Sophon-3: Continual-Learning Deep Models for Non-Stationary Financial Markets

Hans Dahlström Adam Tittenberger Ted Björling
Finforge Research

Public technical research manuscript — Version 1.0 (first public release), July 2026

Abstract

Financial markets are non-stationary: predictive relationships decay, investable universes change, and research backtests diverge from live execution. We argue that financial AI for such markets should be organized as a *causal graph of continually learning decision nodes* rather than as a single static predictor, and we introduce **Sophon-3**, Finforge's family of continual-learning models. Sophon-3 is a node-graph *substrate*: nodes are causal operators composable into *arbitrary directed acyclic topologies* — sequential chains of any length, many-to-one, and one-to-many — of which the chain evaluated here, *Asset Warehouse* → *Screener* → *Forecast* → *Make Portfolio*, is the canonical instantiation. We formalize the substrate with filtration-based admissibility — proved per node for any such graph — and define evaluation as a *frozen procedure with adaptive state*, evaluated out-of-sample against a dated architecture freeze.

Our central finding is that most of the value measured under this frozen protocol comes from the graph's composition. We evaluate three live agents (deployed instances of the graph; two trade the Nasdaq-100, one the S&P 500) by comparing each agent with an ablated variant of itself through the paired daily difference of their active returns (portfolio return minus benchmark return), and we pool the three difference series before inference. Removing the Screener costs a pooled 45 percentage points of active return per year (pp/yr; $t = 4.6$, significant in each agent individually), and removing the forecast-magnitude position sizing costs a pooled 36 pp/yr ($t = 3.8$, significant in each agent): the deep component's demonstrated contribution is cross-sectional bet-sizing. Adaptation matters where selection happens: re-searching the Screener's rules monthly adds a pooled 23 pp/yr over rules fixed once at inception ($t = 3.4$). Two adaptive mechanisms show no measurable value at current sample size, and we report this directly: retraining the forecaster continually is statistically indistinguishable from a single fixed checkpoint (+4 pp/yr, 95% CI $[-5, +13]$), and adaptive ensemble weighting is indistinguishable from an equal average of members (-0.0 ± 0.6 pp/yr). These paired, design-frozen results are the paper's primary evidence; no single live result is.

As out-of-sample corroboration, a single real-money brokerage account returned an active return of approximately +12.9% against a total-return Nasdaq-100 benchmark over the design-frozen window of roughly 41 trading days (27 April – 24 June 2026). We state plainly that a window this short is not statistically conclusive on its own; its role is real-fill evidence that the system operates as designed, not a verdict on edge. The architectural claim — that nodes compose into arbitrary causal graphs whose components adapt at different time scales — is general; the empirical evidence in this paper is for the canonical four-node chain, and we scope it accordingly.

Scope, disclosure, and version. This is the first of three companion papers, covering the *thesis*, the leakage-free formalism, the *ablations*, and the live proof-of-concept; companion papers cover deterministic live-vs-backtest *parity* and screener configuration with the pre-registered *selection* of the live agents. Implementation-sensitive details (data-vendor identifiers, deployed member identities, production rule paths) are abstracted; execution venues are named, since broker identity is publicly verifiable and not a source of edge. This is **Version 1.0**, the first public release (July 2026); a revision scheduled for August 2026 adds the diagnostics enumerated in §17. Every figure is computed from frozen internal runs, never estimated; the live figures in §13 are real, single-account brokerage results, stated exactly.

1 Introduction

The central problem in systematic investing is not prediction but prediction under changing data-generating mechanisms, incomplete observability, market feedback, operational constraints, and delayed evidence about whether a signal was useful. A strategy that performs well in a static backtest can fail when the regime changes, the eligible universe drifts, costs rise, or live execution diverges from research.

Finforge's thesis is that financial AI should be modeled as a causal, continually learning decision graph rather than a frozen predictor or a static rule. By a *frozen predictor* we mean a model whose parameters are fixed after an estimation period and then deployed unchanged; by a *static rule*, a fixed heuristic with constant thresholds — both common in quantitative practice. Periodically refit models sit between these poles; the claim is not that refitting is unknown, but that adaptation should be systematic across every component of the decision graph and auditable under a frozen evaluation procedure (§10). Universe selection, neural forecasts, ensemble weights, and portfolio sizing all adapt, while the graph stays auditable under point-in-time data. The canonical chain

$$\text{Asset Warehouse} \rightarrow \text{Screener} \rightarrow \text{Forecast} \rightarrow \text{Make Portfolio} \tag{1}$$

expresses the natural workflow: define what is eligible, reduce the universe, then spend forecasting compute on a smaller candidate set. The motivating hypothesis is deliberately testable rather than universal: causal continual-learning graphs are a more appropriate architecture for non-stationary markets than once-trained models or fixed rules. This is aligned with the Adaptive Markets Hypothesis [6] and with concept-drift work in which the feature-to-target mapping changes over time [2].

Contributions.

1. A formal, filtration-based definition of leakage-free continual learning for financial decision graphs, with evaluation defined as freezing the *procedure* rather than the model state.
2. A causal adaptive forecast ensemble whose weights depend only on past realized losses (no-future-bleed), and an ablation suite that isolates the contribution of screening, neural forecasting, and adaptive weighting.
3. An evidence framework that dates the architecture freeze and separates results by provenance, evaluated at the final tradable-portfolio level and including a real-money out-of-sample window.

Scope of claims. This paper does not assert that any model beats all markets in all regimes, that positive backtests prove future profitability, or that continual learning eliminates market risk. The claim is architectural and methodological, and the live track record is explicitly short and not yet statistically conclusive (§14).

1.1 Claim hierarchy and evidentiary status

Because claims of different strength coexist in this paper, we state their status up front. Throughout, an *agent* (equivalently, a *flow*) is one deployed instance of the Sophon-3 graph: a configured strategy trading a single benchmark universe; the four live agents are introduced in §11. An *ablation* removes or freezes one component of an agent; effects are paired daily differences of active return (agent minus variant), and *pooled* means the three agents' difference series are averaged before inference. Methods are specified in §11; evidence tiers in §10.

Table 1: Claim hierarchy and current evidentiary status (Version 1.0). A reading guide; the statistics behind each row are defined and reported in §11.

Claim	Status	Evidence	Not yet established
Leakage-safe adaptive graph	Strong	Filtration admissibility and a frozen evaluation procedure (§4, §10)	External reproducibility rests on the companion parity paper
Screener adds value	Strongest empirical	Remove-Screener ablation: 45 pp/yr pooled, significant in all three agents (§12)	Full search-space multiplicity ledger (companion selection paper)
Forecast magnitude adds sizing value	Strong	Forced equal-weight sizing ablation: 36 pp/yr pooled, significant in all three agents (§12)	Mechanism not yet isolated; may proxy volatility/dispersion (§9)
Continual Screener adaptation adds value	Moderate	Static-Screener ablation vs. rules fixed once at inception: +23 pp/yr pooled (§12)	Short design-frozen confirmation
Continual forecast retraining adds alpha	Not yet proven	Static-Forecast ablation: 95% CI spans zero, $[-5, +13]$ pp/yr (§12)	Requires a longer clean forward sample or larger panel
Adaptive ensemble weighting adds return	Null so far	Equal-weight ensemble ablation, -0.0 ± 0.6 pp/yr (§12)	Retained as regime insurance; leak diagnostic pending

2 Technical Thesis and Differentiators

Thesis. Markets are adaptive systems; financial AI should therefore be a causal, continually learning decision graph whose components adapt at different time scales while preserving point-in-time correctness.

Table 2: Core differentiators of Sophon-3.

Differentiator	Mechanism	Why it matters
Composable node graph	Causal operators composable into any DAG (chains, fan-in, fan-out); the four-node chain is one instantiation	Expresses many workflows without rewriting the system
Screener-first adaptation	Rolling explore/exploit search over interpretable rule families	Selection itself learns continually; monthly rule re-search is the largest measured adaptive contribution (+23 pp/yr, §11)
Walk-forward forecasting	Repeated train/predict chunks with strict validation	Treats changing regimes as the default case
No-future-bleed ensembles	Weights resolved strictly before current losses are observed	Prevents same-period loss leakage in adaptive weighting

3 The Sophon-3 Node Graph

Each node is a causal operator producing a timestamped output and an updated internal state, and nodes compose into *any causal directed acyclic graph*: arbitrary-length sequential chains, many-to-one fan-in (a node consuming several parents), and one-to-many fan-out (a node feeding several children). The four-node chain below is the canonical instantiation evaluated in this paper; it matches the natural investment workflow — define eligible assets, reduce the universe, estimate timing and sizing, then construct the portfolio — but the substrate is not limited to it.

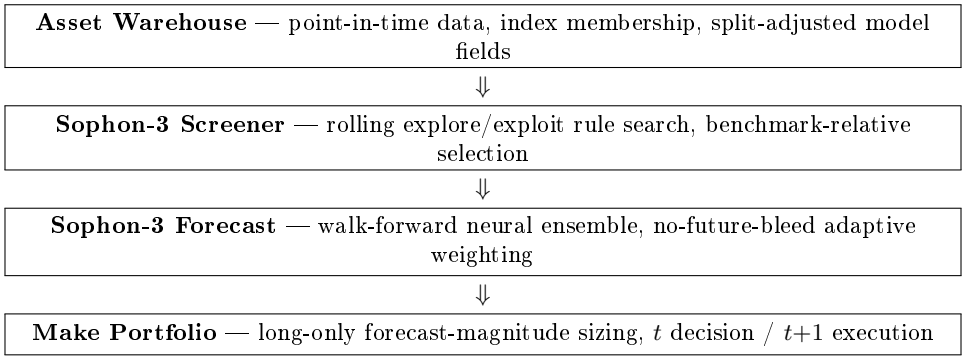

Figure 1: Canonical Sophon-3 graph. The Asset Warehouse emits eligible symbols and point-in-time features; the Screener a timestamped selection map; the Forecast node ensemble rows, weights, and realized diagnostics; Make Portfolio target weights, trades, and NAV.

All nodes share an operational substrate (run-until caps, deterministic seeds, checkpointed state, idempotent writes). A live stop-and-go run is a segmented execution of the same ordered operator sequence as the one-shot historical replay. The parity invariant, in one line: fixing the data snapshot, the frozen procedure Π, deterministic seeds, and a checkpoint surface sufficient to reconstruct node state, the one-shot replay and the segmented live run produce identical decision-relevant artifacts (selections, forecasts, weights, portfolio rows) up to declared tolerances; the proof, the artifact surface, and the regression suite that enforces it are the subject of the companion parity paper.

4 Formal Problem Statement

Let U_t be the investable universe at decision time t, and $\mathcal{F}_t$ the sigma-field generated by all information observable no later than t: closed price bars, available fundamentals, published news, index membership, corporate actions, upstream node outputs, and persisted graph state. A forecast is admissible if $\hat{r}_{i,t,h} = f(x_{i,\le t}, \theta_t) \in \mathcal{F}_t$; a selection set if $S_t \subseteq U_t,\ S_t \in \mathcal{F}_t$; a portfolio if $w_t = g(S_t, \hat{r}_t, c_t, \Gamma) \in \mathcal{F}_t$.

Node-graph admissibility. For a DAG $G = (V, E)$ with each v a causal operator and $\mathrm{pa}(v)$ its parents,

$$Y_t^v = f_v\Big(Y_{\le t}^{\mathrm{pa}(v)}, X_{\le t}, \theta_t^v\Big), \qquad \theta_{t+1}^v = A_v\Big(\theta_t^v, Y_{\le t}^{\mathrm{pa}(v)}, X_{\le t}, Y_t^v\Big). \tag{2}$$

The graph is admissible if $Y_t^v \in \mathcal{F}_t$ for every node and every timestamp — the mathematical form of *no lookahead*. Because $\mathrm{pa}(v)$ ranges over a node's parents, this condition is stated for an arbitrary causal DAG: it already covers many-to-one fan-in and one-to-many fan-out, and the admissibility result does not depend on the specific chain. The only structural requirement is that the graph remain a causal DAG — feedback across timesteps is admissible (a forecast at t may feed selection at $t+1$), but a same-timestamp cycle, in which information at t would depend on itself, is not.

Frozen procedure, adaptive state. Sophon-3 does not freeze parameters during evaluation; it freezes the procedure $\Pi = (G, D, U_t, B, C, H, M, \Omega, S)$ — graph, data contract, universe rule, benchmark, portfolio/cost constraints, horizons, metrics, search spaces, and seed/state rules. Node states θ_t^v evolve only through the update functions specified by Π. The dated freeze of Π (§10) is what makes a clean out-of-sample window possible.

Objective. The objective is benchmark-relative. With $r_{p,t}$ the net portfolio return after modeled costs and $r_{b,t}$ the benchmark return on the same dividend basis (§5.1), active return is $a_t = r_{p,t} - r_{b,t}$, and we report risk-adjusted active metrics (information ratio, active drawdown) and forecast diagnostics: the rank information coefficient (rank IC) and NDCG@K, a top-K ranking-quality score, both used operationally in §11.

5 Data and Universe Protocol

Data contract. Every node reads a timestamped contract: a record enters $\mathcal{F}_t$ only if its observable timestamp is $\leq t$ — a fundamental at its filing/release time, news at publication, a price bar only after it closes. Timestamp semantics are part of Π and stored with each run manifest.

Point-in-time index membership. Eligible assets at t are restricted to constituents of the index at t: $U_t^I = \{i : i \in I_t\}$. This prevents a backtest from selecting names that were not index members at the decision time, and it materially distinguishes Sophon-3 from survivorship-biased backtests.

Prices. The warehouse stores raw and split-adjusted fields; model-facing features use forward split-adjusted prices to avoid discontinuities, while display maps back to familiar prices.

5.1 Benchmark and return convention (symmetric within each domain)

The benchmark convention is the most common source of illusory active return in long-only equity research, so we are explicit.

Basis. Active return is computed on a *symmetric* dividend basis: portfolio and benchmark share the same treatment within each evaluation domain. Simulated backtests do not credit dividends to the portfolio, so they are compared to the *price* index (ex-dividend on both sides); the live brokerage account does receive dividends, so its NAV is total-return and is compared to the *total-return* index (§13). Symmetry, not a particular index, is the requirement.

Cross-basis comparability. A price/price active return and a total-return/total-return one differ only by $(d_p - d_b)$, the held book's dividend yield minus the index yield. For the growth-tilted books evaluated here, whose yield is at or below the Nasdaq-100's ∼0.7–0.9%/yr, price-basis figures overstate active return by at most that gap — bounded, and immaterial next to the double-digit effects of §11; the holdings-weighted differential is measured in the August 2026 revision. Ablation deltas are unaffected on any basis: the benchmark cancels in the paired difference.

6 Adaptive Universe Selection: Screener

The Screener produces a timestamped selection map $S_t = \{i \in U_t : \phi_{\gamma_t}(x_{i,\leq t}) = 1\}$. It is not a return forecaster but a continual-learning selection node. On rolling *fully historical* windows $E_j = [a_j, b_j]$, a guided search (sequential model-based optimization over a constrained, interpretable rule space) proposes filter combinations under a risk-adjusted active-return utility with breadth and sector/concentration guardrails. Guided search is a deliberate choice over the two

obvious alternatives: an exhaustive grid is intractable and turns every cell into a multiple-testing trial, while folding selection into the neural network end-to-end would make it unauditable — keeping the search guided and the rules interpretable and separable both shrinks the multiplicity that the correction in §14 must absorb and lets the ablations isolate the Screener's contribution. Selection is then strictly forward: for a decision time t the Screener uses the latest window with $b_j < t$ and chooses the combination whose *forecasted* next-period utility is highest, so the rule applied at t depends only on information observable before t and never re-uses the best in-sample combination. Selecting the top *in-sample* utility combination would be a winner's-curse; forecasting next-period utility instead trades in-sample fit for out-of-sample honesty, a discipline whose payoff is measured by the static-Screener ablation (§11) rather than asserted here. The utility predictor in this exploit step is a deep forecaster run inference-only from its own pre-trained checkpoint (trained on data through 31 December 2022, and distinct from the Static-Forecast baseline checkpoint), so the Screener's continual learning is the monthly rule re-search over a fixed predictor — no per-window training and no additional training-time leakage surface, which attributes the Static-Screener delta (§11) to re-search alone. Because the exploit layer is itself a predictive model, its accuracy bounds its value; the information coefficient of forecasted vs. realized next-period combo utility is reported as a first-class diagnostic in the companion selection paper. The configuration-level search that produced the live agents — which rule families and data fields each agent screens on, the search budget, and the pre-registered choice — is the subject of the companion selection paper. Within this paper, the Screener's selection means its monthly re-choice of a rule combination, and the node is treated at the level the ablations require: an adaptive selector that reduces the universe before forecasting.

7 Walk-Forward Neural Forecasting

The Forecast node estimates forward signals for the incoming (Screener-selected) universe. With M heterogeneous deep members — transformer-style, interpolation/decomposition-style, and recurrent/temporal-convolutional — member m emits $\hat{r}^{(m)}_{i,t,h}$ and the ensemble is $\hat{r}^{\text{ens}}_{i,t,h} = \sum_m w_{m,t}\hat{r}^{(m)}_{i,t,h}$.

Walk-forward loop. At cursor t, each member trains on a rolling window ending at t and predicts the next chunk: $[t-W, t] \to$ train, $[t+1, t+h] \to$ predict. The default horizon is one step at the configured cadence (day or week), and the rebalance cadence is induced by that horizon — a daily forecast yields daily portfolio decisions, a weekly forecast weekly ones — so cadence is a stated property of the graph rather than a hidden parameter, and turnover, capacity, and the multiplicity count are defined relative to it. Run-until and last-closed-bar caps ensure no fit or prediction reads beyond the decision horizon. Whether continual retraining adds tradable value over a strong frozen checkpoint is an empirical question, not an assumption; the Static-Forecast ablation (§11) addresses it directly. Determinism across the batched and live-resume paths is enforced at the seed, member, and execution level and is detailed in the companion parity paper.

Strict validation. The node requires a nonzero validation slice for any scorable training window. Silent fallback to zero validation is forbidden because it disables early stopping and produces deterministic overfitting — a leading cause of strong-backtest, dead-live outcomes. Sparse assets are pruned rather than truncating the global window, and configurations that cannot support a valid validation slice raise a run-level error rather than training unvalidated.

8 No-Future-Bleed Adaptive Ensembles

For portfolio construction, cross-sectional ordering matters more than point calibration, so the member loss is a rank loss $\ell_{m,t} = 1 - \text{NDCG@}K_{m,t}$ computed over the eligible universe. The

weighting scheme is an exponentially-weighted-forecaster (Hedge-style) rule from online learning [1]: softmax weights over negative exponentially weighted moving-average (EWMA) losses, chosen against the standard alternatives — equal-weight, a single best member, or a static blend — because member skill is regime-dependent under non-stationarity. The ablation verdict is reported in §11: realized adaptive weights track an equal member average so closely that the performance difference is a precise null (-0.0 ± 0.6 pp/yr pooled), so adaptivity is retained as regime insurance at approximately zero measured cost, and the *demonstrated* property of this node is the no-future-bleed discipline below, whose value is quantified by a leak diagnostic scheduled for the August 2026 revision. Each member maintains an EWMA loss state $L_{m,t}$ from which a softmax weight proposal is formed and then smoothed; the production profile recalibrates on a 21-bar cadence with a 126-bar lookback, hard-coded so there is no configuration path to a leak-prone policy.

No-future-bleed order. The essential convention is that the weight at t uses only losses strictly before t:

$$w_{m,t} \text{ is computed from } L_{m,t-1}, \text{ never from } L_{m,t}. \tag{3}$$

At each timestamp the node (1) resolves weights from history before t, (2) emits the forecast for t, and (3) only then updates loss state from the realized outcome at t. Thus realized outcomes at t affect $w_{m,t+1}$ or later, never $w_{m,t}$ — and this holds identically in one-shot and live-resume execution. Per-member leave-one-out health scores (does removing the member help or hurt the ensemble?) provide a model-by-model audit rather than a black box.

9 From Signals to Portfolios

Make Portfolio converts the selected universe and forecast magnitudes into long-only weights. The book is unlevered and there is no portfolio-level optimizer,[1] a direct consequence being that generated portfolios are always fully invested (cash weight $\leq 0.1\%$). Sizing is by absolute forecast magnitude: the raw size is $m_{i,t} = |\hat{\pi}_{i,t}|$ and the long-only weight is $w_{i,t} = m_{i,t} / \sum_{j \in S_t} m_{j,t}$, so a -5% and a $+5\%$ forecast both contribute a 5% weight — the forecast's *magnitude* drives position size while its *sign* is not used directionally. If all magnitudes collapse the portfolio falls back to equal weighting over the eligible set. This is an empirical choice, quantified by the forced equal-weight sizing ablation (§11): forcing the same agent's weights to equal shows magnitude sizing adds $+23$ to $+43$ pp/yr of active return across the three panel agents ($+36$ pp/yr pooled, $t = 3.8$; significant in all three) — the *sizing effect is confirmed*. Positive-only sizing (keeping only long-signed forecasts) is mixed — it degrades the outcome in two of three flows (significantly in one, by 18 pp/yr) and improves it in one — so the absolute-magnitude default is retained on net evidence rather than uniform superiority. What remains a *hypothesis* is the mechanism: a member trained on returns emits large-magnitude forecasts for high-dispersion names regardless of sign, so $|\hat{\pi}_{i,t}|$ plausibly proxies volatility among the momentum-screened survivors rather than conviction about direction. Two diagnostics, reported separately rather than assumed here, would confirm or refute the mechanism: whether $|\hat{\pi}_{i,t}|$ tracks realized volatility/dispersion, and whether magnitude sizing beats sizing by a naive volatility proxy on the same selected set — the test of whether the neural magnitude carries sizing information a simple estimator does not. Stated as a program, the Forecast node's contribution decomposes into four tests: *directional* (does sign predict relative return?), *ranking* (does forecast rank predict cross-sectional return rank?), *magnitude* (does $|\hat{\pi}|$ predict dispersion or volatility?), and *sizing* (does neural magnitude beat naive proxies on the same set?). This release establishes the sizing test; the ranking diagnostics and the directional and magnitude tests are scheduled for the August 2026 revision

[1] No mean–variance or covariance step, no volatility targeting or stop-loss logic, and no cash-buffer management; weights follow directly from the forecasts.

(§17). A decision made at the close of session t is implemented on session $t{+}1$ by a market order placed approximately five minutes before the official close, in highly liquid index constituents; end-of-day reference data come from a commercial market-data provider. Target quantities follow from the intended weights, $q^{\text{target}}_{i,t+1} = w_{i,t}V_t/p_{i,t+1}$.

Execution objective: weight tracking, not a price target. Live execution targets the model's intended portfolio *weights*, not a dollar or price target: the objective is to make realized post-trade weights equal the intended weights and to minimize *weight drift*. For a weight-defined portfolio this is the first-order execution-quality metric, and post-trade reconciliation on the live account targets exactly this quantity; a quantified drift study over the account's full reconciliation history accompanies the August 2026 revision (§17).

Slippage and execution timing. We do not claim price slippage is zero, and targeting weights does not by itself reduce it: slippage is a property of the trades, and two strategies that produce the same weight path incur the same execution cost however those weights were derived. Slippage is small here because of the *trade profile* — per-asset daily rebalances are small, the names are among the most liquid U.S. large-caps, and order sizes are retail-scale — so cost per day is bounded by turnover × per-trade price error.[2] What weight-denomination buys is *comparability*: backtest and live are measured on the same object — target weights held over close-to-close returns — so execution frictions enter the live-vs-backtest gap only through that bounded turnover term and one *timing basis*: fills placed ≈5 minutes before the official close strike realized returns and the close-to-close benchmark a few minutes apart, small and roughly unbiased for mega-caps, and named separately from slippage so it is not double-counted. *Explicit commission* is a documented venue property: the base profile is Alpaca, commission-free on U.S. equities since 2018 and used for both backtest and live; the system also simulates Interactive Brokers' commissioned U.S. schedule, so broker choice is explicit and the edge can be tested under nonzero fees (§12). All of this holds at retail size and re-enters as market impact as participation rises (§12); because the live account trades real fills, the live result in §13 already carries the real execution cost incurred over the window.
Live and backtest execution produce identical decision artifacts under the conditions established in the companion parity paper, on which the live result below relies.

10 Evidence Framework: Provenance Tiers

No performance figure in this paper silently spans evidence of different quality. Every result is anchored to two dated events and classified into one of four evidence tiers; the timeline and the relative-wealth paths below show where each region lies.

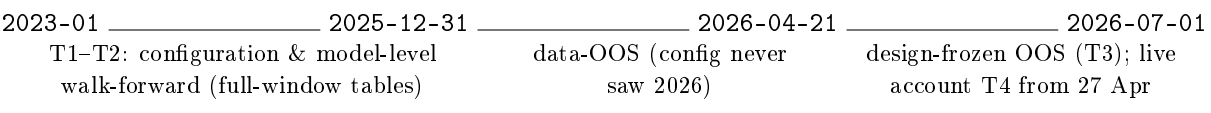

Figure 2: Evidence timeline. Full-window results span all three regions and are labeled T1–T2; only figures right of 21 April 2026 are design-frozen.

Configuration boundary (31 December 2025). Node configuration and tuning used data only through 2025, so all of 2026 is out-of-sample at the *data* level: nothing about the system was chosen with knowledge of 2026.

[2] Illustrative, not measured: with 20% one-way daily turnover and a 2 bp average half-spread-plus-impact on mega-cap names, daily cost $\approx 0.20 \times 0.02\% = 0.004\%$, i.e. roughly 1 pp/yr — an order of magnitude below the smallest pooled ablation effect of §11. The measured turnover and execution table accompanies the August 2026 revision (§17).

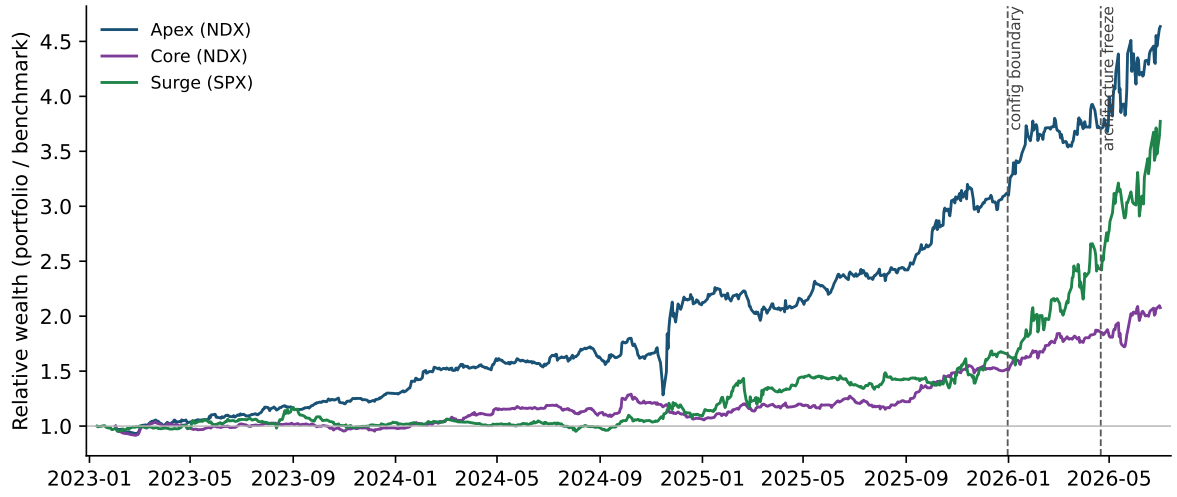

Figure 3: Relative wealth (portfolio NAV / benchmark index, price basis, normalized at inception) for the three live forecast-bearing agents over the full window. Provenance: T1–T2 left of the freeze line; the design-frozen (T3) slice lies right of 21 April 2026. A rising line means the agent is outperforming its benchmark; overlapping-window caveats of §14 apply to any rolling statistic read from these paths.

Architecture freeze (21 April 2026). The evaluation procedure Π (§4) was fixed on this date, so everything afterward is additionally out-of-sample at the *design* level. The freeze is anchored by configuration and checkpoint manifests timestamped at or before 21 April 2026; a public, immutable hash commitment referencing those manifests accompanies the August 2026 revision, making the freeze externally provable rather than asserted.

Four tiers. Each tier removes one way a result could still have been fitted. **T1** is the in-sample backtest through 2025: model-level walk-forward discipline applies, but both data and design were available when the system was configured. **T2** is model-level walk-forward evidence that lies past the data boundary but before the design freeze: no model saw its own future, but the architecture could in principle still have been shaped by this period. **T3** is the design-frozen window after 21 April 2026: neither data nor design could have been fitted to it. **T4** adds real capital and real fills on top of T3. Table 3 summarizes what each tier does and does not control for.

Table 3: Evidence-provenance tiers. No reported figure spans tiers without an explicit seam label.

	Tier	**Controls for**	**Does *not* control for**
T1	In-sample backtest (≤2025)	—	leakage, model and design selection
T2	Walk-forward OOS (model-level)	per-step model leakage	design selection
T3	Design-frozen forward OOS (≥21 Apr 2026)	data *and* design selection	real frictions; short window
T4	Real-money live (≥27 Apr 2026)	all the above + real fills	capacity (single small account)

Reading rules. A 126-bar rolling information ratio observed on 24 June 2026 reaches back to roughly December 2025 — before the freeze — and is therefore a **T2** object, not a design-frozen one. We report it as such and do not conflate it with the post-freeze (T3) window. Two distinct post-freeze windows appear in this paper and are labeled separately: the ablation evaluation

slice (21 April – 1 July 2026, ≈49 trading days) and the real-money account window (27 April – 24 June 2026, ≈41 trading days).

11 Evaluation Protocol and Ablations

What carries the scientific claim. The real-money record (§13, reported after these results) shows the *system* operates; it does not isolate *why*. The claim that continual learning and the graph structure are what help is carried by the ablation suite, which is backtest-computable under the frozen protocol. We state this so the live curve is not over-read.

Baselines. All baselines are Finforge node graphs sharing the same warehouse, benchmark convention, run caps, and Make Portfolio logic: benchmark-only; equal-weight universe (Assets → Make Portfolio); Screener-only (Assets → Screener → Make Portfolio); Forecast-only (Assets → Forecast → Make Portfolio); and full Sophon-3.

Panel and methodology. Four agents trade live, positioned by risk class (high/medium/low). Three contain a Forecast node — Sophon Apex (NDX), Sophon Core (NDX), and Sophon Surge (SPX) — and these constitute the ablation panel; the fourth is a screener-only graph, and is therefore itself a live, running instance of the Screener-only baseline above. The wider internal forward-OOS set from which the four were drawn is the subject of the companion selection paper. Each panel agent is a full walk-forward run from early 2023 through 1 July 2026. Every ablation is compared to its own agent as a *paired daily difference series* of active returns (agent minus ablated variant, common dates), never as two independent estimates. Full-window results are the annualized mean of the difference with a 21-day moving-block bootstrap 95% CI; a panel estimate averages the three flows' daily differences before bootstrapping, which accounts for cross-flow correlation (procedure in Appendix B). The post-freeze slice (21 Apr – 1 Jul, ≈49 trading days) is a *directional* check only. Two disclosures: (i) with 21 paired tests at the 95% level, one or two isolated significances are expected by chance — the evidence is cross-flow consistency, not single cells; (ii) the three flows' active returns correlate at $\bar{\rho} = 0.31$, so the panel carries $N_{\text{eff}} \approx 1.9$ effective flows, not three. All ablations are full from-2023 walk-forward re-runs. Variants that alter only the portfolio-weighting step (forced equal-weight sizing, equal-weight ensemble, positive-only sizing) leave upstream selections and forecasts unchanged by construction — the graph is deterministic and those components do not feed back into training or selection — so they isolate the weighting step exactly.

A decomposition, and its confound. The three forecast-related rows satisfy an exact identity per flow: (agent − forced-EW) + (forced-EW − no-forecast) = (agent − no-forecast). The first term — the sizing effect — is $+42.5/+23.3/+41.3$; the residual is *negative* in all three flows ($-14.7/-18.2/-33.0$), netting to the node totals $+27.8/+5.1/+8.3$ (all terms computed on the three series' common dates, so the identity is exact). The residual conflates the forecast's influence on *selection* with a *cadence* difference (a Forecast node induces daily decisions; the screener-only variant rebalances at the Screener's cadence) and is not yet separated; the cadence ablation is the experiment that splits it.

Verdict, in three tiers. *Demonstrated* (pooled CI excludes zero): the Screener's contribution (+45 pp/yr), forecast-magnitude sizing (+36 pp/yr), and monthly Screener re-search over rules fixed at inception (+23 pp/yr). *Inconclusive* (CI spans zero): continual forecast retraining (+4 pp/yr, $[-5, +13]$), positive in two of three agents and negative in all three over the short post-freeze slice; the frozen checkpoint was trained on data through December 2022 only, so it never saw any part of the evaluation window and the comparison is clean over the full window. Positive-only sizing is likewise inconclusive (mixed signs). *Precise null* (tight CI at zero): adaptive

Table 4: [T1–T2 full window; T3 directional check] Ablation results on the three-agent panel. Δ = paired annualized active-return difference, agent minus ablated variant (positive = component helps), in pp/yr over the full window (early 2023 – 1 Jul 2026); * = per-flow 95% block-bootstrap CI excludes zero. Post-freeze = sign check over 21 Apr – 1 Jul 2026 (geometric window basis); $k/3$ means the component helps in k of the three agents over that slice.

Ablation (definition)	Apex	Core	Surge	Pooled [95% CI]	Post-freeze
Remove Screener	+52.6*	+26.3*	+55.1*	+45.4 [+30, +60]	helps 3/3
Static Screener (rules searched once at inception, Jan 2023)	+23.6	+4.6	+38.9*	+22.7 [+11, +34]	1/3
Forced equal-weight sizing (same flow, weights forced equal)	+42.5*	+23.3*	+41.3*	+36.2 [+21, +50]	helps 3/3
Remove Forecast (node removed; cadence and selection path differ, see text)	+27.8*	+5.1	+8.3	+13.6 [+5, +23]	helps 3/3
Static Forecast (single checkpoint trained on data through Dec 2022; inference-only over the full window)	+13.0	−4.1	+3.5	+4.1 [−5, +13]	0/3
Equal-weight ensemble	+0.1	+0.5	−0.7	−0.0 [−0.7, +0.5]	null
Positive-only sizing	+6.7	−7.9	+18.2*	+5.8 [−2, +17]	mixed

ensemble weighting (-0.0 ± 0.6 pp/yr). We retain the adaptivity as low-cost insurance against member-skill drift; the mechanism's demonstrated property is the no-future-bleed update order, whose leak diagnostic is scheduled for the August 2026 revision. Where a variant matches the full system we report the null. The claim that survives is that composition and selection-layer adaptation carry the measured edge, and the deep component's demonstrated value is its sizing signal rather than its ongoing retraining.

12 Empirical Results

How to read these results. The full-window tables below span the configuration boundary and the architecture freeze (T1–T2) and are not clean design-frozen out-of-sample evidence; the design-frozen slice is §13 and the post-freeze column of §11. The live account is real but, at ≈41 trading days, not statistically significant on its own. The deep component's demonstrated value is forecast-*magnitude sizing*, not directional forecasting alpha. The formalism is topology-general; the empirical evidence covers one topology. Nothing here is evidence of institutional capacity.

All benchmark-relative figures use the symmetric basis of §5.1: price/price for the simulated full-window results below, total-return/total-return for the live account (§13). The main table is reported per agent flow (the three live forecast-bearing agents); these full-window figures span the configuration boundary and freeze (T1–T2 provenance, §10) and are therefore not a clean out-of-sample window — the design-frozen slice is reported separately in §13 and §11.

Capacity (framework). The base case is consumer-scale long-only trading; the live account's size means it has no market impact and is silent on capacity. The zero-slippage base case is a property of this regime — small orders in highly liquid large-caps executed at the close — rather than a universal claim. With A assumed AUM, turnover $\Delta w_{i,t}$, and average daily dollar volume $\mathrm{ADV}_{i,t}$, a participation proxy is $\rho_{i,t}(A) = A|\Delta w_{i,t}|/\mathrm{ADV}_{i,t}$; as A grows, participation rises and

Table 5: [T1–T2 mixed provenance] Main per-agent results, full window (early 2023 – 1 July 2026), price-index benchmarks (price symmetry, §5.1). CAGR = compound annual growth rate; Sharpe computed with $r_f = 0$; MDD is the portfolio's own maximum drawdown; Active MDD is the maximum drawdown of the relative-wealth path (portfolio ÷ benchmark, Figure 3) — the deepest peak-to-trough underperformance versus the benchmark. DJIA is not covered by this panel.

Flow	Series	CAGR	Vol.	Sharpe	IR	MDD	Active MDD
Apex (NDX)	NDX price index	30.9%	20.2%	1.43	—	−22.9%	—
Apex (NDX)	Full Sophon-3	105.2%	39.3%	2.02	1.74	−29.5%	−28.7%
Core (NDX)	NDX price index	30.9%	20.2%	1.43	—	−22.9%	—
Core (NDX)	Full Sophon-3	62.2%	29.2%	1.80	1.41	−23.8%	−18.3%
Surge (SPX)	SPX price index	20.9%	14.9%	1.35	—	−18.9%	—
Surge (SPX)	Full Sophon-3	77.6%	35.5%	1.80	1.50	−24.7%	−18.7%

modeled slippage must be reintroduced, so net performance as a function of A and participation limits is the boundary this analysis characterizes and is left to future stress-testing.

13 Live Track Record (Real Money)

This reports a personal brokerage account of one of the authors (H.D.; Alpaca, account ending 6414), funded with real capital and trading **Sophon Core** — the medium-risk agent of the ablation panel (§11) — since 27 April 2026. It is Tier-4 evidence: the most verifiable in the paper, stated exactly, and small and short, so treated accordingly.

Table 6: [Tier 4] Real-money live result: author account trading Sophon Core. Single account, real capital, real fills. Window 27 April – 24 June 2026 (58 calendar days; ≈41 trading days), entirely within the design-frozen window.

Quantity		Value
Portfolio NAV, start (27 Apr 2026)		$1,772
Portfolio NAV, end (24 Jun 2026)		$2,174
Portfolio return over window		+22.7%
Benchmark (Nasdaq-100), normalized start		$1,772
Benchmark, price-index end	$1,924	(+8.6%)
Benchmark, total-return end	≈$1,926	(+8.7%)
Active return vs. total-return Nasdaq-100		**+12.9%**
Active return vs. price-index Nasdaq-100		+13.0%

Interpretation and limits. The window lies entirely after both the configuration boundary and the architecture freeze, so neither data nor design could have been fit to it, and it uses real broker fills. It is *not* statistically conclusive: ≈ 41 trading days is about 0.16 years, and the standard error on an information ratio scales as $\sqrt{1/\text{years}}$, so even a very high IR does not clear conventional significance over this window (§14). It is a single small account with no market impact, hence silent on capacity. The account runs the same Sophon Core configuration as the panel's Core series in §11; differences from that simulated series arise from real fills, the dividend basis (a real brokerage credits dividends, so the live NAV is portfolio-total-return and is reported against the total-return benchmark, §5.1), and the slightly different window (27 April–24 June here vs. the panel's 21 April–1 July slice).

Supporting exhibits. Daily-series exhibits for this account — NAV and active-return path, realized active volatility, active max drawdown, the window IR with its confidence interval, and

a quantified realized-vs-intended weight-drift study (§9) — together with the live panel's rolling-IR charts and the cross-agent active-return correlation matrix, are scheduled for the August 2026 revision (§17).

14 Statistical Inference

We are deliberate about power because the live window is short and the architecture is adaptive, which inflates researcher degrees of freedom.

Power of a short window. The t-statistic on an information ratio is approximately $t \approx \mathrm{IR} \times \sqrt{\text{years}}$. Over ≈ 41 trading days (≈ 0.16 yr), even $\mathrm{IR} = 3$ gives $t \approx 1.2$. We therefore do not present the live record as significant on its own.

Overlapping windows. A daily-sampled 126-bar rolling IR is heavily autocorrelated: a smooth multi-month curve is approximately *one* independent block, not many confirmations. We present it as evidence of stability, not repeated significance.

Cross-agent correlation. "Four agents beat the benchmark" is four pieces of evidence only if their *active* returns are weakly correlated. With average pairwise correlation $\bar{\rho}$, the effective number of independent bets is $N_{\text{eff}} \approx N/[1 + (N-1)\bar{\rho}]$. We report the correlation matrix and N_{eff} so "all four" is not over-counted; for the three-flow ablation panel the realized values are $\bar{\rho} = 0.31$ and $N_{\text{eff}} \approx 1.9$ (§11), and the four-agent live-panel matrix accompanies the August 2026 revision.

Multiple testing. Multiplicity arises from materially distinct graph variants, search spaces, universes, cost assumptions, and cadences — not from each online retrain. For the ablation suite the inference is *paired*: each variant is compared to its own agent as a daily difference series, block-bootstrapped (21-day blocks), with a panel estimate that averages flows before bootstrapping; 21 paired tests at 95% imply one or two chance significances, so conclusions rest on cross-flow consistency and pooled CIs rather than isolated cells, and the realized panel correlation ($\bar{\rho} = 0.31$, $N_{\text{eff}} \approx 1.9$) is reported alongside. Selection-stage multiplicity — the number of configurations searched before the live agents were fixed — is owned by the companion selection paper, which reports the accounting for that search; this paper's inference is confined to the ablation layer, as reported above.

15 Failure Modes

Underperformance is classified by observable degradation mode rather than a single speculative cause: forecast rank collapse (rank IC/NDCG falls to a benchmark-like baseline); screener over-compression (very few names selected repeatedly); turnover explosion; ensemble concentration (one member captures most weight); data sparsity (coverage falls); cost sensitivity (gross-positive, net-negative); regime mismatch (active drawdown concentrated in one window); and capacity stress (performance decays with AUM). The final standard is the tradable outcome after costs; node-level diagnostics distinguish model, screener, portfolio, data, and capacity failures.

16 Related Work

The paper's motivation rests on two anchors, one from finance and one from machine learning. The Adaptive Markets Hypothesis treats markets as competitive ecologies in which predictive relationships decay as they are exploited, so adaptation is a requirement rather than a refinement [6]; the machine-learning literature names the same phenomenon *concept drift* — the

mapping from features to targets changes over time [2]. We use the term *continual learning* in the operational sense of causal rolling adaptation of a deployed system; in machine learning the term also names a literature on sequential task learning under stability–plasticity constraints [7], and we do not claim contributions to that problem. Large-scale evidence that nonlinear learners add value in return prediction is established [3]; Sophon-3 is complementary to that predictor-level result, contributing a *system-level* architecture — selection, weighting, sizing, and execution composed around the predictors — rather than a new predictor class.

Two mechanisms and one protocol are used directly. The ensemble's softmax-over-EWMA-losses weighting is an instance of the exponentially weighted average forecaster (Hedge) from prediction with expert advice [1], applied with the strict no-future-bleed update order of §11; the member loss is built on NDCG top-K ranking quality [5]. The statistical protocol follows the backtest-overfitting canon: significance judged against an honestly counted number of trials [4]; the corresponding accounting for the selection-stage search is the subject of the companion selection paper (§14).

17 Scope of This Release and the August 2026 Revision

Version 1.0 deliberately limits its scope to what has been computed and verified: the formalism, the three-flow ablation suite with paired statistics, the populated per-agent results, and the real-money proof-of-concept. The August 2026 revision adds, in one place rather than as scattered promises: the leak-prone weighting diagnostic (quantifying what no-future-bleed removes); the cost stress under Interactive Brokers' commissioned schedule; the cadence-matched ablations that separate the Remove-Forecast residual; a turnover, execution, and realized weight-drift table; the forecast-ranking diagnostics (rank IC, NDCG@K, coverage, effective ensemble size) and the directional and magnitude tests of §9; the live account's daily-series exhibits (active-return path, window IR with confidence interval, active drawdown) and the live panel's rolling-IR and correlation exhibits; the holdings-vs-index dividend-yield differential; a public hash commitment anchoring the 21 April freeze; and external heuristic baselines (momentum top-K and a random-screener placebo at matched breadth and turnover). Companion papers on live/backtest parity and on screener configuration and pre-registered agent selection are in preparation.

18 Limitations

1. Non-stationarity remains unresolved; Sophon-3 adapts to it, it does not remove it.
2. The live record is short and single-account: real and out-of-sample, but not statistically conclusive, and silent on capacity.
3. Backtest evidence is conditional on the data contract, universe, broker profile, cost model, and graph family.
4. Point-in-time claims are only as reliable as the timestamp semantics behind them.
5. Zero *commission* is venue-specific: it reflects Alpaca's documented U.S.-equity schedule and would not hold at a commissioned broker or in other fee regimes (the system simulates Interactive Brokers for this reason).
6. Zero *slippage* is regime-specific: it holds for small orders in highly liquid large-caps executed at the close, and dissolves as AUM grows; it is not an institutional-capacity claim.
7. Adaptive systems carry research degrees of freedom; the frozen procedure and dated freeze are the controls, not a claim of their absence.
8. Continual forecast *retraining* is not yet shown to add active return over a single strong pre-trained checkpoint (+4 pp/yr pooled, CI $[-5, +13]$); it is retained as regime insurance, and resolving an effect of this magnitude requires a materially larger panel or window.
9. The ablation panel is three correlated flows (two NDX, one SPX; $\bar{\rho} = 0.31$, $N_{\text{eff}} \approx 1.9$)

with no DJIA coverage; per-flow significances are interpreted through pooled estimates and cross-flow consistency.

10. The post-freeze ablation slice is ≈49 trading days and is directional evidence only; for the monthly-adapting Screener it contains only ≈2 re-selection events.
11. The architecture is topology-general, but the empirical evidence is for a single topology — the canonical four-node chain on equity-index universes — not for the space of admissible graphs.

19 Conclusion

Sophon-3 is not a static alpha model but an adaptive financial AI architecture: a causal node graph in which universe selection, neural forecasting, ensemble weighting, and portfolio construction update under a frozen procedure, with an evidence framework that dates the architecture freeze and separates results by provenance. The ablation suite locates the value: the Screener contributes +45 pp/yr pooled and forecast-magnitude sizing +36 pp/yr, each significant in all three agents; monthly Screener re-search adds +23 pp/yr over rules fixed at inception; continual forecast retraining is inconclusive at current sample size, and adaptive ensemble weighting is a precise null. We report the null and inconclusive results together with the positive ones; the credibility of the positive results depends on it. The empirical claim is evaluated at the tradable-portfolio level, includes a real-money out-of-sample window that is encouraging but not yet conclusive, and is conditional on the declared graph, data contract, total-return benchmark convention, and statistical-correction protocol. Deterministic live-vs-backtest parity and the pre-registered selection of the live agents are established in the two companion papers.

Appendix A: Reproducibility Anchors

Dated events. Configuration data boundary: 31 December 2025. Architecture freeze (Π): 21 April 2026, anchored by internal configuration/checkpoint manifests timestamped at or before that date. **Windows.** Full ablation window: early 2023 – 1 July 2026 (≈855–870 trading days per flow). Post-freeze ablation slice: 21 April – 1 July 2026 (≈49 trading days). Real-money window: 27 April – 24 June 2026 (≈41 trading days). **Panel.** Three live forecast-bearing agents (Apex/NDX, Core/NDX, Surge/SPX); the fourth live agent is screener-only. **Benchmark basis.** Symmetric dividend treatment (§5.1): price indices for simulated backtests, total-return for the live account. **Ablation classes.** All ablations are full from-2023 walk-forward re-runs; weighting-only variants leave upstream selections and forecasts unchanged by construction (§11). **Parity acceptance surface** (companion parity paper): canonical-hash equality of input snapshot and universe membership; timestamp-symbol equality of Screener selections; tolerance-bounded equality of member outputs, final forecast rows, and ensemble weights; exact or tolerance-bounded equality of Make-Portfolio rows; canonical serialization of checkpoint state. **Selection.** The live agents were fixed at the 21 April freeze; the selection rule, search budget, and the count of considered and rejected candidates are disclosed in the companion selection paper.

Appendix B: Bootstrap Procedure for Paired Differences

For each agent–variant pair, let d_t be the daily difference of active returns on the pair's common dates ($n \approx 855$–870). Confidence intervals use a moving-block bootstrap with block length $\ell = 21$ trading days (≈one month, spanning the ensemble's weighting cadence and typical return autocorrelation): overlapping blocks $d_j, \ldots, d_{j+\ell-1}$ are drawn uniformly with replacement and concatenated to length n; the annualized mean $252 \cdot \bar{d} \cdot 100$ is recorded; this is repeated $B = 50{,}000$ times and the 2.5th and 97.5th percentiles reported. For the pooled estimate, the three agents' d_t are averaged date-wise *first* and the identical procedure is applied to the averaged series; averaging before resampling preserves the cross-agent correlation inside each block, so the pooled interval reflects $N_{\text{eff}} \approx 1.9$ rather than treating the

agents as independent. t-statistics are $\bar{d}/(s_d/\sqrt{n})$ on the same series.

Appendix C: Notation

$\mathcal{F}_t$: information available no later than t. U_t: eligible universe; I_t: point-in-time index constituents; S_t: Screener-selected set. $\hat{r}_{i,t,h}$: forecast for asset i at t over horizon h. $w_{i,t}$: portfolio weight; $L_{m,t}$: EWMA rank-loss state for member m. Π: frozen evaluation procedure; G: node graph.

References

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[2] J. Gama et al. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44, 2014.

[3] S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

[4] C. R. Harvey, Y. Liu, and H. Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

[6] A. W. Lo. The Adaptive Markets Hypothesis. *J. Portfolio Management*, 30(5):15–29, 2004.

[7] G. I. Parisi et al. Continual lifelong learning with neural networks. *Neural Networks*, 113:54–71, 2019.