

# Walk-Forward Factor Selection with Meta-Selection

Methodology for the JKP Characteristics Trading Factor challenge

Run of 2026-09-13 19:58 UTC

## 1. Overview

This submission is a long-short US equity strategy built from the characteristics in the CTF panel. It holds no view that was not derivable from data available at the time it was formed. Its distinguishing feature is that the *methodology* itself - which lookback window, how many characteristics, which selection rule, which weighting rule - is chosen by a walk-forward meta-layer rather than by the author. A fixed, symmetric menu of 108 configurations is simulated out of sample, and at each rebalance the strategy adopts the ones with the best track record *up to that date*. Nothing in the pipeline is tuned to the evaluation period, because no human ever chose among the configurations.

The run described here used 1,369,324 stock-months covering 12,325 stocks over 864 months (1951-12-31 to 2023-11-30), and 402 listed characteristics, of which 402 had enough history to be usable. It produced 1,361,635 stock-month weights across 140 rebalances. Rebalancing is monthly: a weight is published for every stock in every month.

## 2. The four stages

### Stage 1 - Characteristic portfolios

For every characteristic and every month, stocks are percentile-ranked cross-sectionally and given weight  $w_i = \text{rank}_i - 0.5$ , which is long the high end and short the low end with a mean near zero. The characteristic's long-short return is the weighted average  $LS = \sum w_i r_i / \sum |w_i|$ . A month needs at least 20 stocks with a valid rank, or the entry is missing; a characteristic needs at least 60 valid months overall, or it is dropped. The result is a months-by-characteristics matrix of realized long-short returns.

**The convention that governs everything downstream:** the value stored at month  $t$  is computed from `ret_exc_lead1m`, the return over month  $t+1$ . It is therefore *not* known at the end of month  $t$ . It becomes known at the end of month  $t+1$ .

### Stage 2 - Walk-forward simulation of a configuration menu

A fixed menu is defined once, before any results are seen, as the full cross of four dimensions:

- **lookback window:** 10y, 20y, expanding
- **number of characteristics held:** 5, 10, 20
- **selection rule:** **Sharpe** (top-N by annualized Sharpe over the window); **ClusterSharpe** (average-linkage clustering on  $1 - |\text{corr}|$  of long-short returns, then the best-Sharpe member of each of the top N clusters, which stops the portfolio filling up with five flavours of the same effect); **ConsistentSharpe** (a characteristic's *worst* Sharpe across three equal contiguous sub-periods of the lookback, which ranks down anything that earned its full-window Sharpe in one lucky stretch)
- **weighting rule:** **EW** (equal weight); **RankSharpe** (proportional to Sharpe rank among those selected); **MeanVar** ( $w \propto \Sigma^{-1} \mu$  with 30% shrinkage of the covariance toward its scaled identity, clipped at zero and normalized, falling back to equal weight if degenerate); **InvVol** ( $w \propto 1/\sigma$ , which uses no mean return at all and so cannot tilt toward a characteristic that merely had a good run)

That is  $3 \times 3 \times 3 \times 4 = 108$  configurations. The cross is deliberately complete rather than curated: a hand-picked menu would smuggle the author's hindsight back in, and no configuration in it was ever chosen for having performed well. Every 6 months each configuration selects its characteristics and weights, then earns their realized long-short returns until the next rebalance. Each configuration therefore builds up a genuine out-of-sample record, not a backtest

of itself.

### Stage 3 - Meta-selection

At each rebalance every configuration is scored by the annualized Sharpe ratio of its *own out-of-sample returns so far*, on an expanding window. Configurations with at least 120 months of track record are eligible, and the top 3 are blended equally. Until any configuration reaches that threshold, the strategy holds an equal blend of all active configurations - an uninformative prior, not a guess. In this run 120 of 140 rebalances were meta-selected and the remainder used the equal-weight prior.

This is the point of the design. If one were to backtest 108 configurations, pick the best over 1990-2023 and report its 1990-2023 Sharpe, that number would be inflated by having used the answer to choose the model. Here the configuration in force in any month was selected from performance strictly before it, so the reported track record is a record of decisions that were actually available at the time.

### Stage 4 - Stock-level weights

The governing rebalance's chosen configurations are blended equally, and each configuration's characteristic weights are mapped back to stocks using month- $t$  ranks only:

$$w_i = (1/K) \sum_{\text{cfg}} \sum_k f w_k \cdot (\text{rank}_{i,k} - 0.5) / \sum_j |\text{rank}_{j,k} - 0.5|$$

Because the mapping is the same linear-rank rule Stage 1 used, each configuration's stock portfolio replicates its characteristic-level return, so the statistics the meta-layer selected on describe the portfolio actually held. Output is (id, eom, w) for every non-zero weight.

### 3. How each stage avoids look-ahead

A decision taken at the end of month  $rp$  may use any return realized by then. Long-short row  $rp-1$  is realized over month  $rp$ , so it is known at the end of month  $rp$ ; row  $rp$  is realized over month  $rp+1$  and is not. Every lookback slice in Stages 2 and 3 therefore ends at  $rp-1$ , and the code contains no slice that ends later.

| Stage             | Decision taken at                 | Latest information it may read             |
|-------------------|-----------------------------------|--------------------------------------------|
| 1                 | n/a (measurement, not a decision) | row $t$ is a $t+1$ return, by construction |
| 2 - selection     | end of month $rp$                 | LS rows $\leq rp - 1$ month                |
| 2 - weighting     | end of month $rp$                 | LS rows $\leq rp - 1$ month                |
| 3 - meta          | end of month $rp$                 | config OOS rows $\leq rp - 1$ month        |
| 4 - stock weights | end of month $t$                  | month- $t$ characteristic ranks only       |

Table 1. Where each stage's information cut-off sits.

#### Every use of the forward return

`ret_exc_lead1m` is the prediction target. It appears in exactly three places in the model script, all of them inside Stage 1, and in one place in the evaluation code that is not part of `main()`:

- **Column selection and NaN filter.** The forward return is pulled alongside the characteristics and rows without one are dropped. A characteristic portfolio's return can only be measured over stocks that have a return, so this is measurement, not signal. It never enters a rank or a weight.
- **The Stage 1 numerator.**  $LS_t$  uses the month- $t$  forward return against month- $t$  ranks. This is the definition of a factor return and the reason the stored value belongs to month  $t+1$ . Every downstream stage is offset by one month precisely to respect it.
- **Evaluation only.** The stand-alone runner merges the forward return onto the published weights to score them. It runs after `main()` has returned and cannot influence any weight.

The forward return is never ranked, never used as a characteristic, and never read by Stages 2, 3 or 4. Check V0 below confirms it is absent from the characteristic list and that no characteristic is a near-copy of it.

### 4. Results

| Period    | Ann. Sharpe (gross) | Months |
|-----------|---------------------|--------|
| 1960-1989 | 2.463               | 360    |
| 1990-2003 | 3.132               | 168    |
| 2004-2013 | 0.751               | 120    |
| 2014-2023 | 0.764               | 119    |
| 1990-2023 | 1.723               | 407    |

Table 2. Annualized gross Sharpe of the published weights. Returns are gross of transaction costs; a first-order net adjustment is  $\text{net} = \text{gross} - 2 \times \text{turnover} \times \text{cost per side}$ .

| Measure                                        | Value |
|------------------------------------------------|-------|
| Average monthly one-way turnover               | 0.162 |
| Annualized volatility of monthly gross returns | 1.26% |
| Mean gross leverage (sum of $ w $ per month)   | 0.390 |

|                                   |           |
|-----------------------------------|-----------|
| Months with a published portfolio | 840       |
| Stock-month weight rows           | 1,361,635 |

Table 3. Portfolio characteristics.

## 5. Verification

The following checks were run against this exact script and this exact dataset. Two of them, V2 and V5, are the properties the challenge tests automatically.

| Check | What it tests                                                                                                                          | Result |
|-------|----------------------------------------------------------------------------------------------------------------------------------------|--------|
| V0    | Forward return absent from the characteristic list; highest  corr  with any characteristic 0.0374                                      | PASS   |
| V1    | Stages 2-3 rerun on a one-month-staler long-short matrix; 1990-2023 Sharpe 1.723 → 1.675                                               | PASS   |
| V2    | main() on data truncated at 1990-01-01 reproduces every earlier decision: 0 differing meta-choices, 0 differing selections, 477,697 we | PASS   |
| V3    | Stock-level return reproduces the characteristic-level return; correlation 1.000000, max difference 2.6e-1                             | PASS   |
| V4    | Volatility and Sharpe within plausible ranges; no long-period Sharpe above 4                                                           | PASS   |
| V5    | Two consecutive runs byte-identical; SHA-256 8c45fdd2adddd72b...                                                                       | PASS   |

Table 4. Verification results.

## 6. Reproducibility

There is no randomness anywhere in the strategy: no sampling, no shuffling, no random initialization, and no dependence on the iteration order of an unordered container. No external data is read; everything comes from the arguments to main(). The design constants - rebalance every 6 months, blend the top 3, 120 months of track record before a configuration is eligible, 60 valid months before a characteristic is usable, 30% covariance shrinkage - were fixed a priori and not revisited after seeing results.

| Component                  | Version / digest                                    |
|----------------------------|-----------------------------------------------------|
| Python                     | 3.12.10                                             |
| numpy                      | 2.5.3                                               |
| pandas                     | 2.2.3                                               |
| scipy                      | 1.18.1                                              |
| Platform                   | Windows-11-10.0.26200-SP0                           |
| strategy.py SHA-256        | be14392810c89b7b4b66a9726b3b1edb5560ba11946b70f0... |
| ctff_chars.parquet SHA-256 | d7d07cc10c5acda1455aa772bf4f60dc3a33f05ced9940f5... |

Table 5. Exact environment. requirements.txt pins these versions.

## 7. References

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *Journal of Finance* 78(5). The characteristics dataset.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric Portfolio Policies. *Review of Financial Studies* 22(9). The linear-rank weighting used here is a close cousin.