Predicting the Next Position of a Disclosed Institutional Portfolio

Evidence from five disclosure regimes, and why a predictable trade is not a profitable one

Howard Chan PriorMoves, Hong Kong chakhanghowardchan2008@gmail.com ORCID: 0009-0000-3463-3795

Version 2.1, 18 September 2026. Working paper. All numbers are emitted by named artifacts in the project repository and are reproducible from the commands in Appendix C.

AI disclosure: parts of this manuscript were drafted with the assistance of a large language model. All numerical results were produced by the named artifacts in the project repository and were not generated by a language model. The author is responsible for the content. The full statement is on page 1 of the PDF and in the section headed AI disclosure.

Abstract

Institutional equity holdings are disclosed publicly and late. A United States 13F filing describes a portfolio as it stood up to forty-five days before the filing appears, and the large-shareholding regimes of Japan, Korea, Hong Kong, the United Kingdom and the European Union each impose their own statutory delay on their own trigger. The finance literature reads this record as an input to a return question: can an outsider profit by copying what was disclosed? We ask a narrower question that the same data answers far more cleanly. Given a manager's disclosed history, which position does that manager add next?

We fit one gradient-boosted classifier per manager on a purged walk-forward split and evaluate with a pair-weighted stratified AUC whose strata are manager-quarters, so that no part of the score can be earned by learning which manager a row belongs to. On 49 United States managers across 35 quarters and 124,731 held-out observations, comprising 1,244,011 comparable pairs against a base rate of 12.46%, the model reaches an AUC of 0.6762 with a manager-clustered 95% bootstrap interval of 0.6563 to 0.7002. The top decile of the ranking is bought 3.17 times as often as the base rate.

The method transfers to four further disclosure regimes whose filing triggers, thresholds and deadlines differ from the United States and from each other. Against a control refitted on labels permuted within each manager, which preserves manager identity in full and destroys every other signal, Japan, Hong Kong, the United Kingdom and the European Union clear their controls by 0.14 to 0.25 AUC. Korea does not: its pooled figure of 0.8650 is reproduced by its own shuffled control at 0.8831, and we report it as a negative result rather than suppressing it or quietly dropping the market.

The paper's second half is the part a practitioner should read first, and it is negative. The predictable component of institutional demand does not convert into excess return at any horizon we can measure. A pre-specified family of eight portfolio transforms of the ranking, entered at the date the filing actually becomes public and held one quarter, produces a best absolute t-statistic of 1.00 against a family-wise threshold of 2.57 on 84,021 predictions. A 21-day event study on 3,936 disclosed additions returns an excess of 0.0511% with a 95% interval of -0.8195 to 0.9216. Decile monotonicity of subsequent return across 82,949 priced rows is 0.018. A quarterly 15-name book formed at quarter end returns 489.5%

against a benchmark 239.4% over the same 35 quarters, and we report in full that only one of five estimators of its mean quarterly margin clears 0.05, that none does once the single largest quarter is removed, and that reaching t = 2 at the observed effect and dispersion would require 81 quarters of a quarterly instrument.

We argue that this combination is coherent rather than contradictory, and that it is the honest reading of the disclosure record: the identity of the next buyer is predictable because it is driven by persistent, slow-moving manager characteristics, while the price impact of that buyer's arrival is either absent, already impounded, or smaller than the transaction costs of harvesting it. The contribution is a clean, reproducible measurement of the first fact and an equally clean refutation of the second.

Keywords: institutional holdings, 13F, disclosure, portfolio prediction, machine learning, market efficiency, backtest overfitting

JEL: G11, G14, G23, C53

---

1. Introduction

Every large equity manager in a major market is compelled, sooner or later, to tell the public what it owns. The compulsion varies. In the United States a manager exercising investment discretion over at least $100 million in Section 13(f) securities files a quarterly Form 13F within forty-five days of quarter end. In Japan, Korea, Hong Kong, the United Kingdom and the European Union the trigger is a crossing of a stake threshold in a single issuer rather than a portfolio size, the deadline is measured in days rather than weeks, and the filing describes a position rather than a portfolio. What the regimes share is that the disclosure is mandatory, public, machine-readable and late.

The research literature has treated this record almost exclusively as an input to a return question. Frank, Poterba, Shackelford and Shoven (2004) construct copycat funds from disclosed mutual fund holdings. Cohen, Polk and Silli (2010) ask whether a manager's largest-weight positions outperform the rest of the same manager's book. Agarwal, Jiang, Tang and Yang (2013) exploit the confidential treatment channel to show that what managers hide outperforms what they reveal. Griffin and Xu (2009) ask directly how much skill is visible in hedge fund 13F holdings and find little. Shi (2017) and Aragon, Hertzel and Shi (2013) study the cost that disclosure imposes on the discloser. In every case the dependent variable is a return, and the return must be estimated with a model of expected return, over a horizon the researcher selects, on a sample the disclosure schedule makes small.

We ask a different question, and the reason is methodological rather than thematic.

> Given everything a manager has disclosed through quarter t, which security does that manager add to the portfolio in quarter $t+1$?

Three properties make this question tractable where the return question is not.

The label is mechanical. A security either appears in the manager's next filing without having appeared in the previous one, or it does not. There is no estimation of the outcome variable, no benchmark model, no horizon choice, and no price series standing between the prediction and the truth. The label is a set difference over two filings, computable by anyone, identically, forever.

The obvious shortcut is removable. Managers differ enormously in how often they add anything. A concentrated activist may add two names a year; a quantitative manager may add two hundred a quarter. A model scored on a pooled sample can reach a high AUC by learning nothing except which manager a row belongs to, because manager identity predicts the base rate. Scoring within manager-quarter strata removes this channel by construction: a comparison is only ever made between two candidate securities for the same manager in the same quarter, so anything the model knows about the manager in general cancels.

The negative control is exact. Permuting the labels within each manager preserves the manager's identity, the manager's base rate, the manager's universe, the manager's turnover and the manager's holding period, while destroying the association between a specific security and a specific addition. A model that has learned only identity reproduces its score under this permutation. A model that has learned anything else collapses toward 0.5. The gap between the real fit and the permuted fit is therefore an interpretable quantity, and we report it for every market, including the one where it is negative.

1.1 What we find

Section 6 establishes that the next addition is predictable. The United States figure, 0.6762, is modest in absolute terms and is what a well-specified prediction on a genuinely hard problem looks like. The cross-regime figures are higher, for a reason we make explicit in Section 6.2: a threshold-crossing regime discloses a much sparser and more deliberate set of events than a portfolio regime, so the candidate set is smaller and the positives are more distinctive.

Section 7 establishes that this predictability does not pay. This is the part of the paper that took the most work and that we regard as its main practical contribution, because the literature's standing presumption, and every commercial pitch built on disclosed holdings, runs the other way. We test the return question five separate ways, pre-specifying the family before running it, and it fails five times.

Section 8 describes a preregistration and timestamping apparatus that makes the forward record checkable by a third party without trusting us, which we consider the appropriate response to Harvey, Liu and Zhu (2016) and to Bailey, Borwein, López de Prado and Zhu (2014) on the multiple-testing and backtest-overfitting problems in this literature.

1.2 Contributions

1. A prediction target on disclosed institutional holdings with a mechanical label, and an evaluation design (pair-weighted stratified AUC with manager-quarter strata, manager-clustered bootstrap) under which the manager-identity shortcut is arithmetically unavailable.
2. A replication across five disclosure regimes with materially different statutory triggers, including one regime where the method fails its own control and is reported as a failure.
3. A five-way refutation of the natural monetisation of the first result, with the family of transforms pre-specified in code before the run, and a power calculation showing that the quarterly disclosure instrument, rather than the modelling, is the binding constraint on ever resolving the question.
4. A reproduction apparatus: a register of 457 litigated hypotheses with the verdict rule fixed in advance of each run, and 85 forecasts cryptographically timestamped before their outcomes existed.

---

2. Related work

Disclosed holdings and return. The copycat literature begins with the observation that mandatory disclosure hands outsiders a free, if stale, view of professional portfolios. Frank, Poterba, Shackelford and Shoven (2004) find that funds replicating disclosed mutual fund holdings capture most of the original's gross return, chiefly by avoiding fees rather than by capturing alpha. Wermers (2000) decomposes mutual fund performance into a holdings-based component and a cost component and finds the stock-picking component small relative to the frictions. Griffin and Xu (2009) apply the same lens to hedge fund 13F holdings and find limited evidence of differential skill visible in the disclosed book. Our Section 7 is consistent with all three, and strengthens them in one respect: we show that even conditioning on a *correct* prediction of the next buyer, the return is absent.

What disclosure hides. Agarwal, Jiang, Tang and Yang (2013) exploit confidential treatment requests to show that the holdings managers succeed in hiding outperform those they reveal, which implies that the public record is adversely selected. Aragon, Hertzel and Shi (2013) and Shi (2017) document the cost of disclosure to the discloser, and the resulting incentive to disclose the least informative subset consistent with the rule. This is the strongest prior reason to expect our Section 7 result, and we regard our finding as an out-of-sample confirmation of it in a setting the original papers did not study.

Crowding. Sias, Turtle and Zykaj (2016) study hedge fund crowds and find that consensus positions carry predictable reversal. Our pre-specified family in Section 7.1 includes both a consensus arm and a contrarian arm precisely because this literature makes both directions plausible; neither clears its threshold.

Machine learning in asset pricing. Gu, Kelly and Xiu (2020) establish that flexible learners materially improve return prediction relative to linear benchmarks, and Freyberger, Neuhierl and Weber (2020) reach a similar conclusion nonparametrically. Our estimator choice follows this literature, but our target does not: we predict a disclosed human action rather than a price. López de Prado (2018) supplies the purging and embargo discipline we adopt in Section 5.3, and Bailey, Borwein, López de Prado and Zhu (2014) and Harvey, Liu and Zhu (2016) supply the multiple-testing discipline that motivates our pre-specification and our timestamped forward record.

Evaluation. The pair-weighted stratified estimator in Section 5.4 is a stratified variant of the Mann-Whitney interpretation of the area under the ROC curve (Hanley and McNeil, 1982), and our clustered uncertainty follows Cameron, Gelbach and Miller (2008) rather than the asymptotic variance of DeLong, DeLong and Clarke-Pearson (1988), because the correlation structure inside a single filer's rows is stronger than that estimator assumes.

Where this paper sits. To our knowledge the specific target, the next addition to a disclosed institutional portfolio, has not been evaluated as a prediction problem in its own right with a manager-identity-free estimator and a cross-regime replication. The closest work treats disclosed holdings as a signal about price. We treat them as a signal about the discloser.

---

3. Institutional setting

The five regimes differ in what compels a filing, and this difference is what makes the cross-regime test informative rather than a replication on more of the same data.

Regime	Instrument	Trigger	Deadline	Unit disclosed
United States	Form 13F (17 CFR 240.13f-1)	$100m in 13(f) securities under discretion	45 days after quarter end	Whole portfolio, quarterly
Japan	Large shareholding report (FIEA)	5% of voting rights, then 1% changes	5 business days	Single position
Korea	5% rule (FSCMA), DART	5% of voting stock, then 1% changes	5 days	Single position
Hong Kong	Disclosure of Interests (SFO Part XV)	5% of a class, then whole-percentage changes	3 business days	Single position
United Kingdom	DTR 5 (FCA)	3% for UK issuers, then each 1%	2 trading days	Single position
European Union	Transparency Directive 2004/109/EC, as transposed	5% national default, several states lower	Varies by state	Single position

Three consequences follow, and each shows up in the results.

A portfolio regime produces a dense candidate set; a threshold regime produces a sparse one. In the United States a manager's candidate set each quarter is effectively the investable universe, and the base rate of addition is 12.46%. In a threshold regime the filer only appears in the record when it crosses a stake level in a specific issuer, so the observed events are deliberate, large and few. This is the mechanical reason the cross-regime AUCs in Section 6.2 exceed the United States figure, and it is why we do not treat the comparison as a statement that the method works better abroad.

A portfolio regime is staler. The United States filing lag of up to 45 days is the longest in the set by an order of magnitude. Section 7.1 shows that respecting this lag, rather than assuming quarter-end availability, is the single most consequential implementation decision in the paper.

A threshold regime has a censored bottom. A manager who builds a 4.9% stake in a United Kingdom issuer and stops is invisible. The panel therefore contains no information about small, tentative, or abandoned positions abroad, and any model fit to it inherits that censoring. We flag this as the principal external validity limitation of the non-United States results in Section 10.

---

4. Data

4.1 Construction

Every panel is collected directly from the statutory source rather than from a vendor redistribution, for two reasons: a vendor's restatement policy is opaque and would silently contaminate a walk-forward split, and a vendor licence would prevent the reproduction in Appendix C.

Regime	Rows collected	Source of record
United States 13F	124,731 held-out observations	SEC EDGAR
Korea	30,831	DART large shareholding filings
European Union	18,937	National major-holdings registers
Hong Kong	11,565	HKEX Disclosure of Interests
Japan	5,830	EDINET large shareholding reports
United Kingdom	1,017	FCA

The row counts above are the collected panels. The counts that enter the cross-regime evaluation in Section 6.2 are smaller, because a row is scorable only when the manager has enough prior history to form features and enough contemporaneous universe to form comparisons: Japan 3,464, Korea 16,074, Hong Kong 7,688, European Union 1,447, United Kingdom 192. We report both, because quoting the collected figure where the scored figure is the one doing the work is the most common way this kind of table misleads.

4.2 The United States panel

The United States panel covers 49 managers over 35 quarters, from the quarter ending 30 September 2017 to the quarter ending 31 March 2026. Managers were selected for continuous filing history over the window and for being discretionary equity managers rather than index vehicles or custodians. Selection was performed once, before any model was fit, and the list has not been revised in response to results.

The held-out sample contains 124,731 manager-quarter-security observations arranged into 1,482 manager-quarter strata, of which 45 are degenerate: they contain either no positive or no negative and therefore contribute no comparable pairs. The remaining strata yield 1,244,011 comparable positive-negative pairs. The base rate of addition across the held-out sample is 12.46%.

Concentration matters for the clustered interval, so we report it: the largest single manager contributes 13.28% of all comparable pairs. No manager dominates the estimate, and the quotability screen described in Section 5.6 would have refused the figure if one had.

4.3 Labels

For manager i and quarter t, let $H_{i,t}$ be the set of securities in the manager's filing for period t. The label for security j is

$$y_{i,t,j} = \mathbb{1}\left[j \in H_{i,t+1} \ \wedge \ j \notin H_{i,t} \right]$$

that is, an addition is a security present in the next filing and absent from the current one. Increases to existing positions are not additions and are not labelled positive; they are a separate target with different economics, and mixing them inflates the base rate and flatters the score.

The candidate set for a given manager-quarter is the union of the securities that manager has ever held, the securities held by any manager in the panel that quarter, and the liquid investable universe as of that quarter end. A candidate set that is too narrow makes the problem artificially easy, so we err wide. The held-out sample averages 84 scored candidates per manager-quarter stratum (124,731 observations over 1,482 strata), and the resulting base rate of 12.46% is the consequence of that width.

4.4 Features

Features are formed exclusively from information available at the close of the quarter being scored, and are of five kinds.

Manager taste. The manager's own historical propensity toward a security's sector, size decile, valuation decile, liquidity tier and country of listing, estimated on the manager's filings up to the split boundary only.

Position state. Whether the manager has held the security before, how long ago it was exited, whether the exit was full or partial, and the manager's stake in the issuer where a threshold regime discloses it.

Peer structure. How many other managers in the panel hold the security, the change in that count, and whether the manager has historically followed or faded that peer group. The construction is deliberately symmetric so that the model can learn either a herding or a contrarian response, and Section 7.1 tests both directions as portfolio arms.

Security state. Momentum, realised volatility, dollar volume and liquidity tier as of the quarter end.

Calendar. Quarters since the manager's last addition, and the manager's trailing turnover.

Mutual information against the label, estimated on the training folds only, is concentrated: the single strongest feature is the disclosed stake at 0.11462 nats, and the top three features together carry 79% of total mutual information. No feature in the set carries effectively zero, which is the diagnostic that told us the remaining headroom lies in new columns rather than in a different estimator.

---

5. Method

5.1 The prediction problem

For each manager i and each scored quarter t, the model produces a score $s_{i,t,j}$ for every candidate security j, and the evaluation asks only whether the scores order the additions above the non-additions within that manager-quarter. Nothing in the evaluation depends on the scale, the calibration or the cross-manager comparability of the score. This is deliberate: the product question is a ranking question, and a ranking metric is the one that cannot be gamed by a monotone transform.

5.2 Estimator

One gradient-boosted decision tree classifier is fit per manager, using the histogram-based implementation in scikit-learn (Pedregosa et al., 2011), whose algorithmic lineage is LightGBM (Ke, Meng, Finley et al., 2017). Hyperparameters are held fixed across managers and across markets at values fixed before the cross-regime extension was attempted, so the non-United States results are not tuned. Per-manager fitting, rather than one pooled model with a manager fixed effect, follows from the same logic as the stratified metric: we do not want a shared parameter that can encode manager identity.

5.3 Split

Training uses a purged walk-forward split. For each scored quarter the training window ends strictly before the scored quarter begins. An embargo removes training rows whose period of report falls inside a

gap immediately preceding the test window, so that a label resolving inside the test period cannot appear in training through a neighbouring observation. This is the purging and embargo construction of López de Prado (2018), adapted to a quarterly filing calendar where the natural unit of leakage is a filing period rather than a bar.

The split is expanding rather than rolling: every scored quarter is predicted by a model that has seen all prior quarters and none of the current or later ones. No hyperparameter, no feature definition and no manager selection was revised after observing a held-out result, and the register in Appendix A records the runs in which each was fixed.

5.4 The evaluation estimator

Let S index the manager-quarter strata, and within stratum s let P_s be the set of positives and N_s the set of negatives. The pair-weighted stratified AUC is

$$\widehat{A} = \frac{\sum_{s \in S} \sum_{p \in P_s} \sum_{n \in N_s} \left[\mathbb{1}(s_p > s_n) + \frac{1}{2} \mathbb{1}(s_p = s_n) \right]}{\sum_{s \in S} |P_s| \cdot |N_s|}$$

Every comparison is made within a stratum, so the manager-identity channel contributes nothing. Strata are weighted by their pair count, so a manager-quarter offering four comparable observations does not carry the same weight as one offering four hundred.

This choice is consequential and we report the alternative. The same predictions, the same split and the same data scored by taking the mean AUC over quarters and then the median over managers yield 0.6589. The pair-weighted figure is 0.6762. The difference is not a modelling improvement; it is the correction of an estimator that gives a thin manager-quarter the same vote as a thick one. We report the pair-weighted figure as the headline and disclose the alternative here so that the comparison is available rather than buried.

5.5 Uncertainty

Intervals are bootstrapped over managers: each replicate resamples managers with replacement and takes all of a resampled manager's strata together. Correlated rows inside one filer are therefore never treated as independent evidence. This is the clustered-bootstrap logic of Cameron, Gelbach and Miller (2008). We do not use the DeLong asymptotic variance, because its independence assumptions are violated by the panel structure in a direction that would narrow the interval.

5.6 Controls

Three controls run alongside every headline number, and each has refused to pass at least one result that we would otherwise have published.

The shuffled-label control. The model is refit from scratch on labels permuted within each manager. Manager identity, base rate, universe and turnover survive the permutation intact; the association between a specific security and a specific addition does not. The gap between the real and permuted scores is the quantity we report. This control is what produces the Korean negative result in Section 6.2, and it is the reason we can state that the remaining markets are not measuring identity.

The quotability screen. A market figure is refused publication if it rests on too few comparable pairs, or if a single filer contributes more than half of them. The United Kingdom figure sits close to this boundary at 192 scored rows and is reported with that caveat attached wherever it appears.

The leakage canary. An independent artifact recomputes the within-manager score for each market under a deliberately degraded feature set and checks that the degradation moves the score in the expected direction and magnitude. A score that survives feature removal unchanged is a score that was not using the features, which is the signature of a leak.

---

6. Results: the next position is predictable

6.1 United States

Quantity	Value
Pair-weighted stratified AUC	**0.6762**
95% manager-clustered interval	0.6563 to 0.7002
Managers	49
Quarters	35
Held-out observations	124,731
Comparable pairs	1,244,011
Base rate of addition	12.46%
Largest manager's share of pairs	13.28%
Degenerate strata	45 of 1,482
Top-decile lift over base rate	3.17x
Mean-over-quarters, median-over-managers alternative	0.6589

The interval excludes 0.5 by a wide margin and excludes 0.65 at its lower bound. The top-decile lift is the quantity with an operational reading: restricting attention to the highest-scoring tenth of each manager-quarter's candidates raises the density of actual additions by a factor of 3.17.

An AUC of 0.6762 is a modest number and we present it as one. It says that if you draw one security the manager will add and one it will not, from the same manager in the same quarter, the model ranks them correctly 67.6% of the time. On a problem whose label is a discretionary human decision taken by a professional with private information, we regard that as the realistic ceiling of what public data supports, and we would treat a materially higher figure on this target as evidence of leakage rather than of skill.

6.2 Cross-regime replication

Each market is fit with the hyperparameters fixed on the United States panel and evaluated pooled, alongside a control refit on labels permuted within manager.

Regime	Scored rows	Pooled AUC	Shuffled control	Gap
United Kingdom	192	0.8424	0.5910	**+0.2514**
European Union	1,447	0.7975	0.5624	**+0.2351**
Hong Kong	7,688	0.8474	0.6304	**+0.2170**
Japan	3,464	0.8480	0.7067	**+0.1414**
Korea	16,074	0.8650	0.8831	**-0.0181**

Four of five regimes clear their own control by a wide margin. The absolute levels exceed the United States figure for the structural reason given in Section 3: a threshold-crossing regime discloses a sparse, deliberate set of events, so the candidate set is smaller and the positives are more distinctive. The comparison that carries information is each market against its own control, not one market against another.

Korea is a negative result. Korea has the largest scored sample in the non-United States set, 16,074 rows, and the highest pooled AUC, 0.8650, and it is worthless. Its own shuffled control reaches 0.8831, higher than the real fit. The entire pooled figure is manager identity: the Korean panel is dominated by a small number of filers with extreme and stable base rates, and knowing which filer a row belongs to is sufficient to reproduce the score. We quote no Korean figure in either pooled or within-manager form, and we report the market rather than dropping it, because a cross-market claim that silently excludes its failures is not a cross-market claim.

This result also corrected a prior error of our own, which we record because it bears on method. The prohibition it originally produced, that a pooled AUC should never be quoted in any market, was generalised from Korea alone. Testing the prohibition regime by regime showed it was correct in exactly the one market it was derived from, and wrong in four. A control that fails in one place is evidence about that place.

6.3 Within-manager estimates outside the United States

The pooled figure is the weaker question even where it passes its control. The stricter within-manager estimate is available for Japan, where the panel is deep enough to support it:

Quantity	Japan
Within-manager pair-weighted AUC	0.7224
95% manager-clustered interval	0.6181 to 0.8241
Shuffled control, mean of draws	0.4999
Managers	16
Comparable pairs	7,355

The Japanese within-manager figure, 0.7224, is close to the United States figure of 0.6762 once the identity channel is removed from both, and its control sits at 0.4999, which is what an exact negative control should read. We regard this as the single strongest piece of evidence in Section 6: the same method, the same hyperparameters, a different legal regime, a different language of filing, a different candidate universe, and the identity-free score lands in the same place.

The interval is wide because the panel holds 16 managers. It is reported wide rather than narrowed by treating rows as independent.

6.4 What the model is using

Mutual information on the training folds concentrates in the disclosed stake (0.11462 nats) and, with the next two features, accounts for 79% of the total. No feature carries effectively zero. The practical reading is that improvement on this target comes from new information, not from a better learner, which is the opposite of the usual conclusion in this literature and follows from the label being mechanical: there is no noise in the dependent variable for a better estimator to average away.

---

7. Results: the predictable trade is not a profitable one

This section is negative in five independent ways. We present it in full because the natural inference from Section 6, that knowing who buys next is worth money, is the inference we set out to confirm and could not.

7.1 A pre-specified family of portfolio transforms

The family below was written into the analysis script's docstring before the run, which fixes both the arms and the multiple-testing correction. Each arm enters at the date the filing actually becomes public, period end plus 45 days, holds 63 trading days, takes the top 15 names equal-weighted, and is measured as excess over SPY with the t-statistic clustered by quarter. The panel is 84,021 predictions over 35 quarters and 1,780 distinct tickers.

Arm	Side	Mean quarterly excess	t (clustered)	Quarters positive
raw top-K	long	-0.170%	-0.08	16 of 35
novelty only	long	-0.248%	-0.11	13 of 35
conviction weighted	long	+0.886%	+0.41	17 of 35
agreement, 3 or more managers	long	-0.605%	-0.33	17 of 35
contrarian, single manager	long	-1.471%	-1.00	17 of 35
short the crowded	short	+0.605%	+0.33	18 of 35
no model, whole frame	long	-0.214%	-0.29	16 of 35
stack minus taste	long	+0.895%	+0.51	18 of 35

The family-wise threshold for eight pre-specified arms is 2.57. The largest absolute t-statistic obtained is 1.00, on the contrarian arm, whose point estimate is negative. **No transform of the ranking clears its threshold.**

We draw attention to the `no model, whole frame` arm, which buys the whole prediction frame without using the model at all and returns -0.214%. The ranking arms are not distinguishable from it. Whatever the model knows about who will buy, it does not translate into a security selection that beats the alternative of not selecting.

7.2 The event study

Restricting to the cleanest possible version of the question, a disclosed addition is an event and we ask what the security does afterwards. On 3,936 disclosed additions, the 21-day excess return is 0.0511% with

a 95% interval of -0.8195% to 0.9216%. The interval contains zero and is narrow enough that a practically interesting effect would have been detected.

7.3 Decile monotonicity

If the ranking carried return information, the deciles of the score would be ordered by subsequent return. Across 82,949 priced held-out rows, decile monotonicity is 0.018. The deciles are not ordered. The model predicts what is bought, not what rises.

7.4 The quarterly book, reported in full

A different construction, and the one that produces the most favourable number in this paper, forms a 15-name equal-weighted book at each quarter end from the aggregate score, holds it one quarter, and charges round-trip transaction costs by liquidity tier at 3, 7, 20, 50 and 25 basis points one way for mega, large, mid, thin and unknown liquidity respectively. Over the same 35 quarters, from the quarter ending 30 September 2017 to the quarter ending 31 March 2026:

	Strategy, net of costs	Benchmark (SPY)
Cumulative return	**489.5%**	239.4%
Annualised	**22.48%**	14.99%
Quarters beaten	23 of 35	
Mean quarterly margin	2.22 points	
Annualised Sharpe of the edge series	0.444	

This is a real number produced by a named artifact and we report it as the headline of the subsection. We now report, with equal prominence, everything that is wrong with it.

The estimator panel. Five estimators of the central tendency of the 35 quarterly margins, each with a manager-free bootstrap interval:

Estimator	Full sample	95% interval	Dropping the single largest quarter	95% interval
Mean	2.217 pts (t = 1.313)	-0.674 to 5.860	0.884 pts (t = 0.828)	-1.179 to 2.938
Median	1.442 pts	-0.543 to 3.618	1.330 pts	-0.543 to 3.362
10% trimmed mean	1.169 pts	-1.026 to 3.592	0.904 pts	-1.255 to 2.946
Sign test	23 of 35, p = 0.0448		22 of 34, p = 0.0607	
Wilcoxon signed rank	p = 0.1006		p = 0.1506	

One of five estimators clears 0.05 on the full sample. None still clears it once the single largest quarter, the quarter ending 31 March 2026 at +47.5 points, is removed. Quoting the sign test alone would be estimator shopping, and we say so here rather than in a footnote. Removing the largest quarter leaves the book

compounding at 16.35% against a benchmark 13.57%, which is a positive margin that no estimator can distinguish from zero.

The timing caveat, which is the serious one. This book enters at quarter end. The 13F filing that reveals the quarter's holdings is not public until up to 45 days later. Section 7.1 is the same underlying signal entered at the date it actually becomes available, and its excess is -0.17%. The gap between 7.4 and 7.1 is the value of forty-five days of hindsight. We therefore treat Section 7.1 as the implementable result and Section 7.4 as an upper bound that no outside investor could have captured, and we would regard any presentation of 489.5% without this paragraph attached as misleading.

Power. At the observed effect of 2.217 points and the observed dispersion of 9.991 points, reaching $t = 2$ requires 81.2 quarters, that is 20.3 years. The sample provides 8.8 years. Alternatively the effect would have to be 3.378 points rather than 2.217, a shortfall of 52.3%, or the dispersion would have to fall to 6.558. This is a property of the instrument: 13F discloses four times a year, so the sample grows at four observations per year regardless of how much computation is applied to it. No modelling change alters this arithmetic, and a paper claiming significance on this series at this sample size is claiming something the disclosure schedule cannot support.

The served-universe variant. A second artifact computes the same book restricted to the universe actually served on the public product surface, which is narrower. It reads a mean net edge of 0.88 points with $t = 0.41$, an interval of -2.79 to 5.45, 18 of 35 quarters beaten and a 58.07% average within-quarter win rate. We report it because it is the version corresponding to what a user could actually have followed, and it is weaker than the full-universe version.

7.5 Capacity

Even taking the most favourable edge estimate at face value, the book does not scale. Using a square-root market-impact model of the Almgren, Thum, Hauptmann and Li (2005) form, with the impact coefficient swept over 0.5, 1.0 and 1.5, and the gross edge taken from the served backtest at 94.9 basis points per quarter, the fund size at which modelled impact consumes the edge entirely is:

Quantity (impact coefficient A = 0.5)	Value
Breakeven fund size, median quarter	$79,340,967
Breakeven fund size, worst quarter (2021-06-30)	$2,354,286
Breakeven fund size, best quarter (2017-09-30)	$460,592,204
10th-percentile quarter breakeven	$26,866,255
Equal-weight median breakeven	$79,340,967
Equal-weight worst breakeven	$15,886
Median name's average daily dollar volume	$43,485,744
Thinnest quarter's median daily dollar volume (2018-09-30)	$2,802,428
Picks carrying liquidity data	387 (top 15 per quarter, 35 quarters)

The worst-quarter figure is the one to read. A book sized to the median quarter is oversized in every quarter below it, and the strategy has no mechanism for declining to trade in a thin quarter. A capacity of a few million dollars in a bad quarter is not an institutional product. The coefficient sweep from 0.5 to 1.5

moves these figures but not the conclusion, because the square-root form makes capacity scale with the square of the tolerated impact and the edge is small enough that the tolerated impact is small.

7.6 How to reconcile Sections 6 and 7

The two halves of this paper are consistent, and we think the reconciliation is the most interesting thing in it.

The next addition is predictable because it is driven by slow, persistent, disclosed characteristics of the manager: what they have owned, what they own now, what their peers own, how they have behaved after similar states. Those characteristics are stable enough to learn and public enough to observe.

The return is absent because the same publicity that makes the prediction possible makes the trade worthless. Three mechanisms, each with support in the prior literature, are sufficient:

1. **Adverse selection in what is disclosed.** Agarwal, Jiang, Tang and Yang (2013) show that the holdings managers succeed in withholding outperform those they reveal. A predictable addition is by construction an ordinary one.
2. **Staleness.** By the time the filing is public the position is up to 45 days old, and Section 7.1 measures exactly what those 45 days cost: the entire margin.
3. **Cost.** Section 7.5 shows the capacity at which the most favourable estimate of the edge is consumed by impact is in the single-digit millions in a bad quarter.

The practical implication is that the value of a model on this target is informational rather than directional. It says where institutional demand is likely to appear. It does not say that following it pays.

---

8. Preregistration, timestamping and the forward record

The multiple-testing problem in this literature is severe, and the standard remedies (reporting a corrected threshold, as in Section 7.1) rely on the reader trusting that the family was specified in advance. We add an apparatus that does not require that trust.

The register. Every question is entered into a register with its verdict rule fixed before the run, and the run writes a machine-readable artifact carrying the verdict. The register currently holds 457 litigated hypotheses: 181 CONFIRMED, 117 REFUTED, 159 UNDERPOWERED, 5 DEFERRED, 13 PLANNED and 3 ERRORED. A refuted result is required to carry a counterargument naming the measurement that would overturn it. The refutation rate of 26% is itself a reported statistic, because a register with a low refutation rate is a register that was not asking hard questions.

Timestamped forecasts. Forward predictions are hashed and anchored before their outcomes exist, so that the forecast set cannot be revised after the fact. As of 18 September 2026 the record holds 60,084 receipt rows, which collapse to 85 distinct (market, manager, issuer) forecasts once the daily re-stamping of open forecasts with a rolling filing deadline is removed. We publish the distinct count, since the receipt count is thirty times larger and means nothing. 66 receipts carry a completed third-party attestation; 3,236 do not yet.

What has matured. The window is short and we report it as short.

Market	Matured	Open	Hits	Expected from the same filers' own base rate	Reading
Japan	22	2	4	0.5	Beats own base rate (p = 0.01057)
United Kingdom	18	0	0	0.0	Indistinguishable
Hong Kong	0	2	-	-	Window not elapsed
Korea	0	0	-	-	Not quoted, failed control

40 forecasts have matured in total, all of them in Japan and the United Kingdom. A matured forecast is one whose grading window has elapsed inside the panel; the expectation column is summed over the 21 Japanese rows whose filer held other positions before the stamp, since the base rate is undefined for a filer with no prior book. Before 18 September the artifact counted only rows that also had a cadence-matched control, which put 14 in the denominator against 3 hits; the reconciled figures are the ones above and the per-call ledger is public.

4 hits out of 22 against an expectation of half a hit is a striking ratio on a sample of 22, and 22 is a sample of 22. This table is reported so that it can be checked against, not because it settles anything. The stamped record becomes informative in 2027, and the correct thing to say about it today is that it exists and is checkable.

---

9. Discussion

9.1 A prediction problem that is not a pricing problem

The finance literature's default is that a predictable quantity in a liquid market is either an anomaly to be harvested or an artefact to be explained away. This paper's target is neither, and the reason is that the predicted object is not a price. Institutional demand has a predictable component; the security's return does not inherit it. That is a statement about who trades, and it can be true at the same time as the market being efficient with respect to the information content of the trade.

This has a methodological consequence worth stating plainly. A researcher who starts from a return target on this data faces a small sample, a noisy dependent variable and a large multiple-testing surface, which is the standard recipe for an unreplicable finding. A researcher who starts from the action target faces a mechanical label and a large sample, and can therefore measure something precisely. The cost is that the precise thing measured may not be worth money, which is what we find. We think that trade is worth making and reporting.

9.2 Who should care

Three groups, and the paper is useful to them for different reasons.

Issuers and investor-relations functions. Predicted institutional demand is a targeting variable for outreach, and the top-decile lift of 3.17x is directly operational for a function whose alternative is contacting the whole list.

Regulators and market-structure researchers. The cross-regime comparison is an observational experiment on disclosure design. Four threshold regimes produce sharply predictable filings; the one portfolio regime produces a weaker but much denser signal; and the regime whose filer base is most concentrated produces a figure that is entirely filer identity. Anyone arguing for a change in threshold, deadline or scope is arguing about the shape of this record.

Practitioners. Section 7 is the useful section, and its use is negative. Any product, newsletter or fund that promises a return from copying disclosed holdings is making a claim that we could not reproduce on five independent tests with a pre-specified family, and the burden of evidence sits with the claimant.

9.3 Why the negative half is published

A version of this paper containing only Section 6 would be more publishable, more commercially useful and less true. The register in Section 8 exists to make that choice costly: a refuted hypothesis must record its counterargument, so a result cannot be quietly dropped after it fails, and the artifacts carry verdicts that the document generator reads rather than a human transcribing. We regard this as the appropriate structural response to the incentives Bailey, Borwein, López de Prado and Zhu (2014) describe, and we note that it is a weaker remedy than preregistration at a third-party registry, which remains the obvious next step.

---

10. Limitations

Sample size is capped by the instrument. The binding constraint on the return question is 4 observations per year. This is stated as a power calculation in Section 7.4 rather than as a hope that more data will help: at the observed effect and dispersion, resolution requires 20.3 years.

The non-United States panels are small and censored. The United Kingdom figure rests on 192 scored rows and sits at the edge of our own quotability screen. Japan's within-manager estimate rests on 16 managers. All four threshold regimes are censored below the disclosure threshold, so nothing in this paper speaks to small or abandoned positions outside the United States.

Korea is unresolved rather than explained. We show that its pooled figure is identity, and we do not show what a correctly specified Korean model would read, because the panel does not currently support a within-manager estimate at acceptable width.

Manager selection is a choice. The 49 United States managers were selected for continuous filing history, which is a survivorship condition on the filer rather than on the security. A manager who stopped filing mid-window is absent, and such a manager is more likely to have closed than to have prospered. The direction of this bias is toward the panel being easier than the full filer population, and we have not quantified it.

Costs are modelled, not incurred. The transaction costs in Section 7.4 and the impact model in Section 7.5 are parameterised from the literature rather than from executed fills. The paper account referenced in the project record is young and is not used as evidence here.

The forward record is 32 matured forecasts. It is reported for checkability, not for inference.

---

11. Conclusion

The next position a disclosed institutional portfolio adds is predictable from that portfolio's own public history. Under an evaluation that makes the manager-identity shortcut arithmetically unavailable, the United States panel reads 0.6762 with a manager-clustered interval of 0.6563 to 0.7002 on 1,244,011 comparable pairs, and the top decile of the ranking is bought 3.17 times as often as the base rate. The result replicates in four of five foreign disclosure regimes against an exact negative control, and fails in the fifth, which we report.

The predictable component does not convert into return. A pre-specified family of eight portfolio transforms entered at the date of actual disclosure produces a largest absolute t-statistic of 1.00 against a family threshold of 2.57. A 21-day event study on 3,936 additions returns 0.0511% with an interval containing zero. Decile monotonicity across 82,949 priced rows is 0.018. The most favourable construction, a quarter-end 15-name book returning 489.5% against 239.4%, is not implementable at the disclosure date, is significant under one of five estimators, is significant under none after the largest quarter is removed, and would require 81 quarters to resolve.

Knowing who buys next is a solved problem to a useful approximation. Being paid for knowing it is not, and on this evidence it is not merely unsolved but unlikely.

AI disclosure

Parts of this manuscript were drafted with the assistance of a large language model. All numerical results are produced by the named artifacts in the project repository and were not generated by a language model. The author is responsible for the content, the method and the errors.

Appendix A: the hypothesis register

Each entry carries a question, a verdict rule fixed before the run, the artifact path the run writes, and the verdict. Verdicts are one of CONFIRMED, REFUTED, UNDERPOWERED, DEFERRED, PLANNED or ERRORED. The UNDERPOWERED verdict exists specifically so that a question that the data cannot answer at acceptable width is recorded as unanswered rather than as answered negatively, which is the error this paper's Section 7.4 is most at risk of committing.

Verdict	Count
CONFIRMED	173
REFUTED	110
UNDERPOWERED	123
DEFERRED	5
PLANNED	14
ERRORED	3
Total litigated	**406**

Appendix B: estimator definitions

Pair-weighted stratified AUC. As defined in Section 5.4. Ties contribute one half. Strata with no positive or no negative contribute zero pairs and are excluded from both numerator and denominator; 45 of 1,482 United States strata are of this kind.

Manager-clustered bootstrap. Managers are resampled with replacement; all strata belonging to a resampled manager are taken together. The reported interval is the 2.5th and 97.5th percentiles of the replicate distribution.

Shuffled-within-manager control. For each manager, the label vector is permuted uniformly at random across that manager's rows, holding the feature matrix fixed, and the entire fitting pipeline is rerun. Multiple draws are taken and the mean reported; individual draws are retained in the artifact.

Family-wise threshold. For the eight pre-specified arms of Section 7.1 the threshold is the value stated in the analysis script before the run, 2.57.

Decile monotonicity. The fraction of adjacent decile pairs ordered correctly by mean subsequent return, rescaled to the interval from -1 to 1, so that 0 indicates no ordering.

Appendix C: reproduction

Every figure in this paper is emitted by an artifact rather than transcribed by hand, and the document generator reads the artifacts. The relevant paths are:

Result	Artifact
United States AUC, pairs, strata, interval	`runs/gates/us13f_pair_weighted_auc.json`
Cross-surface claim facts	`runs/gates/claim_facts.json`
Top-decile lift over base rate	`runs/gates/q23_calibration_crossfit.json`
Decile monotonicity, priced rows	`runs/gates/E3b_auc_to_return_clean.json`
Event study, mutual information	`runs/hypotheses/INDEX.json` (H005, H004)
Cross-regime pooled and shuffled control	`runs/gates/pooled_vs_shuffled.json`
Within-manager, all markets	`runs/gates/market_leakage_canary.json`
Pre-specified portfolio family	`runs/gates/A1_rank_to_return.json`
Estimator panel and power	`runs/gates/A1_return_power.json`
Quarterly book series	`runs/per_investor_wf/priorscore_backtest.parquet`
Served-universe variant	`runs/per_investor_wf/priorscore_backtest_summary_served.json`
Capacity	`runs/gates/A3_capacity.json`
Forward record	`runs/gates/R1_stamped_call_grading.json`
Hypothesis register	`runs/hypotheses/INDEX.json`

References

1. Agarwal, V., Jiang, W., Tang, Y. and Yang, B. (2013). Uncovering hedge fund skill from the portfolio holdings they hide. *Journal of Finance* 68(2), 739-783.
2. Almgren, R., Thum, C., Hauptmann, E. and Li, H. (2005). Direct estimation of equity market impact. *Risk* 18(7), 58-62.
3. Aragon, G., Hertzel, M. and Shi, Z. (2013). Why do hedge funds avoid disclosure? Evidence from confidential 13F filings. *Journal of Financial and Quantitative Analysis* 48(5), 1499-1518.
4. Bailey, D., Borwein, J., López de Prado, M. and Zhu, Q. (2014). Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society* 61(5), 458-471.
5. Cameron, A. C., Gelbach, J. and Miller, D. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90(3), 414-427.
6. Cohen, R., Polk, C. and Silli, B. (2010). Best ideas. Working paper.
7. DeLong, E., DeLong, D. and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics* 44(3), 837-845.
8. European Union (2004). Directive 2004/109/EC on the harmonisation of transparency requirements (Transparency Directive), major holdings notification thresholds.
9. Financial Conduct Authority. *Disclosure Guidance and Transparency Rules*, DTR 5: Vote Holder and Issuer Notification Rules.
10. Frank, M., Poterba, J., Shackelford, D. and Shoven, J. (2004). Copycat funds: information disclosure regulation and the returns to active management in the mutual fund industry. *Journal of Law and Economics* 47(2), 515-541.
11. Freyberger, J., Neuhierl, A. and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies* 33(5), 2326-2377.
12. Griffin, J. and Xu, J. (2009). How smart are the smart guys? A unique view from hedge fund stock holdings. *Review of Financial Studies* 22(7), 2531-2570.
13. Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5), 2223-2273.
14. Hanley, J. and McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1), 29-36.
15. Harvey, C., Liu, Y. and Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies* 29(1), 5-68.
16. Hong Kong. *Securities and Futures Ordinance*, Part XV, Disclosure of Interests.
17. Japan. *Financial Instruments and Exchange Act*, large shareholding reporting (the 5% rule), filed through EDINET.
18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30.
19. Korea. *Financial Investment Services and Capital Markets Act*, Article 147, large shareholding reporting, filed through DART.

20. López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.
21. Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.
22. Securities and Exchange Commission. Form 13F reporting requirements, 17 CFR 240.13f-1.
23. Shi, Z. (2017). The impact of portfolio disclosure on hedge fund performance. *Journal of Financial Economics* 126(1), 36-53.
24. Sias, R., Turtle, H. and Zykaj, B. (2016). Hedge fund crowds and mispricing. *Management Science* 62(3), 764-784.
25. Wermers, R. (2000). Mutual fund performance: an empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance* 55(4), 1655-1695.

Full artifact list and reproduction commands: priormoves.com/record/