TRAVIDENCE

D1 — Technical Methodology Paper

Audit-Grade Pre-Validation of Trading Strategies: Five Contributions to the López de Prado Stack

Ignacio Arias · Travidence, LLC

ORCID 0009-0001-7503-3830

Version 1.0 — August 2026

DOI: 10.2139/ssrn.7308022

Abstract

Pre-validation ahead of the prop firm challenge is a structural blind spot of contemporary retail trading. FPFX Tech data on 300,000+ accounts document that 86% of challengers fail to clear the evaluation and only 7% ever receive a payout — systematic evidence of overfitting that existing tools do not resolve at audit grade. Implementation libraries implementing the López de Prado stack (mlfinlab, RiskLabAI) operate as technical infrastructure without issuing certification; the aggregation-only retail consumer offering delivers an aggregate score without gates or procedural signing. Prop trading firms, simultaneously, sit on a structural conflict of interest that prevents them from offering honest pre-validation. This institutional vacuum motivates an independent third-party methodology with audit-grade discipline.

This paper presents five novel contributions to the López de Prado stack: (i) dual n_trials reporting in the Deflated Sharpe Ratio to taxonomize selection bias; (ii) Capital-normalized Calmar that strips margin-leverage inflation; (iii) sub-period schema with monotonic decay detection; (iv) component isolation test for multi-component strategies; (v) vol-on / vol-off MUST-PASS regime taxonomy. The five integrate into the Robustness Score — a hybrid continuous-gate score on [0, 100], inspired by the procedural discipline of the professional auditing industry.

The methodology is applied to three textbook strategies on public data: Golden Cross on CL crude oil futures (2000–2025), RSI-14 on BTC perpetual (2020–2025), and Asian Range Breakout on EUR/USD (2015–2025). None reaches a Robust band — consistent with the documented 86% failure rate. Under the production audit engine (trade-level returns, self-derived p50 regime cut, admissibility floors), CL and BTC close with a Disclaimer of Opinion for insufficient sample (18 and 31 trades), while EUR/USD emits the framework's public end-to-end score: 56 / 100, Limited Robustness, with no gate binding. The gate-vs-continuous behavior against pure aggregation that defines the defensible methodological differentiator of audit-grade pre-validation is exhibited by the repository's NQ Opening Range Breakout case — admitted with 2,424 trades and capped by the regime and subperiod gates; across the five repository case studies the portfolio covers the three public outcomes of the framework: score emitted, gated, and insufficient sample.

Keywords: backtest overfitting; Deflated Sharpe Ratio; Combinatorial Purged Cross-Validation; Probability of Backtest Overfitting; prop firm pre-validation; Capital-normalized Calmar; regime asymmetry; component isolation; audit-grade methodology; Robustness Score.

JEL: G11, G14, G17, C12, C52. **MSC:** 62F03, 62P05, 91G70.

---

§1. Introduction

§1.1 The 86% problem

The prop firm challenger industry operates on an aggregate outcome profile that the academic literature has not fully absorbed into the body of theory on strategy overfitting. Consolidated data on more than 300,000 accounts administered by FPFX Tech, reported by Finance Magnates (2024), document that 86% of challenge participants fail the initial evaluation and only 7% ever receive a

payout at the end of the funding cycle. The gap between the two figures corresponds to participants who pass the challenge but never reach a withdrawal.

The pattern is inconsistent with the hypothesis of benign adverse selection, and consistent instead with selection by overfitting: a majority of challengers enter the desk with strategies whose in-sample metrics are products of the backtest overfitting documented by Bailey, Borwein, López de Prado, and Zhu (2017). The client pays the entry fee, fails, and the firm absorbs the revenue without a payout counterpart.

The problem is not the existence of the challenge nor the arithmetic asymmetry between fees and payouts. The problem is the absence of an independent mechanism that lets the challenger evaluate, before paying the fee, whether the strategy has genuine edge or whether its in-sample metrics are statistical artifacts. The tools available today to the retail challenger are insufficient on dimensions the next section details.

§1.2 The validation gap

Three classes of actor partially occupy the quantitative validation space, but none provides an independent service at audit grade.

Implementation libraries — chiefly `mlfinlab` (López de Prado et al.) and `RiskLabAI` — implement in code the methodology of the Deflated Sharpe Ratio, Combinatorial Purged Cross-Validation (CPCV), and the Probability of Backtest Overfitting (PBO). They are high-quality technical infrastructure but by construction do not issue certification: they produce metrics the analyst then interprets and reports. There is no third party that signs the verdict.

The retail consumer offering delivers low-cost algorithmic analysis, emitting an aggregate score on the [0, 100] scale without hard gates or procedural signing. Pure aggregation is that offering's central methodological choice and at the same time its structural weakness for prop firm challenge pre-validation: a strategy with severe regime asymmetry can accumulate a numerically acceptable score because the favorable dimensions compensate for the unfavorable ones, with no mechanism halting the certification. There is no equivalent of a cliff-edge audit failure.

Prop trading firms themselves are structurally barred from offering honest pre-validation. Their business model depends on the flow of evaluation fees; effective pre-validation would shrink the universe of challengers willing to attempt the desk — and with it the firm's revenue. The conflict of interest is indistinguishable to the client and not resolvable through disclosure: a firm that sells challenge slots cannot simultaneously warn the client that the strategy is not viable.

The resulting vacuum is precise. There is a strip of service between the implementation libraries (which require analyst technical competence and emit no verdict) and the consumer offering (which emits a verdict but without audit-grade discipline). That strip, in other regulated markets occupied by independent audit firms, has no equivalent incumbent in the prop firm pre-validation market.

§1.3 Contribution of this paper

TRAVIDENCE deliberately positions in the strip described. The methodological proposal of this paper rests on three integrated components.

First, five novel contributions to the López de Prado stack, each targeting a specific retail-pre-

validation failure mode: dual `n_trials` reporting in the Deflated Sharpe Ratio to taxonomize selection bias; Capital-normalized Calmar that strips margin-leverage inflation; a sub-period schema with monotonic decay detection; a component isolation test for multi-component strategies; and a vol-on / vol-off MUST-PASS regime taxonomy. Each is detailed in §4.

Second, a hybrid continuous-gate score — the Robustness Score — that integrates the five contributions into a [0, 100] metric, inspired by the procedural discipline of the professional auditing industry. The score's public band (Robust, Conditional Robustness, Limited Robustness, Not Robust, Disclaimer of Opinion) translates the result into an actionable classification for the retail challenger.

Third, three empirical case studies on textbook strategies and public data — Golden Cross on CL crude oil, RSI-14 mean reversion on BTC perpetual, and Asian Range Breakout on EUR/USD — illustrating the methodology in action. The gate-vs-continuous behavior against pure aggregation that defines TRAVIDENCE's defensible methodological moat is exhibited by the repository's NQ Opening Range Breakout case.

§1.4 Roadmap

§2 reviews related work: the canonical López de Prado stack, reference implementations, and the retail consumer offering. §3 presents the TRAVIDENCE methodological summary. §4 details the five novel contributions with their key equations. §5 describes the empirical study design. §6 reports the three case studies in increasing order of centrality to the thesis (CL → BTC → EUR/USD; the last emits the public end-to-end score). §7 contrasts TRAVIDENCE with the consumer offering and the implementation libraries (mlfinlab, RiskLabAI). §8 discusses implications. §9 enumerates limitations. §10 concludes. Formal references and three appendices (mathematical derivations, audit certificates as figures, reproducibility hash) close the document.

---

§2. Related work

§2.1 The López de Prado stack

The body of work by López de Prado and collaborators forms the academic reference framework for honest validation of quantitative strategies. Three components are particularly relevant to the retail pre-validation problem.

The Deflated Sharpe Ratio (DSR) of Bailey and López de Prado (2014) extends the Probabilistic Sharpe Ratio (Bailey and López de Prado, 2012) to correct simultaneously for (a) non-normality of returns via higher-order moments (skewness, kurtosis), and (b) the selection bias induced by multiple evaluation on the same dataset. The DSR formula estimates the probability that the true Sharpe exceeds a reference threshold conditional on the observed Sharpe, the number of hypotheses explored, and the distributional profile of returns. The original paper provides the exact form of the expected maximum Sharpe under the null for `N` independent trials.

Combinatorial Purged Cross-Validation (CPCV), detailed in López de Prado (2018, chapter 12),

addresses the dual problem of temporal leakage and single-partition bias. CPCV generates multiple paths by combining train/test splits with purging and embargo, producing a distribution of Sharpes instead of a single value. The resulting distribution is robust to the common practice of "I picked the hyperparameters that delivered the best Sharpe on the last window" — a specific form of overfitting that plain walk-forward does not detect. Empirical superiority of CPCV over walk-forward and purged k-fold in synthetic controlled environments, measured via DSR and PBO, has been documented by Arian, Norouzi Mobarekeh, and Seco (2024).

The Probability of Backtest Overfitting (PBO), introduced in Bailey, Borwein, López de Prado, and Zhu (2017), provides an additional metric: the probability that the in-sample best strategy will deliver out-of-sample performance below the median. PBO complements DSR and captures a distinct failure mode — not Sharpe inflation but the fragility of ranking among candidates.

These three components, complemented by the rigorous treatment of feature engineering and label leakage in López de Prado (2018), form the "stack" on which TRAVIDENCE builds its five novel contributions detailed in §4.

A scoping remark: Monte Carlo resampling of a trade sequence — common practice in retail validation — quantifies sequence risk conditional on the edge being genuine. Reordering or resampling the same trades cannot detect whether they came from a selection process over N variants: every simulated path fully inherits the selection bias of the input history. Detecting selection-driven overfitting requires DSR/PBO-class corrections applied to the search process, not simulations over its output.

§2.2 Reference implementations: mlfinlab and RiskLabAI

The `mlfinlab` library (Hudson & Thames) implements CPCV, DSR, PBO, and the meta-labeling methods of López de Prado (2018) on a standard Python stack (NumPy, SciPy, pandas). `RiskLabAI` provides independent implementations of a similar subset with additional coverage of extreme value theory and tail metrics. mlfinlab is today a proprietary library of Hudson & Thames (source visible under a commercial license since 2021); RiskLabAI is open-source. Both libraries are high-quality technical tools: the canonical TRAVIDENCE implementation of the DSR was cross-validated against `RiskLabAI` v1.0.8 with bit-identical agreement (maximum absolute difference 0.0e+00 across twelve test cases); the margin-frame Calmar implementation was cross-validated against `empyrical-reloaded` v0.5.12 with bit-identical agreement to machine precision.

The operational function of these libraries, however, differs from that of an independent validation service at audit grade. A library delivers metrics; the analyst interprets and reports. No third party signs the verdict, no pre-specified threshold triggers a binary PASS/FAIL decision for commercial use, no versioning system of the verdict permits retrospective audit of the basis on which it was issued. The gap is at the service layer, not the technical layer.

§2.3 Retail consumer offering: pure aggregation

The most visible operator in the retail consumer strip accepts client backtests (CSV / equity curve formats), applies a panel of dozens of institutionally inspired checks, and delivers an aggregate score on the [0, 100] scale. Micro-transactional pricing supports scope economy at high volume and captures the pure retail market.

The structural weakness of the service for prop firm challenge pre-validation is pure aggregation. The score is composed as a weighted combination of the evaluated dimensions, without hard gates that halt certification when one of the dimensions exhibits a cliff-edge failure. A strategy with severe regime asymmetry — failing every minimum criterion under vol-on conditions, say — can accumulate a numerically acceptable score because the aggregate average compensates with favorable dimensions. The system does not distinguish between continuous weakness and binary structural failure.

This weakness is functional to the consumer business model but inadmissible for a verdict offered as a basis for decisions about deploying at-risk capital (evaluation fees, configuration of funded accounts). Professional audit methodology — inherited from AICPA and IAASB classifications — requires precisely the opposite mechanism: the verdict incorporates mechanisms that modify the opinion regardless of the rest of the evaluation — a pervasive material misstatement, for instance, produces an adverse opinion.

§2.4 Gap statement

To recap: the López de Prado stack provides the technical apparatus; implementation libraries implement it; the retail consumer offering aggregates and commercializes, but without hard gates; prop trading firms are structurally barred; traditional audit firms (Big Four) have not entered the retail quantitative strategy validation market. The specific gap TRAVIDENCE occupies is the combination of (i) technical implementation of the stack at the level of implementation libraries, (ii) a service layer with a verdict signed by a third party, (iii) a hard-gate system calibrated to defend the institutional cliff-edge, and (iv) a public result band that speaks simultaneously to the retail audience and to professional stakeholders (reviewers, B2B counterparties, future audit scrutiny).

---

§3. TRAVIDENCE methodological summary

§3.1 The five-validator architecture

The TRAVIDENCE methodology rests on five validators implemented as independent modules of the validation engine, each targeting a specific failure mode documented in the literature and enriched with the novel contributions detailed in §4.

The `deflated_sharpe_ratio` validator attacks selection bias from multiple evaluation. Given `observed_sharpe`, the return series, and the dual count (`n_trials_strict, n_trials_session`), it returns the probability that the true Sharpe exceeds the expected maximum under the null for `N` trials. The implementation follows the exact form derived in Bailey and López de Prado (2014).

The `capital_normalized_calmar` validator attacks leverage inflation in the historical Calmar. Given the return series, a declared `capital_baseline`, and an annualization factor, it returns the pair (`calmar_margin_frame, calmar_capital_frame`) and the `inflation_ratio` that diagnoses the divergence. The capital frame is the canonical metric for cross-asset comparison; the margin frame is preserved for historical compatibility.

The `subperiod_monotonic_decay` validator attacks temporal instability of the edge. Given the return series, a number of chunks `K` (default five), a configurable metric, and an endpoint-to-endpoint loss threshold, it returns the canonical boolean `passes_no_decay` under AND semantics: the strategy passes only if the metric sequence is not monotonically non-increasing and the controlled endpoint loss is below the threshold.

The `component_isolation` validator attacks portfolio-level overfitting in multi-component strategies. Given aggregate returns, per-component returns, and an interaction tolerance, it decomposes the aggregate Sharpe into the sum of standalone Sharpes plus an interaction residual, then applies leave-one-out attribution to identify the dominant component. The recommended tolerance scales with `|1 − √K|` because of the structural non-additivity of the Sharpe ratio (see §4.4).

The `regime_must_pass` validator attacks regime asymmetry. Given the return series, regime labels (by default vol-on / vol-off according to a rolling-volatility threshold), minimum criteria per regime, and a minimum observation count per regime, it returns the boolean `passes_all_regimes` under binary AND aggregation across regimes. Observational asymmetry between regimes is expected and not anomalous; what the validator penalizes is performance asymmetry, not coverage asymmetry.

The five contributions operationalize the principle of demanding evidence of robustness in each dimension, in line with the López de Prado stack.

§3.2 The Robustness Score — structure of the hybrid continuous-gate

The Robustness Score integrates the outputs of the five validators into a [0, 100] metric, translated into a public result band (see §3.3). The structure is deliberately hybrid: it combines a continuous weighted aggregation of three quality dimensions with a hard-gate system that acts as caps when a dimension exhibits cliff-edge failure.

The three dimensions of the continuous aggregation are:

- **Edge Quality** (**EQ**) — derived from the Deflated Sharpe Ratio in its conservative `dsr_strict` variant via a piecewise-linear mapping function with knots calibrated per asset-class bucket. EQ captures the statistical strength of the edge corrected for selection bias.
- **Risk Profile** (**RP**) — derived from the capital-frame Calmar, stripping the margin-frame inflation that differentiates it from the historical tear-sheet convention. RP captures the quality of the return/drawdown profile in a form comparable across asset classes.
- **Sample & Coverage** (**SC**) — a composite of three components: trade-count adequacy, calendar coverage, and minimum observation count per regime. SC captures the sample credibility of the two preceding dimensions.

Aggregation produces a continuous `base_score`. On top of that base score act three hard gates:

- **Regime gate** — triggered when `regime_must_pass` reports `passes_all_regimes = False`. The most conservative cap (lowest of the three). Rationale: regime asymmetry is the failure mode with the highest predictive power over failure in live accounts, given that the market will switch regimes and the strategy, by construction, will meet its weak side.

- **Sub-period decay gate** — triggered when `subperiod_monotonic_decay` reports `passes_no_decay = False`. Intermediate cap. Rationale: the edge is dying *now*; by the time decay surfaces statistically in the chunks, the strategy is well into its extinction trajectory.
- **Component isolation gate** — conditional on `n_components ≥ 2`, triggered when `component_isolation` reports `passes_isolation = False`. The highest cap of the three. Rationale: the failure is structural but typically refinable; the aggregate strategy may recover through redesign of the combination logic.

When two or more gates fire simultaneously, the most conservative (lowest) cap prevails. The final score is computed as `min(round(base_score), effective_cap)`. The asymmetry between the probabilistic character of the base score (subject to a confidence interval) and the deterministic character of the cap (the gate fired or it did not) is preserved in the report: the 95% CI is computed on the base score; the gate event is reported as a binary fact.

The dimensions, the gate names, the conceptual definitions, and the hybrid continuous-gate logic are the public surface of the methodology. Internal scoring components — dimension weights, gating caps, mapping curve knot positions, and calibration parameters — are held as trade secret. Consumer-facing tier nomenclature and cutoffs (published in Executive Brief D2 §4) are public labels enabling clients to interpret their score; the internal model assigning those labels is proprietary and kept outside this document (see §3 of the TRAVIDENCE internal specification (unpublished)).

§3.3 The public result band

The numerical score translates into a public band: Robust (80–100), Conditional Robustness (65–79), Limited Robustness (50–64), Not Robust (0–49), and Disclaimer of Opinion (N/A, when the sample is insufficient for a verdict). The band is the actionable classification in commercial language that the retail client receives, and at the same time it is the classification traceable to the institutional criterion that the professional stakeholder — reviewer, B2B counterparty, future media scrutiny — needs.

The procedural discipline of the professional auditing body — codified in AICPA (2011) Statement on Auditing Standards No. 122 §AU-C 705 (Modifications to the Opinion in the Independent Auditor's Report) and IAASB (2015) ISA 705/706 (Modifications to the Opinion in the Independent Auditor's Report; Emphasis of Matter Paragraphs) — is the design inspiration for the band system: a signed verdict, with pre-specified criteria, capable of capping the result below what pure aggregation would assign. The TRAVIDENCE verdict does not claim functional equivalence to a financial audit (it is not an opinion on financial statements under GAAP/IFRS); it does deliberately adopt that procedural discipline as a base of legal and reputational defensibility.

---

§4. The five novel contributions

The following five subsections describe TRAVIDENCE's novel contributions to the López de Prado stack. Each subsection follows the same structure: (i) an explicit statement of the novelty relative

to the canonical baseline, (ii) the key equation and operational form, (iii) the practical implication for the Robustness Score verdict. Full derivations are deferred to Appendix A.

§4.1 Dual n_trials reporting in the Deflated Sharpe Ratio

The canonical application of the Deflated Sharpe Ratio in Bailey and López de Prado (2014) operates on a single `n_trials` value declared by the analyst. In practice, post-mortems of discovery runs consistently reveal that the reported value reflects only the *current session's* selection space — the cardinality of the grid effectively traversed in the run that produced the declared hypothesis — while the *true institutional count*, which includes all prior explorations on the same dataset that informed the hypothesis formulation, is typically an order of magnitude larger.

The TRAVIDENCE contribution is to operationalize dual reporting as an implementation invariant. The `deflated_sharpe_ratio` validator requires two parallel inputs: `n_trials_strict` (the cross-session institutional universe of every variant that ever touched data on this dataset) and `n_trials_session` (the subset compared in the final selection of the current run). By construction of the problem, `n_trials_strict ≥ n_trials_session`; the implementation raises `ValueError` if the invariant is violated.

Bailey and López de Prado (2014) derive the exact form of the expected maximum Sharpe under the null:

$$E[\max SR \mid N] \;=\; \sqrt{V[\{SR_n\}]} \cdot \left[(1-\gamma_E)\, \Phi^{-1}\!\left(1 - \frac{1}{N}\right) + \gamma_E\, \Phi^{-1}\!\left(1 - \frac{1}{Ne}\right) \right]$$

where $\gamma_E \approx 0.5772$ is the Euler-Mascheroni constant, Φ^{-1} is the inverse of the standard normal CDF, and $V[\{SR_n\}]$ is the variance of the Sharpe estimator across the `N` trials. The DSR statistic is then:

$$\text{DSR} \;=\; \Phi\!\left(\frac{SR_{obs} - E[\max SR \mid N]}{\widehat{\sigma}_{SR}} \right)$$

with $\widehat{\sigma}_{SR}$ the estimator deviation per the Mertens (2002) correction for higher-order moments, building on Lo (2002):

$$\widehat{\sigma}^2_{SR} \;=\; \frac{1 - \gamma_3 \cdot SR_{obs} + \frac{\gamma_4 - 1}{4} \cdot SR^2_{obs}}{T - 1}$$

where γ_3 is the unbiased sample skewness and γ_4 is the Pearson sample kurtosis (not excess; Normal → 3). The canonical implementation uses the `scipy.stats` convention `bias=False, fisher=False` to align exactly with the Mertens (2002) form.

Dual reporting produces two parallel verdicts of the TRAVIDENCE Test 5:

- `dsr_strict ≥ 0.95` → strict PASS.
- `dsr_session ≥ 0.95` → session PASS.

(The 0.95 threshold corresponds to the standard confidence level of the reference paper, not a calibrated TRAVIDENCE constant.)

By construction, `dsr_strict ≤ dsr_session` always (the E[max SR | N] cap grows with N). The diagnostically interesting failure mode is PASS @ session AND FAIL @ strict: the strategy survives only under the analyst's comfortable count and collapses under the institutional count. The gap — the distance between the honest session declaration and the cross-session reality — is precisely the pattern dual reporting makes auditable. The TRAVIDENCE primary verdict uses `dsr_strict`; `dsr_session` remains as informational sensitivity.

The reference implementation lives in the `deflated_sharpe.py` module of the validation engine, function `deflated_sharpe_ratio`. The hybrid variance-estimator policy for V[{SR_n}] (empirical when the caller supplies the array of explored SRs with length ≥ 2; asymptotic Lo (2002) otherwise) is traced in the result's `variance_source` field for auditability.

§4.2 Capital-normalized Calmar — capital frame vs margin frame

The historical Calmar ratio (Young, 1991), in the retail tear-sheet convention, is computed as annualized return over maximum percent drawdown, where the drawdown is reported as a percentage of the rolling peak of the equity curve. In leveraged contexts — futures with margin requirements, retail FX under ESMA 30:1, crypto perpetual with cross-margin — the rolling peak grows in proportion to the notional the trader controls, not the capital actually at risk. The resulting Calmar is systematically inflated by the leverage factor embedded in margin sizing, and Calmar values reported across asset classes are not comparable.

The TRAVIDENCE contribution is to separate two Calmar frames explicitly and declare the capital frame as canonical for cross-asset comparison. The `capital_normalized_calmar` validator requires `capital_baseline` declared and strictly positive. The TRAVIDENCE convention (v1.3) defines `capital_baseline` as the trader's real capital at risk (margin posted), not the notional:

$$\text{capital\_baseline} \;=\; \frac{\text{nominal\_capital}}{\text{leverage\_factor}}$$

Over that definition, the two Calmar frames are computed as:

$$\text{Calmar}_{\text{margin}} \;=\; \frac{r_{ann}}{|DD_{t^*}/P_{t^*}|}, \qquad \text{Calmar}_{\text{capital}} \;=\; \frac{r_{ann}}{|DD_{t^*}/\text{capital\_baseline}|}$$

where P_{t^*} is the rolling peak at the worst-drawdown event, DD_{t^*} is the dollar drawdown at that event, and r_{ann} is the geometric annualized return computed over the geometric equity curve $E[t] = \text{capital\_baseline} \cdot \prod_{i \le t}(1 + r_i)$.

The ratio between the two frames provides a diagnostic of the inflation magnitude:

$$\text{inflation\_ratio} \;=\; \frac{\text{Calmar}_{\text{margin}}}{\text{Calmar}_{\text{capital}}} \;=\; \frac{P_{t^*}}{\text{capital\_baseline}}$$

derived algebraically by cancellation of the numerator and of the shared DD_{t^*}. `inflation_ratio ≠ leverage_factor` in general: the two magnitudes coincide only if equity grew by

a factor of `leverage_factor` over the period AND the worst drawdown occurred at the end. In practice, with unit position sizing and limited holding periods, observed inflation ratios sit between 1.0× and 5.5× even under nominal 30× leverage. The divergence is useful validator output, not defect: it measures the *real* divergence between frames at the worst event, not the context's nominal leverage.

The institutional threshold for capital-frame Calmar is private and calibrated per asset class; it applies exclusively to `Calmar_capital`. A strategy with high `Calmar_margin` that does not accompany it with a `Calmar_capital` clearing that bar is rejected at the institutional block regardless of the margin-frame value.

The canonical implementation lives in the `capital_normalized_calmar.py` module of the validation engine, function `capital_normalized_calmar`. The margin frame is preserved for compatibility and diagnostic; the capital frame is the TRAVIDENCE-novel metric feeding the Risk Profile dimension of the Robustness Score (see §3.2).

§4.3 Sub-period schema with monotonic decay detection

The edge of a strategy operating on financial data is not timeless. Regimes change, participants mutate, microstructure adjusts; an edge present in the discovery period may have decayed before the backtest result itself materializes. The common retail practice of reporting aggregate metrics over the full history hides that temporal trajectory. A strategy with aggregate `PF = 1.8` may conceal a fall from `PF = 2.5` in the earliest period to `PF = 1.1` in the most recent; the arithmetic mean does not capture that fact.

The TRAVIDENCE contribution is to operationalize the monotonic decay test as the canonical binary invariant, generalizing the historical 3-period schema of the methodology into a K-chunk equal-split form applicable uniformly over any evaluation period. The `subperiod_monotonic_decay` validator splits the return series into `K` contiguous chunks of equal length (default `K=5`), evaluates a configurable metric (Sharpe by default) on each chunk, and emits the canonical boolean `passes_no_decay` under locked AND semantics:

$$\text{passes\_no\_decay} \;=\; \neg\,\text{is\_weakly\_monotonic\_decay} \;\wedge\; (\text{overall\_decay\_pct} < \theta)$$

where θ is the endpoint-to-endpoint loss threshold (default 20%), `is_weakly_monotonic_decay` is true when the sequence $m_1, m_2, \ldots, m_K$ is non-increasing, and `overall_decay_pct` is $(m_1 - m_K)/m_1$. The AND semantics are deliberate: an OR operator would let through strategies strictly decaying but superficially, whereas AND demands simultaneously the absence of non-increasing monotonicity and a controlled endpoint drop.

A relevant technical point is the display convention for the `overall_decay_pct` field. The formula is mathematically exact but produces visually misleading values when m_1 is small, near zero, or negative. In case 02 (BTC RSI mean reversion), for instance, the per-chunk Sharpe sequence is $[0.39, 0.98, -0.12, 1.01, 1.71]$ — the edge strengthens net from the first to the last chunk, with a one-chunk dip in the middle — and the computed `overall_decay_pct` is -332.04%: mathematically correct, visually confusing. The canonical boolean `passes_no_decay = True` provides the correct answer independent of the magnitude of m_1. (Sequence from the current re-audited engine; §6.2 reports the frozen original-verdict sequence — see the note at the top of §6.)

The TRAVIDENCE convention for client-facing reports is: always show the full `metric_per_subperiod` series (directly readable), report the boolean `passes_no_decay` as the verdict, and suppress the raw percentage when $|m_1| < 0.10$ (replacing it with `N/A — m_1 too small for meaningful percentage` on the certificate).

The canonical implementation lives in the `subperiod_decay.py` module of the validation engine, function `subperiod_monotonic_decay`. The historical 3-period schema (pre-discovery, discovery, post-discovery with date anchoring and multi-level flags) is preserved as the recommended instantiation for discovery runs with expanded reporting, fed by the validator's binary verdict.

§4.4 Component isolation — the arithmetic floor `|1 − √K|`

Multi-component strategies — ensembles of two or more independent signals combined through a signal aggregator or portfolio — are particularly vulnerable to a specific failure mode that univariate tests applied to the aggregate do not detect. The operational metaphor is: four individually weak components, none with edge statistically distinguishable from noise, can produce an aggregate with an apparently healthy Sharpe by sheer combination noise. Profit Factor, Calmar, and DSR applied to the aggregate may all look healthy; yet the fragility is structural, the edge is not replicable, it is not robust to minor changes in combination logic, and it does not survive institutional deployment.

The TRAVIDENCE contribution is the SR decomposition plus leave-one-out attribution as a structural primitive, complemented by explicit calibration of the interaction tolerance against the deterministic arithmetic floor of the problem.

The decomposition operates on the inputs `strategy_returns` (aggregate returns) and `component_returns` (mapping name → standalone returns). The canonical procedure is:

$$\begin{aligned} SR_{full} &= \text{annualized\_sharpe}(\text{strategy\_returns}) \\ SR_i^{\text{standalone}} &= \text{annualized\_sharpe}(\text{component\_returns}[i]) \\ \rho_{\text{int}} &= SR_{full} - \sum_i SR_i^{\text{standalone}}, \qquad \rho_{\text{int,pct}} = \frac{\rho_{\text{int}}}{SR_{full}} \end{aligned}$$

Leave-one-out attribution identifies the dominant component:

$$\Delta_i = SR_{full} - \text{SR}(\text{strategy\_returns} - \text{component\_returns}[i]), \qquad \text{dominant} = \arg\max_i |\Delta_i|$$

The canonical verdict is passes_isolation $= (|\rho_{\text{int,pct}}| \leq \text{interaction\_tolerance})$.

The calibration of `interaction_tolerance` is where the subtlety lies. Under additivity (`r_full = Σ_i r_i`) and independence (IID components with mean μ and variance σ^2), a direct calculation gives:

$$SR_{full} = \sqrt{K} \cdot \frac{\mu}{\sigma}, \qquad \sum_i SR_i^{\text{standalone}} = K \cdot \frac{\mu}{\sigma}, \qquad \frac{\rho_{\text{int}}}{SR_{full}} = 1 - \sqrt{K}$$

The naive baseline $\sum SR_i \approx SR_{full}$ thus **fails** with a deterministic floor of magnitude $|1-\sqrt{K}|$ even under perfect additivity and independence. The floor is not genuine interaction; it is the arithmetic signature of the fact that the Sharpe is a mean/std ratio and is, by construction, non-additive.

The recommended calibration of `interaction_tolerance` per K is structured as the arithmetic floor $|1 - \sqrt{K}|$ plus a finite-sample buffer. The buffer absorbs sample slack in the estimated Sharpes plus mild aggregator non-linearity not attributable to structural overfitting; a strategy whose $|\rho_{\text{int,pct}}|$ exceeds the recommended tolerance for its K is consuming the buffer with real interaction and should be reviewed. The numeric per-K tolerances and the buffer magnitude are calibrated and documented internally per the Disclosure Boundary Policy.

K (components)	Arithmetic floor |1 − √K|	Recommended `interaction_tolerance`
2	0.414	(internal — see Disclosure Boundary Policy)
3	0.732	(internal — see Disclosure Boundary Policy)
4	1.000	(internal — see Disclosure Boundary Policy)
5	1.236	(internal — see Disclosure Boundary Policy)

The canonical implementation lives in the `component_isolation.py` module of the validation engine, function `component_isolation_test`. For single-component strategies (`K=1`), the test reports `N/A — single-component strategy` and does not affect the Robustness Score verdict. For `K ≥ 2`, the `component_isolation` gate of the Robustness Score (§3.2) is activated when `passes_isolation = False`.

§4.5 MUST-PASS regimes with vol-on / vol-off taxonomy

The last of the five validators attacks regime asymmetry, a failure mode that aggregate metrics structurally hide. A strategy operating well in low-volatility regimes and catastrophically in high-volatility regimes may accumulate an acceptable aggregate Sharpe if the time fraction in each regime is functional to the backtest's directional bias. Silent averaging across regimes is invisible to the client and to the reviewer without explicit breakdown.

The TRAVIDENCE contribution is to operationalize the MUST-PASS test with binary AND aggregation across regimes as the canonical invariant, keeping the 4+2 vol-on/vol-off taxonomy as the recommended institutional profile. The `regime_must_pass` validator operates on `returns` indexed by datetime, `regime_labels` aligned to it, a minimum criterion per regime and per metric (structure and cutoffs private, unpublished), and `min_observations_per_regime` (the minimum per-regime observation count below which the regime is treated as a conservative FAIL).

The pass rule is strict conjunction:

$$\text{passes\_all\_regimes} \;=\; \bigwedge_{r \in R} \text{passes\_per\_regime}[r] \;\wedge\; \big|\{r : n_r < \text{min\_obs}\}\big| = 0$$

Any failing regime, any regime with insufficient observations, and any criterion returning NaN counts as global FAIL. There is no cross-regime averaging; the validator's contract is AND.

Each per-regime criterion carries an implicit comparison direction — a floor the metric must meet or exceed, or a ceiling it must not exceed — which the implementation preserves in an auditable, unambiguous way. Outputs preserve the user's key verbatim, enabling exact round-trip and disambiguation when the same base metric name appears under two simultaneous directions in the same regime.

The canonical registry supports seven internally computed metrics: `sharpe`, `sortino`, `calmar` (margin frame; for capital frame use the specialized validator), `dd_pct`, `return`, `win_rate`, and `profit_factor`. The recommended minimum composition for the institutional profile is 4 vol-on regimes + 2 vol-off regimes covering the study's pre-discovery + discovery period. The concrete per-asset-class taxonomy is detailed in Appendix A.

A relevant technical point is the expected observational asymmetry between regimes. Most assets spend more time in one regime than in the other, depending on the chosen threshold and the asset's history; the vol-on regime is typically minority. The asymmetry is neither a defect of the test nor an anomaly: it is structural. The validator treats regimes with `n < min_observations_per_regime` as conservative FAIL to avoid an optimistic verdict based on insufficient sample; above that floor, the asymmetry does not compromise the statistical validity of the AND-across-regimes test — the minority regime simply has metrics estimated with greater variance, which is honest information about the strategy's exposure to that regime.

The canonical implementation lives in the `regime_must_pass.py` module of the validation engine, function `regime_must_pass`. The `regime` gate of the Robustness Score (§3.2) is activated when `passes_all_regimes = False` and constitutes, conceptually, the highest-severity gate in the three-gate system: regime asymmetry is the failure mode with the highest predictive power over failure in live accounts, given that the market will switch regimes and the strategy, by construction, will meet its weak side.

---

§5. Empirical study design

§5.1 Why three textbook strategies on public data

The objective of the three case studies reported in §6 is not to discover alpha. It is to validate the TRAVIDENCE methodology as a filter. Three textbook strategies, widely documented in retail trading literature (Chan, 2013; Clenow, 2013) and derived from ideas with decades of visibility, are adversarial test cases of the filter: if the TRAVIDENCE methodology were lax, one of the three would reach a Robust band and the audit-grade claim would be refuted by counterexample. A scoping remark on methodological honesty: the v1.0 calibration (§9.3) was built in part from a synthetic population derived from these same three cases, so its value here is illustrative of the mechanism, not an independent test of strictness. That independent test, outside the calibration population, is provided by the repository's NQ Opening Range Breakout case (Appendices B–C). If none of the three reaches a Robust band honestly — consistent with the documented 86% failure rate — the methodology sustains its claim; under the production-engine re-audit the three resolve

into two Disclaimers of Opinion for insufficient sample and one public Limited Robustness score with no gate binding (re-audit note, §6).

The choice of textbook strategies also addresses a transparency constraint. The present document can disclose the full strategy specification, the exact data used, the backtest implementation, and intermediate results, because none of that information constitutes private intellectual property. The contrast with any validation using proprietary strategies or private data is structural: that reporting would be necessarily incomplete.

§5.2 Firm/fund separation policy

TRAVIDENCE, as an independent validation service firm, maintains an explicit operational policy of separation between any internal strategies a shareholder or consultant may run as a private investor and the strategies evaluated by the validators published in this document or used in any institutional material. The policy translates into the following case-selection rule: no TRAVIDENCE-published study uses strategies from the proprietary fund of any employee, shareholder, or advisor; the three cases in this paper are synthetic in the sense of being built on public specifications (textbook) applied to public series, not private. The policy protects the institutional independence of the verdict and avoids the structural conflict of interest that affects prop trading firms themselves.

§5.3 Data sources, periods, and assumptions

The three cases use public series accessible via standard libraries. The table below summarizes the sources:

Case	Strategy	Period	Data source	Simulated tier
01	Golden Cross — CL futures	2000–2025	yfinance, `CL=F` daily back-adjusted	T2 (Audit Verified)
02	RSI(14) mean reversion — BTC perpetual	2020–2025	ccxt Binance `BTC/USDT:USDT` 1d	T2
03	Asian Range Breakout — EUR/USD	2015–2025	dukascopy-python `EUR/USD` 1h	T2

In all three cases, tier T2 corresponds to "Audit Verified" in the TRAVIDENCE evidence hierarchy (§0.1 of the canonical methodology). In commercial production, a T1 validation ("Verified Audit") requires a read-only connection via the broker's API; the public series used here are accepted for the reported studies but would not qualify for B2B institutional engagements in their current form.

The vanilla backtest cost assumptions are zero across the three cases (zero commission, zero spread, zero slippage). The choice is deliberate: the goal of the study is to stress the TRAVIDENCE methodology, not to estimate realistic P&L. Any realistic-cost version — assuming approximately $10 USD round-trip per trade in CL, typical exchange fees in BTC perpetual, and typical retail spread + commission in EUR/USD — would degrade the metrics and push each score further from

the Robust band; the zero-cost version is the most favorable side to the strategy and thus the most demanding test for the filter.

Position sizing across the three cases is unit (1 unit per trade, no compounding or pyramiding). The interpretation of "one unit" varies by operational context (1 CL contract in futures, 1 nominal BTC unit in perpetual, 1 standard lot of EUR/USD in retail FX), and the `capital_baseline` for the `capital_normalized_calmar` validator is adjusted accordingly via the convention capital_baseline = nominal_capital/leverage_factor documented in §4.2 and applied in §6.

---

§6. Three case studies

Re-audit note (2026-06-10, production engine). The validator walkthroughs in §6.1–§6.3 are the frozen record of the original audit, produced under the now-retired per-bucket regime calibration and bar-level observation counting (historical note, Appendix A §A.5). The production-engine re-audit — trade-level returns, the self-derived p50 regime cut, and enforcement of the admissibility floors — supersedes the final verdicts as follows. Case 01 (CL) and Case 02 (BTC) close with **Disclaimer of Opinion — Insufficient Evidence**: their trade-level samples (n_trades = 18 and 31) sit below the admissibility floor. Case 03 (EUR/USD) emits the framework's first real public end-to-end score: **56 / 100, Limited Robustness, no gate binding**. The gate-demonstration role formerly carried by EUR/USD passes to repository Case 04 (NQ Opening Range Breakout): declared insufficient by the original bar-counted audit, it is admissible under trade-level counting (n_trades = 2,424) and is capped by the [**regime, subperiod**] gates — binding-gate behavior on an admitted sample. Repository Case 05 (XAU/USD Bollinger reversal) is likewise gated [regime, subperiod]. Across the five repository case studies the portfolio covers the three public outcomes of the framework: score emitted (03), gated (04, 05), insufficient sample (01, 02). The certificates in Appendix B and the reproducibility hashes in Appendix C correspond to the re-audited verdicts. Per the methodology note in internal spec §12.3: audit findings update the auditor's view; auditors do not alter audits to match prior expectations.

§6.1 Case 01 — Golden Cross on CL crude oil (2000–2025)

Strategy description

Simple moving-average crossover on the daily settlement of the continuous WTI crude oil contract (CL). Long entry when the SMA(50) crosses above the SMA(200); exit when the SMA(50) crosses below. Long-only, no stops, no position sizing beyond one unit per trade. The 50/200 crossover is the focal convention of retail trend-following literature going back at least to the 1970s (Brock, Lakonishok, and LeBaron, 1992).

Underlying economic hypothesis

The hypothesis is trend persistence in commodity markets driven by structural supply/demand inertias (geopolitical shocks, OPEC quotas, capacity decisions of the shale cycle) and by slow-moving

macro dynamics (global GDP, USD strength) (Hong and Yogo, 2012). A long-term moving average acts as a low-frequency filter that participates in major bull markets (2003–2008, 2009–2014, 2020–2022) while remaining flat during compressed-range regimes. The hypothesis satisfies the five PFP-MP criteria (Prior, Falsifiable, Parsimonious, Mechanistic, Prespecified) detailed in §0.3 of the canonical methodology.

Vanilla backtest results

The backtest yields 18 trades over 25 years with a win rate of 44.4%, average winner of +36.4%, average loser of –8.9%, profit factor of 3.26, total return of 244%, geometric annualized return of 5.1%, annualized volatility of 25.4%, annualized Sharpe ratio of 0.32, Sortino ratio of 0.46, classical Calmar ratio of 0.083, and maximum percent drawdown of –62.2%. Average trade duration of 288 days. The trade density is typical of trend-following strategies with multi-year holdings.

Application of the TRAVIDENCE validators

Figures from the original frozen verdict; not reproducible against the current `metrics.json` *— see the §6 note.*

The `deflated_sharpe_ratio` validator with `n_trials_strict = n_trials_session = 2` (the two SMA windows, no grid search) returns `dsr_strict = 0.862`. The institutional threshold is 0.95; the Test 5 verdict is marginal FAIL due to sample insufficiency relative to Sharpe noise. Selection bias is low by construction (small hypothesis space), but the magnitude of the observed Sharpe is also insufficient to clear the threshold.

The `capital_normalized_calmar` validator with `nominal_capital = $50,000` and `leverage_factor = 9×` (typical CME margin for CL) returns `capital_baseline = $5,555.56`. The margin-frame Calmar is 0.083 and the capital-frame Calmar is 0.015, with `inflation_ratio = 5.34×`. The institutional threshold for capital-frame Calmar fails in both frames. The high inflation ratio (the highest of the three cases) reflects that the equity peak over the 25-year period grew considerably above the baseline before the worst drawdown.

The `subperiod_monotonic_decay` validator with `K=5` chunks over the 25 years returns the per-chunk Sharpe sequence $[0.78, 0.49, -0.24, -0.10, 0.41]$. The verdict `passes_no_decay = False` because `overall_decay_pct = 47.0% > 20%`, gate triggered: the strategy lost roughly half the Sharpe between the first and the last chunk.

The `regime_must_pass` validator splits the period's observations between the vol-on regime (1,073) and the vol-off regime (5,135) — classified by the strategy's own annualized volatility — and tests each regime separately against the commodities bucket's pre-registered minimum criteria: a Sharpe floor and a drawdown ceiling each regime must satisfy. The vol-on regime records a Sharpe of 0.3 and a maximum drawdown of 46.4%; the vol-off regime, a Sharpe of 0.4 and a drawdown of 71.0%. Neither regime reaches the bucket's Sharpe floor nor stays under its drawdown ceiling, so `passes_all_regimes = False`. Because the rule is strict conjunction (AND across regimes, with no averaging between regimes), one failing regime — here both fail — is enough for the regime gate to trigger.

The `component_isolation` validator does not apply (single-component strategy).

Robustness Score and verdict

Two hard gates trigger simultaneously (regime and sub-period); the most conservative cap prevails. The resulting band is Not Robust — `capped_by: regime` recorded in the verdict. The strategy, in its current textbook form, is not viable for retail prop firm deployment. (Original frozen verdict; see the re-audit note at the top of §6.)

Figure 1 (Appendix B; PDF at `../../../../case_studies/01-golden-cross-cl/audit_certificates/01-cl-golden-cross-audit-certificate-en.pdf`) reproduces the corresponding full audit certificate.

Interpretation

The Golden Cross on CL is a classic case of a strategy with a plausible mechanism — trend persistence in commodities is well documented — but with operational metrics insufficient to clear an audit-grade filter. The aggregate Sharpe is low, and sub-period asymmetry reveals material decay between the first half of the sample (before the shale revolution had reshaped market structure) and the second (when commodity regimes became more volatile and less trended). The strategy exemplifies the "correct mechanism, calibration insufficient for audit grade" failure mode.

§6.2 Case 02 — RSI(14) mean reversion on BTC perpetual (2020–2025)

Strategy description

RSI(14) with Wilder smoothing approximated via rolling mean on the daily close of BTC/USDT:USDT (perpetual contract). Long entry when RSI(14) < 30 and no position open; exit when RSI(14) > 50. Long-only, no stops, unit position sizing. The window (14) and the entry threshold (30) are values from Wilder's (1978) original framework; the mid-line exit (50) is a later convention.

Underlying economic hypothesis

Crypto markets exhibit periodic forced-liquidation cascades — flushes of leveraged longs in the perpetual market — that drive the RSI to extreme oversold readings when price deviates substantially from the normal volatility range. Documented events during the sample period: March 2020 COVID drop, May 2021 Elon-tweet correction, June 2022 Luna collapse, November 2022 FTX collapse. Mean reversion follows as funding rates normalize and dip buyers re-enter. The strategy captures the recovery leg without attempting to predict the cascade. The hypothesis satisfies the five PFP-MP criteria.

Vanilla backtest results

The backtest yields 31 trades over 5 years with a win rate of 54.8%, average winner of +8.9%, average loser of −5.5%, profit factor of 1.96, total return of 80.3%, annualized return of 7.9%, annualized volatility of 30.7%, Sharpe ratio of 0.41, Sortino ratio of 0.57, classical Calmar ratio of 0.16, and maximum percent drawdown of −47.9%. Average trade duration of 13 days.

Application of the TRAVIDENCE validators

Figures from the original frozen verdict; not reproducible against the current `metrics.json` *— see the §6 note.*

The `deflated_sharpe_ratio` validator with `n_trials_strict = n_trials_session = 3` returns `dsr_strict = 0.61`. Clear FAIL relative to the institutional threshold 0.95: the combination of a modest Sharpe and a three-trial selection space (the three Wilder parameters) produces an expected maximum under the null that the observed Sharpe does not clear.

The `capital_normalized_calmar` validator with `nominal_capital = $30,000` (median BTC over the period) and `leverage_factor = 20×` (typical crypto perpetual cross-margin) returns `capital_baseline = $1,500`. The margin-frame Calmar is 0.242 and the capital-frame Calmar is 0.233, with `inflation_ratio = 1.04×`. The ratio is the lowest of the three cases, reflecting that the worst drawdown occurs in early sample (when peak ≈ baseline) — the strategy decays with a deep DD near the start of the period before recovering. Both frames fail the institutional threshold for capital-frame Calmar.

The `subperiod_monotonic_decay` validator with `K=5` returns the sequence $[0.17, 0.72, 0.29, 0.62, 1.10]$. The verdict `passes_no_decay = True`: the strategy *strengthens* its edge over the period. The reported `overall_decay_pct` of –540.76% is mathematically correct but misleading because of small m_1 (0.17); the canonical boolean provides the interpretable answer. The strategy improved substantially between the first and the fifth chunk.

The `regime_must_pass` validator classifies the sample into the vol-on regime (50 observations, above the per-regime minimum-observation count) and the vol-off regime (1,914) by the strategy's own annualized volatility, and tests each regime against the bucket's pre-registered Sharpe floor and drawdown ceiling. The vol-on regime records a Sharpe of 0.2 — below the floor — and a maximum drawdown of 14.5%, within the ceiling: it passes the drawdown criterion but fails the Sharpe one. The vol-off regime records a Sharpe of 0.6, above the floor, and a drawdown of 31.2% that exceeds the ceiling: it passes Sharpe but fails drawdown. Each regime fails at least one criterion, so AND across regimes is False and the regime gate triggers.

Robustness Score and verdict

The triggered regime gate yields band Not Robust, with `capped_by: regime` recorded in the verdict. (Original frozen verdict; see the re-audit note at the top of §6.)

Figure 2 (Appendix B; PDF at `../../../../case_studies/02-rsi-mean-reversion-btc/audit_certificates/02-btc-rsi-mean-reversion-audit-certificate-en.pdf`) reproduces the corresponding full audit certificate.

Interpretation

The BTC case is informative for three reasons. First, it illustrates that `passes_no_decay = True` (the strategy does not decay; it improves) is a *necessary but insufficient* condition for a Robust band: other gates may trigger even when the sub-period test passes. Second, it demonstrates correct handling of observational asymmetry between regimes — vol-on with 50 obs vs vol-off with 1,914 obs — without the asymmetry compromising verdict validity when both regimes sit

above the minimum floor. Third, the example is operationally relevant for crypto investors: mean-reversion strategies on perpetuals with multi-day holding periods tend to exhibit the "apparent edge strengthening in early sample post-launch" pattern combined with severe drawdowns in vol-off, given that most of the post-2020 perpetual sample is dominated by vol-off regimes. The TRAVIDENCE methodology captures that profile correctly.

§6.3 Case 03 — Asian Range Breakout on EUR/USD (2015–2025) — end-to-end public-score case

Strategy description

Breakout of the Asian session range during the London session open on EUR/USD. Sessions in UTC: Asian 00:00–07:00, London 07:00–12:00, NY 12:00–17:00. Per trading day: compute `high_asian = max(high in 00:00–07:00)` and `low_asian = min(low in 00:00–07:00)`. During the London session, monitor the first breakout of the Asian range: enter long if the bar's high exceeds `high_asian`, enter short if the low breaks below `low_asian`. Mandatory exit at the NY close (17:00 UTC). No stops in the vanilla version. The strategy is widely documented in retail FX literature and prop firm educational material going back at least to 2010.

Underlying economic hypothesis

EUR/USD volume during the Asian session is dominated by Asian arbitrage flows and small retail activity; the pair is typically range-bound between 00:00 and 07:00 UTC. When London desks open at 07:00 UTC, ECB-related positioning and European institutional flows establish the day's directional bias. The breakout of the Asian range captures the dominant intra-day flow direction with mean follow-through to the NY close. The hypothesis satisfies the five PFP-MP criteria.

Vanilla backtest results

The backtest yields 2,584 trades over 10 years with a win rate of 59.6%, average winner of +0.33%, average loser of –0.28%, profit factor of 1.76, total return of 771%, annualized return of 15.5%, annualized volatility of 5.4%, Sharpe ratio of 2.67, Sortino ratio of 4.48, classical Calmar ratio of 2.38, and maximum percent drawdown of –6.5%. Average trade duration of 8.6 hours. The numbers are substantially stronger than those of the two preceding cases and, in isolation, would suggest a strategy with a high-band candidate profile.

Application of the TRAVIDENCE validators

Figures from the original frozen verdict; not reproducible against the current `metrics.json` *— see the §6 note.*

The `deflated_sharpe_ratio` validator with `n_trials_strict = n_trials_session = 3` (the three session windows) returns `dsr_strict = 1.0`. Clean PASS: the observed Sharpe is high enough and the selection space small enough to clear the institutional threshold with margin.

The `capital_normalized_calmar` validator with `nominal_capital = $100,000` (1 standard lot) and `leverage_factor = 30×` (retail FX ESMA) returns `capital_baseline = $3,333.33`. The margin-frame Calmar is 2.38 and the capital-frame Calmar is 1.66, with `inflation_ratio = 1.44×`. The margin-frame Calmar clearly clears the institutional threshold; the capital-frame Calmar also clears it, with a narrower margin. This is the central divergence the TRAVIDENCE dual frame makes visible: a strategy that would appear "strong edge" in margin frame (2.38) clarifies its real profile as "credible edge but no longer clearly strong" in capital frame (1.66) — the 44% re-framing is operationally relevant information for the fundability decision.

The `subperiod_monotonic_decay` validator with `K=5` returns the sequence $[2.59, 1.78, 3.03, 3.67, 2.32]$. `passes_no_decay = True`: the strategy does not decay; it oscillates around a high Sharpe without a monotonic trajectory.

The `regime_must_pass` validator is where the case reveals its critical profile. The 203 vol-on observations record a Sharpe of 3.2 and a maximum drawdown of 4.1%: the vol-on regime clears both the bucket's Sharpe floor and its drawdown ceiling comfortably. The 3,589 vol-off observations record a Sharpe of 2.7 — also above the floor — but a drawdown of 5.4% that does not meet the bucket's drawdown criterion: the vol-off regime passes the Sharpe criterion and **fails** the drawdown one. That failure is enough for AND across regimes to be False and the regime gate to trigger.

Robustness Score and verdict — the gate-vs-continuous behavior

The continuous aggregation of the three dimensions produces a profile sitting in the mid-range "ambiguous" zone of the consumer band ladder. In a pure-aggregation system — the pure-aggregation model — that profile would likely be reported as a favorable intermediate band, with a recommendation for minor revision. The strategy would obtain a favorable or lukewarm-favorable classification.

In the TRAVIDENCE hybrid continuous-gate system, the triggered regime gate replaces the continuous profile with the gate's cap level: band Not Robust, `capped_by: regime`. The qualitative transition from a continuous-favorable profile to Not Robust is the gate-vs-continuous behavior that constitutes TRAVIDENCE's defensible methodological moat. The client receives a clear binary signal: the strategy is not viable for retail prop firm deployment in its current form, regardless of the fact that the continuous quality of dimensions EQ and RP would place it in a more favorable range. (Original frozen verdict; see the re-audit note at the top of §6.)

Figure 3 (Appendix B; PDF at `../../../../case_studies/03-asian-range-breakout-eurusd/audit_certificates/03-eurusd-asian-range-breakout-audit-certificate-en.pdf`) reproduces the corresponding full audit certificate.

Interpretation — why EUR/USD anchored the original audit

The EUR/USD case was the original audit's anchor case for three convergent reasons; under the production-engine re-audit (re-audit note, §6) the regime gate no longer binds and the case's current role is the end-to-end public-score demonstration, while the gate-demonstration role passes to the repository's NQ ORB case.

First, it is the case where continuous aggregation and aggregation with hard gates produce quali-

tatively distinct verdicts. In the CL and BTC cases, the continuous profile sits structurally below the threshold even before the gates trigger, and the Not Robust verdict is robust to the choice of aggregation scheme. In EUR/USD, the continuous profile is marginally favorable and the decision depends exactly on whether the system incorporates hard gates. It is the case that separates the TRAVIDENCE methodology from the consumer aggregator methodology.

Second, the case illustrates that a *marginal* failure of a criterion in a *single* regime is enough to trigger the most severe gate. The vol-off regime passes the Sharpe criterion comfortably and does not meet the drawdown criterion for its regime. The AND-across-regimes rule does not contemplate the magnitude of the failure; the gate fires and the cap applies. The inflexibility is functional to the audit-grade design: a strategy with detectable asymmetry is a strategy the market will eventually expose, and the methodology must reject it before deployment, not soften the criterion as a function of the smallness of the margin.

Third, the case is operationally representative of the strategy profile that arrives at prop trading firms with the conviction of its viability: Sharpe 2.67, profit factor 1.76, controlled aggregate drawdown, plausible mechanism. Without the regime dimension explicitly evaluated, the strategy passes every internal filter the challenger would have applied on their own. Detection comes from external, audit-grade methodology with an independent third party. This is exactly the profile that motivates TRAVIDENCE's value proposition to the retail client.

---

§7. Comparison vs existing tools

§7.1 Feature-by-feature table

Table 1 contrasts TRAVIDENCE with the consumer-aggregator category and the reference implementation libraries on dimensions critical to the prop firm pre-validation use case. The ✓ and ✗ marks reflect each column's typical offering as of the publication date (v1.0).

Table 1. Compared capabilities: TRAVIDENCE vs. the consumer-aggregator category and the implementation libraries (mlfinlab, RiskLabAI).

Capability	TRAVIDENCE	Consumer aggregator (category)	mlfinlab	RiskLabAI
Deflated Sharpe Ratio implemented	✓	✓	✓	✓
Dual `n_trials` strict / session reporting	✓	✗	✗	✗
Capital-normalized Calmar (dual frame)	✓	✗	✗	✗

Capability	TRAVIDENCE	Consumer aggregator (category)	mlfinlab	RiskLabAI
Sub-period monotonic decay with AND semantics	✓	✗	✗	✗
Component isolation with $\lvert 1 - \sqrt{K} \rvert$ floor	✓	✗	✗	✗
MUST-PASS regimes vol-on / vol-off	✓	✗	✗	✗
CPCV (Combinatorial Purged Cross-Validation)	partial (Phase 5 upstream)	✗	✓	✓
PBO (Probability of Backtest Overfitting)	partial (Phase 5 upstream)	✗	✓	✓
Hard-gate system (cliff-edge audit)	✓	✗	n/a	n/a
Public band family with Disclaimer of Opinion for insufficient sample	✓	✗	n/a	n/a
Score [0,100] delivered as certificate	✓	✓	✗	✗
Trade-secret calibration per asset class	✓	partial	✗	✗
Certification signed by third party	✓	✗	✗	✗
Bilingual ES + EN as deliverable	✓	✗	✗	✗
Transparent commercial pricing tiers	✓	✓	n/a	n/a
Validator source code accessible	Full public specification (§4 + Appendix A)	✗	visible (commercial license)	✓

The consumer-aggregator column describes the category's typical offering as of the publication date, not a specific product.

Cells marked `n/a` correspond to products whose nature (technical library) does not contemplate the evaluated feature (commercial delivery to the client, signed certification). The partial check on CPCV and PBO for TRAVIDENCE reflects that the two techniques are contemplated as upstream steps of the Phase 5 orchestrator feeding the Robustness Score; tight integration with the validator body is the difference with respect to libraries implementing them as standalone functionality.

§7.2 Discussion of differentiators

An honest reading of Table 1 is that TRAVIDENCE does not invent quantitative validation methodology. The López de Prado stack (DSR, CPCV, PBO) was invented by López de Prado and collaborators; `mlfinlab` and `RiskLabAI` implemented it as reference libraries (RiskLabAI open-source; mlfinlab under a commercial license). TRAVIDENCE does three distinct things: enrich the stack with five novel contributions (§4), operationalize it as an audit-grade service with third-party certification (a service layer absent in the libraries), and calibrate the hard-gate system that functionally distinguishes TRAVIDENCE delivery from consumer-aggregator delivery.

The five novel contributions are orthogonal to the base stack. Dual `n_trials` reporting (§4.1) extends the standard DSR without invalidating it; the capital-frame Calmar (§4.2) extends the traditional Calmar without replacing it; the K-chunk sub-period schema (§4.3) generalizes the 3-period reporting convention without replacing aggregate analysis; component isolation with the `|1 − √K|` floor (§4.4) is a structural primitive complementary to univariate tests on the aggregate; the MUST-PASS vol-on / vol-off taxonomy (§4.5) instantiates López de Prado's general multi-regime validation principle with an operationalizable binary protocol. Each of the five is fully specified in this document (§4 and Appendix A) and is replicable by any third party from the published equations.

TRAVIDENCE's moat is not a technical secret about the methodology (which is publishable, as the present document demonstrates). The moat is institutional: (i) the trade-secret calibration of the Robustness Score per asset class (knots, gate caps, band thresholds), validated empirically on an initial synthetic population and with planned quarterly recalibration on anonymized data accumulated in production; (ii) the third-party signature on the verdict, with procedural discipline inherited from the professional auditing body; (iii) the institutional brand built on the track record of verdicts that survive over time. The conceptual equivalent in another market is the difference between publishing GAAP/IFRS (the technical apparatus, public) and operating a Big Four audit firm (the service layer, institutional). TRAVIDENCE is positioning in the second.

---

§8. Discussion

§8.1 Implications for retail prop firm pre-validation

The immediate application of the TRAVIDENCE methodology is pre-validation of strategies before the challenger pays the evaluation fee. Under the documented 86% challenge failure profile,

the expected calculation for the challenger is: a fee on the order of USD 300–600 per attempt for a standard 50,000–100,000 USD account, against a 14% probability of passing the challenge and a 7% probability (relative to the total) of receiving an effective payout. Expected value before pre-validation is negative for the vast majority of strategy profiles. Honest pre-validation, at a price materially below the fee and with an audit-grade verdict, trims the strategy universe to those with a reasonable probability of clearing the challenge, transforming the challenger's expected calculation.

The retail product's economics must be compatible with the challenge fee's order of magnitude. TRAVIDENCE's pricing is calibrated to sit below the typical evaluation fee but above the micro-transactional level of the consumer aggregator offering. The reason is structural: audit grade requires a review process that does not scale to micro-transactional prices. The strip between the two extremes — scalable yet institutional — is the natural commercial space of the product. Current pricing at travidence.com.

§8.2 Implications for B2B white-label pre-screening

The second application is B2B: tier-2 prop trading firms (smaller than the top-tier firms) facing a substantial cost in funded accounts when challengers pass the evaluation but never reach a payout. For those firms, integrating TRAVIDENCE as a pre-evaluation filter reduces exposure to accounts that will pay the challenge fee but structurally not complete the cycle. The B2B firm obtains three concurrent benefits: (i) a reduction in the cost of funded accounts that never reach a payout, (ii) competitive differentiation against tier-1 firms via the independent pre-validation seal, (iii) externalization of the technical validation decision to a third party whose conflict of interest is structurally bounded.

The commercial structure is Institutional Engagement, with custom scope and technical onboarding for API integration. Per-bucket calibration by asset class (futures vs forex/CFD vs crypto vs equities) preserves the trade-secret of the knots while providing the B2B client with a service configured to its challenger universe.

§8.3 The Big Four analogy

The conceptual parallel between TRAVIDENCE and the Big Four financial audit offering is not merely rhetorical; it is structural. The Big Four firms (Deloitte, EY, PwC, KPMG) operate in the financial assurance market as an independent third party that signs opinions (Unqualified / Qualified / Adverse / Disclaimer) on financial statements under procedural standards (GAAP/IFRS) and auditing standards (SAS/ISA). Their market value derives from three attributes: independence (rules restricting the services that could create conflict with the verdict), standards (procedural discipline is codified and retrospectively auditable), and reputation (the verdict track record sustains the service premium).

TRAVIDENCE proposes itself as the conceptual equivalent in the quantitative strategy assurance market, with three contextual adjustments: (i) the audited object is the strategy's validation methodology, not financial statements (the analogy is to process assurance, not financial outcome); (ii) procedural standards are the TRAVIDENCE methodology documented in this paper and in `docs/methodology/methodology-generalized.md`, inspired by the procedural discipline of AICPA (2011) SAS-122 / IAASB (2015) ISA 705 of the professional auditing body; (iii) the retail-

plus-B2B-tier-2 target market is structurally underserved by the Big Four incumbents, which have not considered the retail strategy pre-validation segment relevant. The vacant institutional strip is precisely the opportunity the TRAVIDENCE model articulates.

---

§9. Limitations

§9.1 Synthetic strategies only

Owing to the firm/fund separation policy (§5.2), the three cases in the present document are textbook strategies on public data. The TRAVIDENCE methodology has not been applied in this document to proprietary strategies of any real client nor to strategies from the proprietary fund of any TRAVIDENCE employee or advisor. Validation of the methodology on real-world client strategies happens in commercial production and accumulates as a post-launch track record; the present paper is the methodological base, not the track-record evidence. The separation is deliberate and is maintained as an institutional invariant.

§9.2 ES + EN languages only

Version 1.0 of the document exists in Spanish and in English, with professional translation between the two. Other languages relevant to the retail target market (Brazilian Portuguese, French, Mandarin) are deferred to future versions. The dual ES + EN delivery is the minimum viable set for the initial TRAVIDENCE market (Spanish-speaking LATAM + global anglophone institutional audience).

§9.3 v1.0 calibration on a synthetic population

Initial calibration of the Robustness Score v1.0 — mapping-function knots, gate cap levels, band thresholds — is performed on a synthetic population built from the three case studies in this document plus a simulated population of variants of the same strategies under controlled perturbations. That calibration constituted the initial `calibration_vintage` v1.0.0 / 2026-Q2; the current vintage documented in the metadata of the audit certificates is v1.1.0 / 2026-Q3 (recalibration of the per-regime must-pass criteria, 2026-07-05). Quarterly recalibration with anonymized real data accumulated in production (the "Layer 3" of the institutional defense stack, referenced in the internal spec) is protocol starting in the quarter after launch. Until then, issued verdicts carry the calibration version explicitly and are retrospectively auditable under that version.

---

§10. Conclusion

Audit-grade pre-validation of quantitative strategies is an institutional strip vacant between the implementation libraries of the López de Prado stack and the retail consumer aggregator offering. TRAVIDENCE articulates that strip through three integrated components: five novel contributions to the López de Prado stack (dual `n_trials` reporting in DSR, Capital-normalized Calmar, sub-period schema with monotonic decay detection, component isolation with the `|1 − vK|` floor, and vol-on / vol-off MUST-PASS regime taxonomy); a hybrid continuous-gate score — the Robustness Score — with a public result band inspired by the procedural discipline of AICPA SAS-122 / ISA 705; and a transparent commercial profile with pricing tiers designed to coexist with the typical prop firm challenge fee.

Three textbook case studies on public data (Golden Cross on CL, RSI(14) mean reversion on BTC perpetual, Asian Range Breakout on EUR/USD) illustrate the application of the filter; none reaches a Robust band. Under the production-engine re-audit (§6 note), CL and BTC close with a Disclaimer of Opinion for insufficient sample, and EUR/USD emits the public end-to-end score — 56 / 100, Limited Robustness, no gate binding. The gate-vs-continuous behavior that separates the TRAVIDENCE methodology from the consumer aggregator methodology — and constitutes the product's defensible methodological moat — is exhibited by the repository's NQ Opening Range Breakout case, admitted with 2,424 trades and capped by the regime and subperiod gates; across the five repository case studies the portfolio covers the three public outcomes: score emitted, gated, and insufficient sample.

Immediate future work includes (i) SOC 2 Type I certification as a procedural assurance layer over the firm's own operation (gate post-launch at 6–12 months), (ii) signature of LOIs with tier-2 prop firms for B2B white-label integration, and (iii) periodic recalibration of the Robustness Score with anonymized real data accumulated in production starting in the quarter after launch. Construction of the track record of verdicts that survive over time — the TRAVIDENCE equivalent of the Big Four audit jurisprudence body — is the medium-term axis of institutional defense.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used Claude (Anthropic) to assist with the drafting and language of the manuscript, under the author's direction. The methodology, analysis, design decisions, and conclusions are the author's own. After using this tool, the author reviewed and edited the content and takes full responsibility for the content of the publication.

Disclaimer

TRAVIDENCE is an independent validation service. It is not financial advice or an investment recommendation. Past performance does not guarantee future results; no validation eliminates the risk of loss.

References

American Institute of Certified Public Accountants (AICPA). (2011). Statement on Auditing Standards No. 122 — *Statements on Auditing Standards: Clarification and Recodification*. AU-C Section 705: Modifications to the Opinion in the Independent Auditor's Report. New York: AICPA.

Arian, H., Norouzi Mobarekeh, D., and Seco, L. (2024). Backtest Overfitting in the Machine Learning Era: A Comparison of Out-of-Sample Testing Methods in a Synthetic Controlled Environment. *Knowledge-Based Systems*, 305, 112477. doi:10.1016/j.knosys.2024.112477.

Bailey, D. H., Borwein, J. M., López de Prado, M., and Zhu, Q. J. (2017). The Probability of Backtest Overfitting. *Journal of Computational Finance*, 20(4), 39–69. doi:10.21314/jcf.2016.322.

Bailey, D. H., and López de Prado, M. (2012). The Sharpe Ratio Efficient Frontier. *Journal of Risk*, 15(2), 3–44. doi:10.21314/JOR.2012.255.

Bailey, D. H., and López de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *Journal of Portfolio Management*, 40(5), 94–107.

Brock, W., Lakonishok, J., and LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *Journal of Finance*, 47(5), 1731–1764. doi:10.1111/j.1540-6261.1992.tb04681.x.

Chan, E. (2013). *Algorithmic Trading: Winning Strategies and Their Rationale*. Hoboken: John Wiley & Sons.

Clenow, A. (2013). *Following the Trend: Diversified Managed Futures Trading*. Hoboken: John Wiley & Sons.

Finance Magnates. (2024, September). Exclusive: Only 7% of 300,000 Prop Trading Accounts Achieved Payouts. https://www.financemagnates.com/forex/analysis/exclusive-only-7-of-300000-prop-trading-accounts-achieved-payouts/ (Proprietary dataset from FPFX Tech; Justin Hertzberg, founder and CEO of FPFX Tech, cited.)

Hong, H., and Yogo, M. (2012). What Does Futures Market Interest Tell Us About the Macroeconomy and Asset Prices? *Journal of Financial Economics*, 105(3), 473–490. doi:10.1016/j.jfineco.2012.01.005.

International Auditing and Assurance Standards Board (IAASB). (2015). ISA 700 — Forming an Opinion and Reporting on Financial Statements; ISA 705 — Modifications to the Opinion in the Independent Auditor's Report; ISA 706 — Emphasis of Matter Paragraphs and Other Matter Paragraphs in the Independent Auditor's Report. IFAC.

Lo, A. W. (2002). The Statistics of Sharpe Ratios. *Financial Analysts Journal*, 58(4), 36–52. doi:10.2469/faj.v58.n4.2453.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Hoboken: John Wiley & Sons.

Mertens, E. (2002). Comments on the Correct Variance of Estimated Sharpe Ratios in Lo (2002, FAJ) When Returns Are IID. Working paper, University of Basel.

Politis, D. N., and Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313. doi:10.1080/01621459.1994.10476870.

Wilder, J. W. (1978). *New Concepts in Technical Trading Systems*. Greensboro, NC: Trend Research.

Young, T. W. (1991, October). Calmar Ratio: A Smoother Tool. *Futures*.

Appendix A — Mathematical derivations

A.1 Variance of the Sharpe estimator (Mertens 2002, building on Lo 2002)

For a return series $\{r_1, r_2, \ldots, r_T\}$ with mean μ, deviation σ, skewness γ_3, and Pearson kurtosis γ_4, Mertens (2002) derives, building on Lo (2002):

$$\widehat{\sigma}^2_{SR} \;=\; \frac{1 - \gamma_3 \cdot SR_{obs} + \frac{\gamma_4 - 1}{4} \cdot SR^2_{obs}}{T - 1}$$

where $SR_{obs} = \mu/\sigma$ is the sample Sharpe (not annualized). The derivation starts from the square of the moment ratio under iid, not-necessarily-normal returns and applies the delta method over the transformation $f(\mu, \sigma) = \mu/\sigma$. The TRAVIDENCE extension implements the formula with the `scipy.stats` convention `bias=False, fisher=False` for bit-identical alignment with the original Mertens (2002) form.

A.2 Exact form of E[max SR | N] (Bailey & López de Prado 2014)

The original paper derives the expected maximum Sharpe under the null of N independent trials:

$$E[\max SR \mid N] \;=\; \sqrt{V[\{SR_n\}]} \cdot \left[(1 - \gamma_E)\, \Phi^{-1}\!\left(1 - \frac{1}{N}\right) + \gamma_E\, \Phi^{-1}\!\left(1 - \frac{1}{Ne}\right) \right]$$

where $\gamma_E \approx 0.5772156649$ is the Euler-Mascheroni constant. The derivation is the convex combination of the $1 - 1/N$ and $1 - 1/(Ne)$ percentiles of the asymptotic distribution of the maximum of N normal draws, with weights $(1 - \gamma_E)$ and γ_E respectively. The Gumbel asymptotic approximation $\sqrt{2 \log N} - \gamma_E / \sqrt{2 \log N}$ overestimates this exact form by ~9–19% (N = 10–1000; the deviation shrinks as N grows); the canonical TRAVIDENCE implementation uses the exact form to preserve the test's declared conservatism.

A.3 Algebraic identity of Calmar `inflation_ratio`

Starting from the canonical definitions:

$$\text{Calmar}_{\text{margin}} = \frac{r_{ann}}{|DD_{t^*}/P_{t^*}|}, \qquad \text{Calmar}_{\text{capital}} = \frac{r_{ann}}{|DD_{t^*}/\text{capital\_baseline}|}$$

the ratio is:

$$\frac{\text{Calmar}_{\text{margin}}}{\text{Calmar}_{\text{capital}}} = \frac{r_{ann}/(DD_{t^*}/P_{t^*})}{r_{ann}/(DD_{t^*}/\text{capital\_baseline})} = \frac{DD_{t^*}/\text{capital\_baseline}}{DD_{t^*}/P_{t^*}} = \frac{P_{t^*}}{\text{capital\_baseline}}$$

with cancellation of r_{ann} and of the shared DD_{t^*}. The conclusion is that `inflation_ratio` is the ratio between the rolling peak at the worst-drawdown event and the declared capital baseline; it coincides with `leverage_factor` only in special cases that practice rarely satisfies.

A.4 Component isolation floor `|1 − √K|`

Under additivity ($r_{full} = \sum_i r_i$) and independence (IID components with mean μ and variance σ^2), the aggregate has mean $K\mu$ and variance $K\sigma^2$ (variances add under independence). By construction:

$$SR_{full} = \frac{K\mu}{\sqrt{K}\sigma} = \sqrt{K} \cdot \frac{\mu}{\sigma}$$

and the sum of standalone Sharpes is:

$$\sum_i SR_i^{\text{standalone}} = K \cdot \frac{\mu}{\sigma}$$

therefore:

$$\frac{\rho_{\text{int}}}{SR_{full}} = \frac{SR_{full} - \sum_i SR_i}{SR_{full}} = 1 - \frac{K}{\sqrt{K}} = 1 - \sqrt{K}$$

The floor is deterministic under additivity and independence and grows monotonically with K. The calibration of `interaction_tolerance` must accommodate this floor plus a buffer for finite-sample slack.

A.5 MUST-PASS taxonomy per asset class

The vol-on / vol-off cut is **derived from the strategy's own series**: it is the `p50` (median) of that series' rolling annualized volatility (`vol_window = 21`, strict `>`), computed in-engine with no per-asset-class calibrated value. Because a median cut is self-referential it is annualization-invariant (the annualization factor scales the rolling vol and its quantile alike and cancels) and yields a ~50/50 split by construction, so both regimes carry non-degenerate observational populations (above the floor `min_observations_per_regime`) without any tuning. There is **no** per-bucket `vol_threshold` and **no** `vol_threshold` field in the validator output — the cut is a public framework rule, not a

calibrated trade-secret value. *(Historical note: the three case studies in this paper were produced under the earlier, now-retired per-bucket calibration, whose per-bucket volatility cuts are trade-secret and not published; their frozen observation counts reflect that retired design, not the current derived rule.)*

Appendix B — Audit certificates as figures

The four audit certificates — the three corresponding to the case studies in §6.1–§6.3 plus the additional repository case referenced in the §6 re-audit note — are included as figures of the document. Binary inclusion of the PDF figures in the final rendered paper is performed at compile time by the render pipeline (pandoc/LaTeX). The markdown source-of-truth references the PDFs by relative path; canonical versions are maintained in `case_studies/*/audit_certificates/` with version v1.1.0 (calibration vintage 2026-Q3).

- **Figure 1.** TRAVIDENCE audit certificate for Case 01 — Golden Cross on CL futures (2000–2025). Path: `../../../../case_studies/01-golden-cross-cl/audit_certificates/01-cl-golden-cross-audit-certificate-en.pdf`. Final band: Disclaimer of Opinion — Insufficient Evidence.

TRAVIDENCE

Independent strategy validation

STRATEGY	CL Crude Oil Golden Cross (50/200 SMA)
ASSET CLASS	Commodity Futures — CL Crude Oil
PERIOD COVERED	2000-08-23 to 2025-05-16
CERTIFICATE ID	TRV-E8A6-E8E8
SCORED UNDER	Robustness Score v1.1.0 (2026-Q3)
ISSUED	2026-08-17 04:42 UTC

ROBUSTNESS SCORE

N/A

Disclaimer of Opinion

VERDICT

This strategy cannot be certified under TRAVIDENCE standards: n_trades = 18 below the admissibility floor Generate additional trade data and resubmit.

GATE STATUS

GATE	STATUS	EFFECT
Regime Must-Pass	PASSED	—
Subperiod Monotonic Decay	PASSED	—
Component Isolation	PASSED	—
Combined		**Final outcome: no gates triggered**

TRAVIDENCE

Independent strategy validation

VALIDATOR DETAIL

Deflated Sharpe Ratio

DSR variant	DSR	n_trials	E[max SR] (annualized)
Strict	0.964	2	0.111
Session	0.964	2	0.111

Capital-Normalized Calmar

Frame	Calmar	Notes
Margin frame	0.295	baseline = 5,555.56 (leverage 9x)
Capital frame	0.105	inflation ratio = 2.808

Subperiod Monotonic Decay

Field	Value
Status (no decay)	FAIL
Metric	sharpe
Per-subperiod (5 chunks)	0.57, 0.54, -0.72, -0.42, 0.37
Max consecutive drop	234.59%
Overall decay	34.49%

Regime Must-Pass

Regime	Status	min Sharpe	max DD	n obs
vol_off	FAIL	nan	0.00%	5205
vol_on	FAIL	1.130	17.36%	338

Component Isolation: N/A — single-component strategy

METHODOLOGY REFERENCES

Bailey, D. H., & López de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality.* Journal of Portfolio Management 40(5).

Lo, A. W. (2002). *The Statistics of Sharpe Ratios.* Financial Analysts Journal 58(4): 36-52.

Young, T. W. (1991). *Calmar Ratio: A Smoother Tool.* Futures Magazine, October.

AICPA SAS-122 / IAASB ISA 705 & 706. *Modifications to the Opinion in the Independent Auditor's Report.*

Reproducibility Hash (full):
`sha256:e8a6e8e8fd238f82fa65da2c9eb2d20f1e8a9d8635b489e8ad9c752966d1de8c · hash schema 2`

v1.1.0 sigma limitation note: This release uses conservative zero-sigma handling for confidence interval propagation. Single-realization Sharpe and Calmar inputs have sigma=0 by convention; full propagation through the delta method is deferred to a future release. CIs shown reflect this conservative bound; tighter intervals require bootstrap or block-bootstrap resampling.

Calibration disclosure: This certificate was issued under v1.1.0 (2026-Q3), which represents a best-achievable calibration of the Robustness Score scoring system. The calibration is conservative by design: strategies receiving Not Robust may include some that could be classified Limited Robustness under refined calibration. Subsequent recalibration with production data is planned as a quarterly protocol.

Generated by the TRAVIDENCE audit engine. Scored under v1.1.0 | calibration 2026-Q3.

Issued by Travidence, LLC · Wilmington, Delaware, USA

Travidence, LLC maintains technology professional liability and cyber (Tech E&O/Cyber) insurance in force.

- **Figure 2.** TRAVIDENCE audit certificate for Case 02 — RSI(14) mean reversion on BTC perpetual (2020–2025). Path: `../../../../case_studies/02-rsi-mean-reversion-btc/audit_certificates/02-btc-rsi-mean-reversion-audit-certificate-en.pdf`. Final band: Disclaimer of Opinion — Insufficient Evidence.

TRAVIDENCE

Independent strategy validation

STRATEGY	BTC Perpetual RSI 14 Mean Reversion
ASSET CLASS	Crypto Perpetual — BTC/USDT
PERIOD COVERED	2020-01-02 to 2025-05-19
CERTIFICATE ID	TRV-6A8E-058C
SCORED UNDER	Robustness Score v1.1.0 (2026-Q3)
ISSUED	2026-08-17 04:42 UTC

ROBUSTNESS SCORE

N/A

Disclaimer of Opinion

VERDICT

This strategy cannot be certified under TRAVIDENCE standards: n_trades = 31 below the admissibility floor Generate additional trade data and resubmit.

GATE STATUS

GATE	STATUS	EFFECT
Regime Must-Pass	PASSED	—
Subperiod Monotonic Decay	PASSED	—
Component Isolation	PASSED	—
Combined		**Final outcome: no gates triggered**

TRAVIDENCE

Independent strategy validation

VALIDATOR DETAIL

Deflated Sharpe Ratio

DSR variant	DSR	n_trials	E[max SR] (annualized)
Strict	0.691	3	0.380
Session	0.691	3	0.380

Capital-Normalized Calmar

Frame	Calmar	Notes
Margin frame	0.789	baseline = 1,500.00 (leverage 20x)
Capital frame	0.572	inflation ratio = 1.379

Subperiod Monotonic Decay

Field	Value
Status (no decay)	PASS
Metric	sharpe
Per-subperiod (5 chunks)	0.39, 0.98, -0.12, 1.01, 1.71
Max consecutive drop	112.66%
Overall decay	-332.04%

Regime Must-Pass

Regime	Status	min Sharpe	max DD	n obs
vol_off	FAIL	nan	0.00%	1228
vol_on	PASS	1.569	14.67%	596

Component Isolation: N/A — single-component strategy

METHODOLOGY REFERENCES

Bailey, D. H., & López de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality.* Journal of Portfolio Management 40(5).

Lo, A. W. (2002). *The Statistics of Sharpe Ratios.* Financial Analysts Journal 58(4): 36-52.

Young, T. W. (1991). *Calmar Ratio: A Smoother Tool.* Futures Magazine, October.

AICPA SAS-122 / IAASB ISA 705 & 706. *Modifications to the Opinion in the Independent Auditor's Report.*

Reproducibility Hash (full):
`sha256:6a8e058c61179010c5be84e68f846e758acd4de1453d37b9b1539b317b5babc6 · hash schema 2`

v1.1.0 sigma limitation note: This release uses conservative zero-sigma handling for confidence interval propagation. Single-realization Sharpe and Calmar inputs have sigma=0 by convention; full propagation through the delta method is deferred to a future release. CIs shown reflect this conservative bound; tighter intervals require bootstrap or block-bootstrap resampling.

Calibration disclosure: This certificate was issued under v1.1.0 (2026-Q3), which represents a best-achievable calibration of the Robustness Score scoring system. The calibration is conservative by design: strategies receiving Not Robust may include some that could be classified Limited Robustness under refined calibration. Subsequent recalibration with production data is planned as a quarterly protocol.

Generated by the TRAVIDENCE audit engine. Scored under v1.1.0 | calibration 2026-Q3.

Issued by Travidence, LLC · Wilmington, Delaware, USA

Travidence, LLC maintains technology professional liability and cyber (Tech E&O/Cyber) insurance in force.

- **Figure 3.** TRAVIDENCE audit certificate for Case 03 — Asian Range Breakout on EUR/USD (2015–2025). Path: `../../../../case_studies/03-asian-range-breakout-eurusd/audit_certificates/03-eurusd-asian-range-breakout-audit-certificate-en.pdf`. Final band: Limited Robustness, public score 56 / 100, no gate binding. End-to-end public-score case — the framework's first real non-gated score emission.

TRAVIDENCE

Independent strategy validation

STRATEGY	EUR/USD Asian Range Breakout
ASSET CLASS	Forex Major — EUR/USD
PERIOD COVERED	2015-01-01 to 2025-05-19
CERTIFICATE ID	TRV-AEC6-AD46
SCORED UNDER	Robustness Score v1.1.0 (2026-Q3)
ISSUED	2026-08-17 04:42 UTC

ROBUSTNESS SCORE

56 / 100

Limited Robustness

VERDICT

The strategy shows continuous quality below the Robustness threshold and exhibits material weaknesses across one or more validators. Structural rework required before challenge.

GATE STATUS

GATE	STATUS	EFFECT
Regime Must-Pass	PASSED	—
Subperiod Monotonic Decay	PASSED	—
Component Isolation	PASSED	—
Combined		**Final outcome: no gates triggered**

TRAVIDENCE

Independent strategy validation

VALIDATOR DETAIL

Deflated Sharpe Ratio

DSR variant	DSR	n_trials	E[max SR] (annualized)
Strict	1.000	3	0.260
Session	1.000	3	0.260

Capital-Normalized Calmar

Frame	Calmar	Notes
Margin frame	2.360	baseline = 3,333.33 (leverage 30x)
Capital frame	1.728	inflation ratio = 1.366

Subperiod Monotonic Decay

Field	Value
Status (no decay)	**PASS**
Metric	sharpe
Per-subperiod (5 chunks)	3.17, 2.11, 3.54, 4.46, 2.74
Max consecutive drop	38.49%
Overall decay	13.42%

Regime Must-Pass

Regime	Status	min Sharpe	max DD	n obs
vol_off	**PASS**	3.762	1.85%	1343
vol_on	**PASS**	2.782	4.83%	1343

Component Isolation: N/A — single-component strategy

METHODOLOGY REFERENCES

Bailey, D. H., & López de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality.* Journal of Portfolio Management 40(5).

Lo, A. W. (2002). *The Statistics of Sharpe Ratios.* Financial Analysts Journal 58(4): 36-52.

Young, T. W. (1991). *Calmar Ratio: A Smoother Tool.* Futures Magazine, October.

AICPA SAS-122 / IAASB ISA 705 & 706. *Modifications to the Opinion in the Independent Auditor's Report.*

Reproducibility Hash (full):
`sha256:aec6ad462f11f53609e946864136edfb953551bcca892906aa713243e1d2db5d · hash schema 2`

v1.1.0 sigma limitation note: This release uses conservative zero-sigma handling for confidence interval propagation. Single-realization Sharpe and Calmar inputs have sigma=0 by convention; full propagation through the delta method is deferred to a future release. CIs shown reflect this conservative bound; tighter intervals require bootstrap or block-bootstrap resampling.

Calibration disclosure: This certificate was issued under v1.1.0 (2026-Q3), which represents a best-achievable calibration of the Robustness Score scoring system. The calibration is conservative by design: strategies receiving Not Robust may include some that could be classified Limited Robustness under refined calibration. Subsequent recalibration with production data is planned as a quarterly protocol.

Generated by the TRAVIDENCE audit engine. Scored under v1.1.0 | calibration 2026-Q3.

Issued by Travidence, LLC · Wilmington, Delaware, USA

Travidence, LLC maintains technology professional liability and cyber (Tech E&O/Cyber) insurance in force.

Note on the "Gate Status" panel (Figures 1–3): in a Disclaimer of Opinion verdict, gate evaluation never occurs — the sample-admissibility check stops the flow before any composite score exists for a gate to cap. The panel reports that non-binding state; the underlying per-validator results, two of which are not satisfied in Cases 01 and 02, are reported in the "Validator Detail" table on the same page. The certificate template lacks a third visual state and renders non-binding as "PASSED"; the same two-state limitation applies to the Component Isolation row when the test reports N/A for single-component strategies (Figures 1–3). This presentation limitation is registered for correction in a future release of the certificate template.

- **Figure 4.** TRAVIDENCE audit certificate for Case 04 (repository) — NQ Opening Range Breakout (2015–2025), evaluated under the current production engine. Path: `../../../../case_studies/04-opening-range-breakout-nq/audit_certificates/04-nq-opening-range-breakout-audit-certificate-en.pdf`. Final band: Not Robust — binding gate [regime, subperiod]. Binding-gate demonstration case referenced in the §6 re-audit note.

TRAVIDENCE

Independent strategy validation

STRATEGY	NQ Opening Range Breakout (ORB)
ASSET CLASS	Equity Index Futures — E-mini Nasdaq-100 (NQ)
PERIOD COVERED	2015-01-02 to 2025-12-31
CERTIFICATE ID	TRV-DBCF-E8D7
SCORED UNDER	Robustness Score v1.1.0 (2026-Q3)
ISSUED	2026-08-17 04:42 UTC

ROBUSTNESS SCORE

capped by the regime gate

Not Robust

VERDICT

The strategy fails to maintain performance across both volatility regimes (high and low). Score capped by the regime gate. Strategy is NOT deployable in current form for prop firm challenges. Significant rework recommended before re-validation.

GATE STATUS

GATE	STATUS	EFFECT
Regime Must-Pass	FAILED	binding cap
Subperiod Monotonic Decay	FAILED	non-binding cap
Component Isolation	PASSED	—
Combined		**Final outcome: capped by the regime gate**

TRAVIDENCE

Independent strategy validation

VALIDATOR DETAIL

Deflated Sharpe Ratio

DSR variant	DSR	n_trials	E[max SR] (annualized)
Strict	0.066	48	0.670
Session	0.462	3	0.253

Capital-Normalized Calmar

Frame	Calmar	Notes
Margin frame	0.043	baseline = 5,555.56 (leverage 9x)
Capital frame	0.041	inflation ratio = 1.032

Subperiod Monotonic Decay

Field	Value
Status (no decay)	FAIL
Metric	sharpe
Per-subperiod (5 chunks)	-1.15, 0.16, -0.28, 1.25, 1.15
Max consecutive drop	276.72%
Overall decay	nan%

Regime Must-Pass

Regime	Status	min Sharpe	max DD	n obs
vol_off	FAIL	0.011	19.66%	1425
vol_on	FAIL	0.326	40.53%	1424

Component Isolation: N/A — single-component strategy

METHODOLOGY REFERENCES

Bailey, D. H., & López de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality.* Journal of Portfolio Management 40(5).

Lo, A. W. (2002). *The Statistics of Sharpe Ratios.* Financial Analysts Journal 58(4): 36-52.

Young, T. W. (1991). *Calmar Ratio: A Smoother Tool.* Futures Magazine, October.

AICPA SAS-122 / IAASB ISA 705 & 706. *Modifications to the Opinion in the Independent Auditor's Report.*

Reproducibility Hash (full):
`sha256:dbcfe8d71e2b03d508a926047dbe061b6f2453f7835277d06757137060996ccb · hash schema 2`

v1.1.0 sigma limitation note: This release uses conservative zero-sigma handling for confidence interval propagation. Single-realization Sharpe and Calmar inputs have sigma=0 by convention; full propagation through the delta method is deferred to a future release. CIs shown reflect this conservative bound; tighter intervals require bootstrap or block-bootstrap resampling.

Calibration disclosure: This certificate was issued under v1.1.0 (2026-Q3), which represents a best-achievable calibration of the Robustness Score scoring system. The calibration is conservative by design: strategies receiving Not Robust may include some that could be classified Limited Robustness under refined calibration. Subsequent recalibration with production data is planned as a quarterly protocol.

Generated by the TRAVIDENCE audit engine. Scored under v1.1.0 | calibration 2026-Q3.

Issued by Travidence, LLC · Wilmington, Delaware, USA

Travidence, LLC maintains technology professional liability and cyber (Tech E&O/Cyber) insurance in force.

Appendix C — Reproducibility hash and version metadata

Item	Value
Paper version	1.0
Canonical methodology version	v1.5 (internal TRAVIDENCE document, unpublished)
Robustness Score version	v1.1.0
Calibration vintage	v1.1.0 / 2026-Q3
Validation results schema version	v1.3
Source repository version	See `CITATION.cff` (`version` field) at repo root
Case 01 reproducibility hash (schema 2)	`sha256:e8a6e8e8fd238f82fa65da2c9eb2d20f1e8a9d8635b489e8ad9c752966d1de8c`
Case 02 reproducibility hash (schema 2)	`sha256:6a8e058c61179010c5be84e68f846e758acd4de1453d37b9b1539b317b5babc6`
Case 03 reproducibility hash (schema 2)	`sha256:aec6ad462f11f53609e946864136edfb953551bcca892906aa713243e1d2db5d`
Case 04 reproducibility hash — repository (schema 2)	`sha256:dbcfe8d71e2b03d508a926047dbe061b6f2453f7835277d06757137060996ccb`
Key dependency versions	Python 3.11+; numpy>=1.26; scipy>=1.13; pandas>=2.1; arch>=6.3 (Stationary Bootstrap; Politis and Romano, 1994); statsmodels>=0.14. Detailed lock in the repo's `uv.lock`. (Additional reference: a one-time cross-validation check against `mlfinlab`>=2.0 and `empyrical-reloaded`==0.5.12 on 2026-05-20 — not pinned dependencies in `uv.lock`.)

The four reproducibility hashes come from the Robustness Score output's `reproducibility_hash` field in `case_studies/*/results/metrics.json` and are verified as invariants of the run with the same inputs over the same validator commits. Changes in any of the validators produce different hashes; the field is used for procedural traceability after the verdict is issued.

Revision note (2026-08): the table above reflects the current schema-2 hashes (calibration v1.1.0, hash schema 2, spec §9.4). Earlier versions of this document cited the schema-1 hashes (engine ≤ v0.8.x), which remain valid verification anchors for certificates issued under that schema.

End of document.