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

Andrew Brilliant
Independent Researcher
Sapporo, Japan
andrew.brilliant@protonmail.org
ORCID: 0009-0004-8024-5442

Article type: Technical Note (Methods)

Abstract

Modern computational methods across scientific domains achieve precision through iterative refinement. This precision regime creates opportunities for refined evaluation methods: as measurement uncertainties decrease while parameter dimensionality remains fixed, statistical significance becomes more easily obtained through combinatorial search. Traditional hypothesis testing could benefit from additional discrimination when nearly any simple relationship can achieve sub-sigma agreement by chance.

We propose a seven-criteria framework emphasizing temporal convergence through pre-registration. The core innovation: pattern predictions are established with timestamping, then tracked against future data releases for directional convergence or stability as precision improves. This requirement provides robust protection against retroactive fitting—reducing susceptibility to common biases including data selection and post-hoc hypothesis adjustment. Combined with six supporting criteria (scale invariance, compression, statistical agreement, mathematical simplicity, independent validation, theoretical viability), temporal tracking enhances discrimination when statistical tests alone could benefit from additional tools.

We demonstrate framework operation using lattice QCD quark mass ratios—deliberately selected as the hardest test case (N=3 parameters at 2% precision, maximum combinatorial coincidence risk). The **Diagnostic Pattern** $2(m_d/m_u)^3 \approx m_s/m_d$ achieves 0.16σ statistical agreement yet self-falsifies through directional divergence: as uncertainties improved 37%, central values converged toward 2.162 rather than the predicted 2.154, with statistical significance doubling from 0.075σ to 0.16σ. This demonstrates successful filtering of numerical coincidence despite passing traditional validation.

The framework's discriminatory capability is validated through historical test cases: the Gell-Mann-Okubo relation correctly passes all criteria, demonstrating that physically meaningful patterns survive multi-criteria evaluation. Framework value is methodology-independent—we demonstrate filtering through failed patterns, not to advocate specific

physics. Initial thresholds serve as community starting points; the contribution is establishing systematic, pre-registration-based standards for pattern evaluation in any domain where computational precision outpaces dimensional growth.

Keywords: Temporal convergence, Pattern evaluation, Pre-registration, Computational methodology, High-precision measurement, Statistical validation, Parameter space search, Numerical coincidence

1 Introduction

1.1 Universal Challenge: Precision Outpacing Dimensionality

Computational methods across scientific domains share a common trajectory: iterative refinement improves measurement precision faster than dimensional growth expands parameter spaces. This precision regime creates opportunities for refined evaluation methods. When uncertainties decrease to 2–3% relative precision while parameter count remains at three to six values, combinatorial search through simple formulas almost guarantees discovery of statistically significant relationships—whether physically meaningful or coincidental.

The challenge manifests across domains. Particle physics: three light quark masses at 2% precision enable countless ratio tests. Cosmology: six CMB parameters from Planck enable systematic formula mining. Condensed matter: critical exponent relationships at phase transitions. Fundamental constant metrology: searching for time-variation or inter-constant relationships at 10^{-9} precision. Machine learning: hyperparameter optimization exploring vast hypothesis spaces where spurious correlations achieve significance. In each case, discrimination might be enhanced through additional methodological tools when search space exceeds parameter dimensionality by orders of magnitude.

Computational fluid dynamics provides a concrete industrial parallel. Navier-Stokes equations cannot be solved analytically in turbulent regimes, requiring iterative approximation through mesh refinement. Critically, CFD validation employs temporal stability: solutions must converge across both spatial refinement and time-stepping iterations. A result matching experimental data at one mesh resolution and timestep is insufficient—convergence must persist across independent temporal and spatial scales. Our framework applies identical methodology to empirical pattern evaluation: statistical agreement at one measurement vintage is insufficient. Patterns must demonstrate temporal stability as experimental precision improves across independent data releases. The added temporal dimension distinguishes genuine structure from numerical artifacts in both domains.

We demonstrate a solution using particle physics quark masses specifically because N=3 represents the methodological worst case. With only three parameters, nearly any algebraically simple formula achieves statistical agreement by chance. If our framework successfully filters coincidences in this regime while preserving historically validated patterns (Koide formula, Gell-Mann-Okubo relations), it provides discrimination in any higher-dimensional context.

The core methodological innovation: *temporal convergence through timestamped predictions*. Pattern predictions are established with timestamping, then tracked against future data releases for directional convergence or stability as precision improves. This requirement

provides robust protection against retroactive fitting—reducing susceptibility to common biases including data selection and post-hoc hypothesis adjustment. When combined with traditional validation criteria, temporal tracking transforms statistical consistency from more easily obtained to discriminatory.

1.2 Historical Precedent and Methodological Gap

Empirical mass relations have historical precedent in revealing physical structure before theoretical understanding emerges. The Gell-Mann-Okubo formula [4, 5] related hadron masses through what was later understood as SU(3) flavor symmetry breaking, identified empirically before the quark model existed. The Koide formula [3] describes charged lepton masses with remarkable precision despite lacking theoretical derivation after four decades. These historical successes demonstrate that empirically robust patterns can guide theoretical development. However, the field still lacks methodology for distinguishing potentially meaningful patterns from numerical artifacts in the precision era. Traditional particle physics methodology was optimized for discovery physics: predict new particle, build detector, confirm or exclude. Understanding mathematical structure of already-known parameters requires different standards.

1.3 Cross-Disciplinary Opportunity

Lattice QCD precision improvements create opportunities for productive empirical phenomenology. Current uncertainties enable discriminatory tests while remaining above the noise floor where combinatorial coincidences overwhelm signal. This regime may persist for approximately a decade as systematic improvements continue.

Explicit evaluation criteria can facilitate cross-disciplinary contributions by providing clear operational targets. The framework separates empirical validation (criteria 1–6) from theoretical explanation (criterion 7), enabling researchers from computational backgrounds to contribute validated patterns that theorists can subsequently investigate.

This work proposes one framework emphasizing systematic evaluation to complement existing peer review processes. Community discussion and iterative refinement can help establish whether these specific seven criteria and associated thresholds provide useful operational standards as experimental precision enables discriminatory phenomenology.

1.4 Motivating Example: When Statistical Significance Proves Insufficient

Framework development was motivated by observing patterns that pass traditional statistical validation yet appear physically questionable. Consider algebraic unification of two consistently reported lattice QCD ratios: $m_s/m_d \approx 20$ and $m_d/m_u \approx 2.16$. These published values approximately satisfy:

$$2(m_d/m_u)^3 \approx m_s/m_d \tag{1}$$

With FLAG 2024 central values $m_d/m_u = 2.162 \pm 0.050$ and $m_s/m_d = 20.0 \pm 0.5$ [1], the pattern predicts $2(2.162)^3 = 20.22$ versus measured 20.0, a deviation of 0.16σ.

We emphasize that this **Diagnostic Pattern** serves purely as a methodological calibration tool—a relationship we expect to fail—to demonstrate the framework's filtering capability, not as a physics claim. We intentionally selected this pattern not as physics claim but as methodological demonstration. We will voluntarily self-falsify it using two framework criteria, showing how the framework enables filtering before peer review and creating concrete example for authors of high rigor standards.

We term this the **Diagnostic Pattern**, emphasizing its role as methodological tool. The pattern achieves statistical consistency (0.16σ deviation) but lacks discriminatory power with only three light quark masses. Traditional hypothesis testing would not reject this pattern, motivating our multi-criteria framework. Despite agreement with lattice world averages within 0.16σ, scale invariance, and cross-collaboration consistency, directional trends and theoretical considerations suggest physical implausibility. The pattern fails criterion 4 through directional divergence (central values trend away from prediction as precision improves). Additionally, extensive survey of flavor symmetry frameworks revealed no structural precedent for the implied Yukawa texture $2y_d^4 = y_u^3 y_s$. The relationship appears incongruent with established symmetry-breaking patterns to the point of lacking physical sensibility. While this constitutes author judgment rather than rigorous proof of impossibility, the combination of directional divergence and apparent physical implausibility motivates self-falsification.

This double failure (directional divergence at criterion 4 combined with physically implausible structure at criterion 7) motivates self-falsification rather than journal submission. Were the pattern converging despite lacking theory, submission to empirical journals would be appropriate. Were theory identified despite current divergence, extended temporal validation might be warranted. Authors bear responsibility for judging pattern severity: framework criteria enable evaluation, venue selection (high-impact journal, empirical journal, preprint archive, or self-falsification) reflects pattern status across multiple dimensions.

Traditional methodology provides no systematic answer: relying solely on p-values would justify publication despite likely physical incorrectness; relying on subjective judgment enables arbitrary rejection. Neither approach provides adequate evaluation. This motivated framework development.

Related work. Our seven-criteria approach addresses this gap by operationalizing reproducibility principles (pre-registration, data availability, temporal validation) within the precision-measurement domain. It dovetails with established notions of scale invariance and parameter compression, and with contemporary reproducibility practice emphasizing data availability and pre-registered evaluation rules (see recent FLAG and PDG reviews [1, 2]).

[1] FLAG reports $m_s/m_{ud} = 27.3 \pm 0.1$. Combined with $m_u/m_d = 0.48 \pm 0.02$, this yields $m_s/m_d = 20.2 \pm 0.5$. We use the rounded value 20.0 for this methodological demonstration.

2 Methods

2.1 Framework Architecture: Self-Falsification Before Expert Review

The proposed framework addresses complementary objectives: protecting theoretical community resources while enabling legitimate empirical contributions. Current practice offers an opportunity to address both objectives through systematic criteria: explicit standards can enable consistent evaluation while facilitating legitimate empirical contributions. The seven criteria operationalize this through filter pipeline:

Criteria 1–6 (Objective Self-Falsification Gates): Authors demonstrate pattern satisfaction before submission; reviewers mechanically verify without subjective judgment; domain expertise in phenomenology sufficient; patterns failing objective tests filter pre-submission.

Criterion 7 (Theoretical Viability as Benchmark): Only patterns surviving criteria 1–6 warrant theoretical attention. Criterion 7 stratifies patterns by theoretical grounding for theorists developing flavor models. When theory predicts X and framework-validated pattern shows X at sub-σ precision, productive collaboration emerges naturally. Absence of explanation with passing empirical criteria does not constitute automatic rejection; patterns stratify by venue type based on Criterion 7 status.

The framework's computational and temporal requirements self-select for serious contributions: patterns must survive major experimental releases before publication, transforming enthusiasm into evidence through sustained agreement.

2.2 The Seven Criteria with Operational Thresholds

2.2.1 Criterion 1: Scale Invariance Under Renormalization Group Evolution

Mass ratios must be scale-invariant under QCD running (deviations $< 10^{-4}$ across 1 GeV to TeV scales; indicative order of magnitude, community to refine as precision improves). This basic prerequisite eliminates scale-dependent relationships.

2.2.2 Criterion 2: Compression of Degrees of Freedom

Patterns must reduce N parameters to N–1 degrees of freedom through unified constraints (e.g., $2(m_d/m_u)^3 = m_s/m_d$ reduces three masses to two DOF).

2.2.3 Criterion 3: Statistical Agreement at Discriminatory Precision

Patterns must agree with measurements within statistical bounds at discriminatory precision. Community to determine thresholds: $< 1\sigma$? $< 2\%$ relative uncertainty? The key requirement: precision sufficient to distinguish between competing formulations.

2.2.4 Criterion 4: Temporal Persistence

Patterns must (a) be pre-registered via timestamped repository before new data releases, and (b) maintain directional convergence or stability through experimental cycles. Community to refine as release frequency evolves.

Robust Protection Through Pre-Registration:

Pre-registration via timestamped public repositories (Zenodo, institutional archives, or preprint servers) creates an immutable record before new measurements become available. This approach reduces susceptibility to:

- Selective data usage favoring particular vintages
- Post-hoc formula adjustment to match updated values
- Retroactive mining through multiple hypothesis variations
- Retrospective claims following observed convergence patterns

This transforms pattern evaluation from statistical (where combinatorial search can make it difficult to discern physically grounded relationships from numerical coincidence) to temporal (fundamentally resistant to data mining). The only way to pass Criterion 4 is genuine predictive success across independent experimental cycles—precisely the evidence distinguishing structure from coincidence.

First-pass threshold rationale. The operational thresholds proposed throughout this framework (three-cycle minimum, specific tolerance bounds, etc.) represent **initial proposals for community evaluation and refinement**, not definitive standards. The three-cycle minimum provides the shortest sequence with any statistical meaning for detecting monotonic trends—two cycles cannot distinguish trend from noise, while four or more cycles would require over a decade of FLAG releases given current cadence. The slope bound $|S| < 0.5\,\sigma_{\text{avg}}/\Delta t$ ties tolerance to characteristic measurement noise. Researchers in specific domains should refine these thresholds based on their release frequency, precision trajectories, and acceptable false-positive rates. We emphasize these choices as operational starting points inviting community debate, not as rigid requirements.

2.2.5 Criterion 5: Mathematical Simplicity

Patterns should minimize complexity. Community to decide: Kolmogorov complexity threshold? Allow transcendentals? Maximum operations? For now: basic arithmetic, small integers (≤ 5), standard constants (π, e).

2.2.6 Criterion 6: Independent Validation Across Multiple Determinations

Patterns must show consistency across independent experimental/computational approaches.

Threshold: Agreement across ≥ 3 independent collaborations/methods with different systematic uncertainties.

2.2.7 Criterion 7: Theoretical Viability as Benchmark Stratification

Empirical relationships must not be demonstrably incompatible with existing theoretical frameworks. Critically: absence of explanation does not constitute automatic failure. Criterion 7 has three possible outcomes:

PASS (Compatible): Flavor theorist identifies mechanism within existing frameworks producing the pattern.

PASS (Unknown): No known incompatibility with gauge invariance, anomaly cancellation, or existing symmetry structures. Pattern awaits theoretical investigation but is not ruled out. This constitutes legitimate publication; empirically robust observations merit documentation even without explanation.

FAIL (Incompatible at present): Theorist demonstrates pattern violates fundamental constraints through explicit proof.

Patterns in Unknown status merit documentation as empirical benchmarks. The field benefits from a searchable repository of published patterns where theorists can identify observations matching their predictions.

Division of labor: Empiricists validate patterns through criteria 1–6; theorists may provide explanations or identify incompatibilities. Neither group bears obligation to the other. The framework separates empirical validation from theoretical explanation: different expertise, different contributions.

Self-falsification option: Authors may choose to self-falsify at criterion 7 if they survey existing frameworks, find no support, and possess no new theory to propose. This demonstrates intellectual rigor and filters patterns pre-submission. However, such self-falsification is voluntary; a pattern lacking known mechanism but compatible with fundamental constraints remains valid submission.

Application to the Diagnostic Pattern: We surveyed existing flavor symmetry mechanisms and identified no framework naturally producing the Yukawa texture $2y_d^4 = y_u^3 y_s$. We possess no new theoretical framework to propose. We therefore choose to self-falsify the Diagnostic Pattern at criterion 7. We present this pattern not as viable submission but as demonstration of how authors should filter implausible patterns before peer review.

3 Results: Framework Validation Through Test Cases

We demonstrate framework operation through two test cases: an established historical pattern (Gell-Mann-Okubo relation) that should pass, and the Diagnostic Pattern that correctly self-falsifies.

3.1 Gell-Mann-Okubo Relation: Historical Precedent

The GMO relation [4,5] predicted hadron mass relationships before the quark model existed. Framework evaluation (retrospective):

GMO relation demonstrates framework would enable historically important empirical observations even before theoretical understanding emerges. The pattern passed empirical criteria (1–6) and was compatible with fundamental constraints (criterion 7: Unknown), allowing publication that guided subsequent theory development.

Table 1: GMO Relation Evaluation

Criterion	Assessment
1. Scale Inv.	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 SU(3) symmetry structure
6. Independent	PASS - multiple hadron measurements
7. Theoretical	**PASS (Unknown at time)** - emerged later
Status: Would pass BEFORE quark model	

3.2 Diagnostic Pattern: Demonstration of Self-Falsification

The Diagnostic Pattern $2(m_d/m_u)^3 \approx m_s/m_d$ demonstrates framework filtering capability:

Table 2: Diagnostic Pattern Evaluation

Criterion	Assessment
1. Scale Inv.	PASS - preserved under QCD RG 1 GeV–1 TeV
2. Compression	PASS - reduces 3 masses to 2 DOF
3. Statistical	PASS - FLAG 2024: within 0.16σ
4. Temporal	**FAIL** - directional divergence
5. Simplicity	PASS - single equation, integer coefficient
6. Independent	PASS - consistent across ETM, BMW, MILC
7. Theoretical	**FAIL** - no mechanism; no theory
Status: Doubly self-falsified at 4 and 7	

Detailed technical analysis (RG evolution, cross-collaboration comparisons, statistical methodology, temporal tracking) provided in Appendix A.

Key findings:

Criterion 4 (Directional Divergence): Central values moved from $m_d/m_u = 2.16 \pm 0.08$ to 2.162 ± 0.050 (away from predicted 2.154) between consecutive reviews. While both measurements remain within 1σ, the 37% uncertainty reduction caused statistical significance to double from 0.075σ to 0.16σ—measurements are converging toward 2.162, not the predicted 2.154. The critical observation is not just the +0.002 directional movement, but that shrinking uncertainties are converging around a value systematically above the prediction, doubling the statistical significance of the deviation. See Appendix A for detailed interpretation of why shrinking error bars revealing directional offset constitute stronger falsification than traditional statistical disagreement.

Criterion 7 (Absence of Support): Survey of existing flavor frameworks identified no

mechanism producing Yukawa texture $2y_d^4 = y_u^3 y_s$. Absent new theoretical framework to propose, pattern self-falsifies at Criterion 7. Theoretical rescue remains possible if flavor theorist identifies compatible mechanism.

The pattern remains open to two rescue pathways: temporal convergence through future measurements (as Koide achieved over decades), or theoretical rescue through identifying compatible mechanisms. However, theoretical rescue would require mechanisms beyond current flavor symmetry formulations, as no existing framework naturally produces the implied Yukawa texture $2y_d^4 = y_u^3 y_s$. We conclude the pattern is likely a numerical coincidence despite current statistical consistency.

Statistical consistency was achieved trivially. This triviality led us to explicitly exclude Monte Carlo p-values from our framework criteria entirely. When statistical significance becomes computationally trivial, it provides no meaningful filter. With only three light quarks and $\mathcal{O}(10^2$–$10^3)$ possible simple formulas to test, finding statistically significant relationships becomes trivial. This constraint on statistical power makes multi-criteria evaluation essential: any single test is weak, but multiple independent criteria provide discriminatory capability.

The Diagnostic Pattern's double self-falsification validates framework discriminatory capability: patterns pass traditional metrics yet correctly self-falsify through directional persistence and theoretical viability criteria.

4 Discussion

4.1 Framework Operation and Methodological Questions

The Diagnostic Pattern identified two failure points: directional divergence (criterion 4) and absence of theoretical support (criterion 7). Rather than submitting this pattern and asking reviewers to make these determinations, we self-falsify, exemplifying the framework's protective function: authors bear responsibility for evaluation before requesting expert attention. Such self-falsification represents valuable scientific contribution, documenting what does not work with the same rigor as what does.

The Diagnostic Pattern's self-falsification raises questions for community debate: (1) Should theoretical viability be first rather than last? (2) Does scale invariance provide sufficient discrimination? (3) How strict should directional convergence be? (4) What constitutes adequate temporal validation?

The Diagnostic Pattern's dual self-falsification illustrates two distinct rescue pathways with different barriers. *Temporal rescue* requires only future measurement convergence—subsequent FLAG reviews showing m_d/m_u approaching $10^{1/3}$ would rehabilitate the pattern, as Koide's formula achieved through decades of validation. *Theoretical rescue* presents a higher barrier: extensive survey of flavor symmetry frameworks found no precedent for the Yukawa texture $2y_d^4 = y_u^3 y_s$, suggesting rescue would require revolutionary new mass generation models beyond current formulations rather than incremental refinement of existing approaches.

The framework treats these rescue pathways asymmetrically by design. Temporal evolution is monitored through Criterion 4's pre-registration requirement, enabling patterns to

demonstrate convergence across multiple experimental cycles. Theoretical viability stratifies patterns through Criterion 7 into a searchable benchmark database, allowing theorists to identify empirical candidates matching model predictions. Patterns failing Criterion 4 but passing Criterion 7 (Unknown status) merit continued temporal monitoring and documentation. Patterns failing both criteria, as demonstrated here, warrant self-falsification despite current statistical consistency—the framework successfully filters likely numerical coincidences before peer review.

4.2 Historical Context: Complementary Methodologies

Note on historical benchmarks. We cite the Koide relation (1982) and the Gell-Mann-Okubo relation solely as historical benchmarks demonstrating that empirically robust patterns can precede theoretical understanding. We do not evaluate, endorse, or make claims about Koide here. All quantitative analyses and conclusions in this manuscript rely exclusively on FLAG/PDG quark mass-ratio determinations and the Diagnostic Pattern analyzed in Section 3.3.

Contemporary particle physics benefits from multiple complementary approaches. Theoretical frameworks and experimental precision have both advanced substantially, creating opportunities for empirical phenomenology to bridge these domains. Explicit evaluation criteria can help identify robust patterns that warrant theoretical investigation while filtering numerical artifacts efficiently.

4.3 Framework Benefits for Collaborative Research

The framework can facilitate collaboration between computational and theoretical researchers. Framework-validated patterns published in appropriate venues become accessible benchmarks. When theorists develop flavor models, they can check whether their predictions match documented empirical observations, creating natural opportunities for productive collaboration. This enables temporal decoupling: patterns documented today may find theoretical explanation when relevant framework development reaches that parameter space, without requiring simultaneous empirical validation and theoretical explanation from the same authors.

4.4 The Meta-Pattern Horizon

The framework's long-term value may emerge through meta-pattern synthesis: relationships between multiple validated patterns revealing latent structure inaccessible to individual observations. Without systematic documentation—including patterns passing empirical criteria (1-6) but lacking mechanisms—such higher-order correlations remain undiscoverable. Historical precedent demonstrates viability: the Balmer series appeared numerological until spectroscopic multiplicity collectively revealed atomic structure. Our framework transforms isolated observations into a curated corpus where emergent mathematical structure in fundamental parameters becomes systematically tractable.

4.5 Limitations and Community Refinement

This framework represents a proposed starting point requiring community evaluation and refinement. The goal is initiating systematic discussion: *what standards should govern empirical pattern evaluation as precision enables discriminatory tests?* Whether these specific seven criteria and thresholds prove optimal matters less than establishing that explicit, debatable standards must replace subjective evaluation.

5 Conclusion

Recent lattice QCD precision achievements create opportunities for systematic empirical phenomenology in particle physics. As experimental precision enables discriminatory tests of mathematical structure in fundamental parameters, the field would benefit from explicit evaluation methodology to complement existing peer review processes.

We propose seven explicit criteria forming a filter pipeline that enables self-falsification through objective criteria while creating validated benchmarks for theory testing. The Diagnostic Pattern demonstrates intended operation: passing multiple objective criteria yet self-falsifying at directional persistence and theoretical viability. The framework successfully filters potentially spurious patterns while allowing robust observations (exemplified by GMO) to proceed, creating empirical benchmarks for theoretical development.

Future work should focus on: (1) community refinement of criterion thresholds and weighting, (2) application to other fermion mass hierarchies and mixing parameters, (3) theoretical investigation of patterns currently lacking known mechanisms, (4) testing framework utility through practical application, (5) iterative improvement based on community feedback.

This framework represents a starting point for community discussion on systematic evaluation standards as experimental precision enables increasingly discriminatory tests of mathematical structure in fundamental parameters.

Acknowledgments

The author thanks Riccardo M. Pagliarella, Ph.D. for encouragement and valuable discussions. This work would not be possible without the extraordinary precision achieved by the lattice QCD community, particularly major collaborations including FLAG, BMW, MILC, HPQCD, and ETM. The author is grateful to the MILC collaboration for making their QCD running code publicly available.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

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 exact FLAG 2019–2024 ratios and derived Δ_i values for temporal convergence analysis are provided in **S1_data.csv** and mirrored at https://github.com/AndBrilliant/Tempo (commit `135deaf`). Minimal reproduction scripts for generating diagnostic plots are included in the repository. Analysis methodologies and computational procedures are fully described in the text and Appendix A.

A Diagnostic Pattern: Detailed Technical Analysis

This appendix provides comprehensive technical documentation of the Diagnostic Pattern's evaluation through framework criteria. While the pattern self-falsifies (Section 3.3), detailed analysis demonstrates methodology for systematic evaluation.

A.1 Scale Invariance: QCD Renormalization Group Evolution

Implementation of MILC collaboration RG running algorithms with two-loop anomalous dimensions, flavor threshold matching, and 200-step numerical integration. Cross-validated against published tabulated values at 25 reference scales.

Table 3: Scale invariance verification using FLAG 2024 central values

Scale (GeV)	m_d/m_u	m_s/m_d	$2(m_d/m_u)^3$	Deviation
1.0	2.162	20.0	20.22	0.22
2.0	2.162	20.0	20.22	0.22
4.2	2.162	20.0	20.22	0.22
10.0	2.162	20.0	20.22	0.22
91.2	2.162	20.0	20.22	0.22
173.0	2.162	20.0	20.22	0.22
1000.0	2.162	20.0	20.22	0.22

Note: Values remain constant across all scales, confirming scale invariance of mass ratios under QCD running. Deviation of 0.22 corresponds to 0.16σ when propagating uncertainties from $m_d/m_u = 2.162 \pm 0.050$.

A.2 Cross-Collaboration Validation

A.3 Statistical Analysis

For the pattern $2(m_d/m_u)^3 \approx m_s/m_d$:

Table 4: Cross-collaboration consistency analysis

Collaboration	m_d/m_u	Uncertainty	Action	Deviation
ETM	2.15 ± 0.08	3.7%	Twisted mass	0.05σ
BMW	2.17 ± 0.07	3.2%	Stout smearing	0.23σ
MILC	2.15 ± 0.08	3.7%	Staggered	0.05σ
HPQCD	2.14 ± 0.06	2.8%	HISQ	0.24σ
World Average	2.162 ± 0.050	2.3%	Combined	0.16σ

- Using FLAG 2024: $m_d/m_u = 2.162 \pm 0.050$, $m_s/m_d = 20.0 \pm 0.5$
- Left side: $2(2.162)^3 = 20.22$
- Right side: 20.0
- Propagated uncertainty: $\sigma = \sqrt{(6 \cdot 2.162^2 \cdot 0.050)^2 + 0.5^2} \approx 1.42$
- Normalized deviation: $(20.22 - 20.0)/1.42 = 0.16\sigma$

We treat m_d/m_u and m_s/m_d as uncorrelated for simplicity; correlated uncertainty treatment would not alter the methodological conclusion.

A.4 Temporal Persistence Analysis

Table 5: Temporal evolution of m_d/m_u determinations

Review	m_d/m_u	Rel. Uncert.	Distance from $10^{1/3}$
FLAG 2019	2.16 ± 0.08	3.7%	$+0.006$ (0.07σ)
FLAG 2024	2.162 ± 0.050	2.3%	$+0.008$ (0.16σ)
Direction: **Away from prediction**			
Uncertainty improvement: 37.5%			
Criterion 4 verdict: **FAILS** (directional divergence)			

Central value moved away from predicted 2.154 while uncertainty tightened. Under directional convergence requirement, this constitutes self-falsification despite remaining within statistical bounds.

A.5 Interpreting Temporal Falsification

Traditional statistical evaluation would note that both FLAG 2019 and FLAG 2024 measurements remain well within 1σ of the predicted value $m_d/m_u = 10^{1/3} \approx 2.154$, seemingly indicating robust temporal stability. However, this perspective misses critical information revealed by simultaneous error bar reduction and directional movement.

Why Error Bar Dynamics Matter: Between FLAG 2019 and FLAG 2024, relative uncertainty decreased by 37.5% (from 3.7% to 2.3%), representing substantial systematic and statistical improvements by the lattice QCD community. This precision gain creates an

effective "zoom in" on the true value. When error bars shrink while measurements remain stable around a predicted value, this provides strong validation (as observed with Koide's formula over decades). Conversely, when error bars shrink while central values converge around a *different* location than predicted, this reveals a systematic offset that broader uncertainties previously obscured.

Statistical Significance Evolution: The Diagnostic Pattern demonstrates the problematic scenario. Despite both measurements remaining within 1σ, statistical significance actually *increased*:

- FLAG 2019: Measured 2.16, predicted 2.154, uncertainty 0.08 $\rightarrow (2.16 - 2.154)/0.08 = 0.075\sigma$
- FLAG 2024: Measured 2.162, predicted 2.154, uncertainty 0.050 $\rightarrow (2.162 - 2.154)/0.050 = 0.16\sigma$

The deviation in units of uncertainty *doubled* from 0.075σ to 0.16σ despite only 0.002 absolute movement. This occurs because the denominator (uncertainty) decreased faster than any convergence of the central value toward prediction. The measurements are converging toward approximately 2.162, not the predicted 2.154.

Contrast with Physical Patterns: When genuine physical relationships exist, precision improvements reveal convergence *toward* predicted values, with statistical significance improving (deviations decreasing in units of σ). The Diagnostic Pattern exhibits the opposite behavior: precision improvements reveal convergence around the wrong value, causing statistical significance to *worsen* (deviations increasing in units of σ).

Physical Interpretation: This temporal behavior suggests the pattern's current statistical consistency results from limited discriminatory power at 2-3% precision rather than underlying physical mechanism. The relationship $2(m_d/m_u)^3 = m_s/m_d$ appears to be a numerical near-coincidence: close enough to pass tests at current precision, but systematic improvements reveal the offset. Projecting forward, continued uncertainty reduction to sub-1% precision would likely expose statistically significant disagreement, revealing the coincidental nature before theoretical investment occurs.

Framework Operation: Criterion 4 intentionally flags this scenario for self-falsification. The requirement for directional convergence or stability prevents patterns from reaching advanced review stages when precision improvements systematically reveal offsets. This filtering occurs despite formal statistical consistency, demonstrating how multi-criteria evaluation provides discrimination beyond hypothesis testing alone. The Diagnostic Pattern's temporal behavior, combined with absence of theoretical support (Criterion 7), motivated voluntary self-falsification rather than submission.

References

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

[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).