configurations, or analysis frameworks?
- Should validation require agreement across different lattice actions, or is agreement across different fitting procedures within the same action sufficient?

Empirical context: FLAG 2024 averages draw on results from BMW, MILC, HPQCD, ETM, RBC/UKQCD, and other collaborations employing different lattice discretizations, fermion actions, and analysis methods. A pattern consistent across these independent determinations has survived tests that a systematic artifact would likely fail.

Proposed starting point: Agreement across ≥ 3 independent collaborations or methods with demonstrably different systematic uncertainties.

2.7. Criterion 7: Theoretical Viability

Patterns must not be demonstrably incompatible with established physics. This criterion differs from the others: it requires theoretical rather than empirical assessment.

Three possible outcomes:

- **PASS (Compatible):** Mechanism identified within existing frameworks.
- **PASS (Unknown):** No known incompatibility. Pattern awaits theoretical investigation.
- **FAIL (Incompatible):** Pattern contradicts established constraints through explicit proof.

Critical: Absence of mechanism does not constitute failure. Patterns may persist in Unknown status indefinitely; this is legitimate.

Open questions for community calibration:

- What constitutes "demonstrable incompatibility"? A pattern that conflicts with the Standard Model at tree level is clearly incompatible; a pattern that requires non-trivial BSM physics occupies grayer territory.
- Who bears the burden of proof? Should proponents demonstrate compatibility, or should critics demonstrate incompatibility? The asymmetry matters for patterns in Unknown status.
- How should patterns be classified when theoretical assessment is contested among experts?

Empirical context: The Koide formula remains in Unknown status despite decades of attention. Multiple theoretical mechanisms have been proposed; none have achieved consensus. This has not diminished the formula's empirical standing: temporal convergence and theoretical explanation are independent tracks.

Division of labor: Empiricists validate through criteria 1–6. Theorists assess criterion 7. Neither bears obligation to the other. This separation allows empirical pattern documentation to proceed without waiting for theoretical explanation, while ensuring that patterns contradicting established physics are appropriately flagged.

On irreducibility: This criterion resists algorithmic operationalization by design. Theoretical compatibility requires expert judgment that cannot be reduced to mechanical rules. This is not a weakness—the framework is a tool for human evaluation, not a replacement for it. A fully automated system would lack the essential feedback loop: community deliberation, contested interpretations, and evolving theoretical understanding. Criterion 7 ensures that human judgment remains central to the evaluation process.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

3. Illustrative Application

To demonstrate how the criteria operate in practice, we apply them to historical cases. We emphasize that alignment between criteria and historical outcomes is expected by construction—criteria were partly informed by examining patterns that persisted. This provides illustration, not validation.

3.1. Patterns That Persisted

Koide Formula [3]: $(m_e + m_\mu + m_\tau)/(\sqrt{m_e} + \sqrt{m_\mu} + \sqrt{m_\tau})^2 = 2/3$

Criterion	Assessment
1. Scale Invariance	PASS – ratio form, leptons do not run
2. Compression	PASS – reduces 3 masses to 2 DOF
3. Statistical	PASS – agrees within $< 0.1\sigma$ consistently
4. Temporal	PASS – persistent through 4 decades of improving precision
5. Simplicity	PASS – single equation, simple constants
6. Independent	PASS – multiple independent measurements
7. Theoretical	PASS (Unknown) – no mechanism identified, none ruled out

Gell-Mann–Okubo Relation [4,5]: Predicted hadron mass relationships before the quark model existed.

Criterion	Assessment
1. Scale Invariance	PASS – hadronic scale relationship
2. Compression	PASS – relates multiple hadron masses
3. Statistical	PASS – agreed with measurements
4. Temporal	PASS – validated by subsequent data
5. Simplicity	PASS – simple symmetry structure
6. Independent	PASS – multiple hadron measurements
7. Theoretical	PASS (Unknown $\rightarrow$ Explained) – SU(3) mechanism emerged later

These cases illustrate the progression from empirical documentation through temporal validation to theoretical explanation. GMO demonstrates that patterns can legitimately persist in Unknown status until theory catches up.

3.2. Patterns That Diverged

Historical patterns that showed initial promise but diverged as precision improved:
Nambu (1952) [7]: $m_\mu/m_e = 3/(2\alpha)$

Criterion	Assessment
3. Statistical (1952)	Initially $\sim 1\sigma$
3. Statistical (2024)	Now $> 20\sigma$ deviation
4. Temporal	FAIL – $T = -1$, diverged catastrophically

Lenz (1951) [8]: $m_p/m_e \approx 6\pi^5$

These are not failures of methodology; they represent reasonable exploration given available precision. They were superseded by improved measurement, not refuted by argument. The temporal convergence criterion would correctly identify both patterns as diverging, filtering them before theoretical investment.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

Criterion	Assessment
3. Statistical (1951)	Initially $< 1\sigma$
3. Statistical (2024)	Now $\sim 50\sigma$ deviation
4. Temporal	FAIL – $T = -1$, diverged catastrophically

3.3. Diagnostic Pattern: A Spurious Light Quark Relation

To illustrate how the framework operates in the present quark mass regime, we consider a deliberately spurious relation built from two familiar ratios. FLAG 2024 quotes [1]

$$\frac{m_d}{m_u} \approx 2.16, \qquad \frac{m_s}{m_d} \approx 20, \tag{1}$$

so it is straightforward to invent simple algebraic combinations that appear to agree numerically. One such toy relation is

$$2\left(\frac{m_d}{m_u}\right)^3 \approx \frac{m_s}{m_d}. \tag{2}$$

We refer to this as the *Diagnostic Pattern*. It is introduced as a worked example only. It is not proposed as a candidate mass relation.

Using FLAG 2024 central values at 2 GeV,

$$\frac{m_d}{m_u} = 2.162 \pm 0.050, \qquad \frac{m_s}{m_d} = 20.0 \pm 0.5, \tag{3}$$

the relation predicts $2(2.162)^3 = 20.22$ on the left-hand side versus 20.0 on the right-hand side. Propagating uncertainties with m_d/m_u and m_s/m_d treated as uncorrelated We treat the two ratios as uncorrelated for simplicity; correlated uncertainties would not change the qualitative conclusion. gives a combined standard deviation $\sigma \approx 1.4$, so the deviation is

$$\Delta \equiv \frac{20.22 - 20.0}{\sigma} \approx 0.16\sigma. \tag{4}$$

By usual static standards the agreement looks more than adequate. At face value, (2) passes Criterion 3.

The pattern also passes several other criteria in mechanical fashion:

- It is constructed from ratios at a fixed renormalization scale, so it is preserved under QCD running to very high accuracy (Criterion 1).
- It reduces three masses to two degrees of freedom (Criterion 2).
- It is mathematically very simple (Criterion 5).
- The underlying ratios are stable across multiple lattice collaborations (Criterion 6).

If evaluation stopped at this stage, (2) would look like a promising empirical pattern.

The framework regards it as spurious for two distinct reasons.

First, temporal behaviour already points in the wrong direction. FLAG 2019 and FLAG 2024 both report values compatible with $m_d/m_u \approx 2.16$, but the trend in central values and uncertainties is unfriendly to (2). A representative summary is As lattice precision improved, the central value moved

Table 1. Illustrative temporal evolution of m_d/m_u relative to the value $10^{1/3} \approx 2.154$, the value required for the relation to hold exactly.

Review	m_d/m_u	Uncertainty	Distance from $10^{1/3}$
FLAG 2019	2.160 ± 0.080	3.7%	$+0.006$ (0.08σ)
FLAG 2024	2.162 ± 0.050	2.3%	$+0.008$ (0.16σ)

slightly away from the implied target while the error bar shrank by roughly forty percent. In absolute

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

terms the shift is small, but in units of the uncertainty the disagreement doubled. At current precision this is not decisive, and we do not claim falsification. It is enough to classify the temporal signal as unfavourable. Under Criterion 4 the pattern receives a provisional $T = -1$ flag: future FLAG updates are expected to increase the tension, not reduce it. Since this example is retrospective, it does not satisfy the pre-registration requirement of Criterion 4; it is included purely as a demonstration of temporal behaviour.

Second, the relation is incongruent with existing flavour model building. Interpreted at the level of Yukawa couplings, (2) corresponds schematically to

$$2\, y_d^4 \approx y_u^3\, y_s, \tag{5}$$

which does not resemble textures generated by known symmetry structures. Standard frameworks built from Froggatt–Nielsen type charges, non Abelian family symmetries, or texture zero schemes do not naturally produce quartic down couplings balanced by a cubic up factor and a single strange factor. The issue is not lack of a specific model, it is the apparent mismatch between the algebraic structure of (2) and the kinds of couplings that current mass generation mechanisms tend to produce. We therefore regard it as incompatible in spirit with established flavour constructions and assign it a failing status under Criterion 7.

Together these two points mark (2) as a useful diagnostic example rather than a viable pattern. It is constructed from numbers that every practitioner knows ($m_s/m_d \sim 20$, $m_d/m_u \sim 2.16$); it passes the usual static checks; it would be trivial to publish in an informal context. The framework makes concrete the intuition that such relations are spurious. Criterion 3 admits them. Criterion 4 and Criterion 7 reject them. This is the intended operation of the filter in the hardest regime: an $N = 3$ light quark sector where coincidences are common and theoretical guidance is limited.

3.4. Summary

In this limited historical sample: all patterns that persisted show $T \geq 0$; all patterns that diverged show $T = -1$. This alignment is expected, as criteria were informed by these outcomes. The test of the framework is prospective: will future patterns classified as $T = +1$ persist, and will those classified as $T = -1$ diverge?

3.5. Acknowledging the Limitation

We address a fundamental limitation directly: the historical record provides few worked examples, and those that exist have already converged or diverged. We cannot draw rigorous comparisons because outcomes are known, criteria were partly informed by examining them, and the precision landscape has changed. Leptons in 1982 are not quarks in 2024.

We do not attempt to resolve this through careful hedging. The framework cannot be validated retrospectively. It can only be validated prospectively, by tracking patterns from this publication forward and observing whether classifications prove predictive.

This is the design, not a limitation to excuse. The contribution is proposing explicit, trackable criteria where none existed. Whether those criteria prove useful is an empirical question that future data will answer.

3.6. Framework Precedence

One might ask why this paper proposes evaluation criteria without demonstrating them on new physics. The answer is structural: framework papers necessarily precede the contributions they enable. The criteria proposed here are intended for application by any researcher; their utility will be judged by whether the resulting evaluations prove consistent and predictive across independent applications.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

4. Prospective Validation

4.1. Why Historical Cases Cannot Validate

We explicitly decline to "test" this framework against historical cases (Koide, GMO, $m_\mu/m_e = 3/(2\alpha)$, etc.).

Such testing would be:

- **Post-hoc:** The criteria were partly informed by examining historical outcomes.
- **Cross-regime:** Historical cases involve leptons with 35,000$\times$ greater discriminatory power.
- **Survivorship-biased:** We know which patterns persisted; failed patterns are largely unrecorded.

Historical cases motivate the problem. They demonstrate that implicit evaluation produced certain outcomes. They cannot validate whether explicit criteria would improve upon implicit evaluation.

4.2. Prospective Tracking

Validation is prospective by design.

Starting from this publication:

- Patterns evaluated using this framework will be documented publicly
- Classifications and outcomes will be tracked
- The framework itself will be assessed by its own temporal convergence criterion

If the framework requires constant ad-hoc modification to match reviewer intuitions, it shows $T = -1$ (divergence) and should be abandoned. If it stabilizes and proves useful, it earns continued consideration.

Repository: https://github.com/AndBrilliant/TemporalConvergence

4.3. Self-Application

We invite the community to treat this framework as we propose treating empirical patterns: provisionally, with temporal tracking.

The criteria proposed here are predictions. Future application will test them. This is not a limitation to address; it is the design.

5. Discussion

5.1. Philosophical Foundations

The temporal convergence criterion connects to established philosophy of science traditions. Lakatos's distinction between "progressive" and "degenerating" research programmes [9] provides conceptual grounding: programmes are progressive when they successfully predict novel facts, degenerating when they merely accommodate known data. Our temporal convergence criterion operationalizes this distinction: patterns showing convergence have made predictions validated by subsequent measurements, while diverging patterns reveal retrospective accommodation exposed by improving precision.

Mayo's error-statistical framework [11] offers complementary support through the concept of "severe testing": claims warrant belief when they have passed tests with high probability of detecting error if present. Survival through multiple generations of improving experimental precision constitutes increasingly severe testing. A pattern that maintains sub-sigma agreement as uncertainties shrink by factors of two or more has passed a test that coincidences systematically fail.

Finally, the pre-registration movement in psychology [12,13], responding to replication failures traceable to post-hoc hypothesis adjustment, validates our timestamped prediction requirement. The replication crisis demonstrated that flexibility in analysis—choosing statistical tests, endpoints, or hypotheses after seeing data—dramatically inflates false positive rates. Our Criterion 4 extends pre-registration from individual experiments to longitudinal pattern validation across experimental generations. The framework inherits the epistemic benefits that pre-registration provides: eliminating post-hoc adjustment, selective reporting, and hypothesizing after results are known.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

5.2. What Explicit Standards Enable

If this articulation reasonably captures implicit community standards, or provides a useful starting point for refinement, it may offer:

For authors:

- Self-evaluation before submission
- Clear criteria for pattern documentation
- Legitimate publication in "Unknown" theoretical status

For reviewers:

- Explicit criteria to reference
- Reduced case-by-case evaluation burden
- Simplified first-pass filtering

For the field:

- Discussable, refinable standards
- Greater accessibility for researchers from adjacent fields
- Temporal decoupling: patterns documented now may find theoretical explanation later

5.3. Why the Quark Sector Specifically

We focus on light quark masses because this regime represents the methodological worst case:

- $N = 3$ provides minimal discriminatory power
- $\sim$2% precision enables meaningful tests but permits many coincidences
- $\overline{\text{MS}}$ running masses add complexity absent in leptons
- No historical precedent directly transfers

If explicit criteria provide value here, they likely provide value in any higher-dimensional context.

5.4. Future Work and Invitation to the Community

The criteria proposed here emerged from analysis of one domain (empirical patterns in quark mass ratios) and have not been tested across the full range of high-energy physics phenomenology. We lack the domain expertise to calibrate these criteria optimally for all contexts in which they might apply.

We therefore invite researchers with deeper expertise in lattice QCD, flavor physics, and related areas to critique, refine, or extend these criteria. Specific areas where expert input would be valuable include:

- **Threshold calibration:** Are the proposed thresholds for scale invariance (10^{-4}), statistical agreement (1σ), temporal releases (≥ 3), and independent validations (≥ 3) appropriately stringent for practical use? Too strict filters useful patterns; too loose admits coincidences.
- **Convergence quantification:** How should "convergence" be operationalized beyond the qualitative description given here? Researchers with experience in sequential analysis or meta-analytic methods may identify more rigorous formulations.
- **Independence assessment:** What constitutes genuine independence among lattice collaborations? Shared gauge configurations, analysis frameworks, or methodological choices may create correlations that undermine Criterion 6.
- **Domain-specific invariance checks:** Are there additional physical constraints, beyond RG scale invariance, that patterns in specific subfields should satisfy?

The framework is offered as a starting point for discussion, not a final specification. If domain experts find the criteria useful but miscalibrated, recalibration is straightforward; that is the advantage of explicit standards. If the criteria prove fundamentally misguided, explicit rejection with documented reasoning advances the field more than silent dismissal of implicit standards.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

We welcome correspondence, critique, and collaboration. The repository at https://github.com/AndBrilliant/TemporalConvergence provides a venue for community discussion and iterative refinement.

5.5. Limitations

Temporal data requirements: Criterion 4 requires multiple data releases. New patterns cannot be fully evaluated immediately.

Threshold calibration: All proposed thresholds are initial estimates. Community input determines appropriate values.

Theoretical coupling subjectivity: Criterion 7 requires domain expertise and involves judgment.

These are not flaws but acknowledgments. We propose a starting point, not a finished system.

6. Conclusions

Lattice QCD has achieved precision that enables discriminatory tests of quark mass relations for the first time. This new capability creates a new problem: statistical significance is easily obtained through combinatorial search, while genuine structure remains difficult to identify. Historical intuitions calibrated on lepton phenomenology—with 35,000× greater discriminatory power—may systematically underestimate coincidence rates.

We propose explicit evaluation criteria emphasizing temporal convergence: patterns must demonstrate convergence or stability as precision improves, established through timestamped predictions. We offer specific thresholds as starting points for community calibration.

Historical cases motivate this proposal by illustrating the problem. They cannot validate the solution; that validation is prospective by design. The framework will be assessed by its own criterion: if it proves predictively useful and stable, it earns continued consideration; if it requires constant modification, it should be abandoned.

If this proposal generates productive disagreement, it will have succeeded. If it is adopted and refined, better still. If it is ignored, the current system of implicit evaluation continues—but we will have documented what that system is, creating a baseline for future comparison.

The contribution is not claiming to have solved the evaluation problem. The contribution is articulating explicit criteria as a basis for community refinement.

Funding

This research was conducted independently without institutional funding or external support.

Data Availability

All numeric values used in this analysis are derived from publicly available FLAG 2024 [1] and PDG 2024 [2] reviews. The FLAG 2019–2024 ratios and derived values for temporal convergence analysis are provided in `S1_data.csv` and mirrored at https://github.com/AndBrilliant/TemporalConvergence. Analysis methodologies are fully described in the text.

Acknowledgments

The author thanks Riccardo M. Pagliarella for encouragement and valuable discussions. This work would not be possible without the extraordinary precision achieved by the lattice QCD community, particularly FLAG, BMW, MILC, HPQCD, and ETM collaborations.

Conflicts of Interest

The author declares no conflict of interest.

1. Y. Aoki *et al.* (FLAG Working Group), Eur. Phys. J. C **84**, 1263 (2024).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2025 doi:10.20944/preprints202511.0138.v4

2. S. Navas *et al.* (Particle Data Group), Phys. Rev. D **110**, 030001 (2024).
3. Y. Koide, Lett. Nuovo Cim. **34**, 201 (1982).
4. M. Gell-Mann, *The Eightfold Way: A Theory of Strong Interaction Symmetry*, Caltech Report CTSL-20 (1961).
5. S. Okubo, Prog. Theor. Phys. **27**, 949 (1962).
6. A. Bazavov *et al.* (MILC), Phys. Rev. D **98**, 054517 (2018).
7. Y. Nambu, "An Empirical Mass Spectrum of Elementary Particles," Prog. Theor. Phys. **7**, 595 (1952).
8. F. Lenz, "The Ratio of Proton and Electron Masses," Phys. Rev. **82**, 554 (1951).
9. I. Lakatos, "Falsification and the Methodology of Scientific Research Programmes," in *Criticism and the Growth of Knowledge*, eds. I. Lakatos and A. Musgrave, Cambridge University Press, pp. 91–196 (1970).
10. L. Smolin, *The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*, Houghton Mifflin (2006).
11. D. G. Mayo, *Statistical Inference as Severe Testing*, Cambridge University Press (2018).
12. Open Science Collaboration, "Estimating the reproducibility of psychological science," Science **349**, aac4716 (2015).
13. B. A. Nosek *et al.*, "The preregistration revolution," PNAS **115**, 2600–2606 (2018).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.