London Journal of PHYSICS

Volume 3 No 1

# Stratification Criteria for Machine Learning Pattern Discovery in Particle Physics: Preparing for the AlphaFold Moment

**Andrew Michael Brilliant**

Independent Researcher, Applied Dynamics Research, Sapporo, Japan
Email: ab@ad-research.org

June 9, 2026

**Abstract**

Machine learning capabilities are expanding into scientific domains at an accelerating pace. When applied to high-energy physics pattern discovery, they will generate candidates faster than traditional evaluation can absorb.

ML finds patterns in past data—it is inherently post hoc. Whether those patterns reflect structure or coincidence is unknowable at discovery time; this limitation applies equally to human and computational pattern-finding. What differs is scale: ML candidate generation is effectively unbounded, while human evaluation capacity remains fixed. When generation rate exceeds evaluation bandwidth, binary accept/reject degenerates to random sampling. Information-theoretically, the only response that preserves ranking under finite evaluation budget is stratification. By focusing on stratification rather than binary filtering, rule adjustments can be made retroactively, thresholds tuned as results accumulate, and evaluation bandwidth focused on top-ranked candidates.

This paper attempts to codify those criteria, proposing seven computationally evaluable standards for stratifying ML-generated patterns. The goal is not to deliver verdicts but to prioritize which candidates merit pre-registration and longitudinal tracking. The framework preserves the essential paradigm: pattern plus theory equals potentially real physics. Patterns alone, however striking, remain candidates until theoretical understanding arrives.

Making these criteria explicit enables prefiltering at scale while creating a collaborative resource rather than a competitive one. ML capabilities extend what physicists can search while preserving how physicists evaluate. We offer this provisional framework for community calibration, with the goal of developing validation infrastructure before the capability fully arrives.

## 1 Introduction

Particle physics has developed sophisticated standards for evaluating numerical patterns. When someone claims a "striking" relationship among measured quantities, trained physicists know how to respond. They ask about temporal priority: was this form predicted before seeing the data, or discovered by searching? They probe for scale dependence: does the relationship hold across energy scales, or only at one? They demand trial accounting: how many functional forms were attempted before this one succeeded? These questions reflect hard-won disciplinary knowledge about how numerical coincidences arise and how to distinguish them from genuine structure.

 DOI 10.69710/ljp.v3i1.17825

Portions of this evaluative expertise have remained tacit. It transmits through apprenticeship, through referee reports, seminar questions, and advisor feedback, rather than through explicit codification. High-energy physics statistical practice already embodies Mayo's concept of "severe testing" [24, 28] at an operational level, even when practitioners do not consciously invoke the philosophical framework. HEP's procedural apparatus—blind analyses, systematic uncertainty audits, Monte Carlo coverage studies, and independent replication across collaborations—constitutes severe testing in the error-statistical sense (Section 6.2). The standards exist; they have simply never required formal articulation.

This paper is offered as a provisional framework for community calibration, not a finished standard. The criteria proposed here represent one researcher's attempt to make evaluation logic explicit; they are intended as a starting point for discussion, not a prescriptive system. Several thresholds are deliberately left coarse, and the sterile neutrino worked example (Section 5.2.1) would benefit from domain expert calibration of the severity assessments. We invite corrections, counterexamples, and alternative formulations. The value of this document lies not in getting every criterion right on the first pass, but in making the criteria explicit enough to be wrong in identifiable, correctable ways.

Machine learning changes this requirement. ML systems searching high-dimensional parameter spaces will generate pattern candidates faster than traditionally trained physicists can evaluate them [3, 4]. Beyond Standard Model theories can involve over 100 free parameters and configuration spaces of $10^9$ to $10^{12}$ points. ML will find patterns in such spaces. Many patterns. Most will be spurious. The practitioners deploying these systems may not have absorbed, through years of disciplinary apprenticeship, the implicit standards that experienced physicists apply automatically.

Why focus on mass relationships? The problem has a specific structure that computational search is suited to exploit. The observables are few: nine precisely measured fermion masses. The search space is vast: every algebraic relationship among them, every transcendental expression, every group-theoretic construction. And the literature is fragmented: Koide (1982) [10], Barut (1979) [17], Rivero (2005) [18], Brannen (2006) [19], Kocik (2012) [20], each explored a small patch of this space by hand over years. No individual researcher can survey the full combinatorial landscape. This is the same structural bottleneck that made protein folding an ideal target for computational methods [6]: low-dimensional observables, an enormous configuration space, and decades of partial human exploration that collectively covered only a fraction of the possibilities. Mass relationships are chosen here not because they are important (though they are—the Standard Model accommodates masses but does not predict them) but because the problem has the right shape for corpus-scale search to offer a structural advantage over individual researchers.

Koide-like formulas (simple algebraic expressions relating measured masses) represent precisely the pattern type ML systems will generate at scale. They require no physical insight to propose, only numerical search. The four-decade community response to Koide's original observation illustrates the evaluation problem: genuine interest, justified skepticism, and no resolution. This paper uses mass relationships as a worked example not to advocate for any specific formula, but because they cleanly illustrate evaluation challenges ML will amplify.

A second example provides something stronger than historical illustration. The seven criteria were formalized before MicroBooNE [1] published its definitive test of the LSND sterile neutrino anomaly in late 2025. The sterile neutrino anomalies—among the most prominent unresolved questions in neutrino physics over the past two decades—therefore reached experimental resolution after the framework existed, providing a rare case where the criteria can be evaluated prospectively against known ground truth.

This paper attempts to define explicit criteria for this purpose. We propose seven criteria for evaluating ML-generated pattern candidates, designed to operationalize what trained physicists

already know: that pre-data prediction matters, that scale-dependent relations are suspect, that trial factors must be reported, and that temporal convergence (stable or improving agreement as measurement precision increases) functions as a severe test that coincidences systematically fail.

Three features define the approach:

First, the framework addresses a specific decision: which patterns merit pre-registration and longitudinal tracking across future data releases. The output is not "this pattern is correct" but "this pattern is worth timestamping and waiting on."

Second, the criteria are designed to be computationally evaluable where possible, enabling application to large candidate sets. Continuous scores rather than binary verdicts allow stratification across millions of candidates.

Third, the framework explicitly acknowledges a precision regime problem. Current quark mass determinations offer approximately 35,000 times less discriminatory power than charged lepton masses (Section 2.2). This gap means statistical agreement alone, the traditional first filter, provides minimal evidence of structure over coincidence. The look-elsewhere effect [5] that particle physics takes seriously for mass peak searches applies equally to numerical pattern searches; the trial factor problem is severe. Different filtering criteria are needed, and those criteria must be explicit rather than implicit.

The paper proceeds as follows. Section 2 characterizes the pattern validation problem, including the discriminatory power gap that makes quark phenomenology qualitatively different from lepton phenomenology. Section 3 examines a precedent from structural biology where similar infrastructure questions arose. Section 4 presents the seven criteria. Section 5 illustrates the criteria against historical cases. Section 6 addresses implementation and limitations.

## 2 The Pattern Validation Problem

### 2.1 Numerical Patterns in Physics

In 1982, Yoshio Koide observed a striking relationship among the charged lepton masses [10]:

$$Q = \frac{m_e + m_\mu + m_\tau}{(\sqrt{m_e} + \sqrt{m_\mu} + \sqrt{m_\tau})^2} = \frac{2}{3} \tag{1}$$

The formula holds to approximately 0.01%, far better than would be expected by chance. Four decades later, it remains unexplained. No derivation from Standard Model principles exists. Many have tried to extend the relationship to quarks; none have achieved the same precision or persistence.

This isolation is what makes the formula epistemologically stuck, not rejected. No theory predicts it. It predicts nothing else. The formula teases us with the possibility that deeper structure exists in particle masses. It cannot tell us whether that structure is real.

The distinction between "real physics" and "mere formula" is less clear than it might seem. Ampère and others developed descriptive equations for electromagnetism decades before Maxwell unified them into a coherent theory. Superconductivity was observed in 1911; the phenomenological equations proved useful for engineering; the microscopic explanation (BCS theory) arrived in 1957. During those forty-six years, superconductivity was "real physics" despite lacking theoretical derivation. The patterns worked; the explanation waited.

### 2.2 The Discriminatory Power Gap

The precision available for testing patterns varies enormously across observables. This variation has profound implications for how we should evaluate claimed relationships.

 DOI 10.69710/ljp.v3i1.17825

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

**Charged Leptons:**

The dynamic range spans $\tau/e \approx 3{,}500$. All masses are pole masses: static, directly measurable, no scale dependence. Two anchors (electron, muon) are known to extraordinary precision. With the worst precision being the tau at 0.007%, there are approximately **50 million distinguishable positions** across this parameter space.

**Light Quarks:**

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

The dynamic range is only $s/u \approx 43$. All masses are $\overline{\text{MS}}$ running masses. No anchors: all three have comparable relative uncertainties. With the worst precision at 3%, there are approximately **1,400 distinguishable positions**.

**The ratio:** 50,000,000 / 1,400 ≈ 35,000×

This is an order-of-magnitude estimate, not a precision claim; the essential point is robust. Leptons offer roughly $3.5 \times 10^4$ times greater discriminatory power than light quarks. Historical intuitions calibrated on lepton phenomenology systematically underestimate coincidence rates in quark phenomenology. A pattern that would be striking among leptons may be unremarkable among quarks.

This is why the Koide formula's persistence matters: four decades of improving measurements have not degraded the agreement. That temporal survival is what coincidences fail to achieve. We posit that temporal convergence functions as a severe test in the sense formalized by Mayo [24]: a hypothesis gains credibility when it passes tests that would likely reveal flaws if flaws existed.

Beyond precision, the lepton sector enjoys a structural advantage: charged lepton masses run negligibly under QED ($\alpha \approx 1/137$), making pole masses effectively scale-invariant. Quark masses run substantially under QCD ($\alpha_s \approx 0.1$–$0.3$), and relations that hold at one scale may fail at another. This constraint is documented by Xing and Zhang [11] and reflected in the community's skepticism toward quark Koide extensions [12]. The Koide formula evaluated for heavy quarks ($c$, $b$, $t$) at pole masses yields $Q \approx 0.63$, unremarkable and unpublished, illustrating how implicit community filters operate before formal evaluation begins.

This asymmetry reveals a deeper methodological point. The Koide formula represents a minimum viable pattern: complexity of approximately 3 (a ratio of mass sums), exact agreement with 2/3, and effective scale invariance. Extensions to quarks face a structural disadvantage. QCD running requires either specifying a scale or adding correction terms, increasing complexity beyond a formula already classified as numerology. The community's implicit rejection of quark extensions reflects sound methodology: patterns more complex than Koide, with less precision than Koide, operating in a sector with less discriminatory power than leptons, fall below any reasonable threshold.

### 2.3 The Validation Bottleneck

If genuine mass relationships exist among Standard Model observables, improving precision will reveal them. Relationships that "almost work" at 2% uncertainty will either crystallize into stable agreement or dissolve into statistical noise as measurements improve.

This does not make ML-discovered patterns worthless; it means they require validation infrastructure current practice has not developed.

Current peer review processes are designed for modest discovery rates. A typical phenomenology paper might explore $10^4$ to $10^6$ parameter points and report a handful of interesting configurations. If automated systems generate thousands of pattern candidates, each with superficially impressive numerical agreement, reviewers cannot evaluate claims faster than they arrive.

The problem compounds when the same data informs both pattern discovery and evaluation. This correlated error means evaluators tend to accept precisely the spurious patterns the generator produces. Temporal convergence breaks this correlation by using future, independent experimental determinations as an external selection channel.

### 2.4 The Theory Interface

In high-energy physics terminology, machine learning operates in hep-ph (phenomenology): it generates candidates. These candidates become physics when they interface with hep-th (theory): when someone can explain why the pattern holds.

This creates a bottleneck that databases alone cannot solve. Discovering patterns is now easier than knowing what to do with them. Parameter space search can produce dozens of sub-sigma candidates; introducing $\pi$ as a permitted constant explodes the space further. Without stratification criteria, resource allocation becomes guesswork.

The filtering criteria proposed here address this interface. The goal is to identify which patterns merit pre-registration and longitudinal tracking: candidates worth timestamping and waiting on, letting temporal convergence do the discrimination that neither statistical fit nor theoretical intuition can provide alone.

This addresses a specific historical moment. Lattice QCD has recently achieved percent-level precision on hadron masses; each successive FLAG review reports smaller uncertainties. If descriptive relationships exist in quark physics, they might finally be findable. The ability to distinguish genuine patterns from coincidences at scale is the prerequisite for answering the question.

## 3 A Precedent from Structural Biology

In 2021, DeepMind's AlphaFold2 predicted structures for over 200 million proteins [6]. The Protein Data Bank had accumulated approximately 180,000 experimentally determined structures over fifty years. AlphaFold exceeded this by three orders of magnitude in months.

This created an interface problem: ML systems generating candidates faster than domain experts could evaluate them. The structural biology community continues developing frameworks for evaluating which predictions to trust and how to incorporate them into experimental workflows [7]. These validation standards emerged after the capability arrived.

AlphaFold had clear targets: experimentally determined 3D structures against which predictions could be validated. Particle physics pattern discovery lacks this clarity. A machine learning system might identify a relationship among quark masses that achieves sub-sigma agreement with current measurements. What does this mean? Is it genuine? A statistical artifact? A coincidence destined to dissolve with improved precision?

 DOI 10.69710/ljp.v3i1.17825

High-energy physics has the opportunity to prepare such infrastructure in advance. The framework proposed here is one attempt to do so.

## 4 Framework: Seven Criteria

The specific thresholds proposed are initial estimates subject to community calibration. Criteria 1 through 6 address empirical validation accessible to phenomenologists. Criterion 7 addresses theoretical context. Figure 1 summarizes the evaluation cascade.

### 4.1 Criterion 1: Scale Invariance Under Renormalization Group Evolution

Mass ratios should ideally be scale-invariant under QCD running. Scale-dependent relationships require justification: if a formula invokes a specific energy scale, one must explain why that scale is privileged. Otherwise, scale choice becomes a fine-tuning parameter.

**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. Relationships expressed as ratios automatically satisfy scale invariance to the extent that running is flavor-universal at leading order.

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

### 4.2 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.

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

**Proposed threshold:** $N$ parameters reduced to at most $N - 1$ degrees of freedom.

### 4.3 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.

**Proposed threshold:** $< 1\sigma$ deviation, with the recognition that this threshold alone admits many coincidences at current quark mass precision. Criterion 4 provides the primary filter.

### 4.4 Criterion 4: Temporal Convergence

Patterns must demonstrate directional convergence or stability across data releases. This is the core discriminator, operationalizing Mayo's severity criterion [24, 29]: a claim gains credibility only by passing tests that would probably have revealed flaws if flaws existed. Severity is a property of the *testing procedure*, not merely of the numerical outcome, a distinction developed in Section 6.2. Improving experimental precision constitutes a severe test because it has high probability of exposing spurious agreement: coincidences systematically fail it. When error bars shrink, genuine patterns show residuals shrinking or stable. Coincidences show residuals growing: precision exposes the latent offset that broader uncertainties masked.

**Requirements:**

- Pre-registration via timestamped repository before new data releases

 DOI 10.69710/ljp.v3i1.17825

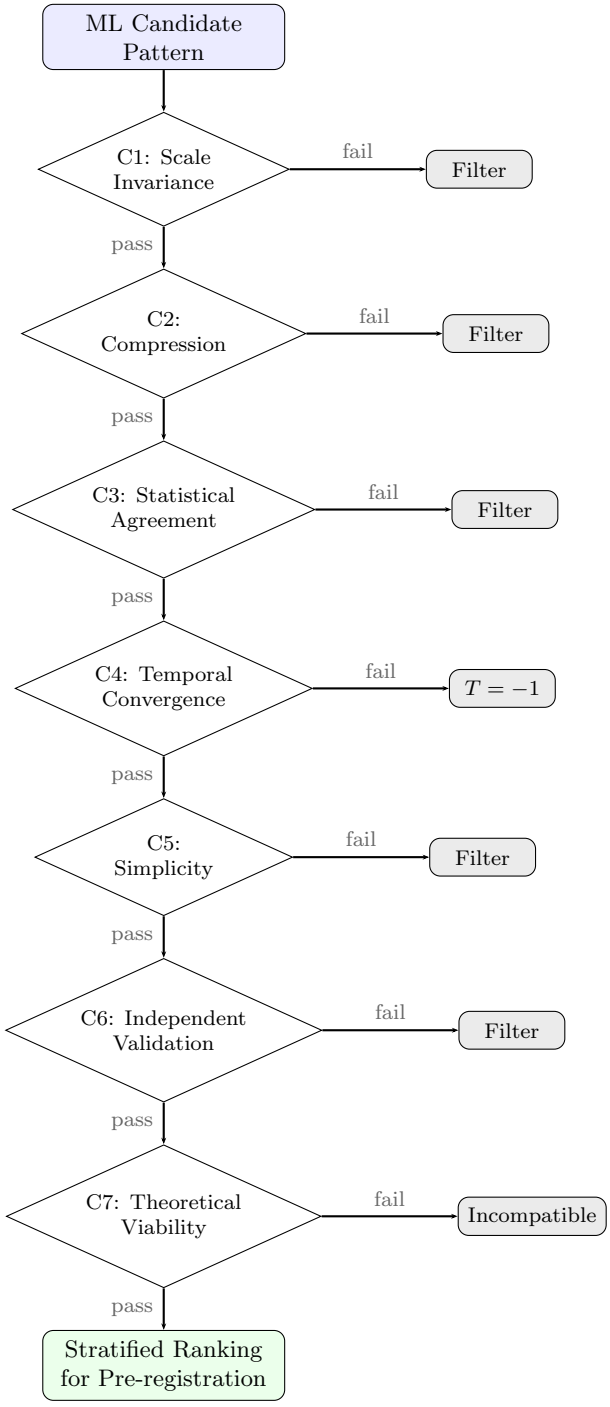


Figure 1: Evaluation cascade for ML-generated pattern candidates. Each criterion acts as a filter; candidates passing all seven receive stratified ranking for pre-registration and longitudinal tracking. Criterion 4 (temporal convergence) is the primary discriminator. The ordering is logical, not strictly sequential: in practice, criteria may be evaluated in parallel where data permits.

- 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, or claim prescience after observing convergence. The only way to pass Criterion 4 is genuine predictive success across independent experimental cycles.

**Empirical context:** The lepton mass relationship has survived four decades of improving measurements. Historical patterns that diverged (Section 5) demonstrate the converse: initial statistical agreement revealed as coincidental when precision improved.

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

### 4.5 Criterion 5: Mathematical Simplicity

Complexity overfits. Simplicity constrains hypothesis space.

**Empirical context:** Unconstrained complexity fits any relationship among any quantities: it explains everything and predicts nothing.

**Proposed threshold:** Basic arithmetic operations, integer exponents $\leq 5$, standard constants ($\pi$, small integers). Formulas requiring evaluation at a specific renormalization scale that is not derivable from the relationship itself incur an additional degree of freedom in the complexity count; a scale-invariant formula is inherently simpler than one requiring a specified $\mu$. This threshold requires sharper operationalization; the principle is sound even if the boundary needs calibration.

### 4.6 Criterion 6: Independent Validation

Patterns must show consistency across independent determinations. A pattern appearing in one collaboration's results but not others may reflect systematic artifacts.

**Empirical context:** FLAG 2024 [8] averages draw on results from BMW, MILC, HPQCD, ETM, RBC/UKQCD, and other collaborations employing different lattice discretizations, fermion actions, and analysis methods.

**Proposed threshold:** Agreement across $\geq 3$ independent collaborations or methods with demonstrably different systematic uncertainties.

### 4.7 Criterion 7: Theoretical Viability

Patterns must not be demonstrably incompatible with established physics. This criterion 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. The lepton mass 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.

## 5 Historical Illustration

To demonstrate how the criteria operate, 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.

A contemporary example offers something stronger. The sterile neutrino anomalies reached experimental resolution after the framework was formalized (Section 5.2.1). Unlike the historical cases above, this constitutes a prospective test: the criteria existed before the ground truth arrived, and were not adjusted after examining the outcome.

### 5.1 Patterns That Converged

**Gell-Mann-Okubo Relation** [13, 14]: Predicted hadron mass relationships before the quark model existed.

| Criterion | Assessment |
|---|---|
| 1. Scale Invariance | PASS |
| 2. Compression | PASS |
| 3. Statistical | PASS |
| 4. Temporal | PASS: validated by subsequent data |
| 5. Simplicity | PASS |
| 6. Independent | PASS |
| 7. Theoretical | Unknown $\rightarrow$ Explained (SU(3)) |

GMO demonstrates that patterns can legitimately persist in Unknown status until theory catches up. The empirical pattern preceded the theoretical explanation by years.

**Lepton Mass Formula** [10]: Passes all empirical criteria, remains in Unknown theoretical status after four decades.

### 5.2 Patterns That Diverged

**Nambu (1952)** [15]: $m_\mu/m_e = 3/(2\alpha)$

Initially $\sim 1\sigma$ agreement. Now $> 20\sigma$ deviation. Temporal convergence: $T = -1$, diverged.

**Lenz (1951)** [16]: $m_p/m_e \approx 6\pi^5$

Initially $< 1\sigma$ agreement. Now $\sim 50\sigma$ deviation. Temporal convergence: $T = -1$, diverged.

These were well-motivated given contemporary precision. They were superseded by improved measurement, not refuted by argument. The temporal convergence criterion correctly identifies both as diverging.

#### 5.2.1 Sterile Neutrinos (Worked Example)

The sterile neutrino anomalies offer a qualitatively different worked example: an experimental anomaly (excess events interpreted as evidence for new physics) that reached definitive resolution in 2025, after the framework criteria were formalized. This temporal sequence is critical: the criteria were not designed to accommodate this outcome, and were not adjusted after reviewing it. The neutrino case therefore serves as a prospective rather than retrospective test of the framework.

An important clarification is necessary. The physics community resolved the sterile neutrino anomaly through its own established practices: independent replication, improved detector tech-

nology, global fits across channels, and sustained skepticism toward claims that could not self-consistently accommodate all available data. The framework proposed here attempts to codify that existing practice into explicit, computationally evaluable criteria. This worked example illustrates that the community's implicit procedures align with the criteria we document, not the reverse. The purpose is not to evaluate past practice but to replicate it outside traditional channels, at the scale that ML-generated candidates will require.

**The anomaly.** In 2001, LSND reported a $3.8\sigma$ excess consistent with $\bar{\nu}_\mu \to \bar{\nu}_e$ oscillations, suggesting a sterile neutrino with $\Delta m^2 \sim 0.2$–$10$ eV$^2$ [32]. In 2018, MiniBooNE reported a $4.5\sigma$ excess in the same channel [33]; combined with LSND, the anomaly reached $6.0\sigma$. Two experiments, different detector technologies, apparently converging on the same signal.

**Application of criteria.** At the point of maximum apparent significance (2018–2020), how would the framework have assessed the sterile neutrino claim?

| Criterion | Assessment (c. 2020) | Notes |
|---|---|---|
| 1. Scale Inv. | Ambiguous | Untested across scales |
| 2. Compression | PASS | 2 params $\to$ 4 anomalies |
| 3. Statistical | PASS | $6.0\sigma$ combined |
| 4. Temporal | Ambiguous | See below |
| 5. Simplicity | PASS | Minimal extension |
| 6. Independent | **FAIL** | ICARUS null [34] |
| 7. Theoretical | **STRESSED** | $> 4\sigma$ tension [35] |

The critical diagnostics were Criteria 6 and 7. While LSND and MiniBooNE appeared to confirm each other, ICARUS found results inconsistent with LSND's preferred parameters [34], and global fits incorporating both appearance and disappearance channels showed persistent internal tension: the mixing parameters required to explain the appearance signal predicted disappearance effects that were not observed [35].

Criterion 4 appeared superficially favorable (two experiments over 17 years), but the two experiments, while at different laboratories (Los Alamos and Fermilab) and different beam energies, shared similar $L/E$ baselines and analogous detector limitations (both relied on Cherenkov-based detection unable to distinguish electrons from single photons). These correlated detection systematics meant that MiniBooNE did not constitute a genuinely independent temporal test in the sense the criterion requires.

**Resolution.** In late 2025, MicroBooNE published the definitive test of the LSND anomaly using liquid argon TPC technology, capable of resolving the electron/photon ambiguity that limited both LSND and MiniBooNE, and simultaneously analyzed two neutrino beams to break appearance–disappearance degeneracies [1]. Result: $\Delta\chi^2 = 0.228$ relative to the three-neutrino model. No evidence for sterile neutrinos in the $\Delta m^2 \sim 1$ eV$^2$ parameter space relevant to LSND.[1]

**What the framework captures.** The sterile neutrino hypothesis *passed* criteria requiring only a plausible model and a significant signal (Criteria 2, 3, 5). It *failed* criteria demanding

[1] We note that KATRIN [2] independently constrains sterile neutrino mixing through tritium $\beta$-decay kinematics in a different region of parameter space ($\Delta m^2 \gtrsim 10$ eV$^2$). While part of the broader sterile neutrino program, KATRIN addresses a distinct anomaly from LSND.

consistency across genuinely independent tests (Criterion 6) and self-consistent theoretical interpretation (Criterion 7). The appearance–disappearance tension visible in global fits by 2020 was the framework's equivalent of a diverging residual.

The sharpest lesson concerns Criterion 4. The LSND→MiniBooNE sequence *appeared* to show convergence, but the two experiments shared analogous detector technology (both relied on Cherenkov detection unable to resolve the electron/photon ambiguity), meaning that MiniBooNE did not constitute a genuinely independent test despite its statistical significance. Genuine temporal convergence requires that each successive test be capable of falsifying the claim. MicroBooNE's liquid argon TPC technology provided the genuinely independent test; the claim did not survive it. Temporal convergence: $T = -1$.

**Methodological note.** Applying the criteria to the state of evidence circa 2020 is retrospective. However, the experimental resolution arrived after the framework was formalized, making the outcome a prospective test: the criteria that would have flagged Criteria 6 and 7 as stressed were not constructed to produce this result. We make no claim that the framework would have prevented the anomaly from receiving attention, nor should it have. The sterile neutrino case serves a different purpose: it tests whether the framework accurately describes how the HEP community already evaluates anomalies. The criteria that flagged the LSND signal as evidentially stressed (Criteria 6 and 7) correspond to concerns that experimentalists and theorists were already raising informally by 2020. The framework did not discover these concerns; it codified them. That the codification produces the same assessment as expert judgment, prospectively confirmed by MicroBooNE's resolution, supports the premise that implicit evaluation standards can be made explicit without distortion. The goal is then to apply this codified description to contexts where ML-generated candidates arrive faster than expert evaluation can absorb them, not to prescribe how domain experts should evaluate evidence, but to document their practice in a form that scales.

### 5.3 The Value of Historical Analysis

In this limited sample: all patterns that persisted show $T \geq 0$; all 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?

We explicitly decline to "validate" this framework against the historical mass-formula cases. Such testing would be post-hoc, cross-regime (leptons offer 35,000× greater discriminatory power), and survivorship-biased. Historical cases motivate the problem; they cannot validate whether explicit criteria improve upon implicit evaluation. The sterile neutrino case is the one partial exception: the experimental resolution arrived after the framework was formalized, providing a prospective rather than retrospective test.

## 6 Discussion

### 6.1 Theoretical Foundations

The temporal convergence criterion implements external selection [25] to break correlated error between pattern generation and evaluation. When the same data informs both discovery and validation, coincidental agreements persist; independent future data provides the separation needed for meaningful tests.

The underlying principle—that genuine results survive repeated, independent testing while

spurious ones do not—is not new. It is arguably the central insight of the philosophy of science. Popper's falsificationism holds that scientific claims gain corroboration precisely by surviving tests that could have refuted them [22]. Lakatos sharpened this into the distinction between progressive research programmes, where "theory leads to the discovery of hitherto unknown novel facts," and degenerating ones that merely accommodate existing data [21]. Mayo's severe testing framework provides the modern statistical formalization: claims gain credibility only by "passing tests that probably would have found flaws, were they present" [24]. The concept that convergence across independent experiments constitutes strong evidence is, in this sense, standard scientific methodology.

What temporal convergence adds is not the principle but its codification as a computationally evaluable criterion for a specific problem: filtering ML-generated pattern candidates at scale. When candidate volume is bounded by human intuition, scientists apply these principles implicitly through peer review, replication, and professional judgment. When ML generation produces candidates faster than human evaluation can absorb, the implicit standard must become explicit and automatable. The $T \in \{-1, 0, +1\}$ scoring, the pre-registration protocol tied to specific data releases (e.g., FLAG reviews), and the integration with the other six criteria constitute the operational contribution—not the philosophical insight that surviving independent tests matters.

A related statistical approach is Efron's false discovery rate (FDR) framework [23], which controls the expected proportion of false positives among rejected hypotheses when testing many candidates simultaneously. FDR addresses multiple testing within a single dataset; temporal convergence addresses a complementary problem: longitudinal evaluation across independent datasets released over time. A pattern may survive FDR correction at a single time point yet diverge under subsequent precision improvements, or conversely, a marginal candidate at one time point may show progressive convergence as data accumulate. The two approaches are complementary, not competing.

The pre-registration movement in psychology [26, 27] provides institutional precedent for such infrastructure.

### 6.2 Error-Statistical Foundations and the Severity Principle

The framework's reliance on severity warrants brief philosophical grounding. Mayo and Spanos [28, 29] developed the error-statistical approach as a meta-methodological principle: a hypothesis $H$ passes a severe test $T$ with data $x$ only if (i) $x$ agrees with $H$, and (ii) the test procedure had a high probability of producing a result that *does not agree* with $H$ if $H$ is false. Severity is a property of the testing *procedure*, not of the numerical output alone [31, 30]. For the purposes of this framework, we require only the claim that improving experimental precision constitutes a severe test of numerical patterns, because coincidences have high probability of being exposed as measurement uncertainty shrinks.

Criterion 4 (Temporal Convergence) operationalizes this principle. Pre-registration before independent data releases ensures that the testing procedure has genuine error-probing capacity: a spurious pattern cannot adjust to new data, and improving precision will expose the latent offset that broader uncertainties masked. The sterile neutrino case (Section 5.2.1) illustrates the failure mode: MiniBooNE's apparent confirmation of LSND was not a severe test in the procedural sense, because analogous detector technology meant the procedure lacked the capacity to discriminate the claimed signal from backgrounds. MicroBooNE's liquid argon technology restored severity by providing genuinely independent error-probing capacity.

### 6.3 Practical Implementation

Existing infrastructure can support pattern documentation without requiring new systems. Zenodo, operated by CERN, provides timestamped DOI assignment for predictions. A researcher identifying a candidate pattern can deposit the formulation before the next FLAG release; the DOI provides immutable pre-registration. Subsequent evaluation against new data constitutes the temporal convergence test.

The lattice QCD community's FLAG collaboration demonstrates how distributed expertise can produce authoritative consensus; similar structures might emerge for pattern evaluation if the need becomes sufficient.

The goal is augmentation, not replacement. Peer review remains essential for theoretical evaluation, methodological scrutiny, and contextual judgment. The criteria proposed here add upstream structure: a pre-filter that reduces the volume reaching human reviewers while increasing the proportion warranting serious attention.

### 6.4 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.

**Domain specificity:** The criteria proposed here are calibrated to particle physics phenomenology, specifically mass relationships testable against lattice QCD and precision measurement data. Adaptation to other domains (collider event classification, dark matter search strategies, cosmological parameter estimation) would require separate development. We do not claim generality beyond the specific use case addressed.

These are not flaws but acknowledgments. We propose a provisional framework, not a finished system. We invite the community to treat this framework as we propose treating empirical patterns: provisionally, with temporal tracking.

## 7 Conclusion

Machine learning will generate particle physics pattern candidates faster than traditional evaluation can absorb them. The structural biology precedent suggests that validation infrastructure is better developed before this capability fully arrives than after.

This framework proposes explicit, computationally evaluable criteria for a specific decision: which patterns merit pre-registration and longitudinal tracking. The output is not "ready for peer review" but "worth timestamping and waiting," letting temporal convergence distinguish genuine structure from coincidence across future data releases.

The goal is augmentation. Peer review remains essential; these criteria add upstream structure that provides a basis for prioritization. The lattice QCD community's decades of precision work are not threatened by pattern-discovery approaches; they are essential to them. Without reliable ground truth, patterns cannot be validated.

This is a starting point, not a solution. If the community finds specific criteria too strict or certain thresholds irrelevant, that discussion is progress. We offer these criteria as a basis for community calibration.

 DOI 10.69710/ljp.v3i1.17825

## 8 Acknowledgements

The author thanks Robert D. Cousins for valuable feedback on an earlier draft, and 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.

## 9 Declarations

**Funding:** This research was conducted independently without institutional funding or external support.
**Competing interests:** The author declares no competing interests.
**Data availability:** All numeric values used in this analysis are derived from publicly available FLAG 2024 [8] and PDG 2024 [9] reviews.
**Code availability:** Analysis methodologies are documented at https://github.com/AndBrilliant/TemporalConvergence.
**Author contributions:** A.M.B. conceived the framework, performed the analysis, and wrote the manuscript.

## 10 Copyright Notice



## References

[1] MicroBooNE Collaboration, "Search for light sterile neutrinos with two neutrino beams at MicroBooNE," Nature (2025), DOI: 10.1038/s41586-025-09757-7.

[2] KATRIN Collaboration, "Sterile-neutrino search based on 259 days of KATRIN data," Nature (2025), DOI: 10.1038/s41586-025-09739-9.

[3] M. Feickert and B. Nachman, "A Living Review of Machine Learning for Particle Physics," arXiv:2102.02770 [hep-ph] (2021, continuously updated).

[4] G. Karagiorgi *et al.*, "Machine learning in the search for new fundamental physics," Nat. Rev. Phys. **4**, 399–412 (2022).

[5] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics," Eur. Phys. J. C **70**, 525–530 (2010).

[6] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583–589 (2021).

[7] M. Varadi *et al.*, "AlphaFold Protein Structure Database in 2024," Nucleic Acids Res. **52**, D431–D438 (2024).

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

[9] S. Navas *et al.* (Particle Data Group), Phys. Rev. D **110**, 030001 (2024).

[10] Y. Koide, "A New View of Quark and Lepton Mass Hierarchy," Lett. Nuovo Cim. **34**, 201 (1982).

[11] Z.-Z. Xing and H. Zhang, "On the Koide-like relations for the running masses of charged leptons, neutrinos and quarks," Phys. Lett. B **635**, 107–111 (2006).

[12] W. Rodejohann and H. Zhang, "Extension of an empirical charged lepton mass relation to the neutrino sector," Phys. Lett. B **698**, 152–156 (2011).

[13] M. Gell-Mann, *The Eightfold Way: A Theory of Strong Interaction Symmetry*, Caltech Report CTSL-20 (1961).

[14] S. Okubo, "Note on Unitary Symmetry in Strong Interactions," Prog. Theor. Phys. **27**, 949 (1962).

[15] Y. Nambu, "An Empirical Mass Spectrum of Elementary Particles," Prog. Theor. Phys. **7**, 595 (1952).

[16] F. Lenz, "The Ratio of Proton and Electron Masses," Phys. Rev. **82**, 554 (1951).

[17] A. O. Barut, "Lepton Mass Formula," Phys. Rev. Lett. **42**, 1251 (1979).

[18] A. Rivero and A. Gsponer, "The strange formula of Dr. Koide," arXiv:hep-ph/0505220 (2005).

[19] C. A. Brannen, "The Lepton Masses," preprint (2006).

[20] J. Kocik, "A note on Descartes and Koide," arXiv:1210.7325 [gen-ph] (2012).

[21] 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).

[22] K. R. Popper, *The Logic of Scientific Discovery*, Hutchinson & Co., London (1959).

[23] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press (2012).

[24] D. G. Mayo, *Statistical Inference as Severe Testing*, Cambridge University Press (2018).

[25] A. M. Brilliant, "Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors," TechRxiv preprint (2026). doi:10.36227/techrxiv.176834656.66652387

[26] Open Science Collaboration, "Estimating the reproducibility of psychological science," Science **349**, aac4716 (2015).

[27] B. A. Nosek *et al.*, "The preregistration revolution," PNAS **115**, 2600–2606 (2018).

[28] D. G. Mayo and A. Spanos, "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction," Br. J. Philos. Sci. **57**(2), 323–357 (2006).

[29] D. G. Mayo and A. Spanos, "Error Statistics," in *Philosophy of Statistics*, Handbook of Philosophy of Science, Vol. 7, Elsevier, pp. 151–196 (2011).

[30] S. C. Fletcher, "Of War or Peace? Essay Review of Statistical Inference as Severe Testing," Philos. Sci. **87**(4), 755–762 (2020).

[31] R. D. Cousins, "Connections between statistical practice in elementary particle physics and the severity concept as discussed in Mayo's Statistical Inference as Severe Testing," arXiv:2002.09713 [physics.data-an] (2020).

[32] A. Aguilar *et al.* (LSND Collaboration), "Evidence for Neutrino Oscillations from the Observation of $\bar{\nu}_e$ Appearance in a $\bar{\nu}_\mu$ Beam," Phys. Rev. D **64**, 112007 (2001).

[33] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), "Significant Excess of Electron-Like Events in the MiniBooNE Short-Baseline Neutrino Experiment," Phys. Rev. Lett. **121**, 221801 (2018).

[34] M. Antonello *et al.* (ICARUS Collaboration), "Experimental search for the LSND anomaly with the ICARUS detector in the CNGS neutrino beam," Eur. Phys. J. C **73**, 2345 (2013).

[35] A. Diaz, C. A. Argüelles, G. H. Collin, J. M. Conrad, and M. H. Shaevitz, "Where Are We With Light Sterile Neutrinos?" Phys. Rep. **884**, 1 (2020).