SCALING PREDICTION AND INVESTMENT IN FINANCIAL MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial machine learning typically follows a predict-then-optimize pipeline: forecast returns, then construct portfolios. We study how model capacity should scale when returns are noisy and cross-sectionally dependent. We derive a finance-specific scaling law that separates diversifiable estimation error from the additional cost induced by common shocks. On a panel of 1,800 U.S. stocks with 153 characteristics, the law raises log-loss R^2 from 0.315 to 0.888 relative to an i.i.d. law and captures finite loss-minimizing model widths. A natural response is to select the loss-minimizing model, but our portfolio results show that this can discard useful information: smaller models achieve lower prediction loss, while larger models generate higher cumulative long–short returns. Together, these results show that financial scaling cannot be reduced to model selection alone; further scaling beyond the loss optimum requires better supervision. We therefore introduce hybrid weak-to-strong (W2S) learning, which combines smaller-model forecasts with realized returns to supervise the larger model. Under conditional orthogonality, prediction loss is quadratic in the mixing weight and admits an interior optimum. Empirically, hybrid W2S reduces prediction loss relative to direct strong learning and achieves higher terminal cumulative long–short returns than both the weak and directly trained strong models.

1 INTRODUCTION

Financial machine learning typically follows a predict-then-optimize pipeline: models forecast returns and then use those forecasts to construct portfolios (Jensen et al., 2026; Kelly et al., 2024; Gu et al., 2021; 2020; Avramov et al., 2023). Greater model capacity can capture richer relationships between asset characteristics and future returns. However, this benefit comes with an estimation challenge: models learn from realized returns, which are noisy and cross-sectionally dependent (Chamberlain & Rothschild, 1982; Bai, 2003). This tension raises the central question of this paper: how can model capacity be used effectively for financial prediction?

Scaling laws provide a natural language for this dependence, relating prediction loss to model capacity and data scale. Standard formulations write this relation as $\mathcal{L}(d, n)$, where d denotes model capacity and n the number of training observations. For a financial panel, $n = NT$, with N assets observed over T dates. In financial panels, however, N and T provide different information: assets diversify idiosyncratic noise, while dates provide new realizations of common shocks (Pesaran, 2006; Bai & Ng, 2002; Kelly et al., 2019). To isolate this distinction, we compare an i.i.d. benchmark with a financial panel model that preserves common shocks:

$$\mathcal{L}_{\text{IID}}(d; N, T) \sim B_d + \frac{S_d}{NT}, \qquad \mathcal{L}_{\text{FIN}}(d; N, T) \sim B_d + \frac{S_d}{NT} + \frac{G_d}{\tau_T}.$$

The financial law contains an additional estimation term, G_d/τ_T, arising from exposure to common shocks. Empirically, the financial law raises log-loss R^2 from 0.315 to 0.888 and captures finite loss-minimizing widths across panel sizes. The optima show that larger models are not uniformly better and that useful capacity varies with cross-sectional and temporal information.

A natural response is to select the model that minimizes prediction loss. Our portfolio results show that this can leave useful information on the table. Across the model pairs we evaluate, the smaller model attains lower prediction loss, while the larger model generates higher cumulative long–short

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

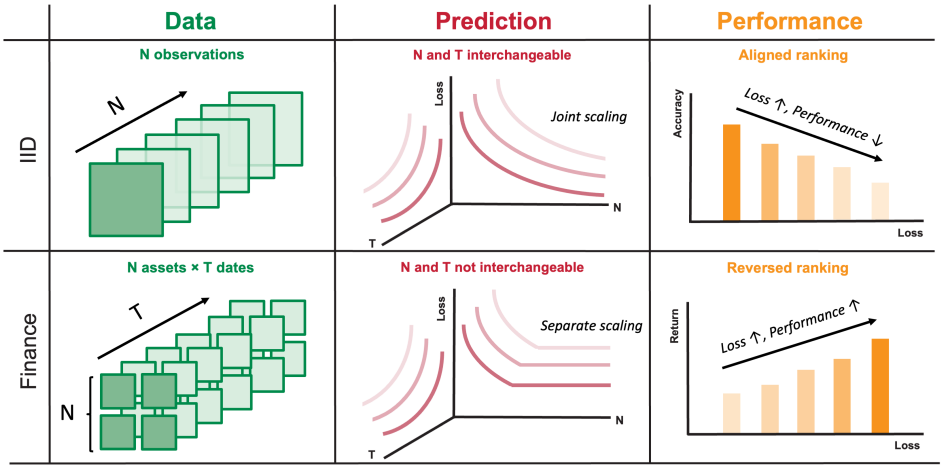

Figure 1: **Financial panels change prediction scaling and model selection.** Unlike IID data, N assets share T dates, so cross-sectional and temporal information are not interchangeable. The larger model can retain a distinct forecast component even when the smaller model minimizes loss, creating the supervision trade-off addressed by hybrid W2S.

returns. We further isolate the larger model's forecast component orthogonal to the smaller model and find that it is less persistent, turns over faster, and is more sensitive to fast-moving characteristics. These results suggest that the lower-loss model does not span all of the predictive information captured by the larger model. Figure 1 summarizes the contrast between i.i.d. and financial panels and previews the resulting supervision problem.

Taken together, these results reveal a key departure from standard scaling: noisy and dependent financial supervision can make the loss-optimal model smaller even when larger models retain distinct predictive information. Beyond the loss-optimal capacity under realized-return supervision, further scaling requires a better supervision signal. The model with lower prediction loss provides a natural source of supervision, motivating weak-to-strong (W2S) learning. Using weak forecasts alone reduces return noise but limits the strong model to the weak signal. We retain realized returns as complementary supervision through the hybrid target

$$Z(\rho) = (1-\rho) f_w(X) + \rho Y, \qquad \rho \in [0,1].$$

The mixing weight balances the denoising benefit of weak supervision against the additional signal available in realized returns. Under conditional orthogonality, this trade-off yields a quadratic prediction risk with an interior optimum. Empirically, the theory-guided weights remain near these optima, and hybrid W2S lowers prediction loss relative to direct strong learning while producing higher terminal cumulative returns than both the weak and directly trained strong models.

2 Preliminaries

This section analyzes how prediction loss scales with cross-sectional breadth N, temporal depth T, and model scale d. Financial dependence determines the effective information in N and T, while model scale governs the approximation–estimation trade-off. Their interaction yields the financial scaling law.

Prediction target and risk. We observe a balanced panel $\mathcal{D}_{N,T} = \{(X_{it}, Y_{i,t+1})\}_{i\in[N],t\in[T]}$, where X_{it} contains asset characteristics and $Y_{i,t+1}$ is the one-period-ahead return. The population prediction target is the conditional mean $f^\star(x) := \mathbb{E}[Y_{i,t+1} \mid X_{it} = x]$. Let P_X denote the population distribution of characteristics, with $\langle f, g\rangle_2 := \mathbb{E}_{X\sim P_X}[f(X)g(X)]$ and $\|f\|_2^2 := \langle f, f\rangle_2$. For a fitted predictor $\widehat{f}$, we measure expected excess prediction loss by

$$\mathcal{R}(\widehat{f}) := \mathbb{E}\left[\|\widehat{f} - f^\star\|_2^2\right], \tag{1}$$

where the expectation is over the training sample and estimator randomness. How this loss responds to additional data depends on the dependence structure of the panel.

Financial dependence. The key departure from i.i.d. sampling is that assets observed at the same date share common return innovations. We represent this dependence as

$$Y_{i,t+1} = f^\star(X_{it}) + b(X_{it})^\top u_{t+1} + \varepsilon_{i,t+1}, \tag{2}$$

where u_{t+1} is a date-level innovation shared across assets and $\varepsilon_{i,t+1}$ is asset-specific noise. We impose $\mathbb{E}[u_{t+1} \mid X_{it}] = 0$, $\mathbb{E}[\varepsilon_{i,t+1} \mid X_{it}, u_{t+1}] = 0$, and $\mathrm{Cov}(u_{t+1}) = \Sigma_u$, so that $f^\star$ remains the conditional-mean target.

The shared innovation makes the two panel dimensions statistically distinct: increasing N adds observations exposed to the same date-level shock, whereas increasing T adds observations across different shocks. The next section derives how this distinction changes the scaling of prediction loss.

A familiar special case is the unrestricted IPCA specification of Kelly et al. (2019):

$$Y_{i,t+1} = X_{it}^\top \Gamma_\alpha + X_{it}^\top \Gamma_\beta F_{t+1} + \varepsilon_{i,t+1}. \tag{3}$$

Write $F_{t+1} = \lambda + u_{t+1}$ gives $b(x) = \Gamma_\beta^\top x$ and $f^\star(x) = x^\top \Gamma_\alpha + b(x)^\top \lambda$, which maps IPCA to (2).

Model scale and prediction space. We index theoretical models by a scalar scale parameter d. For each d, let $\mathcal{F}_d \subset L_2(P_X)$ denote the associated closed linear prediction space and P_d its orthogonal projection. The population-optimal predictor at scale d is

$$f_d^\star := P_d f^\star. \tag{4}$$

The spaces $\mathcal{F}_d$ need not be nested because changing model scale may alter the representation rather than simply add prediction directions. Thus, d indexes a sequence of theoretical prediction models and need not equal the dimension of $\mathcal{F}_d$.

For a predictor $\widehat{f}_d \in \mathcal{F}_d$ trained at scale d, write $\mathcal{L}(d; N, T) := \mathcal{R}(\widehat{f}_d)$. Model scale can affect both how well the prediction space represents $f^\star$ and how accurately the fitted predictor is estimated from finite data. The next section formalizes these components through model-scale rate conditions, first under i.i.d. sampling and then under the financial dependence structure in (2).

3 Financial Scaling: The Cost of Learning from Realized Returns

3.1 IID benchmark

Standard scaling laws summarize training data by a single sample size, so an i.i.d. benchmark treats the NT observations in a panel as interchangeable. Let $f_d^\star := P_d f^\star$ be the population-optimal predictor in $\mathcal{F}_d$. Orthogonality decomposes expected excess prediction loss as

$$\mathcal{L}(d; N, T) := \mathbb{E}\big\|\widehat{f}_d - f^\star\big\|_2^2 = \underbrace{\big\|(I - P_d) f^\star\big\|_2^2}_{B_d} + \underbrace{\mathbb{E}\big\|\widehat{f}_d - f_d^\star\big\|_2^2}_{V_d(N,T)}. \tag{5}$$

Here, B_d is the approximation loss from signal outside the scale-d prediction space, while $V_d(N, T)$ is the cost of estimating the corresponding predictor from finite data. Changing model scale can therefore affect both the represented signal and its estimation cost.

Let $\sim$ denote leading-order equivalence, and let S_d denote the aggregate diversifiable estimation complexity of the scale-d predictor. Under i.i.d. sampling, its estimation cost satisfies $V_d^{\mathrm{IID}}(N, T) \sim S_d/(NT)$. Thus, d indexes the model sequence, while B_d and S_d record its approximation and estimation profiles. A d-coordinate series construction with stable average coordinate variance gives the familiar specialization $S_d \sim \sigma_{\mathrm{div}}^2 d$. The general IID law is

$$\mathcal{L}_{\mathrm{IID}}(d; N, T) \sim B_d + \frac{S_d}{NT}. \tag{6}$$

The two terms capture the approximation–estimation trade-off along the model-scale sequence. For a fixed total sample size NT, the benchmark therefore predicts the same loss regardless of how observations are allocated across assets and dates. Appendix A.1 gives a d-coordinate series construction sufficient for these rate conditions and proves (6).

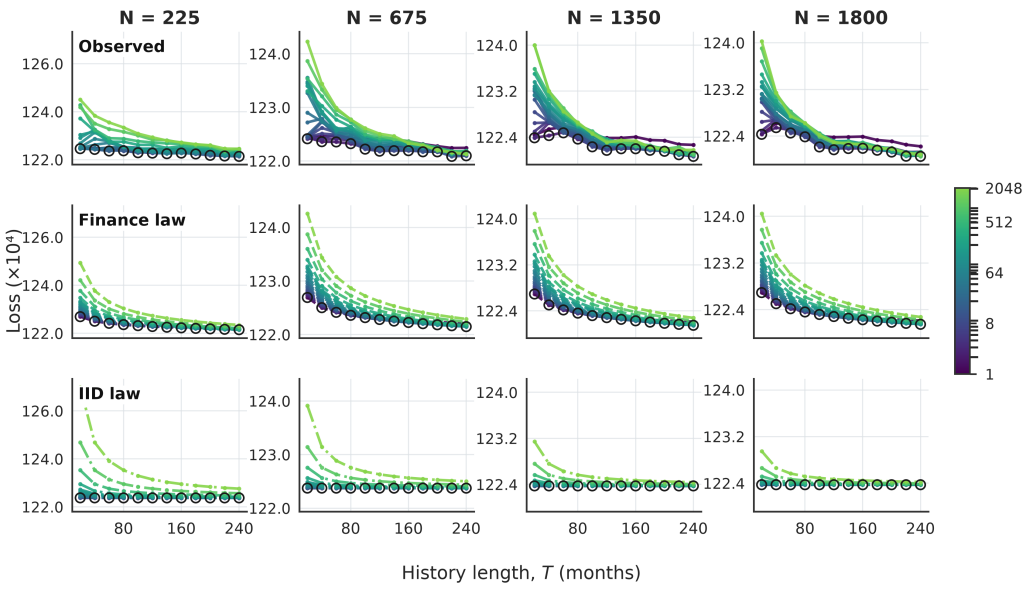

Figure 2: **The finance-specific law captures temporal scaling of prediction loss.** Columns fix N, while rows show observed loss, the finance-specific fit, and the IID fit as T increases. Colors denote width m, and open circles mark the loss-minimizing width.

3.2 Financial scaling law

Financial panels depart from this benchmark because assets observed at the same date share the innovation u_{t+1}. Additional assets help average diversifiable variation but do not provide new realizations of this shared shock; additional dates do both. Appendix A.1 derives the temporal rate from the autocovariance of the common shocks. Absolutely summable autocovariances give $\tau_T = T$, while autocovariances that decay as $h^{-\gamma}$ give $\tau_T = T^\gamma$. Let G_d denote the aggregate exposure of the scale-d prediction space to these shocks. The common-shock estimation cost is G_d/τ_T, while the residual diversifiable component scales as $S_d/(NT)$. Under a square-integrable loading function, G_d is bounded. For nested prediction spaces, it can increase with model scale only toward a finite limit.

Proposition 1 (Financial prediction-loss scaling)**.** *Under Assumptions 8, 10, and 12,*

$$\mathcal{L}_{\mathrm{FIN}}(d; N, T) \sim B_d + \frac{S_d}{NT} + \frac{G_d}{\tau_T}. \tag{7}$$

The additional term changes the scaling implications of the benchmark. Increasing N reduces only diversifiable estimation loss, so its gains flatten when common-shock variation becomes dominant. Increasing T reduces both components by adding new realizations of the shared shocks. Changing d can change approximation loss, diversifiable estimation complexity, and common-shock exposure along the model sequence. The loss-minimizing model therefore need not be the largest available model. This asymmetry between N and T is the main prediction evaluated below. Appendix A.1 derives the coordinate decomposition, derives the temporal rate from primitive covariance conditions, proves that common-shock exposure is bounded under square-integrable loadings, and proves Proposition 1.

Proposition 1 leaves the three model-scale profiles B_d, S_d, and G_d unrestricted. In the empirical analysis, we instantiate the theoretical scale index d with model width m. For $m \geq p$, m is an empirical training-scale index rather than a representational dimension. We write $\mathcal{M}_{\mathrm{obs}}$ for the observed width grid and approximate the three profiles on this grid by

$$B_m \approx Am^{-\kappa}, \qquad S_m \approx \sigma^2_{\mathrm{div}} m, \qquad G_m \approx Qm^{1-\beta}, \qquad m \in \mathcal{M}_{\mathrm{obs}}. \tag{8}$$

These relations parameterize the finite model grid; they are not asymptotic restrictions as $m \to \infty$. In particular, the last relation can describe increasing common-shock exposure over the observed widths while remaining consistent with the theoretical bound on G_d.

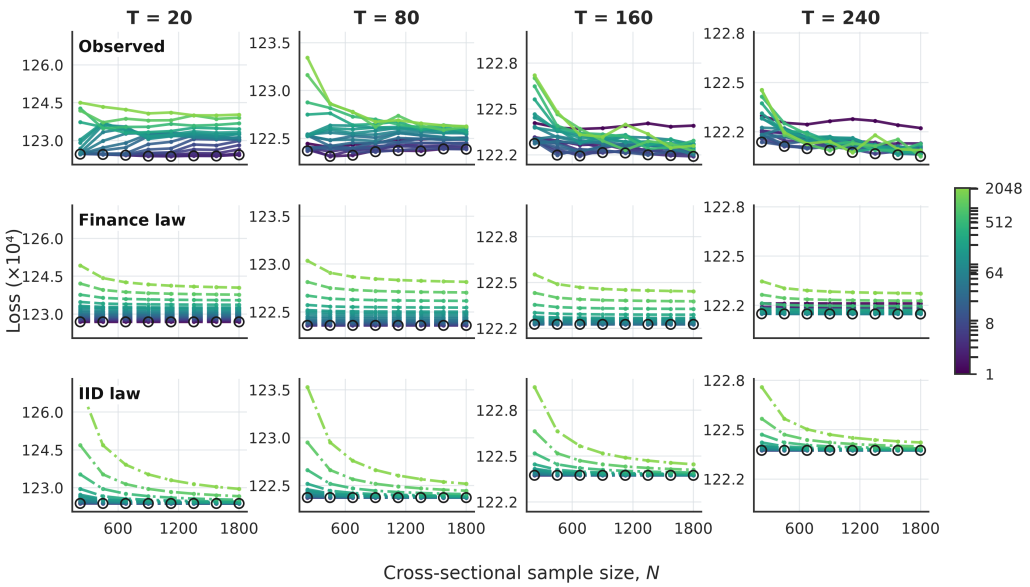

Figure 3: **Prediction gains from larger cross sections flatten quickly.** Columns fix T, while rows show observed loss, the finance-specific fit, and the IID fit as N increases. Colors denote width m, and open circles mark the loss-minimizing width.

3.3 Empirical evidence

We evaluate whether cross-sectional breadth and temporal depth provide interchangeable information using a balanced monthly panel of 1,800 U.S. stocks with 153 characteristics (Jensen et al., 2023). Across rolling samples, we vary N, T, and model width m, which serves as the empirical capacity index; Appendix B.1 reports the full design.

Financial data scale is compositional rather than scalar. Figures 2 and 3 show that increasing N and T produces different loss curves. Gains from additional assets flatten earlier, whereas loss continues to improve with longer histories, consistent with cross-sectional averaging removing idiosyncratic noise while additional dates provide new realizations of common shocks. The loss-minimizing width also shifts with (N, T). Thus, the same nominal sample size NT can imply different prediction loss and different loss-optimal capacity depending on panel composition.

Panel composition reshapes the loss–capacity frontier. Figure 8 evaluates this joint dependence over the full (N, T, m) grid. Relative to the i.i.d. benchmark, the finance-specific law raises log-loss R^2 from 0.315 to 0.888. The law therefore captures not only how loss changes with data scale, but also how the capacity that minimizes loss changes with the composition of that data. Appendix B.8 separately validates the temporal and capacity components of the common-shock term.

The finance-specific law extrapolates to unseen scales. We estimate both laws only on cells with $N \leq 1350$, $T \leq 180$, and $m \leq 96$, and then predict 12 cells beyond all three training boundaries. The finance-specific law reduces log-loss RMSE by 35.3% relative to the IID law and improves R^2_{OOS} from 0.143 to 0.641. Appendix B.2 reports the joint and coordinate-wise holdouts.

4 Why the Loss-Minimizing Model Is Not Enough

4.1 Prediction and portfolio rankings diverge

Loss-optimal capacity reflects an estimation trade-off. Financial scaling produces finite loss-minimizing widths because the estimation cost of additional capacity can offset its approximation gains. If capacity beyond this point were redundant, selecting the loss-minimizing model would lose nothing relevant downstream. We test this implication by carrying the same forecasts into

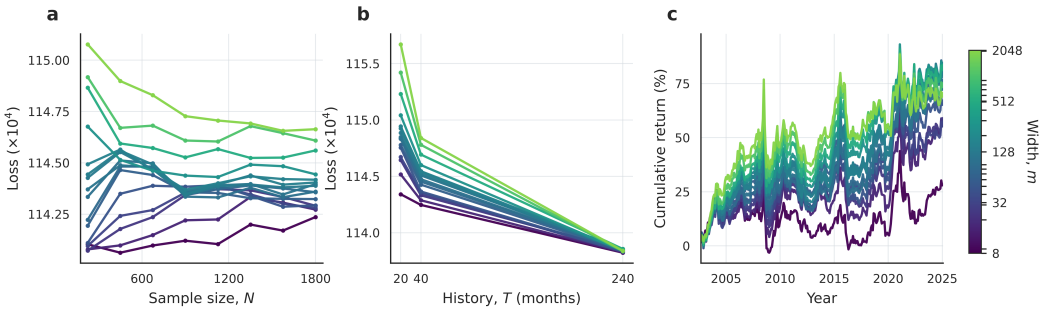

Figure 4: **Prediction loss and portfolio returns scale differently.** Panels **a** and **b** show out-of-sample loss against N and T, with colour indicating model width m. Panel **c** shows the corresponding cumulative D10–D1 returns.

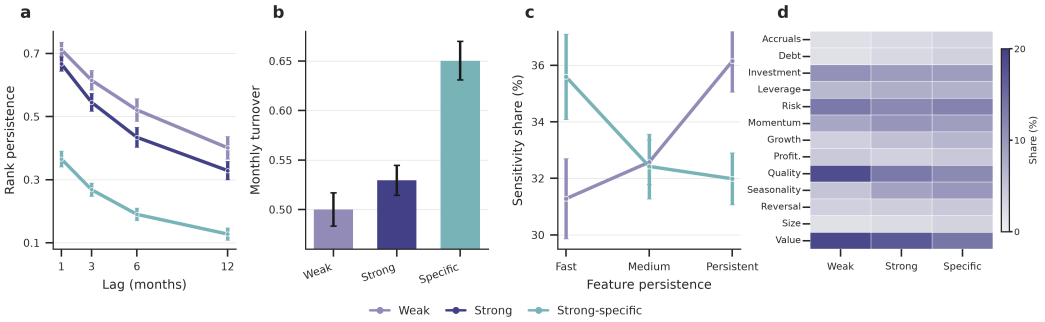

Figure 5: **Model capacity changes forecast persistence and composition.** Results use $N = 900$ and $T = 40$; Strong-specific is the residual of Strong after projection onto Weak. Panels **a–b** report rank persistence and extreme-decile turnover. Panels **c–d** report permutation sensitivity by characteristic persistence and economic theme; error bars in panels **a–c** denote 95% HAC confidence intervals.

equal-weighted D10–D1 portfolios. Figure 4 shows a different capacity ordering: across the model pairs we study, the smaller model attains lower out-of-sample prediction loss, while the larger model generates higher terminal cumulative returns. This divergence shows that capacity beyond the loss optimum retains forecast variation that matters for cross-sectional ranking. What predictive structure appears beyond this loss optimum?

4.2 Larger models retain distinct forecast components

Additional capacity broadens the temporal bandwidth of the forecast. To isolate what the larger model adds beyond the smaller one, we compare a smaller-model ensemble (Weak), a larger-model ensemble (Strong), and the component of Strong orthogonal to Weak (Strong-specific). Appendix B.3 reports the exact construction. Figure 5 shows that Strong-specific varies on a substantially faster timescale than Weak. At 12 months, rank persistence falls from 0.401 for Weak to 0.128 for Strong-specific, while extreme-decile turnover rises from 0.500 to 0.650. The input diagnostics reveal the same separation: Weak is more sensitive to persistent characteristics, whereas Strong-specific places greater sensitivity on fast-moving characteristics.

The separation also extends to economic composition. Weak relies more on value, quality, low-risk, and investment characteristics, whereas Strong-specific is relatively more sensitive to seasonality, profit growth, debt issuance, size, and accruals (Jensen et al., 2023). The agreement between forecast dynamics and input sensitivities shows that additional capacity changes both the frequency and composition of the predictive signal.

Larger models retain predictive structure beyond the loss optimum. The smaller model is better under average prediction loss, while the larger model retains a faster-moving component largely

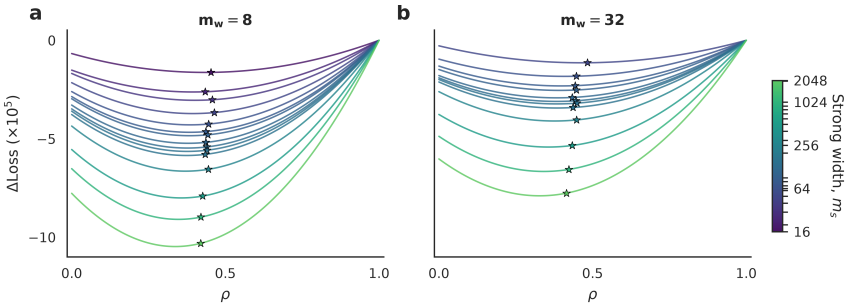

Figure 6: **The affine benchmark illustrates an interior W2S trade-off.** Panels **a** and **b** fix the weak-model widths at $m_w = 8$ and $m_w = 32$, respectively. Each colored curve represents a strong-model width m_s and reports benchmark-implied loss relative to direct strong learning at $\rho = 1$. Stars indicate the mean weights selected independently by the no-look-ahead finance-law estimator.

absent from the smaller forecast. Together with the portfolio reversal, this turns capacity choice into a supervision problem: how can we reduce the estimation cost of the larger model without suppressing the predictive variation added by greater capacity?

5 Hybrid W2S: Balancing Weak and Return Supervision

5.1 Hybrid target and optimal mixing

Hybrid W2S trades weak-supervision error against return-supervision error. Direct strong learning incurs the estimation cost $V_d(N, T)$ characterized by the financial scaling law. Weak-only supervision reduces direct exposure to return noise, but inherits the weak model's discrepancy from the prediction target. Formally, let $f_w \in \mathcal{F}_{d_w}$ be a frozen weak predictor, and let $\mathcal{F}_d$, with $d > d_w$, be the strong prediction space. For a sample of M training pairs (X_ℓ, Y_ℓ), hybrid W2S uses

$$Z_\ell(\rho) := (1-\rho) f_w(X_\ell) + \rho Y_\ell, \qquad \widehat{g}_d(\rho) \in \arg\min_{g \in \mathcal{F}_d} \frac{1}{M} \sum_{\ell=1}^{M} \left[g(X_\ell) - Z_\ell(\rho) \right]^2, \rho \in [0,1]. \quad (9)$$

The endpoints $\rho = 0$ and $\rho = 1$ recover pure W2S and direct strong learning, respectively. Thus, ρ controls whether supervision comes from the weak forecast or the realized return. Let $\widehat{g}_d^w := \widehat{g}_d(0)$ and $\widehat{f}_d := \widehat{g}_d(1)$ denote these two endpoints.

Conditioning on the frozen weak predictor, define the prediction risk as

$$\mathcal{R}(\widehat{g} \mid f_w) := \mathbb{E}\left[\|\widehat{g} - f^\star\|_2^2 \mid f_w \right]. \quad (10)$$

Write their estimation errors as $\eta_d := \widehat{g}_d^w - P_d f_w$ and $\zeta_d := \widehat{f}_d - P_d f^\star$.

To characterize the trade-off analytically, we consider an affine benchmark in which the fitted predictor interpolates between the two endpoints and their estimation errors are conditionally orthogonal. The affine path is exact for estimators that are linear in their training labels. Appendix A.2 proves the result below and relaxes conditional orthogonality, allowing correlated endpoint errors and deriving the corresponding optimal weight. We do not impose this benchmark on the nonlinear learner; its mixing weight is selected by the no-look-ahead empirical estimator described below.

Assumption 2 (Affine-orthogonal benchmark)**.** *Conditional on f_w, the fitted predictor is affine in the mixed target:*

$$\widehat{g}_d(\rho) = (1-\rho)\widehat{g}_d^w + \rho \widehat{f}_d. \quad (11)$$

Moreover,

$$\mathbb{E}[\eta_d \mid f_w] = \mathbb{E}[\zeta_d \mid f_w] = 0, \qquad \mathbb{E}[\langle \eta_d, \zeta_d \rangle_2 \mid f_w] = 0, \qquad \mathbb{E}[\|\zeta_d\|_2^2 \mid f_w] = V_d(N,T). \quad (12)$$

Retaining the approximation error B_d from Section 3.1, define the weak-model discrepancy and weak-endpoint estimation cost as

$$D_d := \|P_d(f^\star - f_w)\|_2^2, \qquad U_d^w := \mathbb{E}[\|\eta_d\|_2^2 \mid f_w]. \quad (13)$$

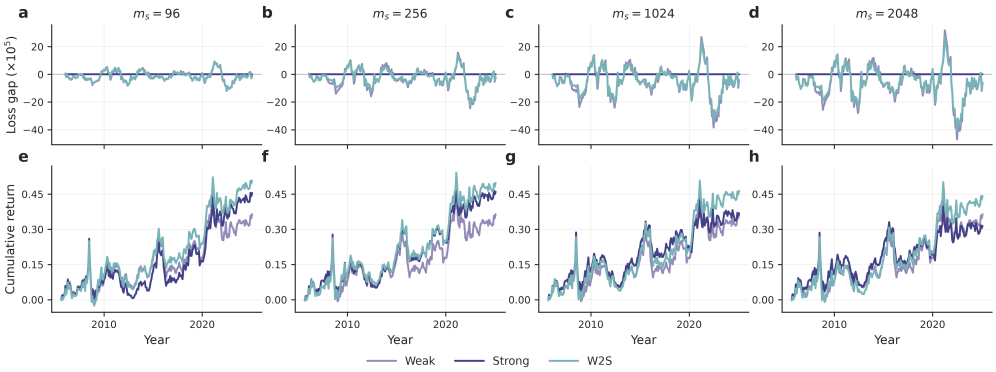

Figure 7: **W2S achieves the highest terminal cumulative return across all four strong-model sizes.** The columns fix $m_w = 32$ and vary $m_s \in \{96, 256, 1024, 2048\}$, comparing Weak, direct Strong, and theory-guided W2S. The top row shows the 12-month rolling OOS loss gap relative to matched Strong, while the bottom row shows cumulative equal-weight D10–D1 returns averaged across the 24 (N, T) cells.

Proposition 3 (Hybrid W2S in the affine-orthogonal benchmark)**.** *Under Assumption 2, the conditional risk is*

$$\mathcal{R}(\widehat{g}_d(\rho) \mid f_w) = B_d + (1-\rho)^2 \underbrace{(D_d + U_d^w)}_{\text{weak-supervision cost}} + \rho^2 \underbrace{V_d(N,T)}_{\text{return-supervision cost}}, \tag{14}$$

If $D_d + U_d^w + V_d(N,T) > 0$, its minimizer on $[0,1]$ is

$$\rho_d^\star = \frac{D_d + U_d^w}{D_d + U_d^w + V_d(N,T)}, \tag{15}$$

and the resulting risk is

$$\mathcal{R}(\widehat{g}_d(\rho_d^\star) \mid f_w) = B_d + \frac{(D_d + U_d^w)\, V_d(N,T)}{D_d + U_d^w + V_d(N,T)}. \tag{16}$$

If both $D_d + U_d^w$ and $V_d(N,T)$ are positive, then $\rho_d^\star \in (0,1)$ and its risk is lower than the risk at either endpoint. If $D_d + U_d^w = V_d(N,T) = 0$, every $\rho \in [0,1]$ attains risk B_d.

The optimal weight connects financial scaling to weak-to-strong supervision. The weak-supervision cost $D_d + U_d^w$ combines the discrepancy left by the weak target with the cost of fitting that target. As this cost rises, $\rho_d^\star$ rises and supervision shifts toward realized returns. The return-supervision cost $V_d(N,T)$ is characterized by the financial scaling law in Section 3. As this cost rises, $\rho_d^\star$ falls and supervision shifts toward the weak forecast. Thus, the same financial data limitations that determine which model capacity can be estimated reliably also determine how that capacity should be supervised.

5.2 Theory-guided mixing-weight selection

The interior supervision trade-off persists beyond the affine benchmark. The closed-form weight in (15) characterizes the stylized affine benchmark rather than the nonlinear learner. For each weak–strong pair and data configuration $c = (d_w, d, N, T)$, we train anchor models at $\mathcal{A} = \{0, 0.5, 1\}$. Using only anchor losses observed through month $t-1$, nonnegative least squares estimates the weak-side cost $\widehat{W}_{c,t-1}$ and return-side cost $\widehat{V}_{c,t-1}$ under the quadratic risk restriction. The continuous weight deployed in month t is

$$\widehat{\rho}_{c,t} = \frac{\widehat{W}_{c,t-1}}{\widehat{W}_{c,t-1} + \widehat{V}_{c,t-1}}. \tag{17}$$

When both estimated costs are zero, we set $\widehat{\rho}_{c,t} = 1$. The strong model is retrained at the selected continuous weight before month-t outcomes are observed. Appendix B.4 gives the NNLS criterion, historical windows, model grids, and evaluation protocol.

Figure 6 shows that the theory-guided weights lie near the interior minima of the benchmark-implied loss curves across model widths. Thus, although the nonlinear learner need not follow the affine path exactly, its selected supervision mix exhibits the same qualitative trade-off between weak and return supervision predicted by the theory.

5.3 Prediction and portfolio performance

Hybrid supervision improves the use of larger model capacity. We compare Weak, direct Strong, and theory-guided W2S while fixing $m_w = 32$ and varying the strong-model width $m_s \in \{96, 256, 1024, 2048\}$. Figure 7 reports the rolling out-of-sample loss gap relative to direct Strong in the upper row and cumulative D10–D1 returns in the lower row. Appendix B.4 gives the full evaluation construction.

The two evaluations reveal complementary gains from hybrid supervision. W2S frequently achieves lower rolling prediction loss than direct Strong, although the advantage varies over time and does not uniformly dominate Weak. The portfolio result is more consistent across capacity: W2S achieves the highest terminal cumulative D10–D1 return for all four strong-model widths. Thus, the benefit of larger capacity need not be recovered by direct return supervision alone; mixing weak and return supervision can improve how that capacity is used.

6 Related Work

Financial machine learning and portfolio construction. Financial machine learning uses flexible models to extract information about expected returns from asset characteristics (Gu et al., 2020; Han et al., 2024; Jensen et al., 2026). Studies of return prediction compare models using out-of-sample measures of forecast accuracy(Gu et al., 2020; Han et al., 2024; Bali et al., 2023). Related work uses predicted returns to rank assets or form portfolios and evaluates the resulting investment performance(Gu et al., 2020; Murray et al., 2024; Jensen et al., 2026). Prediction loss measures forecast accuracy across observations, whereas portfolio performance measures the economic outcome of a particular investment rule; these are distinct evaluations of the same predict-then-optimize pipeline(Kelly et al., 2024; Jensen et al., 2026). We first examine how prediction loss varies with model capacity and financial-panel scale, then evaluate the resulting forecasts under a fixed portfolio rule, without imposing a general relationship between prediction loss and portfolio returns.

Scaling laws under financial dependence. Scaling laws describe how prediction loss changes with model capacity, training-data size, and computational scale (Bordelon et al., 2024; Choshen et al., 2025). Standard formulations typically summarize data scale by a single sample count n(Hoffmann et al., 2022; Bahri et al., 2024). In a financial panel with N assets observed over T dates, the nominal number of asset–date observations is NT (Kelly et al., 2019; Bai, 2003). Common market and factor shocks induce dependence across assets observed at the same date, so adding assets and adding dates reduce different sources of estimation error(Kelly et al., 2019; Bai, 2003; Pesaran, 2006). We therefore study $\mathcal{L}(d, N, T)$, separating approximation loss, diversifiable estimation loss, and the estimation cost generated by common shocks(Bahri et al., 2024).

7 Conclusion

We derive a finance-specific scaling law that distinguishes cross-sectional from temporal information. For the evaluated model pairs, loss and portfolio rankings differ, and model capacities use signals with different persistence. Hybrid W2S mixes weak forecasts with realized returns, eventually improving loss over direct learning and terminal cumulative returns over both endpoints. Several directions remain open. (i) Our scaling law imposes reduced-form rate conditions for diversifiable estimation loss and uses finite-grid power laws to approximate the three model-scale profiles. The closed-form W2S rule is a stylized affine-orthogonal benchmark, while the nonlinear learner's weight is selected by a no-look-ahead finance-law estimator; future work could derive the target-mixing result directly for nonlinear feature-learning models with shared samples. (ii) The observed difference between loss and portfolio rankings does not establish a general link between the two objectives, motivating a theory that connects predictive and investment performance.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, assist in drafting portions of the manuscript, support retrieval and discovery of related work, and assist with the development and verification of mathematical arguments. All AI-assisted material was reviewed and revised by the authors. Suggested references were manually verified against the original sources before inclusion, and all mathematical claims and proofs were independently checked for correctness. We take full responsibility for the final content of this work, including all text, citations, claims, proofs, and other artifacts produced with the aid of generative AI.

REFERENCES

Doron Avramov, Si Cheng, and Lior Metzker. Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619, 2023.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Turan G Bali, Heiner Beckmeyer, Mathis Mörke, and Florian Weigert. Option return predictability with machine learning and big data. *The Review of Financial Studies*, 36(9):3548–3602, 09 2023. ISSN 0893-9454. doi: 10.1093/rfs/hhad017. URL https://doi.org/10.1093/rfs/hhad017.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1982.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10683–10699. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/choshen25a.html.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The review of financial studies*, 33(5):2223–2273, 2020.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

Yufeng Han, Ai He, David E Rapach, and Guofu Zhou. Cross-sectional expected returns: new fama–macbeth regressions in the era of machine learning. *Review of Finance*, 28(6):1807–1831, 11 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

Theis Ingerslev Jensen, Bryan Kelly, Semyon Malamud, and Lasse Heje Pedersen. Machine learning and the implementable efficient frontier. *The Review of Financial Studies*, pp. hhag022, 03 2026. ISSN 0893-9454. doi: 10.1093/rfs/hhag022. URL https://doi.org/10.1093/rfs/hhag022.

Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024.

Bryan T Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019.

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *Advances in Neural Information Processing Systems*, 38:159269–159305, 2026.

Scott Murray, Yusen Xia, and Houping Xiao. Charting by machines. *Journal of Financial Economics*, 153:103791, 2024. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2024.103791. URL https://www.sciencedirect.com/science/article/pii/S0304405X2400014X.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997. doi: 10.1016/S0304-4076(97)00011-0.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

Peter M. Robinson. Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, 23(5):1630–1661, 1995. doi: 10.1214/aos/1176324317.

A Theoretical Proofs

A.1 Financial Scaling of Prediction Loss

This section gives a d-coordinate series construction for the model-scale rate conditions in Section 3. The construction separates unrepresented signal from estimation noise and proves the resulting IID and finance-specific laws. The theoretical scale index d counts coordinates; the empirical analysis treats network width m separately.

A.1.1 Prediction-loss decomposition

For this construction, specialize $\mathcal{F}_d$ to a d-dimensional space with orthonormal basis $\{\phi_1^{(d)}, \ldots, \phi_d^{(d)}\} \subset L_2(P_X)$. Define $\theta_j^{(d)} := \langle f^\star, \phi_j^{(d)} \rangle_2$, so that $P_d f^\star = \sum_{j=1}^d \theta_j^{(d)} \phi_j^{(d)}$, and consider the series estimator

$$\widetilde{\theta}_j^{(d)} := \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \phi_j^{(d)}(X_{it}) Y_{i,t+1}, \qquad \widehat{f}_d := \sum_{j=1}^{d} \widetilde{\theta}_j^{(d)} \phi_j^{(d)}. \tag{18}$$

Lemma 4 (Orthogonal risk decomposition). *For any $\widehat{f}_d \in \mathcal{F}_d$,*

$$\mathbb{E}\big\|\widehat{f}_d - f^\star\big\|_2^2 = \big\|(I - P_d) f^\star\big\|_2^2 + \mathbb{E}\big\|\widehat{f}_d - P_d f^\star\big\|_2^2. \tag{19}$$

For the series estimator in (18),

$$\mathbb{E}\big\|\widehat{f}_d - P_d f^\star\big\|_2^2 = \sum_{j=1}^{d} \mathbb{E}\left[\left(\widetilde{\theta}_j^{(d)} - \theta_j^{(d)}\right)^2\right]. \tag{20}$$

Proof. Write

$$\widehat{f}_d - f^\star = \left(\widehat{f}_d - P_d f^\star\right) - (I - P_d) f^\star.$$

The first term belongs to $\mathcal{F}_d$, while the second belongs to its orthogonal complement. Pythagoras' identity and expectation give (19). For the series estimator,

$$\widehat{f}_d - P_d f^\star = \sum_{j=1}^{d} \left(\widetilde{\theta}_j^{(d)} - \theta_j^{(d)}\right) \phi_j^{(d)}.$$

Orthonormality of the coordinates gives (20). □

The approximation component is $B_d := \|(I - P_d) f^\star\|_2^2$. The theoretical decomposition leaves its variation with model scale unrestricted. The empirical analysis approximates this profile by a power law over the observed width grid.

A.1.2 IID benchmark

We first consider a counterfactual benchmark in which all NT asset–date observations are independent and identically distributed. For coordinate j, define the centered coefficient score

$$\xi_{d,j,it} := \phi_j^{(d)}(X_{it}) Y_{i,t+1} - \theta_j^{(d)}. \tag{21}$$

Because $f^\star(X) = \mathbb{E}[Y \mid X]$ and $\theta_j^{(d)} = \mathbb{E}[\phi_j^{(d)}(X) f^\star(X)]$, each score satisfies $\mathbb{E}[\xi_{d,j,it}] = 0$. Define the aggregate diversifiable estimation complexity of this coordinate model by

$$S_d := \sum_{j=1}^{d} \mathrm{Var}(\xi_{d,j,it}). \tag{22}$$

Assumption 5 (IID variance envelope). *The coordinate scores have finite second moments, and their average variance satisfies*

$$\frac{1}{d} \sum_{j=1}^{d} \mathrm{Var}(\xi_{d,j,it}) \longrightarrow \sigma_{\mathrm{div}}^2 \in (0, \infty). \tag{23}$$

This condition is one specialization of the general component S_d. It requires the average sampling variance per included coordinate to remain stable and implies $S_d \sim \sigma^2_{\text{div}} d$. This is the standard variance envelope behind the $d/(NT)$ estimation rate for series estimators (Newey, 1997).

Lemma 6 (IID estimation rate)**.** *Under IID sampling and finite coordinate-score variances,*

$$V_d^{\text{IID}}(N,T) = \frac{S_d}{NT}. \tag{24}$$

Together with Lemma 4, this gives

$$\mathcal{L}_{\text{IID}}(d;N,T) = B_d + \frac{S_d}{NT}. \tag{25}$$

Under Assumption 5, $S_d \sim \sigma^2_{\text{div}} d$, which recovers the familiar linear series-estimation specialization.

Proof. The coefficient error is

$$\widetilde{\theta}_j^{(d)} - \theta_j^{(d)} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\xi_{d,j,it}.$$

IID sampling and the centering of $\xi_{d,j,it}$ give

$$\mathbb{E}\left[\left(\widetilde{\theta}_j^{(d)} - \theta_j^{(d)}\right)^2\right] = \frac{\text{Var}(\xi_{d,j,it})}{NT}.$$

Summing over the orthonormal coordinates and applying the definition of S_d proves (24). Adding the approximation error from Lemma 4 proves (25), which is equivalent to (6). Assumption 5 gives the stated linear specialization. □

A.1.3 Financial panel

Recall the return decomposition in (2) and define

$$a_j^{(d)} := \mathbb{E}\left[\phi_j^{(d)}(X)b(X)\right], \qquad h_{d,j,it} := \phi_j^{(d)}(X_{it})Y_{i,t+1} - \theta_j^{(d)} - \left(a_j^{(d)}\right)^\top u_{t+1}. \tag{26}$$

Lemma 7 (Financial coordinate-error decomposition)**.** *The coefficient error satisfies*

$$\widetilde{\theta}_j^{(d)} - \theta_j^{(d)} = C_{d,j,T} + H_{d,j,N,T}, \tag{27}$$

where

$$C_{d,j,T} := \frac{1}{T}\sum_{t=1}^{T}\left(a_j^{(d)}\right)^\top u_{t+1}, \qquad H_{d,j,N,T} := \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{d,j,it}. \tag{28}$$

Proof. Add and subtract $\left(a_j^{(d)}\right)^\top u_{t+1}$ inside each asset–date coefficient score. Because this term is shared within a date, averaging over assets leaves $C_{d,j,T}$ unchanged. The remaining centered variation is collected in $H_{d,j,N,T}$. □

We next derive the temporal common-shock rate and characterize its dependence on model scale without imposing unbounded exposure growth. Let $\Gamma(h) := \text{Cov}(u_{t+h}, u_t)$ denote the autocovariance matrix of the mean-zero covariance-stationary process $\{u_t\}$.

Assumption 8 (Temporal covariance)**.** *The common shocks have finite second moments and satisfy one of the following conditions.*

1. Short memory. *The autocovariances are absolutely summable, $\sum_{h=-\infty}^{\infty}\|\Gamma(h)\| < \infty$, and $\Omega := \sum_{h=-\infty}^{\infty}\Gamma(h)$ is positive definite. Set $\tau_T := T$.*

2. Persistent dependence. *For some $\gamma \in (0,1)$ and positive-definite matrix H,*

$$\left\|\Gamma(h) + \Gamma(h)^\top - Hh^{-\gamma}\right\| = o(h^{-\gamma}). \tag{29}$$

Set $\tau_T := T^\gamma$ and $\Omega := H/\{(1-\gamma)(2-\gamma)\}$.

Lemma 9 (Effective temporal information)**.** *Under Assumption 8,*

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} u_{t+1}\right) = \frac{1}{\tau_T}\{\Omega + o(1)\}, \tag{30}$$

where the remainder converges to zero in operator norm.

Proof. Stationarity gives

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} u_{t+1}\right) = \frac{1}{T^2}\left[T\Gamma(0) + \sum_{h=1}^{T-1}(T-h)\{\Gamma(h) + \Gamma(h)^\top\}\right]. \tag{31}$$

Under short memory, multiplying this expression by T and applying dominated convergence yields the long-run covariance Ω. Under persistent dependence, $\sum_{h=1}^{T-1}(T-h)h^{-\gamma} \sim T^{2-\gamma}/\{(1-\gamma)(2-\gamma)\}$. Substitution into (31) yields (30). □

Assumption 10 (Square-integrable factor loadings)**.** *The loading function has finite energy,*

$$\mathbb{E}\|b(X)\|^2 < \infty. \tag{32}$$

For each model scale, define its aggregate common-shock exposure by

$$G_d := \sum_{j=1}^{d} \big(a_j^{(d)}\big)^\top \Omega a_j^{(d)}. \tag{33}$$

Lemma 11 (Bounded common-shock exposure)**.** *Under Assumptions 8 and 10,*

$$\sum_{j=1}^{d} \mathbb{E}\big[C_{d,j,T}^2\big] = \frac{G_d}{\tau_T}\{1 + o(1)\}, \tag{34}$$

and

$$0 \le G_d \le \mathbb{E}\big[b(X)^\top \Omega b(X)\big] \le \lambda_{\max}(\Omega)\mathbb{E}\|b(X)\|^2 < \infty. \tag{35}$$

If the prediction spaces are generated by a nested orthonormal sequence, then G_d is nondecreasing and converges to a finite limit G_∞. If that sequence is complete in $L_2(P_X)$, then $G_\infty = \mathbb{E}[b(X)^\top \Omega b(X)]$.

Proof. Write $\mathrm{Var}(T^{-1}\sum_{t=1}^{T} u_{t+1}) = \tau_T^{-1}(\Omega + R_T)$, where $\|R_T\| = o(1)$ by Lemma 9. Then

$$\sum_{j=1}^{d} \mathbb{E}[C_{d,j,T}^2] = \frac{1}{\tau_T}\left\{G_d + \sum_{j=1}^{d} \big(a_j^{(d)}\big)^\top R_T a_j^{(d)}\right\}.$$

Because Ω is positive definite,

$$\left|\sum_{j=1}^{d} \big(a_j^{(d)}\big)^\top R_T a_j^{(d)}\right| \le \frac{\|R_T\|}{\lambda_{\min}(\Omega)} G_d = o(G_d),$$

which proves (34). For the bound, apply Bessel's inequality componentwise to the vector-valued function $\Omega^{1/2}b(X)$:

$$G_d = \sum_{j=1}^{d} \Big\|\mathbb{E}\big[\phi_j^{(d)}(X)\Omega^{1/2}b(X)\big]\Big\|^2 \le \mathbb{E}\big[b(X)^\top \Omega b(X)\big].$$

The eigenvalue bound gives the final inequality in (35). Under a nested basis, each increment is nonnegative, so G_d increases to a finite limit. Completeness and Parseval's identity give the stated value of G_∞. □

Assumption 12 (Residual averaging)**.** *Along the joint scaling sequence, there exists a nonnegative sequence S_d such that the remaining components in (28) satisfy*

$$\sum_{j=1}^{d} \mathbb{E}\big[H_{d,j,N,T}^2\big] = \frac{S_d}{NT}\{1 + o(1)\}, \tag{36}$$

$$\sum_{j=1}^{d} \mathbb{E}\big[C_{d,j,T} H_{d,j,N,T}\big] = o\left(\frac{S_d}{NT} + \frac{G_d}{\tau_T}\right), \tag{37}$$

The first condition requires residual coefficient scores to average over assets and dates after the common component is removed. The second permits dependence between the common and residual components when their aggregate covariance is asymptotically negligible. Approximate factor models motivate this separation (Bai, 2003; Pesaran, 2006), and characteristic-dependent loadings give the exposures an asset-pricing interpretation (Kelly et al., 2019). The covariance conditions in Assumption 8 cover both short memory and slower information accumulation under persistent dependence (Newey & West, 1987; Robinson, 1995). For the coordinate-series specialization, a stable average residual variance gives $S_d \sim \sigma_{\mathrm{div}}^2 d$. The financial scaling law itself leaves the dependence of S_d on model scale unrestricted.

Lemma 13 (Financial estimation rate)**.** *Under Assumptions 8, 10, and 12,*

$$V_d^{\mathrm{FIN}}(N,T) \sim \frac{S_d}{NT} + \frac{G_d}{\tau_T}. \tag{38}$$

Proof. Equations (20) and (27) give

$$\begin{aligned} V_d^{\mathrm{FIN}}(N,T) &= \sum_{j=1}^{d} \mathbb{E}\big[(C_{d,j,T} + H_{d,j,N,T})^2\big] \\ &= \sum_{j=1}^{d} \mathbb{E}\big[H_{d,j,N,T}^2\big] + \sum_{j=1}^{d} \mathbb{E}\big[C_{d,j,T}^2\big] + 2\sum_{j=1}^{d} \mathbb{E}\big[C_{d,j,T} H_{d,j,N,T}\big]. \end{aligned}$$

Applying Lemma 11 together with (36) and (37) yields (38). □

Proof of Proposition 1. Lemma 4 gives $\mathcal{L}_{\mathrm{FIN}}(d; N, T) = B_d + V_d^{\mathrm{FIN}}(N,T)$. Substitute the estimation rate in Lemma 13. This yields (7). □

A.2 Hybrid Weak-to-Strong Learning

This section proves Proposition 3 from the hybrid target in (9). We first derive a risk identity that allows arbitrary dependence between the pure-W2S and direct endpoints. The proposition follows by imposing the conditional moment restrictions in Assumption 2.

Lemma 14 (Exact risk along an affine W2S path)**.** *Suppose only that the affine relation in (11) holds. Define*

$$\delta_d := P_d(f^\star - f_w), \qquad A_d := \mathbb{E}[\|\eta_d - \delta_d\|_2^2 \mid f_w], \qquad \widetilde{V}_d := \mathbb{E}[\|\zeta_d\|_2^2 \mid f_w], \tag{39}$$

and

$$C_d := \mathbb{E}[\langle \eta_d - \delta_d, \zeta_d \rangle_2 \mid f_w]. \tag{40}$$

Then, for every $\rho \in [0,1]$,

$$\mathcal{R}(\widehat{g}_d(\rho) \mid f_w) = B_d + (1-\rho)^2 A_d + \rho^2 \widetilde{V}_d + 2\rho(1-\rho) C_d. \tag{41}$$

Proof. By the definitions of the endpoint errors,

$$\widehat{g}_d^{\,w} = P_d f_w + \eta_d, \qquad \widehat{f}_d = P_d f^\star + \zeta_d. \tag{42}$$

Substituting (42) into the affine path (11) and subtracting $f^\star$ gives

$$\widehat{g}_d(\rho) - f^\star = -(I - P_d) f^\star + (1-\rho)(\eta_d - \delta_d) + \rho \zeta_d. \tag{43}$$

The first term lies in $\mathcal{F}_d^{\perp}$, while the remaining terms lie in $\mathcal{F}_d$. Its cross-products with those terms therefore vanish pointwise. Squaring (43), conditioning on f_w, and applying the definitions of B_d, A_d, $\widetilde{V}_d$, and C_d proves (41). □

The covariance term in (41) records both conditional bias and dependence between the two endpoint errors. Thus, the quadratic shape survives without conditional orthogonality, although its coefficients and minimizer change.

Corollary 15 (Optimal weight without conditional orthogonality)**.** *Let*

$$Q_d := A_d + \widetilde{V}_d - 2C_d = \mathbb{E}[\|\eta_d - \delta_d - \zeta_d\|_2^2 \mid f_w] \geq 0. \tag{44}$$

If $Q_d > 0$, the risk-minimizing weight on $[0,1]$ is

$$\rho^{\star}_{d,\mathrm{gen}} = \min\left\{1, \max\left\{0, \frac{A_d - C_d}{Q_d}\right\}\right\}. \tag{45}$$

The solution is interior exactly when $A_d > C_d$ and $\widetilde{V}_d > C_d$. In that case, the minimized risk is

$$\mathcal{R}(\widehat{g}_d(\rho^{\star}_{d,\mathrm{gen}}) \mid f_w) = B_d + \frac{A_d\widetilde{V}_d - C_d^2}{Q_d}. \tag{46}$$

If $Q_d = 0$, the two endpoint errors in (43) coincide almost surely, so the risk is constant in ρ.

Proof. Expanding (41) gives

$$\mathcal{R}(\widehat{g}_d(\rho) \mid f_w) = B_d + A_d + 2(C_d - A_d)\rho + Q_d\rho^2. \tag{47}$$

When $Q_d > 0$, the unconstrained minimizer is $(A_d - C_d)/Q_d$. Truncating this value to $[0,1]$ gives (45). It lies strictly between zero and one if and only if $A_d > C_d$ and $\widetilde{V}_d > C_d$. Substitution proves (46). When $Q_d = 0$, (44) implies $\eta_d - \delta_d = \zeta_d$ almost surely, and the affine error in (43) is independent of ρ. □

Proof of Proposition 3. Under the conditional mean restrictions in (12),

$$\begin{aligned} A_d &= \mathbb{E}[\|\eta_d - \delta_d\|_2^2 \mid f_w] \\ &= \|\delta_d\|_2^2 + \mathbb{E}[\|\eta_d\|_2^2 \mid f_w] - 2\langle \mathbb{E}[\eta_d \mid f_w], \delta_d\rangle_2 \\ &= D_d + U_d^w. \end{aligned} \tag{48}$$

The same restrictions give $\widetilde{V}_d = V_d(N,T)$ and

$$C_d = \mathbb{E}[\langle \eta_d, \zeta_d\rangle_2 \mid f_w] - \langle \delta_d, \mathbb{E}[\zeta_d \mid f_w]\rangle_2 = 0. \tag{49}$$

Substituting these identities into (41) proves (14).

Let $H_d := D_d + U_d^w$. When $H_d + V_d(N,T) > 0$, the part of the risk that depends on ρ is

$$q_d(\rho) = (1-\rho)^2 H_d + \rho^2 V_d(N,T). \tag{50}$$

Differentiating gives

$$q_d'(\rho) = -2(1-\rho)H_d + 2\rho V_d(N,T), \qquad q_d''(\rho) = 2\{H_d + V_d(N,T)\} > 0. \tag{51}$$

The unique unconstrained minimizer therefore satisfies

$$\rho_d^{\star} = \frac{H_d}{H_d + V_d(N,T)}, \tag{52}$$

which lies in $[0,1]$ because both costs are nonnegative. Substitution into (50) yields

$$q_d(\rho_d^{\star}) = \frac{H_d V_d(N,T)}{H_d + V_d(N,T)}. \tag{53}$$

Equations (52) and (53) prove (15) and (16).

If $H_d = V_d(N,T) = 0$, (14) equals B_d for every $\rho \in [0,1]$, which proves the proposition's degenerate case.

Finally, if $H_d > 0$ and $V_d(N,T) > 0$, then $\rho_d^\star \in (0,1)$. The endpoint risks exceed the optimized risk by

$$\begin{aligned}\mathcal{R}(\widehat{g}_d(0) \mid f_w) - \mathcal{R}(\widehat{g}_d(\rho_d^\star) \mid f_w) &= \frac{H_d^2}{H_d + V_d(N,T)} > 0,\\ \mathcal{R}(\widehat{g}_d(1) \mid f_w) - \mathcal{R}(\widehat{g}_d(\rho_d^\star) \mid f_w) &= \frac{V_d(N,T)^2}{H_d + V_d(N,T)} > 0.\end{aligned} \tag{54}$$

Hence the interior mixture has strictly lower conditional risk than either endpoint. □

B Empirical Analyses and Implementation Details

B.1 Financial scaling-law estimation

Each month, we retain the 1,800 U.S. stocks with the largest market capitalizations to limit the influence of extreme microcaps. The panel contains 153 characteristics standardized within each cross section and one-month-ahead excess returns (Jensen et al., 2023). We evaluate 294 test months from August 2000 to January 2025. With $p = 153$, the empirical predictor is

$$f_m(x) = \frac{1}{p}\sum_{j=1}^{p}\left[\mathrm{LeakyReLU}\big(WW^\top x; 0.01\big)\right]_j, \qquad W \in \mathbb{R}^{p\times m}. \tag{55}$$

Width, rank saturation, and overparameterization. The predictor adapts the tied-weight reconstruction architecture of Liu et al. (2026). Their representation-scaling analysis maps a larger set of input features into a lower-dimensional hidden space and focuses on hidden dimensions below the input dimension. Our empirical grid spans both $m < p$ and $m \geq p$, which require different interpretations. For $m < p$, increasing m relaxes the constraint $\mathrm{rank}(WW^\top) \leq m$ and can enlarge the set of attainable transformations. For $m \geq p$,

$$\left\{WW^\top : W \in \mathbb{R}^{p\times m}\right\} = \mathbb{S}_+^p, \tag{56}$$

because every positive-semidefinite $p \times p$ matrix admits such a factorization, and adding zero columns extends it to larger values of m. All widths $m \geq p = 153$ therefore share the same attainable set of transformations.

For $m \geq p$, increasing width changes the redundant factorization used in training while the attainable transformation class remains fixed. Width variation in this regime can therefore reflect initialization, finite-step optimization, and implicit regularization. Our experiments do not separate these channels. Persistence across these widths provides a robustness check over alternative overparameterized factorizations.

For each test month t, T denotes the immediately preceding T months, so the rolling training window ends at $t-1$. Within each training month, we use a fixed random seed to order the 1,800-stock universe and retain the first N stocks. The same seeded ordering is used across values of N within that month, so the smaller cross sections are nested within the larger ones. The test cross section remains fixed at all 1,800 stocks for every value of N. We vary $N \in [225, 1800]$, $T \in [20, 240]$, and $m \in [1, 2048]$. All models use the same ten-epoch AdamW protocol. This design yields 1,824 (N, T, m) configurations. We fit the laws to geometric-mean out-of-sample squared loss and parameterize temporal information as $\tau_T = T^\gamma$ for $0 < \gamma \leq 1$. The full-grid estimates summarize loss over both width regimes. Width variation below p captures representation-limited scaling; variation at or above p captures the training effects of redundant factorization.

B.2 Extrapolation to unseen scales

An in-sample fit does not establish that a scaling law can guide resource allocation before the corresponding models are trained. We therefore estimate each law on smaller values of cross-sectional sample size, history length, and model width, and then predict losses at scales that are excluded from estimation. The rolling evaluation months remain fixed, so this experiment isolates extrapolation across (N, T, m).

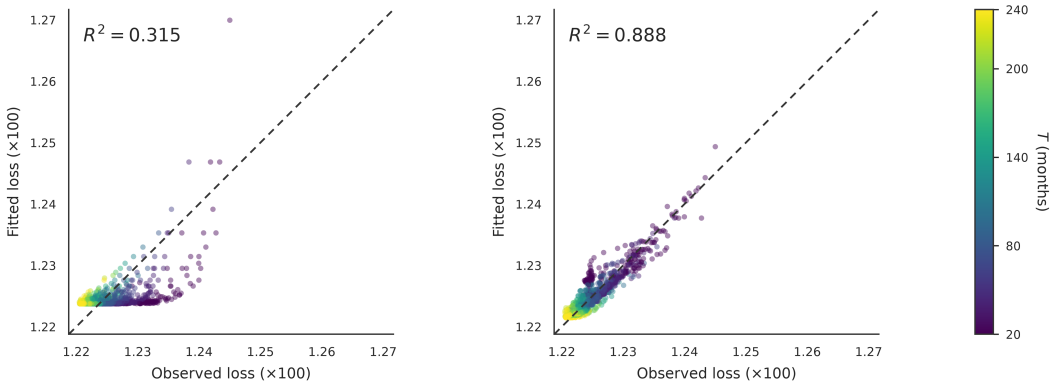

Figure 8: **The finance-specific law fits the observed loss surface more closely.** Each point is one (N, T, m) configuration, color indicates T, and the dashed line indicates equality. The log-loss R^2 is 0.315 for the IID law and 0.888 for the finance-specific law.

Experimental design. The joint training region is

$$\mathcal{G}_{\text{train}} = \{(N, T, m) : N \leq 1350,\ T \leq 180,\ m \leq 96\}, \tag{57}$$

and the joint test region is

$$\mathcal{G}_{\text{test}} = \{1575, 1800\} \times \{200, 220, 240\} \times \{128, 153\}. \tag{58}$$

Every test cell lies beyond the training boundary in all three scaling dimensions. The 12 test losses do not enter parameter estimation, optimization-start selection, or specification selection. We end the joint test at $m = p = 153$, where the attainable rank reaches the input dimension. Larger widths alter the parameterization without expanding attainable rank.

For each specification $M \in \{\text{IID}, \text{FIN}\}$, we estimate

$$\widehat{\theta}_M \in \arg\min_{\theta \in \Theta_M} \sum_{g \in \mathcal{G}_{\text{train}}} w_g \left[\log \overline{L}_g - \log L_M(g; \theta)\right]^2, \tag{59}$$

where $\overline{L}_g$ is the geometric-mean rolling out-of-sample loss in cell g, and w_g is the number of evaluated months. The parameter constraints, deterministic multi-start optimizer, and log-loss objective are identical to those used in the full-grid estimation. We freeze $\widehat{\theta}_M$ before forming $\widehat{L}_g = L_M(g; \widehat{\theta}_M)$ for the held-out cells.

Evaluation metrics. The primary error measure is

$$\text{RMSE}_{\text{OOS}}(M) = \left\{ \frac{1}{|\mathcal{G}_{\text{test}}|} \sum_{g \in \mathcal{G}_{\text{test}}} \left[\log \overline{L}_g - \log \widehat{L}_g\right]^2 \right\}^{1/2}. \tag{60}$$

We also report an out-of-sample coefficient of determination relative to the mean log loss in the corresponding training region,

$$R^2_{\text{OOS}}(M) = 1 - \frac{\sum_{g \in \mathcal{G}_{\text{test}}} (\log \overline{L}_g - \log \widehat{L}_g)^2}{\sum_{g \in \mathcal{G}_{\text{test}}} (\log \overline{L}_g - \overline{\ell}_{\text{train}})^2}, \tag{61}$$

where $\overline{\ell}_{\text{train}}$ is available before the test losses are observed. We do not use the held-out mean loss as the reference forecast because the joint test losses have very little dispersion.

Results. Table 1 reports the results for the 294 rolling test months from August 2000 through January 2025. In the strict joint holdout, the Finance law attains a log-loss RMSE of 0.001423, compared with 0.002198 for the IID law. The Finance specification therefore reduces joint extrapolation error by 35.3%. Its R^2_{OOS} is 0.641, compared with 0.143 for the IID law. The corresponding mean absolute percentage errors are 0.141% and 0.218%, respectively.

Table 1: **The Finance law predicts losses more accurately at unseen scales.** Each row estimates both laws on the indicated smaller-scale cells and evaluates them without refitting. RMSE is computed in log-loss space. The OOS R^2 uses the mean training-region log loss as the reference forecast.

		Log-loss RMSE ↓		RMSE reduction	OOS R^2 ↑	
Held-out scale	Cells (train/test)	IID	Finance	vs. IID	IID	Finance
Joint (N, T, m)	702/12	0.002198	0.001423	35.3%	0.143	0.641
N only	1,080/360	0.002067	0.000770	62.8%	0.148	0.882
T only	1,080/360	0.001983	0.000515	74.0%	0.253	0.950
m only	1,248/192	0.002545	0.001002	60.6%	0.277	0.888

The joint result is not driven by one scaling coordinate. We repeat the exercise by withholding the larger values of N, T, and m separately while retaining the common ranges of the other two coordinates. The N test withholds $N \in \{1575, 1800\}$, the T test withholds $T \in \{200, 220, 240\}$, and the m test withholds $m \in \{128, 153\}$. Relative to the IID law, the Finance law reduces extrapolation RMSE by 62.8% for the N holdout, 74.0% for the T holdout, and 60.6% for the m holdout. The Finance specification also raises R^2_{OOS} in every design. These results show that the common-shock term captures a feature of the loss surface that remains informative beyond the scales used for estimation.

B.3 Information across representation-limited widths

We use 300 rolling out-of-sample months from February 2000 to January 2025 and fix $N = 900$ and $T = 40$. Weak averages forecasts from $m \in \{8, 32\}$, while Strong averages forecasts from $m \in \{64, 128\}$. Let z^W_{it} and z^S_{it} denote these forecasts after standardization within each monthly cross section. We define Strong-specific as the standardized residual from the monthly cross-sectional projection of Strong on Weak:

$$u^S_{it} := \text{Std}_t\left(z^S_{it} - \pi_t z^W_{it}\right), \tag{62}$$

where π_t is the projection slope.

Panel **a** computes same-stock Spearman rank correlations between current forecasts and forecasts lagged by 1, 3, 6, and 12 months. Panel **b** computes monthly turnover for the predicted bottom and top deciles over the common stock universe in adjacent months. For panel **c**, we sort the 153 characteristics into three equally sized fast-moving, medium-persistence, and persistent groups using their measured 12-month same-stock rank persistence, then measure forecast-rank sensitivity by permuting one characteristic at a time. For panel **d**, we use the 13 predefined economic themes in the JKP taxonomy and jointly permute all characteristics within each theme to measure the resulting forecast-rank disruption (Jensen et al., 2023). Error bars in panels **a–c** are 95% HAC confidence intervals computed from the corresponding monthly series.

B.4 Hybrid W2S implementation and evaluation

For Figure 6, we evaluate the model-implied risk over weak widths $m_w \in \{8, 32\}$ and the full grid of strong widths. For each weak–strong pair, we average the risk curve over the rolling evaluation period and normalize it relative to direct strong learning at $\rho = 1$. At test month t, the mixing weight is estimated using information available through $t - 1$ and fixed before deployment in month t.

For Figure 7, we fix $m_w = 32$ and use $m_s \in \{96, 256, 1024, 2048\}$. The upper panels report the 12-month rolling prediction-loss gap relative to direct strong learning, while the lower panels report cumulative returns from the equal-weighted D10–D1 portfolios defined in Section 4.1. All four strong-model specifications allow a higher attainable rank than the weak specification $m_w = 32$. The comparison with $m_s = 96$ remains in the representation-limited regime. The specifications $m_s \in \{256, 1024, 2048\}$ share the full positive-semidefinite transformation class and evaluate whether the benefits of hybrid supervision persist across increasingly redundant factorizations of that class.

B.4.1 FEASIBLE NO-LOOK-AHEAD SELECTION OF THE W2S MIXING WEIGHT

The population decomposition in Proposition 3 motivates the mixing-weight selector. The weak-target discrepancy D_d and the weak-label fitting cost U_d^w enter the optimal weight through their sum. We therefore define the effective weak-side cost

$$W_d := D_d + U_d^w. \tag{63}$$

Let

$$c = (d_w, d, N, T)$$

index a weak–strong pair and its data configuration. For each previously realized test month s and anchor weight

$$a \in \mathcal{A} := \{0, 0.5, 1\},$$

we observe the out-of-sample cross-sectional prediction loss

$$\ell_{s,c}(a) := \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} \left(Y_{i,s+1} - \widehat{f}_{c,s,a}(X_{is}) \right)^2. \tag{64}$$

Each predictor in (64) is trained using only observations available before its test month.

For the decision made at test month t, let $\mathcal{H}_{t,T}$ contain the most recent $v(T)$ realized out-of-sample months strictly before t. We use

$$v(20) = 4, \qquad v(40) = 8, \qquad v(240) = 24.$$

Using only losses in this historical window, we estimate $\widehat{W}_{c,t-1}$ and $\widehat{V}_{c,t-1}$ by

$$\left(\widehat{W}_{c,t-1}, \widehat{V}_{c,t-1}, \{\widehat{\alpha}_{s,c}\}_{s \in \mathcal{H}_{t,T}}\right) \in \arg\min_{\substack{W,V \geq 0 \\ \{\alpha_s\}}} \sum_{s \in \mathcal{H}_{t,T}} \sum_{a \in \mathcal{A}} \left[\ell_{s,c}(a) - \alpha_s - (1-a)^2 W - a^2 V \right]^2. \tag{65}$$

The month fixed effects α_s absorb the component of prediction loss that does not vary with a. They therefore include the approximation term and the realized month-specific loss level. Equation (65) is implemented as nonnegative least squares after demeaning all variables within each historical month.

The resulting observable selection criterion is

$$\widehat{R}_{c,t-1}(\rho) := (1-\rho)^2 \widehat{W}_{c,t-1} + \rho^2 \widehat{V}_{c,t-1}. \tag{66}$$

Any component independent of ρ is omitted because it does not affect the minimizer. The selected continuous mixing weight is consequently

$$\widehat{\rho}_{c,t} = \arg\min_{\rho \in [0,1]} \widehat{R}_{c,t-1}(\rho) = \frac{\widehat{W}_{c,t-1}}{\widehat{W}_{c,t-1} + \widehat{V}_{c,t-1}}, \tag{67}$$

when the denominator is positive. We set $\widehat{\rho}_{c,t} = 1$ when both estimated costs are numerically zero. The final experiment uses the continuous value in (67) rather than rounding it to a discrete grid.

After selecting $\widehat{\rho}_{c,t}$, we train the strong model over its outer rolling training window using the mixed target

$$\widetilde{Y}_{iu}^{\,t} = (1 - \widehat{\rho}_{c,t}) \widehat{f}_{w,t}(X_{iu}) + \widehat{\rho}_{c,t} Y_{i,u+1}. \tag{68}$$

All quantities entering $\widehat{\rho}_{c,t}$ are measurable using information available by the end of month $t-1$. The return realized in test month t is used only to evaluate the selected model and to update the anchor-loss archive for future decisions. During burn-in, the anchor models are still trained and evaluated so that their realized losses populate the archive.

Algorithm 1 Rolling implementation of the W2S selector

Require: $c = (d_w, d, N, T)$, $\mathcal{A} = \{0, 0.5, 1\}$, and $v(T)$

$\mathcal{D} \leftarrow \varnothing$
for test months t in chronological order **do**
  Fit $\{\widehat{f}_{c,t,a} : a \in \mathcal{A}\}$ using data through $t-1$
  $\mathcal{H}_{t,T} \leftarrow \{t - v(T), \ldots, t-1\}$
  if $\mathcal{D}$ contains all losses over $\mathcal{H}_{t,T} \times \mathcal{A}$ **then**
    $Q_t(W, V, \alpha) \leftarrow \sum_{s \in \mathcal{H}_{t,T}} \sum_{a \in \mathcal{A}} [\ell_{s,c}(a) - \alpha_s - (1-a)^2 W - a^2 V]^2$
    $(\widehat{W}_{c,t-1}, \widehat{V}_{c,t-1}) \in \arg\min_{W,V \geq 0} \min_{\{\alpha_s\}} Q_t(W, V, \alpha)$
    $\widehat{\rho}_{c,t} \leftarrow \dfrac{\widehat{W}_{c,t-1}}{\widehat{W}_{c,t-1} + \widehat{V}_{c,t-1}}$
    $\widetilde{Y}^{\,t} \leftarrow (1 - \widehat{\rho}_{c,t}) \widehat{f}_{w,t}(X) + \widehat{\rho}_{c,t} Y$
    Fit and deploy $\widehat{f}^{\,\mathrm{W2S}}_{c,t}$ on $\widetilde{Y}^{\,t}$
  After observing test returns, append $\{\ell_{t,c}(a) : a \in \mathcal{A}\}$ to $\mathcal{D}$

B.5 Portfolio performance of theory-guided W2S

Prediction loss is an intermediate criterion rather than the final asset-pricing outcome. We therefore examine whether theory-guided W2S also improves the cross-sectional ranking of realized returns. The weak model has width $m_w = 32$, and we compare it with direct strong models and matched W2S models of width $m_s \in \{96, 256, 1024, 2048\}$. The evaluation uses the complete scaling grid

$$N \in \{225, 450, 675, 900, 1125, 1350, 1575, 1800\}, \qquad T \in \{20, 40, 240\},$$

which contains 24 (N, T) cells. All forecasts are out of sample and cover August 2005 through January 2025, for a total of 234 test months. The same 1,800-stock universe and realized returns are used for Weak, Strong, and W2S within each test month.

Pooled no-look-ahead selection of the W2S weight. The W2S target combines the weak prediction and the realized training label, with $\rho = 0$ corresponding to pure weak-label distillation and $\rho = 1$ to direct strong-model training. For every deployment month t, we estimate the mixing weight using only the preceding 60 months of rolling out-of-sample losses. At a fixed (m_w, m_s, T), the estimator pools the eight values of N and uses the three observable anchor losses at $\rho \in \{0, 0.5, 1\}$. We fit the finance-specific restriction

$$L_{u,N,T}(\rho) = \mu_{u,N,T} + (1-\rho)^2 A_{d,T} + \rho^2 \left(\frac{a_{d,T}}{N} + b_{d,T} \right) + \varepsilon_{u,N,T}(\rho), \tag{69}$$

where u indexes a historical test month, $A_{d,T} = D_d + U_d^w$ is the cost of anchoring the strong model to the weak target, and $a_{d,T}/N + b_{d,T}$ is the sum of diversified and common estimation costs. Cell–month fixed effects $\mu_{u,N,T}$ are removed by within-cell demeaning, and the remaining three coefficients are estimated by nonnegative least squares. The deployed weight is

$$\widehat{\rho}_t(N, T, d) = \frac{\widehat{A}_{d,T,t}}{\widehat{A}_{d,T,t} + \widehat{a}_{d,T,t}/N + \widehat{b}_{d,T,t}}. \tag{70}$$

The model is retrained at this continuous weight before month-t losses and returns are observed.

Portfolio construction and inference. In each test month and each (N, T) cell, we sort stocks from the lowest to the highest out-of-sample forecast and form ten equal-weight portfolios of 180 stocks. D1 contains the lowest predicted returns, and D10 contains the highest. To retain the same estimand as the pooled scaling results, we first average the stock-level portfolio weights across the 24 cells,

$$\bar{w}_{i,t}^{(j)} = \frac{1}{24} \sum_{c=1}^{24} w_{i,t,c}^{(j)},$$

where j indexes Weak, Strong, or W2S, and then compute realized portfolio returns. The long–short strategy is long D10 and short D1. Table 2 reports monthly mean decile returns, the monthly mean D10−D1 return, its annualized Sharpe ratio, and a Newey–West t-statistic for the mean with six lags. Monthly one-way turnover is

$$\mathrm{TO}_t^{(j)} = \frac{1}{2} \sum_i \left| \bar{w}_{i,t}^{(j)} - \bar{w}_{i,t-1}^{(j)} \right|,$$

where a stock has zero prior weight when it enters the sample. Return statistics use all 234 months, whereas turnover uses 233 month-to-month transitions.

Table 2: **Portfolio performance of theory-guided W2S.** W2S raises the pooled D10−D1 spread relative to the matched Strong model at every strong-model width.

	$m_w = 32$	$m_s = 96$		$m_s = 256$		$m_s = 1024$		$m_s = 2048$	
Portfolio	Weak	Strong	W2S	Strong	W2S	Strong	W2S	Strong	W2S
D1	0.668	0.674	0.651	0.658	0.642	0.679	0.646	0.706	0.652
D2	0.700	0.698	0.680	0.686	0.688	0.720	0.681	0.717	0.701
D3	0.722	0.736	0.708	0.748	0.715	0.742	0.712	0.743	0.722
D4	0.769	0.759	0.762	0.754	0.763	0.762	0.761	0.736	0.755
D5	0.797	0.771	0.783	0.769	0.784	0.773	0.788	0.777	0.774
D6	0.798	0.788	0.825	0.785	0.817	0.778	0.809	0.780	0.801
D7	0.821	0.805	0.811	0.817	0.817	0.801	0.828	0.798	0.823
D8	0.810	0.807	0.812	0.818	0.816	0.802	0.820	0.793	0.816
D9	0.798	0.801	0.806	0.818	0.811	0.812	0.818	0.816	0.819
D10	0.823	0.869	0.868	0.854	0.855	0.837	0.844	0.840	0.841
D10−D1 mean (%)	0.156	0.194	0.217	0.197	0.213	0.157	0.199	0.134	0.189
Sharpe ratio	0.209	0.264	0.294	0.266	0.283	0.212	0.257	0.179	0.243
NW t-statistic	1.05	1.38	1.46	1.44	1.47	1.19	1.32	0.98	1.24
Turnover (%)	77.7	75.2	73.7	75.1	74.6	76.0	75.0	76.0	75.2

Notes: D1 and D10 contain the lowest and highest forecasted returns. Decile entries and D10−D1 means are monthly returns in percent. Portfolio weights are first averaged across the same 24 (N, T) cells as the scaling results. D10−D1 is long D10 and short D1. Newey–West standard errors use six lags. Monthly one-way turnover is $\frac{1}{2} \sum_i |w_{i,t} - w_{i,t-1}|$.

Results. The weak model produces a monthly D10−D1 return of 0.156%. Direct strong training raises the spread at widths 96 and 256, but the spread subsequently falls to 0.157% at width 1024 and 0.134% at width 2048. Thus, increasing width alone does not monotonically improve portfolio ranking. In contrast, W2S produces spreads of 0.217%, 0.213%, 0.199%, and 0.189% across the four strong widths. Relative to the matched direct strong model, these differences are approximately 2.3, 1.6, 4.1, and 5.5 basis points per month, respectively. The larger average gains at widths 1024 and 2048 are consistent with weak anchoring being more useful under more redundant parameterizations.

The improvement comes mainly from the bottom of the forecast distribution. Across all four widths, W2S leaves the D10 return close to that of the matched strong model but lowers the realized return of D1. It therefore improves the separation between predicted losers and winners rather than mechanically raising every decile return. W2S also raises the annualized Sharpe ratio at every width and lowers one-way turnover by 0.5–1.5 percentage points relative to the matched strong model.

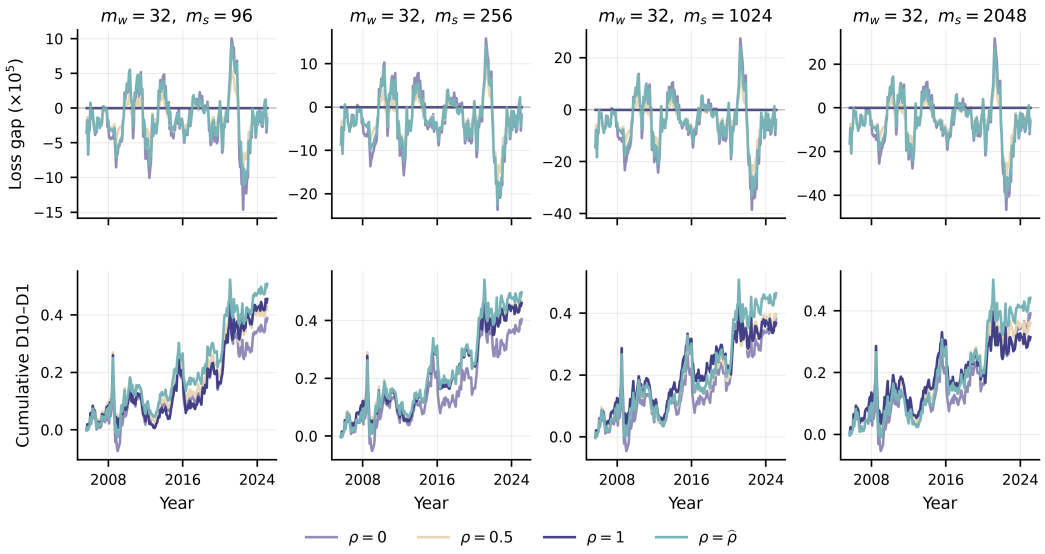

Figure 9: **Fixed and selected W2S mixing weights produce different prediction and portfolio rankings.** The columns fix the weak-model width at $m_w = 32$ and vary the strong-model width over $m_s \in \{96, 256, 1024, 2048\}$. The top row reports the rolling out-of-sample loss gap relative to direct strong training at $\rho = 1$, with negative values indicating lower loss. The bottom row reports realized cumulative D10–D1 returns. Here, $\rho = 0$ is pure distillation into the strong architecture, $\rho = 0.5$ is an equal mixture of weak predictions and realized returns, and $\widehat{\rho}_t$ is selected using only information available before test month t. Results are averaged across the balanced 24-cell (N, T) grid. The selected rule has the highest realized terminal cumulative return at all four widths, although paired return differences are not statistically significant at conventional levels.

B.6 Fixed and Selected W2S Mixing Weights

To distinguish the value of the estimated mixing rule from that of an arbitrary convex combination, we explicitly compare four training targets:

$$\rho \in \{0,\ 0.5,\ 1,\ \widehat{\rho}_t\}. \tag{71}$$

Here, $\rho = 0$ trains the strong architecture using only weak pseudo-labels, $\rho = 0.5$ assigns equal weight to weak pseudo-labels and realized returns, and $\rho = 1$ is direct strong-model training using only realized returns. The final specification uses the time-varying weight $\widehat{\rho}_t$ selected from information available before test month t, as described in Section B.4.1. Thus, $\rho = 0$ is a pure-distillation strong model and should not be confused with the weak model itself.

We fix the weak width at $m_w = 32$ and report strong widths $m_s \in \{96, 256, 1024, 2048\}$. Every method is evaluated on the same 24-cell grid formed by eight cross-sectional sample sizes and three history lengths, using the same rolling test months, stock universes, and training seed within each matched comparison. Because the selector requires 60 prior out-of-sample months, the common evaluation period is August 2005 through January 2025, comprising 234 months. For prediction loss, we average log MSE equally across the 24 (N, T) cells within each month and exponentiate the result, yielding a geometric-mean out-of-sample MSE. For portfolio performance, we first average portfolio weights equally across the same 24 cells and then compute realized returns. The portfolio outcome is the equal-weighted D10–D1 return. Sharpe ratios are annualized, and return t-statistics use Newey–West standard errors with six lags.

In the realized sample, the selected rule produces the highest D10–D1 return and annualized Sharpe ratio among the four specifications at every reported strong width. Relative to direct training, its monthly return differences are 2.26, 1.62, 4.12, and 5.46 basis points for $m_s = 96, 256, 1024, 2048$, respectively. The corresponding Newey–West t-statistics for the paired monthly return differences are 0.61, 0.63, 1.01, and 1.31. The return differences are therefore directionally consistent but not statistically significant at conventional levels.

The selected rule also reduces out-of-sample loss relative to direct training at all four widths. The Newey–West t-statistics for the paired monthly log-MSE differences are -2.25, -2.63, -2.95, and -3.29. Pure distillation, $\rho = 0$, often attains the lowest MSE, whereas $\widehat{\rho}_t$ attains the highest realized portfolio return. This difference is consistent with the two criteria emphasizing different aspects of the forecast: average squared error evaluates the full cross section, whereas D10–D1 returns depend primarily on extreme forecast rankings. Finally, the cell–month mean of $\widehat{\rho}_t$ ranges from 0.273 to 0.349 across strong widths, indicating that the estimated rule does not mechanically collapse to any one of the three fixed benchmarks.

Because $\widehat{\rho}_t$ is estimated from the three fixed anchor weights, this comparison evaluates the deployed selector against those anchors but does not establish that the retrained nonlinear predictor follows an exact affine path in ρ.

B.7 Estimation and Stability of the Financial Scaling Exponents

Estimation. We estimate the empirical loss specification

$$L_{\mathrm{FIN}}(m; N, T) = A m^{-\kappa_{\mathrm{app}}} + \frac{\sigma_{\mathrm{div}}^2 m}{NT} + \frac{Q m^{\delta}}{T^{\gamma}}, \qquad \delta := 1 - \beta, \tag{72}$$

where effective temporal information is not estimated freely at each T; it is restricted to $\tau_T = T^{\gamma}$. For each configuration $g = (N, T, m)$, we first take the geometric mean of the rolling out-of-sample loss and then jointly estimate all coefficients by minimizing

$$\frac{1}{\sum_g w_g} \sum_g w_g \left[\log \overline{L}_g - \log L_{\mathrm{FIN}}(m_g; N_g, T_g)\right]^2, \tag{73}$$

where w_g is the number of test months in cell g. We impose $A > 0$, $\sigma_{\mathrm{div}}^2, Q \geq 0$, $\kappa_{\mathrm{app}} > 0$, $\delta \in [0, 1]$, and $\gamma \in (0, 1]$. We estimate δ directly and report $\widehat{\beta} = 1 - \widehat{\delta}$. Nonnegative least squares at a grid of deterministic exponent values supplies starting points, after which all parameters are jointly optimized by constrained nonlinear least squares.

Confidence intervals. The main estimate uses the single training seed shared by all nineteen widths, 294 rolling test months from August 2000 through January 2025, eight values of N, twelve values of T, and model widths from 1 through 2048. We construct confidence intervals with 500 moving-block bootstrap replications using 12-month blocks. Each replication applies the same resampled month indices to every (N, T, m) cell, thereby preserving the common time-series variation and the dependence across scaling cells. We then reconstruct the geometric-mean loss surface and re-estimate all model parameters. All 500 replications produced valid solutions, and Table 3 reports percentile 95% intervals.

Table 3: **Block-bootstrap estimates of the financial scaling law.**

Parameter	Estimate	95% CI
$\delta = 1 - \beta$	0.163	$[0.063,\ 0.241]$
β	0.837	$[0.759,\ 0.937]$
γ	0.441	$[0.089,\ 0.758]$
log-loss R^2	0.888	$[0.723,\ 0.903]$

The point estimate implies slowly increasing common-shock exposure with model width and sublinear accumulation of temporal information. The bootstrap intervals exclude the boundaries for both exponents, although the interval for γ remains wide. The estimate of δ reaches a boundary in 1.6% of replications, while $\widehat{\gamma}$ reaches a boundary in 0.2%. The approximation exponent remains close to zero, with a 95% interval of approximately $[0, 0.0029]$. We therefore interpret the fitted loss surface and the interior common-shock exponents over the observed grid, rather than treating them as universal asymptotic constants.

Subperiod and grid stability. We evaluate temporal stability on three overlapping windows using the same training seed and complete nineteen-width grid. We separately test grid sensitivity on the main 294-month sample. Besides the complete grid, we use two disjoint alternating grids:

$$\mathcal{G}_A: \quad N \in \{225, 675, 1125, 1575\}, \qquad T \in \{20, 60, 100, 140, 180, 220\},$$
$$\mathcal{G}_B: \quad N \in \{450, 900, 1350, 1800\}, \qquad T \in \{40, 80, 120, 160, 200, 240\}.$$

The alternating grids contain no common N or T values and therefore provide a more demanding check than dropping a single grid point.

Table 4: **Subperiod and grid stability of the Finance law.**

Sample	$\widehat{\delta}$	$\widehat{\beta}$	$\widehat{\gamma}$	log-loss R^2
Panel A. Overlapping 15-year windows				
2000–2015	0.123	0.877	0.273	0.847
2005–2020	0.148	0.852	0.467	0.880
2010–2025	0.181	0.819	0.563	0.888
Panel B. Alternative (N, T) grids				
Full grid	0.163	0.837	0.441	0.888
Alternating grid A	0.203	0.797	0.495	0.878
Alternating grid B	0.134	0.866	0.568	0.928

The fit of the Finance law is stable across these perturbations: log-loss R^2 remains between 0.847 and 0.928. The estimate of β lies between 0.797 and 0.877, while γ lies between 0.273 and 0.568. The subperiod windows overlap and should therefore be interpreted as window-shift robustness rather than as independent structural-break tests. Taken together, the results support the stability of the fitted loss surface, but not a claim that β and γ are precisely estimated or constant structural parameters.

B.8 Component-Level Validation of the Financial Scaling Law

The overall fit of the financial scaling law does not by itself establish that each component has the predicted shape. Different terms may compensate for one another in a joint nonlinear fit. We therefore estimate the three components sequentially and test their model-scale restrictions separately. Throughout this empirical subsection, the theoretical model-scale index is evaluated at $d = m$; thus, every plotted value of d corresponds to an observed factorization width. The grid contains both the representation-limited and rank-saturated regimes, so full-grid width profiles need not represent changes in function-class capacity.

Sequential decomposition. The financial scaling law is

$$\mathcal{L}(d; N, T) = Ad^{-\kappa} + \frac{\sigma_{\mathrm{div}}^2 d}{NT} + \frac{G_d}{T^\gamma}. \tag{74}$$

For each fixed pair (d, T), we first estimate the cross-sectional regression

$$L(d, N, T) = I_{d,T} + \frac{c_{d,T}}{N}. \tag{75}$$

The slope identifies the diversifiable component through $\widehat{S}_{d,T} := T\widehat{c}_{d,T}$. The intercept removes the term that vanishes with $1/N$ and satisfies

$$I_{d,T} = C_d + G_d T^{-\gamma}. \tag{76}$$

For each width, we fit this relation across history lengths while imposing a common temporal exponent γ. This step produces $\widehat{C}_d$ and $\widehat{G}_d$ without imposing their model-scale restrictions.

Figure 10 validates the temporal step of the decomposition. Across most widths, the estimated intercept falls quickly at short histories and then flattens as T increases. A common temporal exponent captures this curvature while allowing the level and magnitude of the temporal component

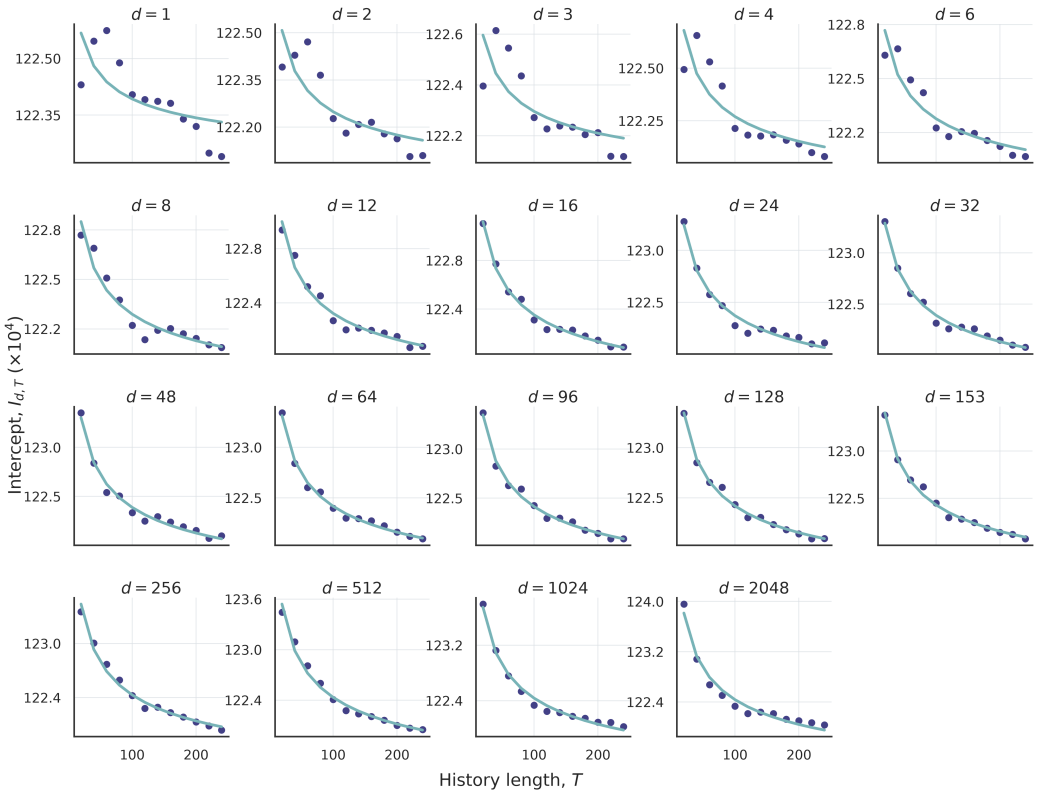

Figure 10: **The loss remaining after cross-sectional diversification declines systematically with history length.** Each panel fixes factorization width $m = d$. Points show the estimated intercept $\widehat{I}_{d,T}$ from the first-stage regression of loss on $1/N$. Curves show the second-stage fit $\widehat{I}_{d,T} = \widehat{C}_d + \widehat{G}_d T^{-\widehat{\gamma}}$. The levels C_d and G_d vary freely across widths, while γ is common to all panels. The sequential fit gives $\widehat{\gamma} = 0.345$ and $R^2 = 0.969$.

to differ across widths. This result supports the $T^{-\gamma}$ specification before imposing any power law in d.

We next evaluate two theoretical model-scale restrictions and one finite-grid approximation:

$$C_d = Ad^{-\kappa}, \qquad S_d = \sigma_{\text{div}}^2 d, \qquad G_d \approx Qd^{1-\beta} \quad \text{for } d \in \mathcal{D}_{\text{obs}}. \tag{77}$$

Confidence intervals are computed from 500 moving-block bootstrap replications with 12-month blocks.

Results. Figure 11 separates the evidence for the three restrictions. Panel **a** gives $\widehat{\kappa} = 0.0015$. The approximation component declines with width, although the estimated decline is very slow over the observed model grid.

Panel **b** provides weaker evidence for the diversifiable-noise restriction. The estimates from different history lengths do not collapse to a common constant, and uncertainty is especially large for small models. Thus, the linear restriction $S_d = \sigma_{\text{div}}^2 d$ receives only partial support.

Panel **c** provides clearer evidence for width-dependent common-shock exposure. The sequential estimates are $\widehat{\beta} = 0.759$ and $\widehat{\gamma} = 0.345$. Since $1 - \widehat{\beta} > 0$, the common-shock component increases at a sublinear rate over the observed width grid. Since $\widehat{\gamma} > 0$, longer histories reduce this component. The component-level evidence therefore supports the common-shock channel most strongly, while the approximation gain is small and the linear diversifiable-noise restriction is less stable. These full-grid profiles are descriptive model-width relationships. For $m \geq p$, they can reflect parameterization and training dynamics rather than further expansion of the attainable transformation class.

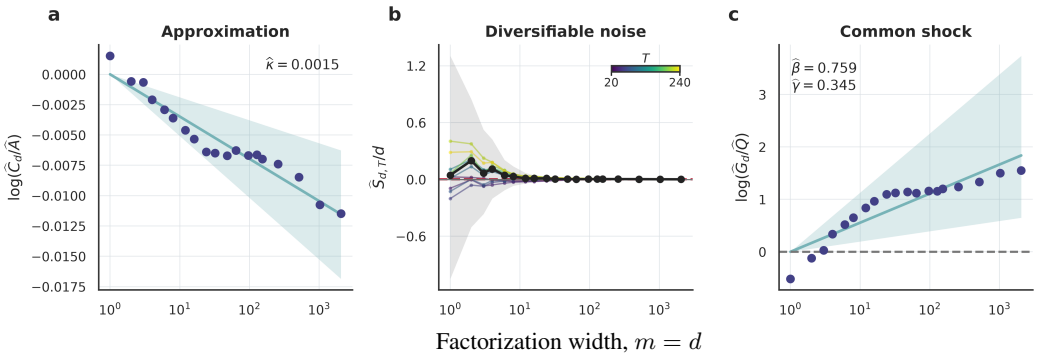

Figure 11: **The common-shock component exhibits the clearest model-width profile, while the diversifiable restriction receives only partial support.** Panel **a** plots the estimated approximation component against model width and fits $C_d = Ad^{-\kappa}$. The common normalization $\bar{A}$ affects the plotted level but not the estimated slope. Panel **b** plots $\widehat{S}_{d,T}/d$ across widths, with colors denoting history length T. The restriction $S_d = \sigma^2_{\text{div}} d$ predicts a constant level across both d and T. Panel **c** plots the estimated common-shock component and fits the finite-grid approximation $G_d \approx Qd^{1-\beta}$ over the observed widths $\mathcal{D}_{\text{obs}}$. The common normalization $\bar{Q}$ affects only the plotted level, and the dashed line shows the width-invariant benchmark $\beta = 1$. Shaded regions report 95% intervals from 500 moving-block bootstrap replications with 12-month blocks.

B.9 Subperiod Stability of the Financial Scaling Law

We examine whether the fitted financial scaling law is specific to one sample period by repeating the complete estimation on three overlapping 15-year windows. Each window uses the common model-width grid and the same data construction, training protocol, and fitting objective as the full-sample analysis. The figures below report temporal scaling, cross-sectional scaling, and observed-versus-fitted loss within each window. Together, they test whether the distinction between cross-sectional breadth and temporal depth remains visible across different market periods.

Early period (2000–2015). Figures 12–14 show the fitted loss surface for the 2000–2015 window. The finance-specific law reproduces the sustained loss reductions from longer histories and the faster flattening of gains from larger cross sections. Its fitted losses also remain close to the observed configurations across model widths.

Middle period (2005–2020). Figures 15–17 repeat the analysis for 2005–2020. The same qualitative asymmetry remains: additional dates produce persistent loss reductions, whereas gains from additional assets flatten more quickly. The finance-specific fit continues to track the observed loss surface across model widths.

Recent period (2010–2025). Figures 18–20 show the corresponding results for 2010–2025. Despite the shift to a more recent sample, the fitted surface preserves the distinct roles of assets and dates and remains close to the observed losses over the model-width grid.

Across the three windows, the finance-specific law attains log-loss R^2 between 0.847 and 0.888. The estimated finite-grid common-exposure exponent β ranges from 0.819 to 0.877, while the temporal exponent γ ranges from 0.273 to 0.563. These results show that the fitted loss surface and its central cross-sectional–temporal asymmetry are stable to shifting the sample window. Because the windows overlap, this exercise should be interpreted as a robustness check rather than as three independent tests or a formal test for structural breaks.

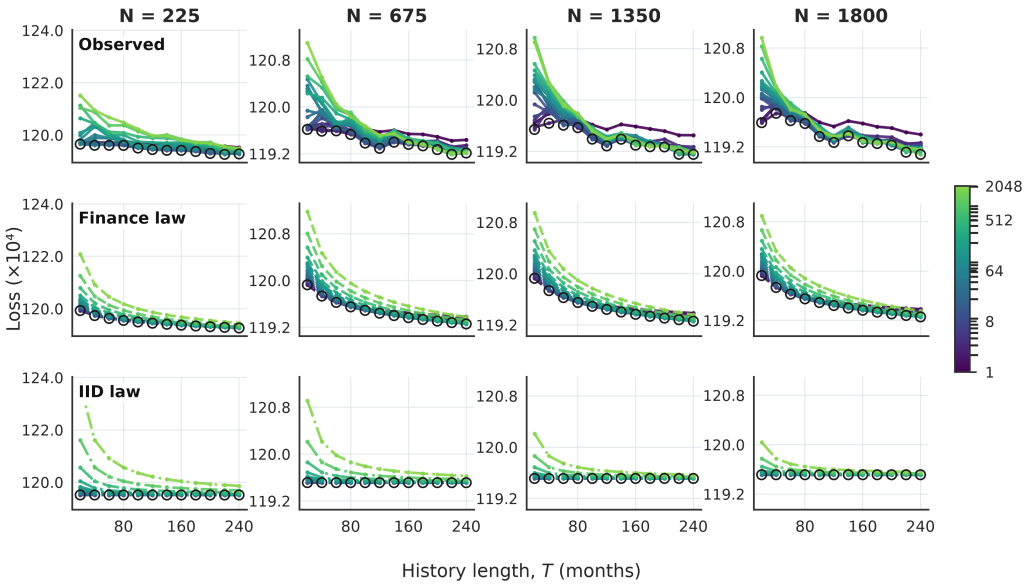

Figure 12: **Temporal scaling remains visible in the 2000–2015 window.** Columns fix N, while rows show observed loss, the finance-specific fit, and the IID fit as T increases. Colors denote width m, and open circles mark the loss-minimizing width.

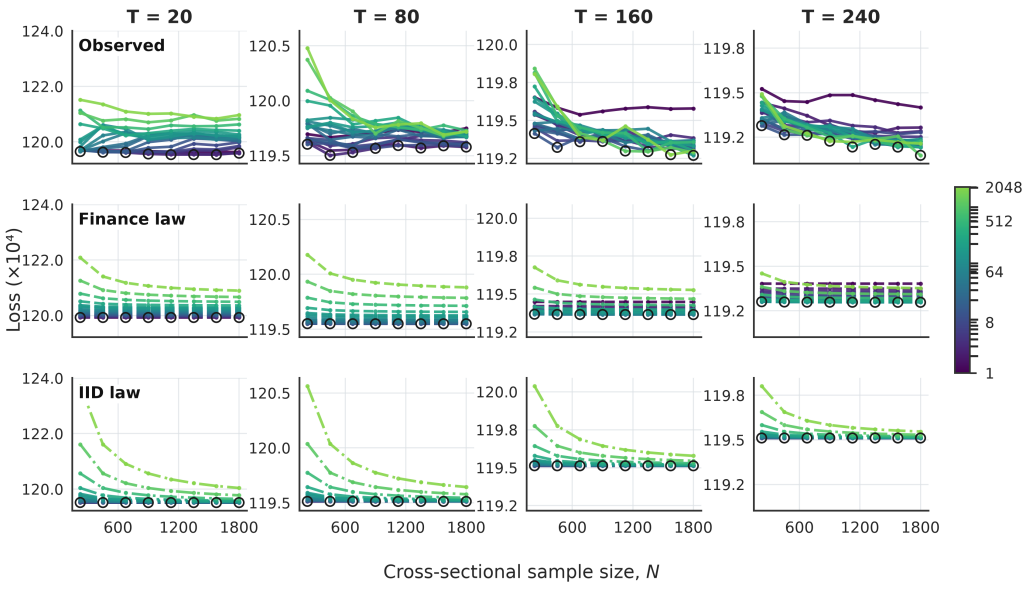

Figure 13: **Cross-sectional gains flatten in the 2000–2015 window.** Columns fix T, while rows show observed loss, the finance-specific fit, and the IID fit as N increases. Colors denote width m, and open circles mark the loss-minimizing width.

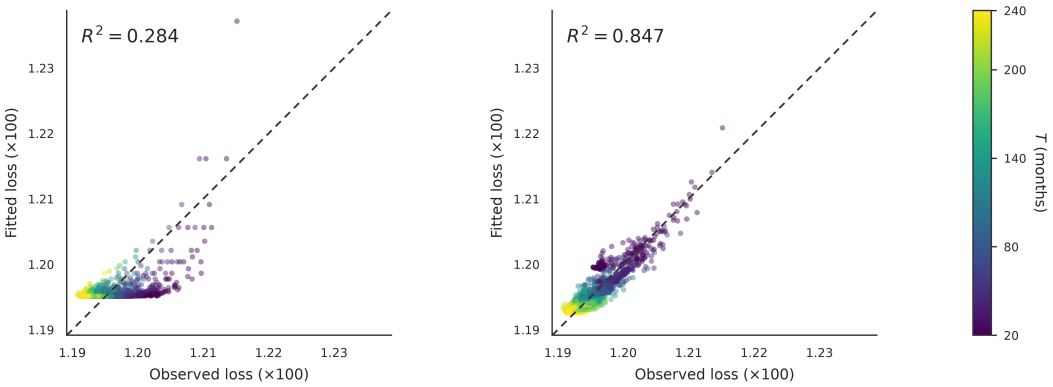

Figure 14: **Observed and fitted losses align in the 2000–2015 window.** Each point represents one (N, T, m) configuration, color denotes T, and the dashed line indicates equality.

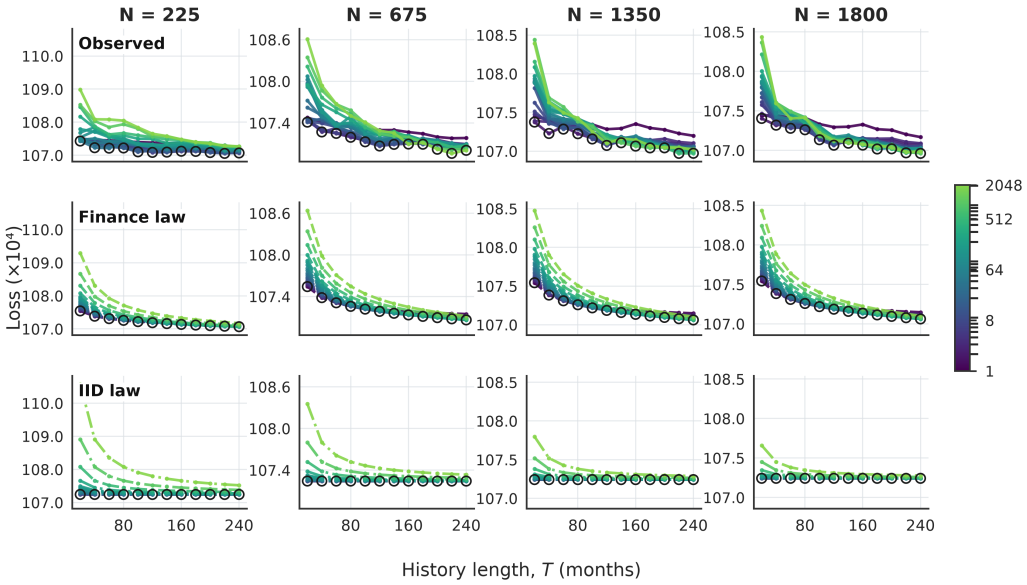

Figure 15: **Temporal scaling remains visible in the 2005–2020 window.** Columns fix N, while rows show observed loss, the finance-specific fit, and the IID fit as T increases. Colors denote width m, and open circles mark the loss-minimizing width.

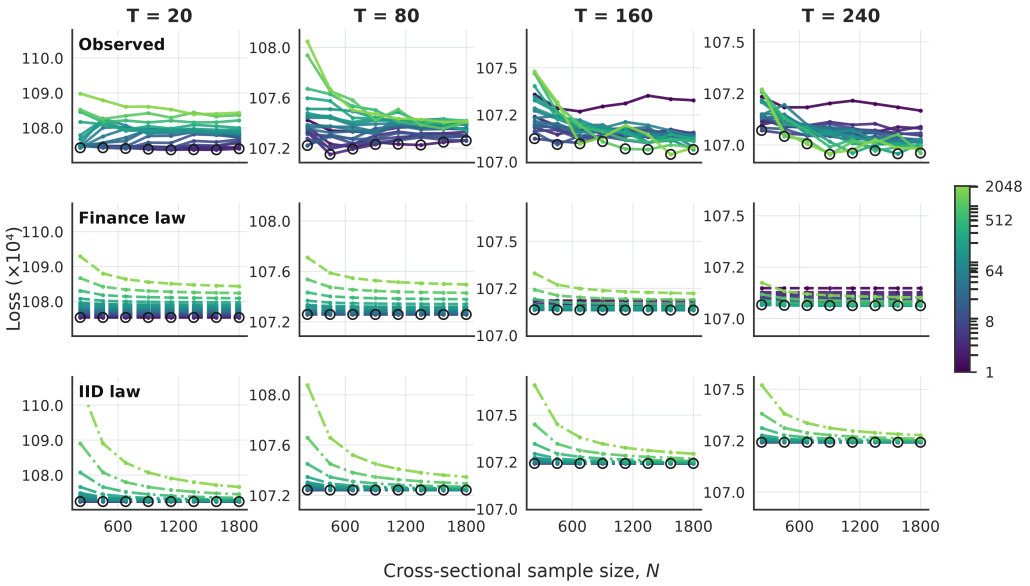

Figure 16: **Cross-sectional gains flatten in the 2005–2020 window.** Columns fix T, while rows show observed loss, the finance-specific fit, and the IID fit as N increases. Colors denote width m, and open circles mark the loss-minimizing width.

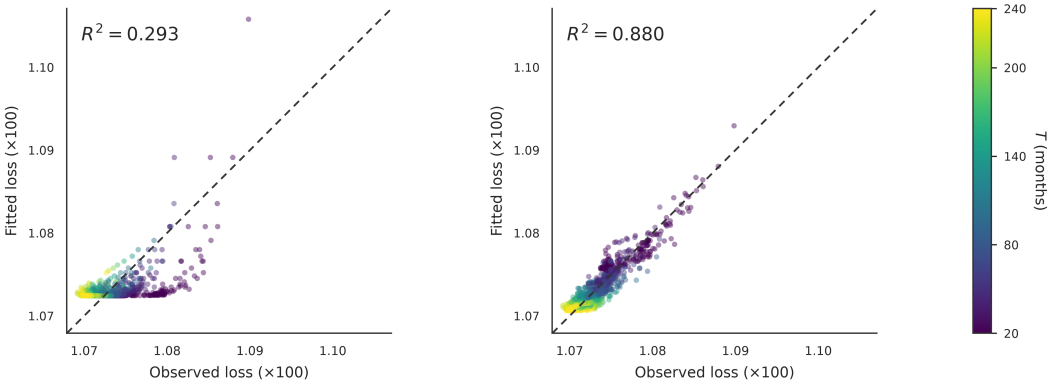

Figure 17: **Observed and fitted losses align in the 2005–2020 window.** Each point represents one (N, T, m) configuration, color denotes T, and the dashed line indicates equality.

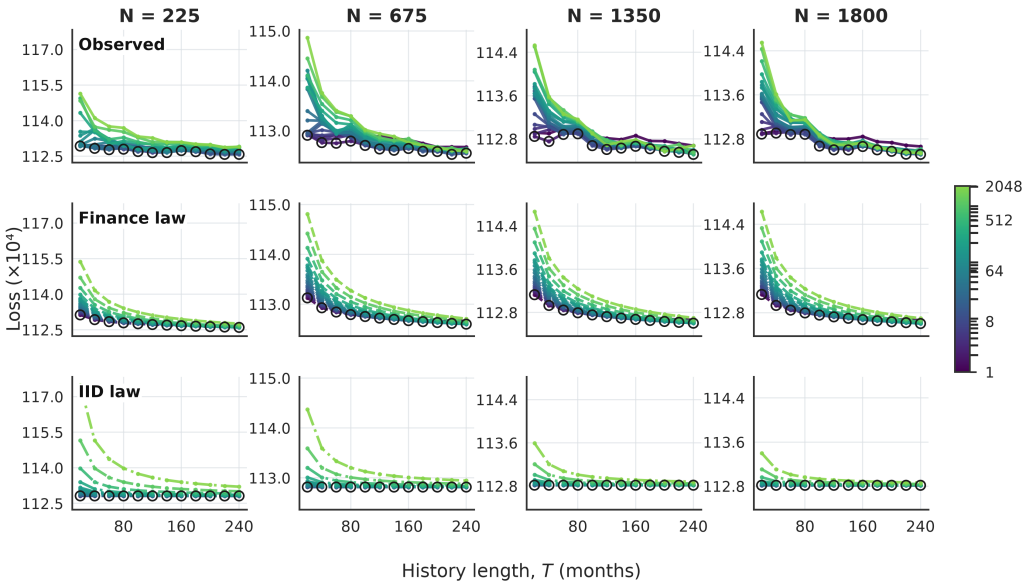

Figure 18: **Temporal scaling remains visible in the 2010–2025 window.** Columns fix N, while rows show observed loss, the finance-specific fit, and the IID fit as T increases. Colors denote width m, and open circles mark the loss-minimizing width.

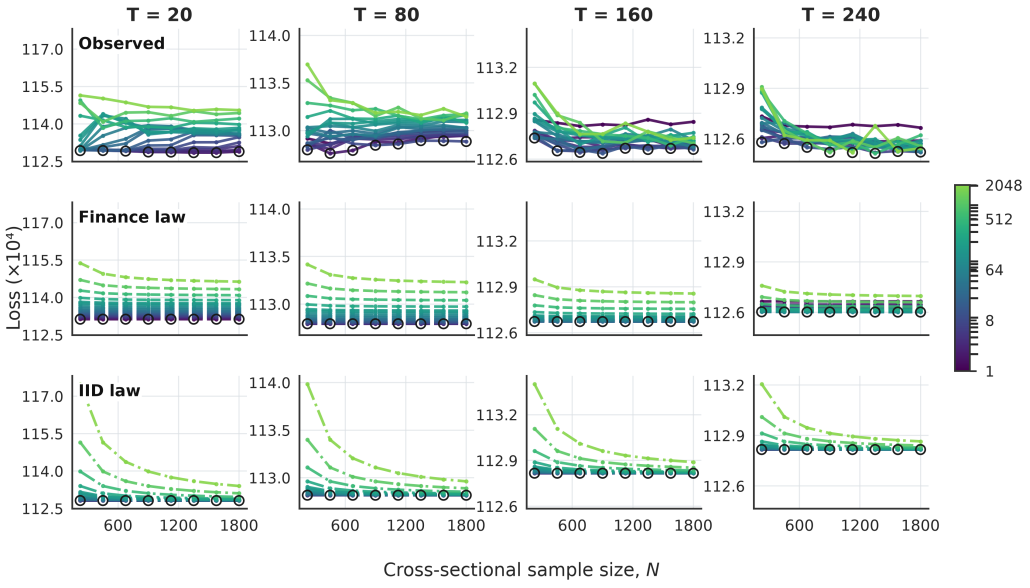

Figure 19: **Cross-sectional gains flatten in the 2010–2025 window.** Columns fix T, while rows show observed loss, the finance-specific fit, and the IID fit as N increases. Colors denote width m, and open circles mark the loss-minimizing width.

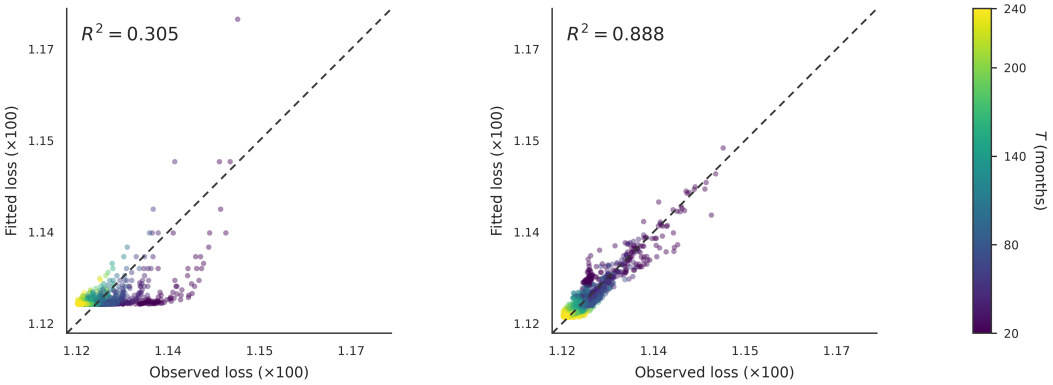

Figure 20: **Observed and fitted losses align in the 2010–2025 window.** Each point represents one (N, T, m) configuration, color denotes T, and the dashed line indicates equality.