Discussion of

"Reviving Anomalies"

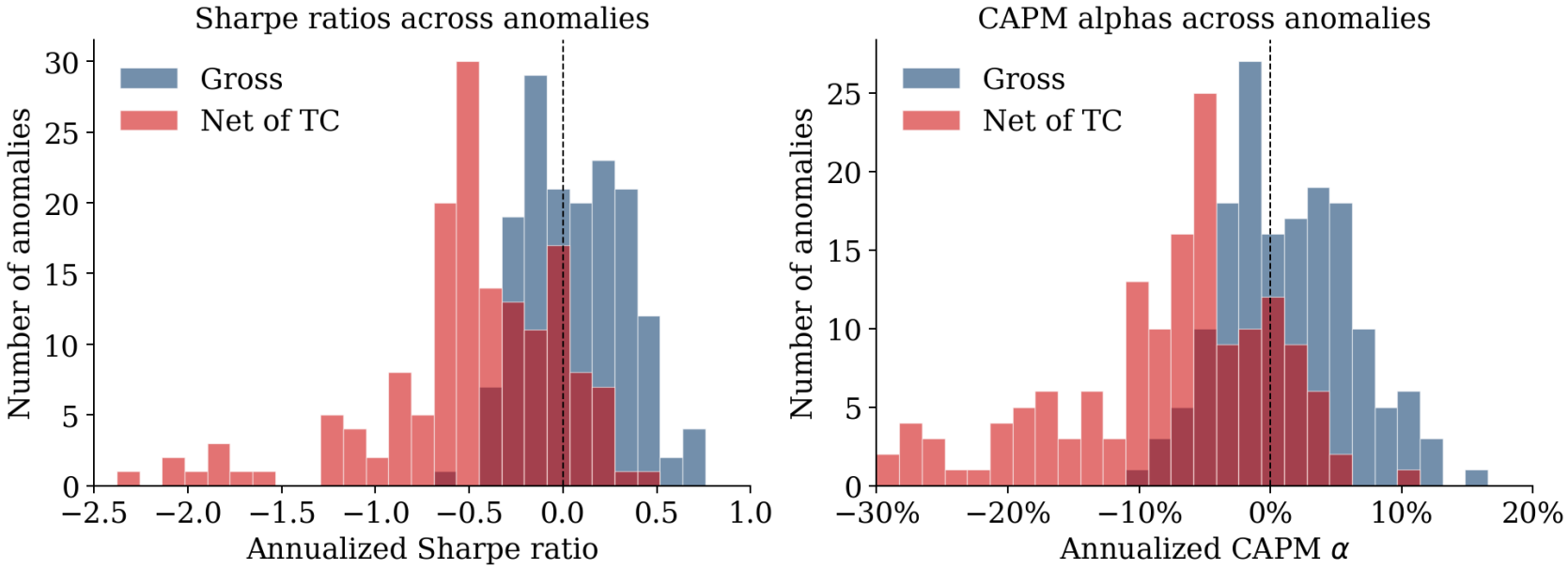
by Beckmeyer, Berg, Wiedemann, and Wortmann.

Daniele Bianchi
Queen Mary, University of London

1st Liechtenstein Workshop on AI in Finance

Most anomalies don't survive transaction costs

Across 159 long–short anomalies (2004–2023), the share with positive Sharpe ratio collapses from **57% (gross)** to **13% (net)**.



Gross vs net managed portfolio returns. Long–short value-weighted anomaly portfolios. Proportional transaction costs (half bid-ask) computed as in Barroso and Detzel (2021, JFE).

This paper

Main idea: replace the standard high-minus-low quintile sort

$$r^{HL}_{c,t+1} = r^{High}_{c,t+1} - r^{Low}_{c,t+1}$$

with a **dependent double sort** on the anomaly characteristic c and the expected net return $E_t[r^{net}_{i,t+1}] = E_t[r_{i,t+1}] - E_t[TC]$:

$$r^{HL,\text{Revived}}_{c,t+1} = r^{High,High}_{c,t+1} - r^{Low,Low}_{c,t+1}.$$

This paper

Main idea: replace the standard high-minus-low quintile sort

$$r^{HL}_{c,t+1} = r^{High}_{c,t+1} - r^{Low}_{c,t+1}$$

with a **dependent double sort** on the anomaly characteristic c and the expected net return $E_t[r^{net}_{i,t+1}] = E_t[r_{i,t+1}] - E_t[TC]$:

$$r^{HL,\text{Revived}}_{c,t+1} = r^{High,High}_{c,t+1} - r^{Low,Low}_{c,t+1}.$$

Two ingredients:

- ↪ $E_t[r_{i,t+1}] = f(c_{i,t})$ from a feed-forward NN on 153 characteristics (Gu, Kelly, and Xiu, 2020).
- ↪ $E_t[TC] \propto w_{i,t}\,\text{AUM} \,/\, V_{i,t}$, scale-dependent (Frazzini, Israel, and Moskowitz, 2018; Jensen, Kelly, Malamud, and Pedersen, 2025).

My comments:

Comments/observations:

#1: What does the ML augmentation actually buy us?

#2: Re-reading the mutual-fund evidence.

#3: Multiple testing and the headline count.

Comment #1: What does the ML augmentation actually buy us?

Comment #1a: The mechanics of the double sort

Two sequential sorts:

	$E_t[r^{net}]$ quintile				
c	L	2	3	4	H
H					**Long**
4					
3					
2					
L	**Short**				

Comment #1a: The mechanics of the double sort

Two sequential sorts:

	$E_t[r^{net}]$ quintile				
c	L	2	3	4	H
H					**Long**
4					
3					
2					
L	**Short**				

The second sort comes from

$$E_t[r^{net}_{i,t+1}] = E_t[r_{i,t+1}] - E_t[TC]$$

N.B: $E_t[r_{i,t+1}] = f(\cdot)$ is conditioned on the **full** 153-characteristic vector — **including c itself**.

Comment #1a: The mechanics of the double sort

Two sequential sorts:

	$E_t[r^{net}]$ quintile				
c	L	2	3	4	H
H					**Long**
4					
3					
2					
L	**Short**				

The second sort comes from

$$E_t[r^{net}_{i,t+1}] = E_t[r_{i,t+1}] - E_t[TC]$$

N.B: $E_t[r_{i,t+1}] = f(\cdot)$ is conditioned on the **full** 153-characteristic vector — **including c itself**.

The two sorts share information:

↪ What is the marginal contribution of the first sort vs. the second?

↪ Is ML here really complementary to c?

Comment #1b: Three missing "benchmarks"

The robustness check in the paper compares to **finer characteristic sorts** — but that is not the relevant counterfactual.

Comment #1b: Three missing "benchmarks"

The robustness check in the paper compares to **finer characteristic sorts** — but that is not the relevant counterfactual.

↪ **Linear $E[r]$ model** (Ridge / Lasso) in the same double sort.
Is the NN's flexibility essential, or does the signal alone suffice?

Comment #1b: Three missing "benchmarks"

The robustness check in the paper compares to **finer characteristic sorts** — but that is not the relevant counterfactual.

↪ **Linear $E[r]$ model** (Ridge / Lasso) in the same double sort.
Is the NN's flexibility essential, or does the signal alone suffice?

↪ **Single sort on $E[r]$ alone**, no characteristic.
Does the anomaly characteristic add anything beyond the ML signal?

Comment #1b: Three missing "benchmarks"

The robustness check in the paper compares to **finer characteristic sorts** — but that is not the relevant counterfactual.

- ↪ **Linear $E[r]$ model** (Ridge / Lasso) in the same double sort.
  Is the NN's flexibility essential, or does the signal alone suffice?
- ↪ **Single sort on $E[r]$ alone**, no characteristic.
  Does the anomaly characteristic add anything beyond the ML signal?
- ↪ **Drop high-$E[TC]$ stocks**, then standard quintile sort on c.
  Is $E[TC]$ doing more than acting as a liquidity screen?

Comment #1b: Three missing "benchmarks"

The robustness check in the paper compares to **finer characteristic sorts** — but that is not the relevant counterfactual.

↪ **Linear $E[r]$ model** (Ridge / Lasso) in the same double sort.
Is the NN's flexibility essential, or does the signal alone suffice?

↪ **Single sort on $E[r]$ alone**, no characteristic.
Does the anomaly characteristic add anything beyond the ML signal?

↪ **Drop high-$E[TC]$ stocks**, then standard quintile sort on c.
Is $E[TC]$ doing more than acting as a liquidity screen?

Without peeling apart these ingredients, we cannot tell what the ML augmentation contributes beyond characteristics or ex-post transaction costs.

Comment #2: Re-reading the mutual-fund evidence

Comment #2a: What does alignment identify?

Alignment with implementable anomalies can capture two things:

- ↪ **Characteristic-conditioning**: holdings that overlap with the corners of the augmented sort.
- ↪ **Raw ML exposure**: holding stocks with high $E_t[r^{net}_{i,t+1}]$.

Comment #2a: What does alignment identify?

Alignment with implementable anomalies can capture two things:

↪ **Characteristic-conditioning**: holdings that overlap with the corners of the augmented sort.

↪ **Raw ML exposure**: holding stocks with high $E_t[r^{net}_{i,t+1}]$.

The 6–8 pp annualized result mixes the two. **Controlling for $\sum_i w_{f,i,t}\, E_t[r^{net}_{i,t+1}]$ would help with identification.**

↪ If alignment survives, Section 4 identifies fund-level use of the *characteristic-conditional* information.

↪ If not, Section 4 identifies funds' overall ML-signal alignment.

Comment #2b: Strengthening the bridge to Section 3

Section 4 currently reads as a parallel exercise: a separate object (alignment), a separate analysis (panel regressions on fund characteristics).

Two ways to tighten the bridge:

- ↪ **Descriptive link**: for each fund, plot realized returns against the realized return of the implementable-anomaly basket at the fund's AUM.
- ↪ **Reframe**: present Section 4 as a *consequence* of the framework rather than *independent* evidence about fund behavior — "if the framework captures something real, we should see it in the cross-section of fund returns."

Either move tightens the link without requiring new tests.

Comment #3: Multiple testing and the headline count

Comment #3: Are significant returns disciplined for MHT?

Table 1 / Figure 2 are count statistics: how many of 153 anomalies are significant at the 5% level for a given fund size.

$\hookrightarrow$ 153 tests $\times$ highly correlated anomalies $\Rightarrow$ the count statistic is not disciplined for multiple testing.

Comment #3: Are significant returns disciplined for MHT?

Table 1 / Figure 2 are count statistics: how many of 153 anomalies are significant at the 5% level for a given fund size.

↪ 153 tests $\times$ highly correlated anomalies $\Rightarrow$ the count statistic is not disciplined for multiple testing.

The natural ask: apply a multiple-testing correction to the headline counts.

↪ E.g., recompute Table 1 with $|t| > 3$, the bar the paper itself cites in Fig. 1 (Harvey, Liu, and Zhu, 2015); or apply Benjamini–Hochberg FDR.

↪ If the augmented counts remain meaningfully larger than the unadjusted baseline, the headline is robust.

Minor comments

↪ **Literature**: Novy-Marx and Velikov (2016) on the trading costs of anomalies, and Novy-Marx (2013) on quality/profitability, are directly relevant and uncited.

Minor comments

- ↪ **Literature**: Novy-Marx and Velikov (2016) on the trading costs of anomalies, and Novy-Marx (2013) on quality/profitability, are directly relevant and uncited.
- ↪ **Clarity on $E[TC]$**: state explicitly that $E[TC]$ enters only via the second sort variable in Eq. 3, not via weighting or stock screening.

Minor comments

- ↪ **Literature**: Novy-Marx and Velikov (2016) on the trading costs of anomalies, and Novy-Marx (2013) on quality/profitability, are directly relevant and uncited.
- ↪ **Clarity on $E[TC]$**: state explicitly that $E[TC]$ enters only via the second sort variable in Eq. 3, not via weighting or stock screening.
- ↪ **TC calibration**: the Frazzini, Israel, and Moskowitz (2018) price-impact estimate is dated (pre-HFT saturation). Sensitivity to halving/doubling Λ would tell us how knife-edge the medium/large-fund counts are.

Summary

Nice paper, I learned a lot.
Looking forward to the next version!