

Benchmark 004: independent reproduction

Could independent research machinery recover the executable behavior of an external channel-breakout notebook and independently recapture its frozen market-data panel?

FULL_REPLICATION_MATCH

INPUT CELLS COMPARED	NUMERIC DISAGREEMENTS	MISSINGNESS DISAGREEMENTS
449,856	0	0
1,704 dates x 66 symbols x 4 fields		

Publication report. Generated from preserved closed research evidence. No strategy replay, retuning, new market-data capture, or re-evaluation was performed for this document.

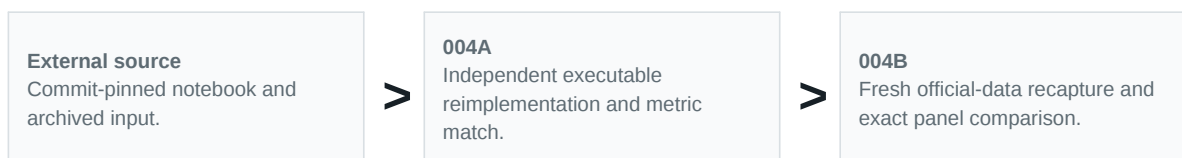
Published 5 September 2026.

Reproduce the executed object, not the cleanest story

The external source was the public *Statistical Arbitrage in Cryptocurrencies* repository, frozen at commit fdb17c13b81ca007bac8ac59b55f14d9088d5a28, specifically the Momentum / Channel Breakout notebook.

Original source: [frozen channel-breakout notebook at the source commit](#)

Benchmark 004A independently reimplemented the executed notebook path in Go. When explanatory prose, comments, and executable code disagreed, the reproduction target followed the code that actually produced the displayed result. The discrepancies were preserved separately as audit findings.



Frozen executable details reproduced

- Long entry from the highest prior 46 daily highs; short entry from the lowest prior 35 daily lows.
- Prior 50-day volume filter at the executable 0.68 quantile.
- Asymmetric 2-day long and 40-day short price exits.
- Reproduced 33-day post-hoc position-age mask, including its non-recursive behavior.
- 35-day sample volatility and executed square-root-of-volatility scaling before normalization.
- Transaction cost equal to 20 bps multiplied by daily portfolio turnover.

Historical reproduction target: the independent evaluator reproduced the source notebook's displayed +287.34% net cumulative return and 1.430 net Sharpe on its historical holdout. These are reproduction targets, not prospective evidence of future profitability.

The archived input was not simply taken on faith

Benchmark 004B froze the exact 66-symbol universe produced by 004A and captured daily high, low, close, and volume fields from the official Binance Spot public REST API. It did not substitute a current universe for the historical one.

PROVIDER REQUESTS 132 / 132 Zero automatic retries	SYMBOLS 66 Frozen before capture	PANEL DATES 1,704 2021-01-01 through 2025-08-31
---	---	--

449,856 field cells

1,704 dates x 66 symbols x high/low/close/volume



Numeric disagreements: 0

Missingness disagreements: 0

Max abs diff: 0

The canonical serialized input SHA-256 matched as well. 004B then reused the evaluator binary sealed by 004A rather than rebuilding it after the independent capture was visible. The same published display metrics were reproduced.

Required match	Outcome
Date grid	PASS
Symbol grid	PASS
Missing / non-missing state for every field cell	PASS
Finite high, low, close and volume values	PASS
Canonical CSV SHA-256	PASS
Published strategy display metrics	PASS

A perfect match did not erase the awkward parts

These issues were recorded before the independent reproduction result was known. Reproduction preserved them rather than laundering them into a cleaner specification after the fact.

Future availability influenced the historical universe The executable notebook required at least 90% non-null close coverage over the full panel through August 2025 before evaluating the portfolio from 2024 onward. Later availability could therefore influence an earlier universe.	Written volatility scaling and executed scaling differed The prose described a 30-period rolling standard deviation, lower exposure to higher-volatility assets, and EWMA smoothing. The executed holdout used a 35-day sample standard deviation, multiplied positions by $\text{sqrt}(\text{std})$, and did not apply the described EWMA smoothing.
A parameter comment did not match the executed value The executable volume threshold was 0.68 (68th percentile), while an inline comment described the 70th percentile.	Spot OHLCV was not a complete short-execution model The source data were Binance Spot prices while the portfolio could take negative weights. Borrow, funding, margin, liquidation, basis, and venue-specific short implementation were outside the published backtest.
Cost wording and executable formula differed The README summarized 20 bps transaction cost "per trade". The executable formula charged 20 bps multiplied by daily portfolio turnover.	The original dynamic universe could not be recreated honestly The source selection used then-current CoinGecko and Binance listings plus an older hourly file. Without historical snapshots, later reconstruction would require guesswork.
Parameter-selection evidence was mixed The notebook reported that every examined parameter set had at least one validation window with negative Sharpe, then selected the final choice using validation information-ratio comparisons before the held-out execution test.	

What FULL_REPLICATION_MATCH means - and what it does not

It does mean	It does not mean
• The specified executable behavior was independently recoverable. • The frozen 66-symbol market-data panel could be independently recaptured from the official provider with exact agreement. • The sealed independent evaluator reproduced the published display metrics on that input.	• Every methodological choice was sound. • The historical return was free from selection effects. • Short-side execution was realistic. • The strategy should be traded or would remain profitable in future data.

Reproducibility answers: can we get the same result from the same defined object?
It does not answer: will this make money in a new period?

What happened next

The next experiment did not modify the reproduced strategy to make it cleaner. Research 005 carried the reproduced executable semantics into a new, prospectively frozen 365-day holdout. That later experiment finished **FALSIFIED** because two predeclared survival criteria failed.

This is why the two terminal outcomes are compatible: Benchmark 004 established reproducibility of a historical object; Research 005 tested whether that same object satisfied a new prospective decision rule.

Selected preserved evidence

The public site exposes a deliberately small set of byte-for-byte text artifacts from the preserved Benchmark 004B terminal bundle. The upstream notebook and archived market dataset are not redistributed; raw provider payloads, binaries, and runner material are also omitted.

Terminal result	98f968533312722e705fdf8441d2df40ad84614f1d44785610f6c6b72745f635
Independent input comparison	263e0652a7197ceea7be26f183a78fdb234f179da3ec563eea5379626c05d9e9
Reproduced evaluation	f036a3631231ace778195203064d190b4562618fab186c7296535816ef34fdb2
Frozen executable specification	4ce2e6571b1ca1241798966bcf45500eb8875dacb614a6c39779984fc6f9442e
Frozen decision specification	fbcc7fc0762f287bc1e87ba884a319728cc67a92b0573f91788a3634c16c8853
Predeclared audit findings	622c37e4d04dfc875d836af96d9e91aca24a470b5f5a73dae1a3f418a3932a49

Evidence state: closed. This report is an editorial presentation layer over already-preserved evidence, not a new research artifact.

Detailed case study: </research/benchmark-004/>