

7458140-ranks-into-returns-neural-learning-to-rank-for-equity-selection-under-real-world

AVTORJI: Lucas Jaccard

POVZETEK: Raziskava obravnava, kako različne re-encodiranje metode učenja in rangiranja vplivajo na vrstni red in vrhunec vloženosti v podjetja. Izpeljana je, da je prehod od rangiranja s centrirano percentilsko rangiravo do inverzne normalne rangirave poveča vrednost vrhunec vloženosti za 0,370 (4,46 t-stat). Tega vpliva deli v dve: 0,224 z pomenom (0,020) se nanaša na skalarni prehod, medtem ko 0,146 (0,065) se nanaša na oblikovni prehod. Raziskava je izvedena na podlagi 144 kakovostnih mer na 1000 največjih ameriških podjetij, z 272 mesečnimi obravnavami od aprila 2003 do novembra 2025.

METODA: Raziskava je izvedena na podlagi 144 kakovostnih mer na 1000 največjih ameriških podjetij, z 272 mesečnimi obravnavi. Različne re-encodiranje metode učenja in rangiranja so preizkusljiva na različnih rangiravah. Raziskava je uporabila gradient-boosted ranker kot kontrolno metodo, ki je vrnila natančno nulto razliko med različnimi rangiravami.

KLJUČNE UGOTOVITVE:

- Prehod od centrirane percentilne rangirave do inverzne normalne rangirave poveča vrednost vrhunec vloženosti za 0,370 (4,46 t-stat).
- Skalarni prehod poveča vrednost vrhunec vloženosti za 0,224 (0,020).
- Oblikovni prehod poveča vrednost vrhunec vloženosti za 1,95% (0,015).
- Vrednost vrhunec vloženosti poveča od 18,51% na 22,14%, medtem ko volatilitet zmanjša od 16,81% na 15,06%.
- Vpliv izvajanja je večji, saj vsako od 12 specifičnih postavk izgublja med 0,45 in 0,56 vrednosti vrhunec vloženosti.

POMEN ZA TRGOVANJE: Raziskava je uporabna za trgovce, saj ukazuje na pomembnost pravilnega rangiranja in re-encodiranja podatkov pri izbiri vključenih podjetij v poročilo. Skalarni in oblikovni prehodi lahko povečajo vrednost vrhunec vloženosti, kar je pomembno za učinkovito trgovino.

Ranks into Returns: Neural Learning-to-Rank for Equity Selection under Real-World Frictions

Lucas Jaccard

23.08.2026

MSc in Finance, Asset & Risk Management
HEC Lausanne, University of Lausanne (UNIL)

Supervisor: Marc-Aurèle Divermois, PhD

Expert: Flavio Simoes da Silva

Disclaimer: The work is the responsibility of the author; in no way does it engage the responsibility of the University, nor of the supervising Professor.

A reproducibility repository is available at:

https://github.com/quantquant-max/Master_thesis

Abstract

A monotone re-encoding of firm characteristics leaves every within-date ordering intact and still changes the book a neural ranker trades. Pooled across five differentiable ranking objectives, moving from a centred percentile rank to an inverse-normal rank is worth +0.370 of annualised gross Sharpe ratio ($t = 4.46$) and +0.307 net. A third arm, carrying the percentile shape at the inverse-normal scale, splits that gap in two. The larger half belongs to the change of scale, +0.224 ($p = 0.020$), which multiplies every rank by a constant and cannot reorder anything. The change of shape at matched scale contributes +0.146 ($p = 0.065$).

The setting is 144 point-in-time characteristics on a top-1000 United States equity universe, 272 consecutive monthly rebalances from April 2003 to November 2025, and 23 purged walk-forward folds. Representation, ranking objective, weighting scheme and traded tail width each vary alone; the universe, the monthly label, the architecture, the five-seed ensemble and the execution rules do not. The training target stays on the percentile encoding, so the contrast runs through the feature channel alone. A gradient-boosted pairwise ranker on the identical surface acts as an invariance control and returns a contrast of exactly zero: its two arms fit the same trees, agreeing on topology, split feature, split gain and every leaf value.

The effect is close to invisible to the statistic this literature usually reports. Mean rank information coefficient moves from 0.0383 to 0.0403, a difference no test separates from zero ($p = 0.44$). What moves is the steadiness of the ordering, whose dispersion falls from 0.1063 to 0.0892 and whose information ratio rises from 0.361 to 0.452. The two steps then act on different quantities. Scale moves IC dispersion and book volatility, leaving both means within their standard errors. Shape moves the annualised return, by 1.95 percentage points ($p = 0.015$), leaving dispersion and volatility within theirs. The gain also belongs to the ensemble rather than to any single fit, since one seed on its own recovers +0.192 of the +0.370.

Annualised return rises from 18.51% to 22.14% and volatility falls from 16.81% to 15.06%. Execution is the larger force. Replayed at the next open under spread, square-root impact, stock-loan and financing costs (plain rules, deliberately left unoptimised), every one of the twelve sweep specifications loses between 0.45 and 0.56 of Sharpe ratio, more than the representation choice adds, and every one reaches its deepest drawdown in 2025 without recovering inside the sample.

Part I states the research questions, positions them in the learning-to-rank, empirical asset pricing and implementability literatures, and fixes the data, protocol and methodology. Part II reports the representation, construction and friction results. Part III

examines subperiods, the drawdown common to every specification, and the alternative explanations that bear on the interpretation.

Keywords: learning to rank; rank information coefficient; inverse-normal ranks; long-short equity; portfolio construction; transaction costs.

JEL classification: C45; G11; G12; G17.

Acknowledgements

I thank my supervisor, Marc-Aurèle Divernois, for guiding this thesis from its first outline to its final form, and for the latitude to pursue the question in my own way.

I also thank Rean Morinaj and Abdul J. Bakari, whose long discussions with me went well beyond this thesis and left a mark on how I think.

Contents

I	Motivation, Theory, and Design	9
1	Introduction	10
1.1	From a good ranking to a good portfolio	10
1.2	Four choices along the chain	11
1.3	Research questions	12
1.4	Contributions	13
1.5	What this thesis does not claim	14
1.6	Structure	15
2	Literature Review and Conceptual Framework	16
2.1	Cross-sectional prediction and the ranking view	16
2.2	Learning to rank	18
2.2.1	Pairwise objectives	18
2.2.2	Listwise objectives	19
2.3	Cross-sectional feature representation	19
2.4	From scores to portfolios	21
2.5	Trading frictions and implementability	21
2.6	Conceptual framework and hypotheses	22
3	Data, Universe, and Research Protocol	25
3.1	Data sources	25
3.2	Panel construction	26
3.3	Investable universe and rebalancing	27
3.4	Label and return conventions	28
3.5	Point-in-time feature surface	28
3.6	Walk-forward protocol and evaluation sample	31
3.7	Inference protocol	33
4	Methodology	38
4.1	Cross-sectional setup and notation	38
4.2	Feature representation treatments	38

4.2.1	What the three arms identify	40
4.3	Ranking objectives	42
4.3.1	Why the target encoding is held fixed	43
4.4	The fixed surface	44
4.4.1	Architecture and optimisation	44
4.4.2	Training length: a fixed budget, and its control	45
4.5	Gradient-boosted benchmark and the invariance argument	45
4.6	A naive benchmark	46
4.7	Linear anchors	47
4.8	From scores to portfolios	48
4.8.1	Sector demeaning and selection	48
4.8.2	Two weighting schemes	48
4.8.3	Leg decomposition	50
4.9	Tail width	50
4.10	Execution regimes and the cost model	50
4.11	Reproducibility and audit trail	53

II Core Empirical Evidence 54

5 Representation and Ranking-Objective Results 55

5.1	Estimands and the three arms	55
5.2	Ordering quality	57
5.3	Economic translation	60
5.3.1	What this means for a practitioner	64
5.3.2	Which channel carries each step	65
5.3.3	The training budget	66
5.3.4	The path, turnover and the common drawdown	68
5.4	Why the two do not look alike	70
5.5	The gradient-boosted invariance control	74
5.6	Calibration against external reference points	75
5.6.1	A naive rule	75
5.6.2	The linear anchors	76
5.6.3	The gradient-boosted ranker as a comparator	77
5.7	Mechanism in the fitted function	78
5.8	Pooling across objectives	79
5.9	What is carried forward	82

6	Portfolio Construction and Tail Width	83
6.1	Width and weighting as treatments	83
6.2	Results across widths	83
6.3	Selection against sizing	86
6.4	The two legs	88
6.5	Choosing a width	90
7	Trading Frictions and Implementability	92
7.1	Replay design	92
7.2	Cost assumptions	93
7.3	The gross-to-net bridge	94
7.4	Capacity	98
7.5	What survives	99
III	Robustness, Failure, and Interpretation	102
8	Robustness, Subperiods, and Failure	103
8.1	Subperiods	103
8.2	The worst common drawdown	106
8.3	The legs under stress	111
8.4	Alternative explanations	114
8.5	Initialisation noise and the seed dimension	116
8.6	What this evidence does not establish	116
9	Conclusion	119
9.1	Answers to the research questions	119
9.2	What the criteria chain shows	123
9.3	Implications for practice	124
9.4	Limitations	126
9.5	Future research	130
	Bibliography	133
A	Feature Surface	140
A.1	Point-in-time availability	152
A.1.1	Where the invariance control puts its weight	154
B	Protocol and Configuration	156
B.1	Walk-forward calendar	156
B.2	Resolved configuration	156

B.3	Gross ceiling and cost scaling	157
B.4	Soft-rank temperature	158
B.5	Hyperparameters	158
B.6	Where each arm sits on the activation curve	159
C	Full Results	163
C.1	Representation by objective	163
C.2	Multiplicity-adjusted inference	165
C.3	Per-seed results	168
C.4	The short-squeeze reading of the terminal episode	169
C.5	Tail-width tables	170
C.6	Gradient-boosted benchmark	173
	Use of generative artificial intelligence	175

List of Tables

3.1	Data sources	25
3.2	Feature families	29
3.3	Fixed research protocol	32
4.1	Representation arms	39
4.2	Objective parameters	43
4.3	Architecture and optimisation	44
4.4	Portfolio-construction parameters	49
4.5	Cost-model parameters	53
5.1	Ordering quality across objectives and representation arms	58
5.2	Inference on the ordering statistics	58
5.3	Representation levels by objective	61
5.4	Paired representation contrasts	61
5.5	Decomposition of the representation contrast	63
5.6	The two steps move different things	72
5.7	Reading the invariance control	74
5.8	The naive benchmark and the linear anchors	76
5.9	Representation mechanism diagnostics	78
5.10	May the objectives be pooled?	80
6.1	Gross performance across tail widths	84
6.2	Representation contrast by weighting scheme and width	87
6.3	Leg decomposition by specification	88
7.1	Gross-to-net bridge	96
7.2	Representation contrast before and after costs	96
7.3	Capacity	98
8.1	Subperiod definitions	104
8.2	Subperiod performance	104
8.3	The benchmark by regime	104
8.4	Anatomy of the worst common drawdown	108

8.5	Factor exposures of the combined book	112
8.6	The decomposition against the stopping date	118
A.1	Admitted characteristic inventory	140
A.2	Point-in-time availability by source	154
A.3	Where the invariance control puts its weight	155
B.1	Where each arm sits on the activation curve	162
C.1	Complete cross of specifications, gross	164
C.2	Complete cross of specifications, net	165
C.3	Inference tiers and multiplicity control	167
C.4	Seed dispersion and the ensemble-size curve	169
C.5	Turnover across the traded-fraction grid	172
C.6	Gradient-boosted control: configuration and invariance audit	174

List of Figures

1.1	Research design pipeline	12
3.1	The purged walk-forward schedule	32
4.1	The three representation arms	39
5.1	Cumulative performance by representation arm	69
5.2	Representation contrasts by objective	69
5.3	Rolling mean rank IC by representation arm	70
5.4	Ranking quality and portfolio quality across the grid	73
5.5	Dispersion of the rank IC and volatility of the book	74
6.1	Performance and turnover across the traded fraction	86
7.1	The gross-to-net bridge	94
8.1	Performance by subperiod and representation arm	106
8.2	Equity curves and the worst common drawdown	107
8.3	Cumulative contribution of each leg	113
8.4	Cumulative return of all ten bins	114

Part I

Motivation, Theory, and Design

Chapter 1

Introduction

1.1 From a good ranking to a good portfolio

The economic content of a cross-sectional equity signal lies in the ordering it imposes on the investable universe. A manager who buys the top fraction of a sorted list and sells the bottom fraction is indifferent to the level of any individual forecast, provided that outperformers are placed above underperformers (Grinold and Kahn, 2000). Empirical practice has organised itself around that observation. Signal quality is commonly judged by the rank information coefficient, the period-by-period Spearman correlation between predicted scores and subsequently realised returns (Coqueret and Guida, 2020), and the machine-learning asset-pricing literature reports portfolio sorts as the economic counterpart of its statistical results (Gu et al., 2020).

The difficulty is that the rank information coefficient and the traded portfolio are separated by a sequence of decisions that receive far less attention than the model itself. A score vector must first be mapped into a selected set of names, then into position sizes, then into a book of a particular concentration, and finally into returns net of the cost of trading it. Each of those steps can amplify, dilute, or reverse a difference that looks small in a full-cross-section metric, and the ranking of competing models is not preserved along the chain: a specification ordering the whole cross-section slightly better may hold a book that is more concentrated, more volatile, or more expensive to turn over. The evidence that trading costs reorder nominally profitable strategies is by now substantial: Korajczyk and Sadka (2004) locate the capital at which momentum exhausts its own liquidity, Novy-Marx and Velikov (2016) show that realistic costs materially reduce, and for high-turnover strategies eliminate, the profitability of documented anomalies, and Chen and Velikov (2023) find little that survives publication and costs together. The size of the wedge is nevertheless contested: the live execution record of Frazzini et al. (2018) places realised costs for a large manager an order of magnitude below several academic estimates. Jensen et al. (2026) push the argument furthest, showing that accounting for

implementation shifts the frontier itself rather than translating it downward.

This thesis takes that transmission chain as its object of study, asking where along the path from the numerical representation of the inputs to net performance an apparent advantage survives or disappears. Four dimensions are varied, and Figure 1.1 shows what is held fixed around them.

1.2 Four choices along the chain

Four decisions structure the empirical work. Three of them are argued over in the literature. The representation, in the empirical asset-pricing applications, is inherited as a preprocessing convention rather than defended, and it is taken first for that reason.

Feature representation. The population target of a ranking objective is invariant to any strictly increasing transformation of the features, but its optimisation need not be: a neural scorer trained by stochastic gradient descent has no reason to be indifferent to the spacing of its inputs, which enters the local curvature of the loss surface and the relative magnitude of the gradients that different regions of the cross-section generate (LeCun et al., 2012). A linearly centred percentile rank places names uniformly on an interval and therefore compresses the extremes relative to the middle. An inverse-normal rank, defined by $z_i = \Phi^{-1}(r_i/(n+1))$, preserves the ordering exactly while stretching the tails. Since the traded book is formed from the tails, a transform that changes nothing ordinally may still change the fitted score geometry precisely where it matters economically.

Those two mappings differ in the size of the numbers they produce as well as in their spacing, so the axis carries a third arm: the uniform spacing at the normal scale, blown up by an analytic constant. It splits their contrast into a pure scale term and a shape term at matched scale, which that pair cannot separate on its own, and a convergence-matched re-estimate repeats the contrast at a second operating point.

Ranking objective. Learning-to-rank losses were developed for information retrieval, where the governing metrics are position-discounted and top-heavy (Järvelin and Kekäläinen, 2002; Burges, 2010). The rank information coefficient is neither: it weights the cross-section symmetrically and penalises an inversion in proportion to the rank distance it spans. Pairwise objectives such as RankNet treat every inversion as equally costly (Burges et al., 2005), whereas LambdaRankIC weights each pair by the change in rank correlation that swapping it would produce (Lin et al., 2026a). Listwise objectives read the entire cross-section at once. These families need not agree, and if the representation effect is genuine it should not be confined to a single loss.

Weighting scheme. Equal weighting within each leg is the most transparent possible mapping from a selected set to a book: it reflects the model’s selection and nothing else. Inverse-volatility sizing is closer to practice and usually improves risk-adjusted performance, but it also mixes the model’s selection with a risk-management rule that is known to help on its own (Barroso and Santa-Clara, 2015; Moreira and Muir, 2017). A representation effect that appears only under inverse-volatility weighting is a different, and weaker, finding than one that survives equal weighting, because the former may reflect an interaction with the sizing rule and not better stock selection. Both schemes are therefore run, with equal weighting as the benchmark, on its strong record against optimised alternatives once estimation error is taken seriously (DeMiguel et al., 2009).

Traded tail width. Narrowing the traded book concentrates the signal but raises idiosyncratic risk, turnover, and execution cost, and it shrinks the number of names across which estimation error can diversify. The width that is best gross therefore need not be the width that is best net.

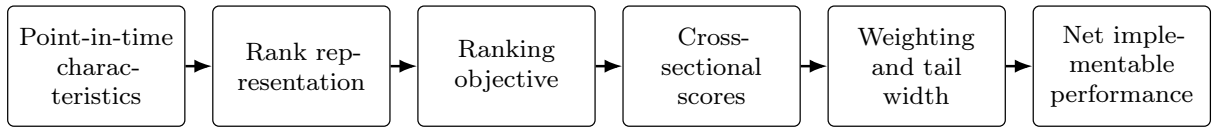


Figure 1.1: The transmission chain studied in this thesis. The panel, label, walk-forward schedule, architecture, seeds, and execution rules are held fixed throughout. The representation, the ranking objective, the weighting scheme, and the traded fraction are varied one stage at a time, so that a difference observed at the end of the chain can be traced to the choice that was varied; Section 1.5 states what that attribution does and does not license.

1.3 Research questions

The thesis addresses six questions, ordered along the transmission chain of Figure 1.1.

- Q1.** Does the numerical mapping applied to within-date feature ranks change the long-short portfolio performance of a neural ranking model, when the ordinal information, the training target, and every other element of the pipeline are held fixed?
- Q2.** Is any such effect a property of one objective, or does it persist across the pairwise and listwise ranking families, and does a gradient-boosted benchmark respond in the way the theory of monotone transformations predicts?
- Q3.** Does the effect survive equal weighting, or does it require inverse-volatility sizing: is the mechanism better stock selection or an interaction with the risk-based weighting rule?
- Q4.** How does the traded tail width affect gross performance, turnover, concentration, and drawdown, and does the preferred width depend on the weighting scheme?

- Q5.** Which combinations of objective, representation, weighting, and width retain economically meaningful performance, judged by the net Sharpe ratio and the capital at which it survives, once an executable entry convention and a capacity-aware cost stack are imposed?
- Q6.** What do the separate long and short legs, the subperiod record, and the drawdown episodes common to every specification reveal about the conditions under which the results hold?

1.4 Contributions

The controlled measurement of feature representation as a modelling choice in cross-sectional ranking is the central contribution. The representation is varied while the target encoding, the feature set, the architecture, the seeds, the folds, and the portfolio rules are held constant, and the contrast is repeated across five ranking objectives, so the finding does not rest on a single loss function, though the five cells share the panel, the folds and the seeds, and are read throughout as correlated readings of one design rather than as independent replications (Section 5.8). The quantity has, to the best of my knowledge, no established magnitude at the portfolio level in the cross-sectional asset-pricing setting, and supplying one is the point of the exercise; Section 2.3 documents how the default encoding is inherited and how rarely the choice is examined.

The gradient-boosted benchmark carries a methodological point. The centred percentile and the inverse-normal encodings are strictly monotone functions of the same within-date uniform rank, so a ranker that splits on thresholds should be indifferent between them up to the discretisation of its split search. That prediction makes the tree a placebo control isolating the residual channels rather than a second test of the same hypothesis; Section 4.5 develops the argument.

The same cell doubles as the study’s only competitive reference point, a role it was not built for and which none of the six questions of Section 1.3 asks. Fitted on the identical surface, at the traded decile under inverse-volatility sizing, the gradient-boosted ranker carries a higher gross Sharpe ratio than every cell of the percentile arm and a lower one than every cell of the inverse-normal arm, while earning a higher return than any cell of either. The neural family changes position relative to it between the two arms while the architecture, the folds and the portfolio rules do not, which is what makes the model-class comparison and the representation result the same finding rather than two. The margin is narrower under the other weighting rule: three inverse-normal cells sit below the control gross under equal weighting, and net the pooled margin changes sign between the two schemes. Section 5.6.3 states the levels, the margin net of costs under both schemes, and the reasons no test is attached to the comparison in either cost regime.

The implementation layer is carried end to end. Frozen score vectors are run through

a deliberately plain execution, including a square-root impact term (Tóth et al., 2011), and the resulting gross-to-net bridge is reported component by component at every tail width. The two legs are reported separately, the sample is decomposed into economically defined subperiods, and the drawdown common to every specification is analysed in its own section.

1.5 What this thesis does not claim

Three limits on interpretation apply throughout Parts II and III.

The evidence comes from one panel, one label, one architecture, and five ranking objectives. What is measured is the effect of the encoding on a five-member rank-averaged ensemble of a single fixed neural surface, estimated on one sample of US common equity, and nothing here establishes that an inverse-normal encoding helps any estimator on any data. The gradient-boosted and linear cells bound the claim from one side, by identifying model families in which the effect does not appear, but they do not extend it to families that were not run. The level comparison against the gradient-boosted cell carries a further limit of its own. It is descriptive throughout, it rests on a single tree configuration that received no search, it is struck at one traded fraction under one weighting rule and its net margin reverses in sign under the other, and that margin is in any case inside what this sample can resolve, so nothing below should be read as establishing that a neural ranker is preferable to a gradient-boosted one in general.

The architecture and optimiser settings follow standard practice for tabular networks and are held fixed across every cell, with no hyperparameter search for any representation arm or any objective. This buys a common operating point on which the treatments can be compared under identical settings, at the cost of reporting no cell at its best attainable configuration, and it leaves open whether the common settings suit the arms unequally. The convergence-matched re-estimate of Table 5.5 answers the sharpest form of that objection, that the fixed budget rather than the encoding produces the contrast, by letting every cell train to its own validation optimum. The contrast survives that change in the respect the objection is about: it keeps its sign in every objective and, gross, about four fifths of its pooled magnitude. It survives it less comfortably in two others. Re-tuning the budget arm by arm makes the two books behave less alike, the standard error of the total rises by more than half in consequence, and the pooled figure clears conventional significance in neither cost regime; and the division of the total between its two steps, which is unchanged gross, comes apart net, where the smaller step is the one left standing. Section 5.3.3 reports both, and reads the widened intervals as a cost of the treatment rather than as evidence against the effect. Whether a full per-cell hyperparameter search would widen or narrow the measured differences remains a question this design does not answer.

The cost model is partly measured and partly assumed, so the net figures are a point estimate under a stated mid-range parameterisation; Section 7.2 identifies which components rest on assumption. The executable entry convention belongs to the same layer. It is required for the net figures to describe a portfolio anyone could have held, and Section 7.3 measures its size directly.

1.6 Structure

The remainder of Part I positions and specifies the study: Chapter 2 reviews the literatures and closes with the hypotheses, Chapter 3 fixes the data, the universe, the label and the protocol, and Chapter 4 defines the treatments, the portfolio construction and the cost model. Part II follows the transmission chain of Figure 1.1: Chapter 5 measures the representation contrast on identical score vectors, so that the criteria of (2.1) can be compared, Chapter 6 varies the traded fraction and the weighting scheme, and Chapter 7 replays the frozen scores through the cost stack. Part III interprets the evidence: Chapter 8 decomposes the sample into regimes, treats the common drawdown as an object of study, and states the alternative explanations, and Chapter 9 answers the six questions and records the limitations. The appendices carry the inventories, the full cross of results, and the reproducibility trail.

Chapter 2

Literature Review and Conceptual Framework

This chapter reviews the five literatures on which the empirical design draws and converts them into testable hypotheses. The coverage extends to what bears on the comparisons of Parts II and III, not to a survey of cross-sectional asset pricing.

2.1 Cross-sectional prediction and the ranking view

The characteristic-based literature has established that a large number of firm-level variables carry information about the cross-section of subsequent returns and that much of that information is shared: Green et al. (2017) count the independent signals among dozens of published characteristics, Freyberger et al. (2020) select nonparametrically which of them add incremental information, and Kozak et al. (2020) and Feng et al. (2020) discipline the resulting factor zoo through shrinkage and through a formal test of each candidate’s marginal contribution, so the zoo requires selection, not accumulation. Machine-learning methods improved on linear benchmarks in this setting mainly by accommodating nonlinearity and interaction, at two levels. At the level of the prediction function, trees and feed-forward networks fit the conditional mean of returns on characteristics directly (Gu et al., 2020). At the level of the conditional factor model, Kelly et al. (2019) instrument time-varying loadings with characteristics but keep the map from characteristics to loadings linear. The later work relaxes that map: Gu et al. (2021) learn the loadings from characteristics through an autoencoder, and Chen et al. (2024) estimate a stochastic discount factor under a no-arbitrage criterion, with the test assets that identify it constructed adversarially and the conditioning states extracted from macroeconomic series. Kelly and Xiu (2023) survey the field. What that record does not settle is which of the flexible families is preferable. Gradient-boosted trees and shallow networks are usually found close to one another on panels of this kind (Gu et al., 2020), and the same

is reported for tabular problems generally (Gorishniy et al., 2021). Section 5.6.3 locates a preprocessing choice that reverses the gross ordering of the two families on a common surface, a reason to fix and state the encoding before any comparison between model classes is read.

The statistical gain is nevertheless modest on the metric usually reported: monthly out-of-sample R^2 values are typically a fraction of one percent even for the strongest specifications (Gu et al., 2020), which is a natural consequence of the low signal-to-noise ratio in individual stock returns rather than a defect of any particular estimator. Kelly et al. (2024) qualify that reading along a dimension this study holds fixed, showing that heavily parameterised models can keep improving portfolio-relevant performance past the point of interpolation; the common ground is that R^2 is a poor guide to economic value and that a specification should be judged on the criterion the portfolio actually pays for.

A parallel literature qualifies these results. McLean and Pontiff (2016) document that published predictors decay once they are known, Harvey et al. (2016) argue on multiple-testing grounds that many published findings are likely false, and Hou et al. (2020) find that a large share of the published cross-section fails to replicate under uniform protocols, although Jensen et al. (2023) argue on Bayesian grounds that the failure rate is substantially lower once multiple testing is handled hierarchically instead of factor by factor. Either way the evidentiary bar for a new claim is high, and the response is the pre-specified protocol of Section 3.7.

Two features of that record motivate the ranking view. First, if the level of a forecast is estimated poorly but its relative position is estimated less poorly, then an estimator trained to minimise squared error is optimising a criterion that is harder than the one the portfolio actually needs. Second, the standard evaluation metric already reflects this: signals are commonly judged by the rank information coefficient, the Spearman correlation (Spearman, 1904) between predicted and realised orderings (Grinold and Kahn, 2000; Coqueret and Guida, 2020).

A small but growing literature acts on that observation. Poh et al. (2021) show that treating strategy construction as a ranking problem, rather than as prediction followed by sorting, improves realised risk-adjusted performance. Lin et al. (2026b) extend the comparison to the standard empirical asset-pricing setting and report that ranking objectives deliver higher Sharpe ratios and shallower drawdowns than forecast-and-sort alternatives, and Lin et al. (2026a) derive the objective that this thesis carries into its neural setting. Adjacent work reaches the portfolio through a different route, optimising the trading objective directly instead of scoring and then sorting: Guijarro-Ordóñez et al. (2025) learn statistical-arbitrage signals end to end, and Simon et al. (2025) parameterise the portfolio policy itself. Ranking sits between pointwise prediction and end-to-end policy learning: it aligns the training criterion with the cross-sectional ordering that a fractile book consumes, while retaining a score vector that can be frozen and replayed under alternative

construction and execution rules. The present design requires that second property.

2.2 Learning to rank

Learning to rank originates in information retrieval, where the relevant metrics are position-discounted, an asymmetry normalised discounted cumulative gain formalises (Järvelin and Kekäläinen, 2002). Methods are conventionally grouped by the granularity of the loss (Liu, 2009): pointwise, where items are scored independently; pairwise, where inverted pairs are penalised; and listwise, where the loss is defined over the entire ordering at once.

2.2.1 Pairwise objectives

RankNet introduced the pairwise logistic formulation that underlies most of the family (Burges et al., 2005). For an ordered pair (i, j) the modelled probability that i outranks j is a logistic function of the score gap, and the cross-entropy against the true preference produces a gradient that depends only on that gap. Every pair enters the sum with the same weight, which makes RankNet closely related to Kendall’s τ and, for that reason, a reasonable if imperfect proxy for a symmetric rank criterion. LambdaRank modifies it by scaling each pairwise gradient by the change in the evaluation metric that would follow from swapping the two items (Burges et al., 2006; Burges, 2010). The construction is heuristic in origin, but Wang et al. (2018) show that it can be interpreted as bound optimisation of a proper likelihood, which is what makes the choice of metric inside the swap weight analytically meaningful. When the metric is normalised discounted cumulative gain, the resulting weight decays with rank position and concentrates gradient at the top of the list. That suits retrieval, and only partly suits a long-short book that trades both ends of the distribution.

Lin et al. (2026a) address this misalignment directly by deriving the change in rank information coefficient produced by swapping two items and using it as the LambdaRank weight. The resulting weight is proportional to the product of the predicted and realised rank separations, so it grows with distance rather than decaying with position, and it distributes learning effort across the whole cross-section. They show that the resulting objective upper-bounds $1 - \text{IC}$ and implement it as a custom objective inside gradient-boosted trees (Chen and Guestrin, 2016). A margin or hinge formulation offers a third pairwise variant, penalising a pair only until it is correctly ordered by a fixed margin and then withdrawing gradient from it; (4.6) states the form carried into the empirical design. These constructions differ in the weight a pair carries and in the surrogate applied to its score gap, and not in the pairwise reduction itself, which predates all of them: Freund et al. (2003) combine pairwise preferences by boosting an exponential loss over inverted

pairs.

2.2.2 Listwise objectives

Listwise losses evaluate an entire permutation, not a set of pairs drawn from it. Soft ranking makes the rank operator itself differentiable and optimises a smoothed version of the evaluation metric directly (Blondel et al., 2020), the listwise construction most closely aligned with a symmetric rank criterion; in the variant used here the hard rank is replaced by a sum of sigmoids of pairwise score differences, yielding a differentiable approximation to Spearman correlation (Qin et al., 2010). RankCosine instead treats the score and target vectors as directions and penalises the angle between them (Qin et al., 2008), which makes the loss scale-invariant in the scores while remaining sensitive to their relative magnitudes; (4.8) states it. These two are the listwise objectives carried into the empirical design.

For the present study the practical distinction is that pairwise objectives are estimated on a sampled subset of pairs, because the full pair set grows with the square of the cross-section, whereas listwise objectives consume the entire cross-section at each date. The five objectives compared here are therefore best read as practical ranking procedures on a common surface rather than as five losses running under one identical estimator, a point made explicit in Section 4.3.

2.3 Cross-sectional feature representation

Cross-sectional rank normalisation is close to universal in the characteristic-based literature. Mapping each characteristic to its within-date rank removes the influence of outliers, makes variables with different units comparable, and neutralises time-varying dispersion, which is why it appears in work as otherwise different as Kelly et al. (2019) and Freyberger et al. (2020). What is far less standardised, and what is almost never justified, is the choice of scale onto which those ranks are mapped. Kelly et al. (2019) divide each stock’s rank by the number of names with the characteristic present and subtract one half, mapping it into $[-\frac{1}{2}, \frac{1}{2}]$. The machine-learning asset-pricing literature that follows inherits the linear form of that mapping but not its interval: Gu et al. (2020) and Gu et al. (2021) map the same ranks into $[-1, 1]$. The first arm of the representation axis below is therefore not an arbitrary comparator but the field’s default spacing, and the question is whether the default is the right one for a model whose traded positions come from the tails.

The two intervals in use are not interchangeable, and the difference between them is of the kind this thesis measures. A rank on $[-\frac{1}{2}, \frac{1}{2}]$ has standard deviation $1/\sqrt{12} \approx 0.289$ and one on $[-1, 1]$ has $2/\sqrt{12} \approx 0.577$, so a study reporting only that it used centred

ranks has not said enough to place its network at a known operating point; the curve in Section B.6 is steep across the range the two conventions span, and the wider of them sits roughly half way to unit scale. Neither is standardised. That the field’s convention is a choice of *scale* as much as of *shape* is what makes the decomposition of Section 5.1 worth striking.

Two mappings are natural. The centred percentile places the i -th of n names at $r_i/(n+1) - 1/2$, so that the cross-section is uniform on an interval. The inverse-normal or normal-score transform places it at $\Phi^{-1}(r_i/(n+1))$, so that the cross-section is approximately standard normal. Both are strictly increasing functions of the same rank, so they carry identical ordinal information and any rank-based population criterion is indifferent between them; Section 4.2 states them formally.

The indifference does not extend to estimation. For a model that partitions the feature space by thresholds, such as a regression tree, a strictly monotone transform maps every candidate split to a corresponding split and leaves the fitted partition unchanged. For a model that forms weighted sums and passes them through smooth nonlinearities, the spacing of the inputs enters the geometry of the loss surface, which is the classical argument for normalising network inputs (LeCun et al., 2012) and part of the motivation for normalisation layers (Ioffe and Szegedy, 2015). Normalisation layers weaken this channel without closing it, and where they sit decides how much of it they close. The architecture used here carries layer normalisation at every hidden block, but the normalisation follows the activation rather than the linear map, and the activation is not positively homogeneous, so a rescaling of the inputs is transformed before it is normalised and cannot be normalised away. Section 4.2.1 states the routes that survive and Section B.6 measures the one that carries them. On tabular data, well-tuned simple baselines remain competitive with specialised architectures and there is still no universally superior solution (Gorishniy et al., 2021), a further reason to treat input representation as a first-order modelling decision.

The same estimation argument, however, cuts both ways. The two encodings differ not only in the shape of their spacing but in the overall size of the numbers they produce, since the normal score carries $\sqrt{12}$ times the cross-sectional standard deviation of the centred percentile, and for an adaptive optimiser under a finite training budget input magnitude alone can move the fitted model. A contrast between the two encodings therefore bundles a claim about tail spacing with a claim about optimisation scale, and the two must be separated before either can be asserted; that separation is what the third arm of this design delivers, an encoding with the uniform shape at the normal scale.

2.4 From scores to portfolios

A fractile sort is the standard selection rule: names are ordered and the extreme fractions are traded, so the ordering within the selected set is irrelevant to an equally weighted book: only membership matters. That observation has a direct consequence for evaluation. Position-discounted metrics such as normalised discounted cumulative gain reward placing the best names earliest within the traded tail, which is not what an equally weighted decile portfolio pays for; membership accuracy at the traded fraction, and the realised returns of the selected names, are more closely aligned with the economics of the book.

Risk management enters through sizing. Equal weighting is the natural neutral benchmark, with a strong record relative to optimised alternatives once estimation error is taken seriously (DeMiguel et al., 2009). Inverse-volatility sizing conditions position size on an estimate of risk and is closely related to the volatility-managed strategies studied by Moreira and Muir (2017) and to the volatility scaling that Barroso and Santa-Clara (2015) apply to momentum. Because such scaling improves risk-adjusted performance for reasons unrelated to signal quality, a model comparison conducted only under inverse-volatility weighting cannot separate better selection from a favourable interaction with the sizing rule; running both schemes at a common gross exposure is what makes that separation possible.

Narrowing the traded fraction raises the average signal strength of the selected names if the model is informative, but it also increases idiosyncratic risk and turnover and reduces the diversification of estimation error, so the fraction that is best gross need not be the fraction that is best net. And the two legs of the resulting book need not behave alike: short positions in a market with a positive unconditional drift face a headwind, momentum-style strategies suffer severe and asymmetric drawdowns during rebounds (Daniel and Moskowitz, 2016), and a combined figure cannot be interpreted without the legs behind it.

2.5 Trading frictions and implementability

The gap between a paper portfolio and a traded one has been measured repeatedly. Korajczyk and Sadka (2004) show that momentum strategies exhaust capacity at economically relevant fund sizes. Trading costs materially reduce, and in many cases eliminate, the profitability of documented anomalies (Novy-Marx and Velikov, 2016), while the live-trading evidence of Frazzini et al. (2018) places realised costs for a large manager below several academic estimates, which sharpens rather than resolves the calibration question. Optimal combinations of characteristics change once costs are included (DeMiguel et al., 2020), and model comparison itself is cost-dependent (Detzel et al., 2023). Chen and Velikov (2023) put the level of what survives at a few basis points a month once effective

spreads, post-publication decay and the modern trading regime are charged together, and attribute some two-fifths of the gross anomaly return to costs alone. Jensen et al. (2026) take the argument to the frontier itself: a cost-aware objective selects different characteristics from a return-forecasting one, so implementability reorders the opportunity set.

The point applies with particular force to machine-learning signals. Avramov et al. (2023) show that the profitability of ML-based strategies is concentrated in exactly the names where implementation is hardest, namely small, illiquid, distressed, and high-limits-to-arbitrage stocks, and that performance deteriorates markedly once such names are excluded or realistic constraints are imposed. That finding motivates two design commitments. The universe is restricted *ex ante* to large, liquid names, where the cost model is credible and the result cannot be an artefact of microcap exposure (Section 3.3), and every comparison is reported gross and net side by side, so that any reordering induced by implementation is visible.

The cost stack used here follows the standard decomposition into spread and commission, market impact, stock-loan fees, and financing, with a square-root impact term on the metaorder evidence of Tóth et al. (2011) and Bouchaud et al. (2018), whose exponent Sato and Kanazawa (2025) confirm stock by stock. The trade-off between that cost and the timing risk of executing more slowly is the subject of Almgren and Chriss (2001), and this thesis optimises neither. Liquidity is itself priced in the cross-section (Amihud, 2002). Two evaluation choices follow. Execution timing must be stated explicitly, since a strategy cannot trade at the close its signal is formed on, and costs must be reported component by component, since the components scale differently with book size, turnover, and concentration.

2.6 Conceptual framework and hypotheses

The literatures above imply a hierarchy of criteria that are frequently conflated. For a fractile long-short strategy, they are related but not equivalent:

$$\begin{aligned}
 & \text{rank correlation over the full cross-section} \\
 \neq & \text{ membership accuracy at the traded fraction} \\
 \neq & \text{ gross portfolio performance after sizing} \\
 \neq & \text{ net performance after frictions.}
 \end{aligned} \tag{2.1}$$

Each inequality in (2.1) corresponds to a stage of Figure 1.1, and each is a place where a comparison between two specifications can change.

Five hypotheses follow, answering the first, second, third and fifth questions of Section 1.3. The width question is descriptive in what it asks of the grid; its closing clause, whether the preferred width depends on the weighting scheme, is answered descriptively

from the grid in Section 6.2 and carries no hypothesis of its own. The diagnostic question carries no hypothesis.

Hypothesis 1 (Representation). Holding the training target, the feature set, the architecture, the folds, the seeds, and the portfolio rules fixed, the numerical mapping applied to within-date feature ranks affects the gross long-short Sharpe ratio, and the inverse-normal mapping, which separates the extremes, outperforms the centred percentile mapping, which compresses them. *Primary tier for the pooled contrast; secondary tier for the five per-objective contrasts.*

Hypothesis 2 (Objective robustness). If Hypothesis 1 reflects a property of the representation rather than of one specification, the sign of the effect is stable across the pairwise and listwise ranking families. *Secondary tier, under the Romano–Wolf stepdown over the five objectives.*

Hypothesis 3 (Model-class invariance). By the threshold-invariance argument of Section 2.3, the gradient-boosted ranker’s measured representation effect should be small relative to that of the neural rankers, since histogram binning can produce only small departures from exact invariance; a contrast comparable in size to the neural one would count against the invariance argument itself rather than be absorbed by it. *No tier: the cell belongs to no inference family, and is read against the pre-specified mapping of Table 5.7 instead.*

Hypothesis 4 (Selection versus sizing). If the representation effect operates through improved membership of the traded tails, it is present under equal weighting. If it is present only under inverse-volatility sizing, it is better read as an interaction with the weighting rule than as selection. *The equal-weighting branch is settled by the primary endpoint re-struck under equal weighting, which carries no family of its own (Section 3.7); the comparison between the two schemes is exploratory tier, Benjamini–Hochberg at $q = 0.10$.*

Gross differences between specifications need not survive execution. That is the setting the last hypothesis is posed in rather than a claim it tests: how much execution removes is measured by the gross-against-net comparisons of the exploratory family, and whether any particular contrast survives is a question for the tier that contrast belongs to.

Hypothesis 5 (Implementability). The width and weighting combination that maximises gross risk-adjusted performance is not necessarily the one that maximises net performance, because turnover and concentration enter the cost stack nonlinearly. *Exploratory tier, Benjamini–Hochberg at $q = 0.10$.*

Hypothesis 1 through Hypothesis 5 are directional statements about economically meaningful magnitudes rather than sharp point predictions; the tests of Part II are two-sided throughout, so the direction is read from the sign of the estimate rather than built

into the test, and the reported p -values are conservative for the directional claim. The tier note appended to each hypothesis states where its constituent contrasts sit in the multiplicity structure of Section 3.7 and how they are adjusted and reported; it does not by itself fix what counts as support for the directional statement above it. Section 3.7 defines the tiers named above, fixes the primary endpoint, and sets the multiplicity correction applied within each. The pre-specification is a discipline internal to the study, fixed before the definitive estimation run rather than lodged in an external registry; the specifications it fixes are not the survivors of a prior search, and Section 8.4 sets out the limits of that discipline.

Chapter 3

Data, Universe, and Research Protocol

This chapter fixes the empirical environment. Everything described here is common to every specification reported in Parts II and III, which is what allows a difference in outcomes to be attributed to the design choices varied in Chapter 4.

3.1 Data sources

The division of labour among the archives listed in Table 3.1 is the conventional one. The availability rules are stated per source in Section A.1, and the configuration recorded with each run is described in Appendix B.

Table 3.1: Source archives. Spans give the coverage of each archive, not the estimation sample; the evaluated window is defined in Section 3.6.

Role	Source	Key, frequency
Price, volume, quotes, delistings	CRSP US Stock Database, daily stock file	PERMNO $\times$ trading day, 1995–2025
Valuation, profitability, leverage	WRDS Financial Ratios, firm level	PERMNO $\times$ month, 1990–2025
Balance sheet and income statement	Compustat North America, quarterly	gvkey $\times$ report date, 1990–2025
Earnings surprise	I/B/E/S Summary History, surprise file	Ticker $\times$ announcement, 1992–2025
Identifier bridge	I/B/E/S ticker-to-company concordance	Ticker $\times$ company, 1993–2025
Factor benchmark	Fama/French 5 factors and momentum, daily	Trading day, 1963–2026

Everything the strategy observes about the market comes from the CRSP US Stock Database at daily frequency: prices, total returns, trading volume, closing bid and ask quotes, shares outstanding, delisting records, and the Standard Industrial Classification

code. CRSP total returns incorporate distributions, so no adjustment-factor reconstruction is required. Much of the later design is keyed on this one file. Market capitalisation and dollar volume drive the universe screens, the closing quotes price the spread term of Section 4.10, and the SIC code is the root of the industry structure, defining the four-digit peer groups of the industry-relative characteristics and, mapped onto the twelve Fama-French industries, the sector label used for the sector demeaning of Section 4.8. The security begin date marks the start of CRSP coverage rather than a listing event, and the listing-age characteristic of Appendix A is accordingly an age within coverage.

The accounting side comes from two archives that differ in their degree of processing. Compustat North America supplies statement-level quarterly data, the balance-sheet stocks such as total assets, the income-statement flows such as quarterly earnings, and the cash-flow items from which the accrual measure is built. The WRDS financial-ratios file supplies pre-computed firm-level ratios, covering valuation, profitability, and leverage, at monthly frequency and dated on their publication dates. The ratios file also carries both the Compustat and the CRSP identifiers, and is therefore the link through which the company-level archives attach to the security-level panel.

Analyst expectations come from the I/B/E/S summary history, whose surprise file sets reported earnings against the consensus forecast. I/B/E/S identifies firms by exchange ticker, and tickers are reused across companies, so surprises are joined through a CUSIP-anchored concordance that discards any ambiguous ticker rather than resolving it, with the residual matched on the exchange ticker where the concordance is silent and that ticker maps to a single company; 93.5% of announcements resolve. Section A.1 states how it is built.

The Ken French library closes the set with the daily five-factor returns (Fama and French, 1993, 2015) and the daily momentum factor (Carhart, 1997). The same series appear twice in the design: first as the regressors behind the residual-momentum and idiosyncratic-volatility construction of Section 3.5, where the market and downside betas are estimated against the value-weighted market factor on returns in excess of the risk-free rate, and later as the benchmark against which factor-adjusted performance is read in Part II.

3.2 Panel construction

Security selection. The panel is keyed on CRSP’s permanent security number, which is stable across ticker, CUSIP and name changes and is assigned per share class, so a company with two listed classes enters as two securities, each ranked on its own capitalisation and liquidity. The investable set is restricted to ordinary common shares of operating companies incorporated in the United States and listed on the New York, American, or Nasdaq exchanges.

Return and price construction. Daily returns are CRSP’s own total return, with the delisting return compounded in on the delisting date, so a security that leaves the panel is carried to the value its holder received. Where the delisting date carries no surviving observation, the payment is compounded onto the security’s last traded close instead, which is the price it is measured against. CRSP records no delisting return at all for roughly one event in twenty. Those are priced on the delisting *action*: mergers and exchange transfers settle at the last price, while securities dropped from the exchange, liquidated without a final payment, or removed for an unrecorded reason are charged -30% , following Shumway (1997). The events that do carry a measured return show the action separating them cleanly. Across the 21,363 delistings in the archive the mean is $+0.9\%$ for mergers and $+1.1\%$ for exchange transfers, against -13.7% for securities dropped from the exchange.

The trailing momentum and moving-average signals, and the horizon returns that define the prediction target of Section 3.4, are functions of a price level. They read an index built by compounding the total return within each security. The ratio of that index at two dates recovers the realised total return between them, and the construction is immune to splits, since a split is a price event and not a return event.

3.3 Investable universe and rebalancing

Rebalancing is monthly at business month-end, and the book formed at a rebalance is held until the next one. The holding period and the label window of Section 3.4 are therefore the same object: months that partition the sample without overlap or gap. At each rebalance the universe is the 1000 largest names by market capitalisation among the securities that satisfy three screens: a closing price above \$5, a trailing 20-day average daily dollar volume of at least \$5 million, and at least 252 trading days of price history, the scale of the longest trailing windows.

The screens serve the cost model. In the liquid large-capitalisation segment the parameterisation of Section 4.10 is credible, short positions are generally borrowable, and a long-short book of the size considered here can plausibly be traded. Studies conducted on the unfiltered cross-section typically derive a large share of their gross performance from micro-capitalisation names where these conditions fail (Novy-Marx and Velikov, 2016; Chen and Velikov, 2023). The screens therefore lower the reported gross performance relative to an unrestricted universe, which is the intended direction of the bias.

The volume floor is a nominal threshold applied across three decades, and it binds unevenly: it constrains the eligible set in the early years of the sample and is slack by the end, where the capitalisation ranking does the selecting. The universe is identical across every specification, so this cannot affect any contrast reported here, but its composition is not constant over the sample.

3.4 Label and return conventions

The prediction target is the return to the next rebalance. Writing t^+ for the rebalance date that follows t , the label of stock i at rebalance date t is

$$y_{i,t} = \frac{P_{i,t^+} - P_{i,t}}{P_{i,t}}, \quad (3.1)$$

computed on the total-return index P of Section 3.2. The label window is the holding period of the book formed at t : one month, which runs between 19 and 23 trading days over almost all of the sample, the exception being September 2001, when the exchanges closed for four days and the month holds 15. The length varies between dates but never within one, and the model consumes only within-date ranks of (3.1), so the variation never reaches the estimator; it appears only in the realised monthly returns.

Two return conventions are used and are never mixed. The *gross* convention measures the close-to-close return from rebalance date t to t^+ and applies no costs. It is an idealised benchmark that assumes execution at the same close at which the signal is formed. The *realistic* convention submits the signal after the close of date t , enters at the open of the next trading day, and exits at the close of t^+ , so that the post-signal overnight move is excluded from the strategy return. The overnight gap is the time an investor needs to run inference and stage execution, and the realistic convention forgoes it. Section 4.10 adds the cost stack to that second convention.

On average the entry step is small, since both legs of a dollar-neutral book earn almost the same overnight drift and the market-wide component cancels, leaving only the differential move of the selected names, and Section 7.3 measures it against the rest of the cost stack.

Both conventions exit at the close of t^+ , so the only differences between them are the entry price and the cost stack. The distance between them in the gross-to-net bridge of Section 7.3 is therefore attributable to implementation and not to a change in the return window being measured. The training label is the gross convention, so (3.1) is the quantity the model is fitted to predict as well as the benchmark against which it is scored.

3.5 Point-in-time feature surface

The feature surface carries 144 firm characteristics, summarised by family in Table 3.2. Two rules govern it: when a datum becomes visible, and how it is normalised.

Information availability. A fundamental or analyst figure becomes visible to the model on its reported availability date when that date falls on or after the fiscal pe-

Table 3.2: Families of the 144-characteristic surface, with the number of columns each contributes. Table A.1 in Appendix A lists every characteristic individually with its definition.

Family	Representative content	n
Total momentum	Cumulative log returns from 252, 126, and 63 trading days before the rebalance to 21 days before it (the strict 12-1, 6-1, and 3-1 windows, so the reversal month is excluded), each with a volatility-scaled variant, together with two horizon spreads that measure the slope of the momentum term structure, the object Novy-Marx (2012) shows is not flat.	8
Residual momentum	The same three horizons, with the same one-month skip, cumulated from the residuals of a rolling regression on the five factors, the momentum factor, and an industry return — extending the three-factor construction of Blitz et al. (2011) — each with a volatility-scaled variant.	6
Trend (MACD and moving averages)	Moving-average convergence-divergence signals at three speeds and five lags, their damped composite, four acceleration terms formed as differences between adjacent lags, anchors relative to the 50- and 200-day averages, and the Kaufman efficiency ratio at three horizons, which measures how directly a given net price move was achieved.	25
Mean reversion: residual shocks	Short-horizon residual returns over 5, 10, and 21 days and their standardised counterparts.	6
Mean reversion: path shape	Worst and best single days, downside dispersion, first-order autocorrelation and sign-flip rate of the residual series; overnight and intraday decompositions; position within the recent range; distance from recent highs and lows; and two state-transition terms measuring the change in reversal state since the previous rebalance.	15
Risk and style	Annualised volatility, idiosyncratic volatility, market and downside betas at 63 and 252 days, return skewness, crash minima, drawdown measures, and the maximum daily return over 21 days, the lottery characteristic of Bali et al. (2011).	14
Liquidity and size	Log market capitalisation, log dollar volume, dollar-volume and turnover shocks, share turnover at 21- and 63-day horizons, Amihud illiquidity at the same two horizons with a time-series standardised form (Amihud, 2002), listing age, and net share issuance at two horizons.	12
Seasonality	Sine and cosine encodings of the rebalance month.	2
Fundamentals	Valuation, profitability, quality, growth, and leverage ratios from both the ratios file and statement-level data, a single accrual measure built from statement-level cash flows rather than from balance-sheet differences (Hribar and Collins, 2002), together with their six- and twelve-month changes and industry-relative versions.	44
Analyst surprise	Standardised unexpected earnings, its absolute and scaled forms, rolling aggregates and changes across four consecutive annual announcements, and the industry-relative version of each.	12
Total		144

riod end. When it does not, or when none is recorded, a conservative fallback applies instead, so an absent or implausible reported date delays an item and never advances it. Ratios delivered on a publication date are dated on that date. Section A.1 states the rules per source, with the publication lag, the staleness windows beyond which an item is withdrawn rather than carried forward, and the peer minimum below which no industry-relative version is formed. These safeguards address the look-ahead that otherwise contaminates characteristic-based research (López de Prado, 2018).

They govern *when* an item becomes visible, however, and not *which* vintage of it is read. Statement items are taken from the standard quarterly file, whose values are revised as later filings restate them, so a figure timed at its original announcement date may carry a number that was corrected afterwards. The study is therefore subject to Compustat restatement bias. The exposure is shared with the machine-learning asset-pricing literature this thesis builds on, which addresses look-ahead through release lags rather than through as-first-reported vintages: Gu et al. (2020) delay monthly characteristics by one month, quarterly items by four and annual items by six, on the same restated source. An as-reported panel would remove the residual; it is not used here, and the direction of the resulting bias is not signed.

Part of the surface measures change rather than level. Twenty-six of the 144 columns are not levels but first differences or horizon spreads, so that a firm whose profitability has just improved to a given level is distinguishable from one that has been at that level throughout. They are formed after the level characteristics, from observed rather than substituted values, and are subject to the same availability rules. Appendix A states their construction.

Cross-sectional normalisation. Every characteristic is converted to a within-date rank and then mapped onto a common scale. The choice of that mapping is the primary treatment of this thesis and is deferred to Section 4.2; what matters here is that the mapping is applied identically to all characteristics within a date. The two seasonality encodings are the sole exception. A sine and cosine of the calendar month are already bounded and equally spaced, and they take twelve values common to the entire cross-section on any given date, so a within-date rank would destroy them. They enter the model on their native $[-1, 1]$ scale in every arm, which puts them beyond the reach of the representation treatment.

Values still missing at that point are replaced by the contemporaneous cross-sectional median before ranking, so an absent observation is placed at the centre of its own cross-section instead of at an extreme. The substitution is neutral by construction, but it has a side effect: a characteristic with a high missing rate acquires a large block of tied names at the median, and tied blocks are spaced differently by the two rank mappings under study. Both the missing rate and the mass of exact zeros among the observed values are

therefore recorded per characteristic on every build, before the substitution, since a zero mass point is a tied block of the same kind arising from the data rather than from the fill. Measured over the evaluation sample the surface is close to complete, and the zero masses that do occur are economically meaningful states rather than artefacts; Appendix A gives both rates and names the file that records them per characteristic. The same tied-block consideration governs which characteristics enter the surface at all. A variable resolving the cross-section into a handful of tied blocks, a counter or a binary flag for instance, would move with the representation treatment through the spacing of those blocks rather than through anything economic, so variables of that kind are not built.

3.6 Walk-forward protocol and evaluation sample

Models are estimated on an expanding walk-forward schedule. Each fold divides the available history into a training block, a validation window, and a test window of the same length that is used neither for fitting nor for selection; the window then advances by a year and the procedure repeats. Validation serves the families that have something to select, the number of boosting rounds of the tree benchmark and the probe of the convergence-matched specification, while the deployed neural fits train for a fixed number of epochs on the training and validation blocks combined under the rule of Section 4.4.2. A purge of one rebalance separates training from validation and validation from test at both boundaries, so that no label used for fitting is realised over a window falling inside the block on which the fit is assessed. No test block is scored until 60 months of history precede it, which keeps the earliest folds from being estimated on a sample too short to be stable. Figure 3.1 draws the geometry that results.

The December 2025 rebalance has no subsequent month in which to earn a return and is therefore not scheduled; every other month of the window is. The evaluation sample is 272 consecutive monthly out-of-sample observations across 23 folds, the last of which is short at eight months and is scored on the months it has. Calendar 2025 enters all headline estimates and is examined separately as a subperiod in Part III, and no period is dropped on the basis of what it turned out to contain. Table 3.3 lists these constants alongside the universe, label, and feature conventions that every specification in Parts II and III shares.

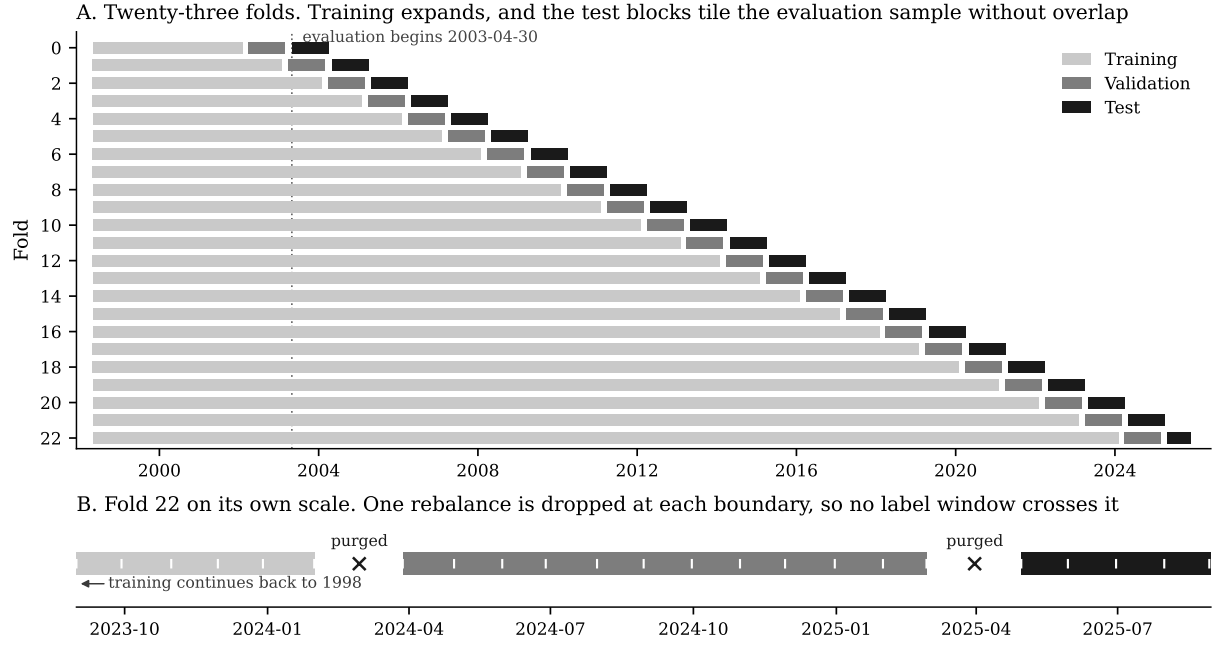


Figure 3.1: Panel A draws all 23 folds. Training expands from a common start date, and the test blocks advance by a year and tile April 2003 to November 2025 without overlapping, which is where the 272 evaluated rebalances come from; the last is short at eight months. Panel B redraws that fold at a scale where the purge is visible, with ticks on the rebalance grid and crosses on the two dropped months, one at each boundary.

Table 3.3: Fixed elements of the research environment. These values are identical for every estimated model specification in Parts II and III; the momentum rule carries no folds or seeds and the linear anchors carry a single seed.

Block	Item	Value
Universe	Cross-section per rebalance	Top 1000 by market capitalisation
	Price floor	\$5
	Minimum average daily dollar volume	\$5 million (20-day trailing)
	Minimum price history	252 trading days
	Rebalance frequency	Business month-end
Label	Horizon	To the next rebalance (one month)
	Gross convention	Close of t to close of t^+
	Realistic convention	Open of $t + 1$ to close of t^+
	Delisting return	Compounded into the daily total return
Features	Characteristics	144
	Normalisation	Within-date rank transform
	Missing values	Neutral centre of transformed cross-section
Splits	Minimum training history	60 months
	Validation window	12 months
	Test window	12 months
	Purge	38 calendar days (one rebalance) at both boundaries
	Folds	23
Sample	Model history begins	1998-04-30
	First evaluated rebalance	2003-04-30
	Last evaluated rebalance	2025-11-28 (return earned to 2025-12-31)
	Out-of-sample rebalances	272
	Seeds per specification	5

3.7 Inference protocol

Several specifications are compared on a single history. The protocol below was fixed before the estimates of the definitive run were read, as a discipline internal to the study; Section 8.4 returns to what that discipline does and does not establish.

Primary endpoint. The study has a single confirmatory endpoint: the equally weighted mean, across the five ranking objectives, of the paired difference in annualised Sharpe ratio between the inverse-normal and percentile arms, written Δ^{repr} and defined in Section 5.1, computed on the 272 monthly out-of-sample portfolio returns. It is struck under the gross convention of Section 3.4 and under the inverse-volatility sizing of Table 4.4. Section 7.3 re-strikes the same quantity under the realistic convention and Section 6.3 re-strikes it under equal weighting; both are readings of the endpoint under a second execution or construction rule rather than further confirmatory tests, and neither belongs to a family of its own. The weighting comparisons of the exploratory tier are a different quantity: they set one scheme against the other, not one arm against the other. Pairing is essential throughout: two specifications compared on this history share the universe, the dates, and the market environment, so their return series are strongly dependent and an unpaired comparison would be badly calibrated.

Descriptive endpoints. Annualised return and volatility, maximum drawdown, Sortino ratio, one-side net turnover, the separate long and short leg contributions, the mean rank information coefficient over the full cross-section and within the top and bottom traded decile, its information ratio, and decile precision on each side. As *levels* these characterise the mechanism and establish no significance beyond a test against the relevant null of no skill, which is reported unadjusted. Where a *contrast* between two arms is formed from one of them and a p -value is printed, that contrast is a member of the exploratory family defined below and is adjusted with it. No position-discounted statistic is among them, for the reason given in Section 2.4.

Testing the primary endpoint. A Sharpe ratio is a nonlinear function of two estimated moments, so a difference of two of them is a function of four. Its standard error therefore requires the delta method rather than the standard error of a mean difference, which is a different estimand and which, as Section 5.3.2 shows, can point the other way on the same data. Writing x_i for the monthly return series of specification i , μ_i and $\gamma_i = \mathbb{E}[x_i^2]$ for its first two moments, and $\text{SR}_i = \mu_i / \sqrt{\gamma_i - \mu_i^2}$, the estimator stacks the four sample moments $(\hat{\mu}_a, \hat{\mu}_b, \hat{\gamma}_a, \hat{\gamma}_b)$ and propagates their joint uncertainty through the

gradient of $\text{SR}_b - \text{SR}_a$,

$$s^2(\widehat{\Delta \text{SR}}) = \frac{1}{T} \nabla' \hat{\Psi} \nabla, \quad \nabla = \begin{pmatrix} -\gamma_a(\gamma_a - \mu_a^2)^{-3/2} \\ +\gamma_b(\gamma_b - \mu_b^2)^{-3/2} \\ +\frac{1}{2}\mu_a(\gamma_a - \mu_a^2)^{-3/2} \\ -\frac{1}{2}\mu_b(\gamma_b - \mu_b^2)^{-3/2} \end{pmatrix}, \quad (3.2)$$

where $\hat{\Psi}$ is the heteroskedasticity- and autocorrelation-consistent long-run covariance of those four moments (Newey and West, 1987), computed with Bartlett weights at the bandwidth $\lfloor 4(T/100)^{2/9} \rfloor$ (Newey and West, 1994), with T the length of the series under test. This is the delta-method expansion of Ledoit and Wolf (2008), implemented with a Bartlett kernel.¹

The p -value comes from a studentised bootstrap with 4000 replications: blocks of six months of the joint series are resampled *circularly* (Politis and Romano, 1992), so that no observation is under-represented at the ends of the sample. Six is not a free choice: the block length is $\text{round}(T^{1/3})$ evaluated at the length of the series under test, and that rule returns six across the full-sample lengths reported here. Each replication is divided by its own (3.2) standard error before the null distribution of the t -ratio is formed. Studentising is what makes the interval second-order accurate, and it is also why an analytic standard error is needed at all: a raw percentile interval has no studentiser.

The convention is exhaustive for that estimand rather than merely typical of it. Every interval and every p -value this thesis reports on a Sharpe difference, pooled or by objective, gross or net, in every chapter, comes from the single implementation these paragraphs describe; no table computes one of its own, and where two tables report the same contrast they report the same interval. The dispersion comparisons of Table 5.6 are a different estimand and carry their own circular block bootstrap on the log variance ratio, described where they are used.

Because (3.2) is less familiar than the classical alternative, every reported contrast additionally carries the closed-form test of Jobson and Korkie (1981) with the correction of Memmel (2003), which is the same delta-method expansion evaluated under an assumption of independent normal returns. The two are reported side by side. The robust variant is the headline because normality is rejected for these return series and the individual series are serially correlated.

¹The expansion is written on the moment estimator $\hat{\sigma}^2 = \hat{\gamma} - \hat{\mu}^2$, which divides by T , whereas every Sharpe ratio and volatility reported in this thesis uses the unbiased $T - 1$ estimator. The two differ by the constant $\sqrt{(T - 1)/T}$, which cancels in the t -ratio, the studentised bootstrap p -value, and the shape of the percentile- t interval; only the point estimate and the interval endpoints move, by 0.18% at $T = 272$. All reported figures use the unbiased convention, so a contrast can be checked against the difference of the two levels it is formed from.

Attribution, not further testing. Where a Sharpe difference is established it is decomposed into the two channels through which it can arise, with σ_a and σ_b the return standard deviations $\sqrt{\gamma - \mu^2}$ of the two specifications,

$$\text{SR}_b - \text{SR}_a = \underbrace{\frac{\mu_b - \mu_a}{\sigma_b}}_{\text{return channel}} + \underbrace{\mu_a \left(\frac{1}{\sigma_b} - \frac{1}{\sigma_a} \right)}_{\text{volatility channel}}. \quad (3.3)$$

The identity is exact, so (3.3) attributes one rejected hypothesis rather than introducing two further ones, and no multiplicity adjustment is owed for reporting where an effect came from. The channels are nevertheless accompanied by descriptive tests, so that the attribution does not rest on point estimates alone: a Newey-West test on the mean return difference for the numerator, and a delta-method test on the log volatility ratio $\frac{1}{2}(\log \hat{v}_b - \log \hat{v}_a)$, where $\hat{v}_i$ is the sample return variance of specification i , against the same $\hat{\Psi}$ for the denominator. The volatility channel is also checked against the classical paired-variance test of Pitman (1939) and Morgan (1939), which reduces to a correlation test on the sum and the difference of the two series and depends on none of the machinery above.

Multiplicity and selection. The study makes many more inferential statements than the representation contrast alone, and a single undifferentiated family would be both conceptually wrong and needlessly punitive. Three tiers are used.

The *primary* family has a single member: the equally weighted mean of Δ^{repr} across the five ranking objectives. It pools one scientific question across the objectives through which it is measured, which is how Section 5.1 reads the evidence in any case; testing a single hypothesis five times and then paying a fivefold penalty for having done so discards information instead of protecting against it. As a single pre-specified quantity it requires no multiplicity adjustment. The pooled contrast is computed from a joint bootstrap in which the months are resampled once and every cell is recomputed on the same resample, so the dependence across objectives is carried by the resampling. That dependence is substantial and cannot be ignored: the five difference series are strongly correlated, so the five objectives amount to materially fewer than five independent replications, and any argument resting on the number of agreeing signs is correspondingly weaker than a count of five would suggest.

The *secondary* family is the five per-objective representation contrasts. It is examined only where the pooled contrast rejects, so that the ordering costs no additional error rate, and adjustment is by the studentised stepwise procedure of Romano and Wolf (2005b), with adjusted p -values computed as in Romano and Wolf (2016) and the general theory given by Romano and Wolf (2005a). Romano-Wolf rather than Holm (Holm, 1979) because the statistics are estimated on the same months and on books that overlap

heavily, and a stepdown taking its critical values from the bootstrap distribution of the maximum statistic exploits that dependence instead of bounding it. The decomposition of the contrast through the scale-matched arm and the convergence-matched re-estimate of Section 5.3.3 are reported unadjusted and outside the three families, on the attribution logic above. No confirmatory claim is made for either step taken on its own. One reading below does rest on a single step: the practical recommendation of Section 5.3.1 is a reading of the pre-specified decomposition rather than a tested result, and that section states the limit where it makes the claim.

The *exploratory* tier collects the remaining *contrasts* that carry a reported conclusion: the sixty tail-width comparisons, the fifteen weighting comparisons, the fifteen gross-against-net comparisons, and seven ordering contrasts: the six of Table 5.2 (the rank information coefficient over the full cross-section and within each traded tail, decile precision on each side, and the information ratio of the coefficient), together with the asymmetry between the two tails, which Section 5.2 reports. That is a family of ninety-seven, adjusted together by the Benjamini–Hochberg procedure at $q = 0.10$ (Benjamini and Hochberg, 1995) and labelled exploratory wherever they appear. They describe; they do not confirm. One property of that family is worth stating before any of its counts are read: the fifteen netting contrasts are near-tautological once costs are positive, and because their p -values are the smallest in the family they take the lowest ranks, where the Benjamini–Hochberg critical value is smallest. Every other member is therefore compared against a larger threshold than it would face in a family without them. Section C.2 reports the family with the netting group set aside, which is the sharper reading, and any member that survives only in the full family should be read as surviving on the family’s composition rather than on its own evidence. The between-objective comparisons of Section 5.8 are deliberately outside it: they are adjusted within their own stepdown and support a modelling choice rather than a reported result, so folding them into this family would widen it without bearing on any conclusion drawn from it.

Four classes of reported statistic are deliberately outside the family, and the reason differs in each case. The scale and shape steps of any contrast are excluded on the attribution logic already stated above. Level statistics tested against a no-skill null, and the per-cell factor alphas of Table 8.5, are excluded because they are not comparisons between specifications and no selection operates among them, every cell being reported; they carry unadjusted p -values and are marked as such. Excluding them is conservative, since their p -values are small enough to occupy the lowest ranks of the step-up and would *raise* the threshold available to every contrast that is a member. Subperiod statistics carry no test at all, here or in Chapter 8, and so cannot be family members; they are read for their pattern across regimes. And the two re-strikes of the primary endpoint named above, under the realistic cost convention and under equal weighting, are outside it because they are not further contrasts: each is the one pre-specified member read under

a second execution or construction rule, computed from the same joint bootstrap at the same seed as the endpoint itself.

The endpoints and families above are fixed before the estimates are read, rather than recovered afterwards through a data-snooping adjustment over the strategy universe (White, 2000). Sharpe-ratio standard errors reported at the level of a single cell account for the non-normality of monthly returns through the skewness and kurtosis correction of Mertens (2002) to the independent-returns expansion of Lo (2002).

What these standard errors do not contain. Every standard error above describes sampling variation in the realised return series; the reported cells average five random seeds before a single backtest is run, so the intervals carry no term for initialisation noise. Because the seeds are shared across arms, an additive seed effect cancels in a paired contrast, and the relevant seed-by-arm interaction is estimated in Section 8.5 from the retained per-seed score vectors. It is not folded into the standard errors reported here; it would add to them in quadrature, so every interval below understates total uncertainty by whatever that component is worth. A representation effect smaller than the spread across initialisations is not treated as a finding.

Interpretation. With 272 monthly cross-sections, adjacent Sharpe ratios cannot be treated as sharply distinguished population parameters. The results are therefore read for the direction and the economic size of differences, for their consistency across objectives, weighting schemes, widths, and subperiods, and for their survival under costs. Small gaps supported by a single comparison are not presented as evidence of superiority.

Chapter 4

Methodology

This chapter defines the treatments and the machinery that maps a trained model into a traded book.

4.1 Cross-sectional setup and notation

Fix a rebalance date and suppress the time index. The cross-section contains n names. Name i carries a raw characteristic vector and a realised label y_i from (3.1). A scoring function f_θ maps transformed characteristics to a scalar score s_i . Let $r_i \in \{1, \dots, n\}$ denote the within-date rank of a characteristic, let $\hat{r}_i$ denote the rank of name i when names are sorted by score, and let $\tilde{y}_i$ denote its rank when sorted by realised label. Ties are broken by index, and both rank vectors are permutations of $\{1, \dots, n\}$.

The evaluation statistic is the rank information coefficient, the Spearman correlation between the two orderings (Spearman, 1904),

$$\text{IC} = 1 - \frac{6 \sum_i (\hat{r}_i - \tilde{y}_i)^2}{n(n^2 - 1)}. \quad (4.1)$$

The reported coefficient averages ranks over ties, and (4.1) is the form it takes when the cross-section carries none.

4.2 Feature representation treatments

Each admitted characteristic is transformed within each date: the values are ranked and the ranks are mapped onto a numerical scale. Writing $u_i = r_i/(n + 1) \in (0, 1)$ for the within-date uniform rank, the two mappings a modeller would choose between are

$$\text{centred percentile: } c_i = u_i - \frac{1}{2}, \quad (4.2)$$

$$\text{inverse-normal: } z_i = \Phi^{-1}(u_i), \quad (4.3)$$

where Φ is the standard normal distribution function. Both are strictly increasing in u_i and therefore carry identical ordinal information. They differ in spacing: (4.2) is uniform on $(-\frac{1}{2}, \frac{1}{2})$, while (4.3) is approximately standard normal and separates the extremes. At $n = 1000$ the most extreme name sits about 3.1 standard deviations from the centre under (4.3) against about 1.7 under (4.2), and the gap between the two most extreme names is roughly eighty times the gap at the centre of the cross-section, whereas under (4.2) every gap is the same.

The arms are set out in Table 4.1 and drawn in Figure 4.1. The two a modeller would choose between differ in the *shape* of the spacing, uniform against convex-in-the-tails, and in the *scale* of the numbers: a centred percentile rank has standard deviation $1/\sqrt{12} \approx 0.289$ by the algebra of the uniform distribution, against approximately one for the inverse-normal, so the two encodings sit a factor of $\sqrt{12} \approx 3.46$ apart in scale. The third arm sets $\zeta_i = \sqrt{12} c_i$, the percentile shape at the inverse-normal scale.

Table 4.1: The three arms of the representation axis. The training target is the centred percentile encoding in every neural arm, so the contrast operates through the feature channel alone; the gradient-boosted control is trained on decile grades, equally fixed across arms. Section 4.2.1 states which statistic the scale-matched constant matches and what the match leaves open.

Cell	Mapping	Range	SD
Percentile	centred percentile, (4.2)	± 0.5	0.289
Inverse-normal	normal score, (4.3)	± 3.1	≈ 1
Scale-matched	$\zeta_i = \sqrt{12} c_i$	± 1.73	1

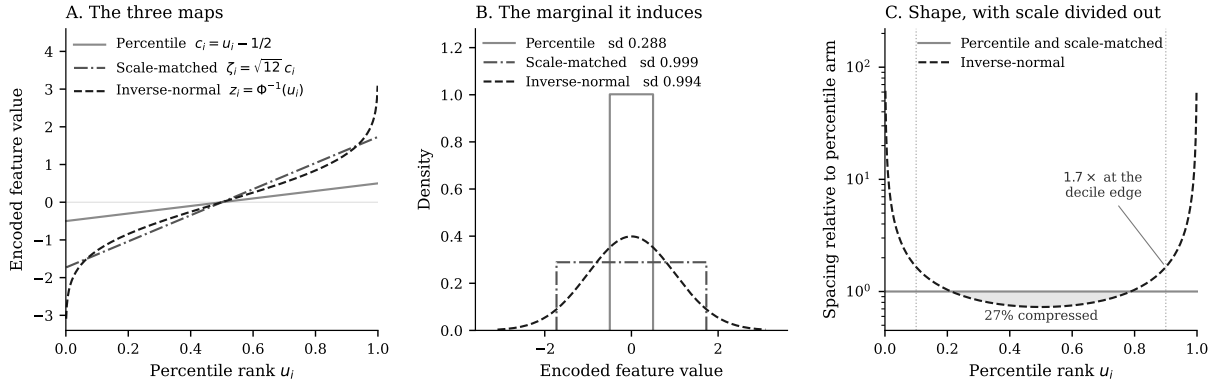


Figure 4.1: Panel A draws the mappings of (4.2), (4.3) and $\zeta_i = \sqrt{12} c_i$ on a tie-free cross-section of 1000 names. All three are strictly increasing, so no pair of them can differ in the ordering it presents. Panel B gives the marginal each one induces: moving from the narrow box to the wide one is the scale step, which leaves the shape alone, and moving from the wide box to the bell is the shape step, taken at a matched standard deviation. Panel C divides each arm’s spacing by its own standard deviation, so the percentile and scale-matched arms fall on one line and only the shape step remains. Measured that way the inverse-normal map compresses the middle 58% of the cross-section, by 27% at the median, and stretches the outer 42%; the decile edges the headline book trades lie in the stretched region, at 1.7 times the percentile spacing.

4.2.1 What the three arms identify

Everything upstream of the mapping is identical across the arms, and each arm is a strictly increasing function of the same within-date rank, so any difference between them is attributable to the numerical presentation of the ranks rather than to a difference in what the estimator was told.

That presentation, however, bundles shape with scale, and the bundle matters because this architecture is not neutral to the magnitude of its inputs. The reason is a property of the hidden block rather than of the training schedule, and it can be stated exactly. Each block applies a linear map with a bias, then the activation, then dropout, then the normalisation. Multiplying every input by a positive constant multiplies the weighted sum entering every first-layer unit by that same constant, exactly and unit by unit, and leaves the bias where it was. What happens next decides whether anything survives. Were the activation positively homogeneous, so that $\phi(\lambda t) = \lambda \phi(t)$ for $\lambda > 0$, the constant would pass through it untouched and the normalisation that follows would divide it out, leaving the block invariant up to the first-layer bias. GELU has no such property. It is close to linear near the origin and bends away from it, so a rescaling of its argument is not a rescaling of its value, and the normalisation downstream restores the size of the block’s output without restoring its shape.

What the input scale therefore sets is how much of the activation’s curvature the first block uses, and Section B.6 measures it at initialisation as the share of a hidden unit’s post-activation variance that the best affine function of its own pre-activation fails to explain. At the percentile scale the share is 3.2%: the first block is within a few percent of an affine map, and a network whose first block is nearly affine is, to that extent, a shallower network. The scale-matched arm reads 19.7% and the inverse-normal arm 19.6%. The channel is confined to one layer, which is what makes it legible: only the first block reads the encoding, every later block reads a normalised output and sits at the same scale whatever the encoding was, the three arms running between 17.5% and 20.4% at blocks two, three and four, so what the percentile encoding costs is one nonlinear layer out of four. Two further routes are open and are smaller. The first-layer bias is not rescaled with the inputs, so its weight relative to the signal falls by the same constant, and it is the reason the block would not be exactly invariant even under a homogeneous activation. Decoupled weight decay acts on the weights rather than on the gradients, so it does not scale with the inputs either.

One route commonly assumed to be open is not. A larger input does not buy a larger optimiser step here, the two arms’ gradient norms falling within 2.6% of one another rather than a factor of $\sqrt{12}$ apart (Section B.6). The channel is the geometry of the block, not the size of the step.

Arguments at this level fix which routes exist and rank them; they do not settle the

size or even the sign of the resulting difference, and the design does not ask them to. A separate concern belongs to the schedule rather than to the architecture: under the fixed training budget of Section 4.4.2 the two arms may sit at different points of their own optimisation paths after the same number of epochs, and a critic can attribute any contrast to that alone. Section 4.4.2 carries the control for it.

A third arm separates the two channels. Multiplying the centred percentile rank by $\sqrt{12}$ changes no ordering and no shape, the spacing remaining uniform, but presents the network with inputs at the inverse-normal scale. The constant is the reciprocal of the encoding’s own standard deviation, so the arm is the field’s default encoding standardised (Section 5.3.1). The total contrast then decomposes exactly into a pure scale term, percentile to scale-matched, and a shape term at matched scale, scale-matched to inverse-normal; Section 5.1 states the decomposition formally and Section 5.3 reports it. If the scale term reproduces the total and the matched-shape term is null, the effect is an optimisation artefact of input magnitude; if the shape term carries it, the tail spacing is doing the work the hypothesis claims. Either outcome is informative, and neither can be steered: the shape effect of the third arm is zero by construction, so no part of a measured shape term can be an artefact of the device.

Calling that arm scale-matched presupposes a statistic on which the match is struck, and the two encodings it sits between can be aligned on more than one. The constant $\sqrt{12}$ equates their cross-sectional standard deviations. Equating their supports instead would require 6.18, since the centred percentile spans ± 0.5 against the ± 3.09 of (4.3) at $n = 1000$. The choice is not innocuous: (5.1) is an exact identity for any positive multiple of c_i , so what the multiple fixes is not whether the two terms sum to the total but where the boundary between them falls.

Second moments are matched for a reason belonging to the second step rather than the first. The third arm exists to make the shape term a comparison at held scale, and an arm aligned on the supports would carry a standard deviation of $6.18/\sqrt{12} \approx 1.78$, so the step from it to the inverse-normal encoding would change the spacing and contract the dispersion by nearly a factor of two at once. That is the confound the third arm was introduced to remove, and matching supports would reintroduce it in the one comparison the hypothesis turns on. The constant also stays clear of the realised panel. Writing $\zeta_i = \sqrt{12} c_i$ uses the ratio of two population standard deviations and carries no n , whereas the ratio of supports is $2\Phi^{-1}(n/(n+1))$, fixed by the value of the map at a single point of the cross-section rather than by its behaviour across the whole of it.

What matching the second moments cannot do is match the extremes, which stay apart by the factor of 1.78 above. That is the content of the hypothesis and not a gap in the design: at a common dispersion, separating the tails and placing the extreme names further out in absolute terms are the same statement, and nothing here attempts to divide them. The division that does depend on the convention is the one between Δ^{scale}

and Δ^{shape} , and a different multiple would move mass between them by an amount this design does not sign, since the scale channel is not established to be monotone in input magnitude over any range. The split is reported throughout as one struck at matched second moments, not as a scale-invariant partition of the total.

4.3 Ranking objectives

Five differentiable ranking objectives are compared. Three are pairwise and are estimated on a bounded sample $\mathcal{P}$ of pairs drawn uniformly within each date; two read the entire cross-section. Within this section, τ denotes a temperature, β a logistic shape parameter, and γ a margin.

RankNet applies a logistic penalty to each sampled pair with unit weight (Burges et al., 2005),

$$\ell_{\text{RN}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \tau \text{softplus} \left(-\frac{s_i - s_j}{\tau} \right). \quad (4.4)$$

LambdaRankIC weights each pair by the change in (4.1) that swapping it would produce (Lin et al., 2026a). Summing over all pairs, oriented so that i is the better name by realised return, and rescaling the sampled sum by $N/|\mathcal{P}|$, where $N = n(n-1)/2$, gives an unbiased estimator of the full objective,

$$\hat{\ell}_{\text{LRIC}} = \frac{N}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \frac{12 |\hat{r}_i - \hat{r}_j| |\tilde{y}_i - \tilde{y}_j|}{n(n^2 - 1)} \log_2(1 + e^{-\beta(s_i - s_j)}). \quad (4.5)$$

The swap weight grows with rank separation instead of decaying with rank position, so gradient is spread across the cross-section rather than concentrated at the top of the list.

Margin ranking replaces the logistic surrogate with a hinge. Because the output scale of a neural scorer drifts during training, scores are standardised within each date, $\hat{s}_i = (s_i - \bar{s})/\sigma_s$, with $\bar{s}$ and σ_s the within-date mean and standard deviation of the scores, so that the margin has a stable meaning,

$$\ell_{\text{hinge}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max(0, \gamma - (\hat{s}_i - \hat{s}_j)). \quad (4.6)$$

Soft-Spearman differentiates (4.1) directly by replacing the hard rank with a smooth count of dominating scores,

$$\tilde{r}_i(\mathbf{s}) = 1 + \sum_{j \neq i} \varsigma \left(\frac{s_j - s_i}{\tau} \right), \quad \varsigma(v) = \frac{1}{1 + e^{-v}}, \quad (4.7)$$

and minimising one minus the correlation between $\tilde{\mathbf{r}}(\mathbf{s})$ and the encoded target, with scores standardised per date so that the soft rank does not saturate (Qin et al., 2010;

Blondel et al., 2020).

RankCosine treats the score vector and the encoded target vector as directions and penalises the angle between them (Qin et al., 2008). Writing $\mathbf{s}^c$ and $\mathbf{y}^c$ for the scores and the encoded target after centring within the date, the latter oriented so that a larger value marks a better name,

$$\ell_{\text{RC}} = 1 - \frac{\langle \mathbf{s}^c, \mathbf{y}^c \rangle}{\|\mathbf{s}^c\| \|\mathbf{y}^c\|}. \quad (4.8)$$

It is invariant to the scale of the scores but remains sensitive to their relative spacing.

Table 4.2 lists the free parameter of each objective. All are conventional default values, fixed before estimation and tuned nowhere, under the fixed-surface rule of Section 4.4.

Table 4.2: Free parameters of the five ranking objectives. Values are common to every representation arm. Scores are standardised within each date for margin ranking and Soft-Spearman, so that the margin and the soft-rank temperature have a stable meaning across dates and across objectives.

Objective	Granularity	Parameter	Value
RankNet	Pairwise, sampled	Temperature τ	1.0
LambdaRankIC	Pairwise, sampled	Logistic shape β	1.0
		Pair reduction	Sum
Margin ranking	Pairwise, sampled	Margin γ	1.0
Soft-Spearman	Listwise, full	Soft-rank temperature τ	1.0
RankCosine	Listwise, full	None	n.a.

The soft-rank temperature governs the allocation of gradient rather than the fidelity of the rank, since (4.7) is monotone in the score at every temperature, so the flatter inherited value is retained; Appendix B reports the measured gradient shares behind that choice.

Gradient clipping is disabled uniformly, as a condition of the cross-objective comparison; Appendix B gives the reason.

4.3.1 Why the target encoding is held fixed

Holding the target at the centred percentile encoding serves a specific purpose. For RankNet and margin ranking the target enters only through the direction of each comparison, so any strictly monotone re-encoding of the target leaves (4.4) and (4.6) unchanged. For LambdaRankIC the label enters through the integer rank $\tilde{y}_i$ computed inside the loss, so (4.5) is likewise invariant to the encoding presented to the network. For these three objectives the feature channel is therefore the only active channel by construction. Soft-Spearman and RankCosine do consume the encoded target numerically, which is why fixing it matters for them.

4.4 The fixed surface

One invariant governs the comparison. The feature set, the architecture and its capacity, the universe and its screens, the walk-forward splits, the seeds and their aggregation, the pair budget and sampling rule, and the portfolio construction are identical across objectives and across representation arms. Only the treatment under study changes. A separate hyperparameter search for each cell might raise individual results, but it would confound the comparison with differences in tuning effort. Every cell therefore sits at a common and conservative operating point. Training length is the one element where an otherwise fixed surface could leak.

4.4.1 Architecture and optimisation

The scorer is a four-hidden-layer feed-forward network of widths 256, 128, 64, and 32 with GELU activations, layer normalisation at each hidden block, dropout after each hidden activation, and a scalar linear head. Weights and biases are left at the framework default for a linear layer, Kaiming uniform (He et al., 2015) at PyTorch’s legacy setting $a = \sqrt{5}$, which reduces to $\mathcal{U}(\pm 1/\sqrt{\text{fan-in}})$ for both. Each block applies its linear map, then the activation, then dropout, then the normalisation, in that order. The order is stated because it is not the only one in use and because Section 4.2.1 turns on it: a normalisation placed after the nonlinearity cannot undo a rescaling the nonlinearity has already acted on. Table 4.3 lists the settings.

Table 4.3: Architecture and optimisation settings, held fixed across every objective and representation arm. The validation rank information coefficient serves as the common stopping criterion only in the convergence-matched specification of Section 4.4.2.

Item	Value
Hidden widths	256, 128, 64, 32
Activation	GELU
Block order	Linear $\rightarrow$ activation $\rightarrow$ dropout $\rightarrow$ layer normalisation
Normalisation	Layer normalisation at each hidden block
Dropout	0.10
Output head	Scalar linear
Initialisation	Framework default, $\mathcal{U}(\pm 1/\sqrt{\text{fan-in}})$ for weights and biases
Optimiser	AdamW (Loshchilov and Hutter, 2019), weight decay 5×10^{-4}
Learning rate	10^{-3} , constant
Gradient clipping	Disabled
Training length	Fixed, 10 epochs on the combined train-validation block
Deployed weights	Final epoch; no checkpoint selection
Queries per batch	4
Sampled pairs per query	32,768, drawn uniformly
Listwise objectives	Full cross-section, no pair sampling
Seeds	0 through 4, averaged at the within-date rank level

4.4.2 Training length: a fixed budget, and its control

Every deployed model trains for exactly ten epochs on the combined training-plus-validation block, at a constant learning rate, and the final-epoch weights are kept. Nothing is selected: no early stopping, no checkpoint reversion, no per-cell budget.

The budget is small by design and by evidence. With four queries per optimiser step, an epoch is between roughly fifteen and eighty steps depending on the fold’s training length, so ten epochs is between roughly one hundred and fifty and eight hundred updates in all. A stopping criterion struck on a twelve-month validation window would select on a noisy argmax rather than on a measured optimum, which is the substantive reason a fixed rule is preferred to one here. A fixed budget buys exact comparability (every cell of every arm trains for the same number of epochs) at the price that the operating point is arbitrary: under the scale difference of Section 4.2.1 no budget can be known to place both arms at the same distance from their own optimum, so equality of epochs must not be read as neutrality of the schedule.

Two things carry that concern. The scale-matched arm of Section 4.2.1 isolates what input magnitude alone does at this budget. A convergence-matched specification then re-estimates every step of (5.1), across all five objectives in all three arms, with the fixed rule replaced by a genuine validation optimum: a probe fit on the training block records the validation curve of the common rank information coefficient with its per-epoch standard errors, the selected length is the *longest* epoch statistically indistinguishable from the optimum (a one-standard-error rule that keeps a noisy twelve-month window from deploying near-untrained models), and the deployed fit trains for that length, rescaled to the combined block, with no floor. If the contrast survives at both operating points, the objection that the budget favoured one arm is closed in both directions; if it flips between them, the effect is a property of the optimisation schedule and is reported as such. Section 5.3.3 reports both.

4.5 Gradient-boosted benchmark and the invariance argument

A gradient-boosted pairwise ranker (Chen and Guestrin, 2016; Friedman, 2001) is run under both representation arms. Its training target is graded rather than continuous: within each date the realised labels are binned into deciles and the ranker consumes the integer grade, which is the form its pairwise objective is defined over. The grading is identical in both arms, so here too the arms differ only in the features. The role of the cell is not to provide a second test of Hypothesis 1 but to serve as a control, and the reason is structural.

The cell nonetheless does a second piece of work, and it is worth separating the two

before either is read. The tree receives the fixed surface of Section 4.4 and hyperparameters that are conventional values for a low signal-to-noise cross-section, fixed once in advance rather than searched. On the one axis where the design permits selection, however, the two families are not treated alike. The number of boosting rounds is chosen fold by fold on the validation rank information coefficient under early stopping, retaining 214 of a permitted 800 rounds in every seed (Table C.6), while the neural cells train for a fixed ten epochs and deploy the final-epoch weights with no selection of any kind (Section 4.4.2). Whatever the level comparison between the two families turns out to be, the asymmetry in fitted capacity runs in the tree’s favour and not in the network’s. The cell is therefore read twice below: as the invariance control of Table 5.7, which is what it was built for and what Hypothesis 3 concerns, and, separately and with no test attached, as the study’s closest thing to a competitor at a common operating point (Section 5.6.3).

A regression tree partitions the feature space by axis-aligned thresholds of the form $x \leq \theta$. Let g be a strictly increasing function and consider re-encoding a feature as $x' = g(x)$. For any threshold θ on the original scale, the threshold $g(\theta)$ on the new scale induces exactly the same partition of the observations, and the converse holds because g is invertible. Both (4.2) and (4.3) are strictly increasing functions of the same within-date uniform rank u_i , so one is a strictly increasing function of the other, namely $z_i = \Phi^{-1}(c_i + \frac{1}{2})$. A threshold-splitting learner is therefore invariant to the substitution, and its fitted ranking should be unchanged up to two residual channels: the finite resolution of the histogram binning used to propose split candidates, and floating-point representation. Both are measured rather than assumed away: Section C.6 compares the two fitted arms field by field across every tree of every seed, and neither channel is active on this data. For the same reason no tree cell is run in the scale-matched arm: a multiplication by a positive constant is strictly monotone, so such a cell would reproduce the percentile cell identically and measure nothing.

The prediction is therefore that the tree contrast is close to zero, and Table 5.7 fixes in advance how each outcome is read.

4.6 A naive benchmark

The tested comparisons of this thesis are internal: they set one modelling choice against another within a family of neural ranking models. That leaves the level of the reported performance uncalibrated, since a Sharpe ratio has no natural scale. Three reference points are therefore carried through the same pipeline so that the fitted models can be read against something: the fixed rule defined below, the two linear anchors of Section 4.7, and the gradient-boosted ranker of Section 4.5. None is a member of any inference family, and none is tested.

The rule is twelve-month momentum with a one-month skip: the formation window

of Jegadeesh and Titman (1993) under the skip-a-month convention of Carhart (1997), matching the momentum factor used in Section 3.5. Writing P for the total-return index of Section 3.2, the score at rebalance t is

$$s_{i,t}^{\text{mom}} = \log \frac{P_{i,t-21}}{P_{i,t-252}}, \quad (4.9)$$

the skip of one month being the standard guard against short-horizon reversal contaminating the signal. Because selection is a sort and sizing operates within the selected set, the encoding of (4.9) is immaterial: any strictly increasing transform of it produces the identical book, which is the invariance argument of Section 4.5 applied to a portfolio rule rather than to an estimator.

The benchmark is not a model and is not treated as one. It has no training, no folds and no seeds, so it is available on every evaluated rebalance from the first, and everything downstream of the score is shared: the rule travels the identical replay path the frozen model scores take.

The level it sets is a floor rather than a fair reading of momentum, for reasons Section 5.6 states where the result is read.

4.7 Linear anchors

Two linear models complete the reference set: ordinary least squares and an elastic net, each fit on the identical feature surface under both representation arms, with the folds, target encoding, and portfolio rules of the fixed surface. Their role is different from the tree's. Ordinary least squares is exactly invariant to a common rescaling of its inputs, since the coefficients absorb the constant; the elastic net's penalty is not, so its inputs are standardised per column inside the fit, which restores the same invariance by construction. The scale channel of Section 4.2.1 therefore reaches neither anchor. Neither is invariant to the shape of the mapping, since replacing uniformly spaced ranks by normal scores changes the design matrix in a way no rescaling of columns or coefficients can undo. A linear contrast between the arms therefore isolates the shape channel for a model with no optimisation race to confound it, complementing the tree, which is blind to both channels. How much of the neural cells' performance a linear model on the same information already delivers is reported in Table 5.8.

Both are deterministic single fits, so they carry one seed rather than five and no ensemble. The elastic net selects its penalty by the common rank information coefficient, the same criterion every other family uses, on a holdout carved from the final fifth of its own training dates; ordinary least squares has nothing to select. Neither anchor is a member of any inference family.

4.8 From scores to portfolios

4.8.1 Sector demeaning and selection

Scores are demeaned within sector at each rebalance, $\tilde{s}_i = s_i - \bar{s}_{\text{sec}(i)}$, using the sector observed on that date, so that a name is selected on its standing within its sector. Names are then sorted by $\tilde{s}_i$ and the extreme fractions are traded. With a traded fraction α per side, the number of names per leg is

$$k = \max(k_{\min}, [\alpha n]), \quad (4.10)$$

where $[\cdot]$ denotes rounding and the floor $k_{\min} = 10$ guards against a degenerate book in a cross-section too small to support the nominal fraction; it binds nowhere on the width grid of (4.16) (Section C.5).

4.8.2 Two weighting schemes

Let $\mathcal{L}$ and $\mathcal{S}$ denote the long and short sets. Under *equal weighting*, every selected name in a leg receives the same magnitude,

$$w_i^{\text{EW}} = \frac{G}{2k} (\mathbb{1}[i \in \mathcal{L}] - \mathbb{1}[i \in \mathcal{S}]), \quad (4.11)$$

where G is the gross exposure of the book, fixed below, so that the book reflects the selected set and nothing else.

Under *inverse-volatility weighting*, the magnitude is inversely proportional to an estimate $\hat{\sigma}_i$ of annualised volatility from a 63-day exponentially weighted estimator, subject to a per-name cap,

$$w_i^{\text{IV}} = \eta_{\mathcal{J}} \min\left(\frac{v^*}{\hat{\sigma}_i}, \bar{w}\right) (\mathbb{1}[i \in \mathcal{L}] - \mathbb{1}[i \in \mathcal{S}]), \quad (4.12)$$

with v^* a sizing constant that cancels in (4.13), $\bar{w}$ a per-name ceiling applied before any book-level rescaling, and $\eta_{\mathcal{J}}$ the leg-level rescaling defined below.

In both cases the book is scaled to a fixed gross exposure G . Writing $\tilde{w}_i$ for the pre-normalisation magnitude, each leg is rescaled separately,

$$\eta_{\mathcal{J}} = \frac{B}{\sum_{i \in \mathcal{J}} \tilde{w}_i}, \quad \mathcal{J} \in \{\mathcal{L}, \mathcal{S}\}, \quad (4.13)$$

with the side budget $B = G/2$ when both sides are populated and $B = G$ when only one is. For the traded book each side therefore carries a gross exposure of $G/2$, the book is dollar-neutral by construction, and the realised gross exposure equals G at every

rebalance. A one-sided book, whether a standalone leg or the long-only comparison of Section 6.4, takes the full G , so a leg struck on its own is twice the size of the same names inside the dollar-neutral book, and its exposure and its financing are read off the same ceiling. Under equal weighting the same normalisation yields (4.11) directly.

A common gross exposure is what makes the two schemes comparable: any difference in performance then reflects how capital is distributed across the selected names, not how much capital is deployed. Because a scalar rescaling leaves the Sharpe ratio unchanged, comparisons of risk-adjusted performance are invariant to this normalisation, while reported returns are stated at a common gross scaling. The level of G does, however, enter the net figures directly, since the financing term of Section 4.10 is charged on the book's debit balance, which at $G = 4$ against a unit of equity is one unit of borrowed cash.

The level itself is set at $G = 4$, two units of gross exposure per side, for the reasons Appendix B gives, and it is the only parameter through which the design sets the level of risk. What sizes a position relative to its neighbours is $1/\hat{\sigma}_i$; what sizes the book is G . No volatility target is imposed anywhere in this design, and the realised risk of each book is an outcome rather than a constraint: the fifteen neural cells run between 14.4% and 17.6% annualised, and under an enforced target they would not. That freedom is what makes the volatility channel of (3.3) an object that can move at all, since a book held at a common realised risk would show none by construction.

Table 4.4: Parameters of the mapping from scores to positions. Values are identical across every objective, representation arm, and weighting scheme, and are fixed before estimation. The traded fraction is the one axis that is varied, over the grid in (4.16).

Block	Parameter	Value
Selection	Sector treatment	Within-sector score demeaning
	Traded fraction α per side, headline	10%
	Minimum names per side $k_{\min}$	10
Sizing	Volatility estimator	63-day EWM, annualised
	Sizing constant v^*	0.15
	Volatility floor, measured names	5th percentile of cross-section
	Volatility fill, missing names	95th percentile of cross-section
	Per-name ceiling $\bar{w}$	5.0
	Gross exposure G	4.0
Accounting	Rebalances per year	12
	Seed aggregation	Mean of five within-date rank vectors

The pre-cost portfolio return and one-side net turnover are

$$R_t = \sum_i w_{i,t} y_{i,t}, \quad \text{TO}_t = \frac{1}{2} \sum_i |w_{i,t} - \tilde{w}_{i,t}|, \quad (4.14)$$

where $\tilde{w}_{i,t}$ is the weight the previous book carries into date t . Nothing is traded between

rebalances, so each position grows with its own return while the book grows with the return of the book as a whole,

$$\tilde{w}_{i,t} = w_{i,t-1} \frac{1 + y_{i,t-1}}{1 + R_{t-1}}, \quad (4.15)$$

and it is this drifted book that the rebalance at t trades away from. Weights are aligned by security across dates, and (4.15) puts the trades that restore a drifted position, and those that close one, into the measured turnover and therefore into every charge levied on traded notional.

4.8.3 Leg decomposition

Long and short contributions are reported separately by replaying the joint weights and splitting (4.14) by the sign of $w_{i,t}$, so that the two contributions add back exactly to the total. A leg run as a standalone portfolio with its own normalisation answers the different question of whether that side is viable on its own, and Section 6.4 reports it under that reading, on the excess basis a one-sided book requires: the additive split has nothing to allocate because the traded book’s net investment is zero, while a standalone leg’s is its full gross notional (Section 4.10). Its return does not add back and is no part of this decomposition.

4.9 Tail width

The traded fraction is varied over

$$\alpha \in \{20\%, 10\%, 5\%, 2.5\%, 1\%\}, \quad (4.16)$$

holding the trained models fixed, so that width effects are not confounded with retraining. The decile is the headline width, fixed in advance with the other parameters of Table 4.4. At $\alpha = 1\%$ on a top-1000 universe the book holds ten names per side, and that configuration is reported as a concentration stress test and not as a recommended operating point, since name-level frictions and idiosyncratic gaps dominate a book of that size. Widths are crossed with both weighting schemes, so that the interaction between concentration and sizing is visible.

4.10 Execution regimes and the cost model

Every reported specification is evaluated under the two conventions of Section 3.4. The gross regime applies no costs and enters at the signal close. The realistic regime enters

at the next open, exits at the close of the next rebalance, and charges four components expressed in units of portfolio return: spread and commission, market impact, stock loan, and financing.

A spread and commission term charges each name half its own quoted spread on the date it trades plus a commission of c^{sp} basis points, both levied on the notional that name trades, which sums over the book to 2TO_t . Market impact follows the volatility-scaled square-root law,

$$c_t^{\text{imp}} = \kappa \sum_i \sigma_{i,t} |\Delta w_{i,t}| \sqrt{\frac{|\Delta w_{i,t}| A}{\text{ADV}_{i,t}}}, \quad (4.17)$$

where $\Delta w_{i,t} = w_{i,t} - \tilde{w}_{i,t}$ is the change in weight the rebalance trades, measured against the drifted book of (4.15), A is managed capital, $\text{ADV}_{i,t}$ is trailing average daily dollar volume, $\sigma_{i,t}$ is the name's daily return volatility (the annualised estimate $\hat{\sigma}_i$ that the sizing rule of (4.12) consumes, divided by $\sqrt{252}$), and κ is a dimensionless prefactor. This is the form the impact literature agrees on: the cost of demanding a given fraction of a day's volume is proportional to the day's typical price movement, so impact scales with the name's own volatility rather than with a common dollar rate, and it grows with the square root of participation. A stock-loan term c_t^{borrow} charges a general-collateral rate on short positions plus a hard-to-borrow premium for names whose average daily volume falls below a threshold. A financing term charges a spread on the margin loan the book actually draws,

$$c_t^{\text{fin}} = \frac{1}{12} \cdot \frac{s}{10^4} \cdot \max(0, L_t - 1), \quad L_t = \sum_{i: w_{i,t} > 0} w_{i,t}, \quad (4.18)$$

where s is the spread over the funding benchmark and the base is the debit balance: long market value less the one unit of equity. A short sale is not funded by a margin loan, because it releases cash rather than absorbing it. Selling S_t units short delivers S_t of proceeds, which are pledged to the stock lender and earn a rebate; what the short book costs is the lender's fee, which the stock-loan term has already charged, and levying a funding spread on the same notional would price it twice. The interest lines of a book at $L_t = S_t = G/2$ against a unit of equity are a payment of $(r_f + s)$ on the debit balance $L_t - 1$ and a receipt of $(r_f - b_t)$ on the S_t of proceeds, where b_t is the realised borrow rate. At $G = 4$ these net to $r_f - s - 2b_t$. The risk-free component is interest on the one unit of equity, which is what an excess-return convention removes, so (4.18) carries the spread alone and no risk-free term enters any return reported here. That netting is exact for the dollar-neutral book every reported specification runs, and only for it. The general statement is that the excess return of any book is $\sum_i w_i (y_i - r_f) = \sum_i w_i y_i - r_f \sum_i w_i$, so the risk-free term is charged on *net* investment and not on gross exposure: it vanishes when the book is dollar-neutral, costs Gr_f on a long-only book at gross G (one unit of

equity forgone plus $G - 1$ borrowed), and is *credited* to a short-only book, which is a net lender of its own proceeds. Gross leverage sets the financing spread and the stock-loan fee, not the risk-free term. The one-sided diagnostics of Section 6.4 are struck on that basis for this reason. Gross, the resulting Sharpe ratio is invariant to G , so a figure computed at a leg's own scale carries over to the same book delevered. Net it is not. Spread and commission are charged on the notional the book trades and the stock-loan fee on the notional it holds short, so both rescale with the book; impact grows with the three-halves power of what is traded, by (4.17); and the financing spread is kinked at a gross of one, below which no debit balance is drawn. A levered one-sided book therefore pays charges that no rescaling of a figure struck at another exposure reproduces, and the net comparison of Section 6.4 is run at both exposures rather than converted from one. Net return is

$$R_t^{\text{net}} = R_t - c_t^{\text{sp}} - c_t^{\text{imp}} - c_t^{\text{borrow}} - c_t^{\text{fin}}. \quad (4.19)$$

Of these parameters the impact prefactor is the one with a substantial literature behind it. Tóth et al. (2011) establish the square-root form across markets and instruments, fitting an exponent near 0.5 and taking the prefactor itself to be of order unity. Sato and Kanazawa (2025) confirm that exponent stock by stock on a complete eight-year survey of one exchange, Bouchaud et al. (2018) survey the evidence for it as an established regularity, and the convention of Grinold and Kahn (2000) sits at the same end of the prefactor range. The base case is set below them, at $\kappa = 0.5$, on live-trading evidence rather than on the metaorder literature: the execution record of Frazzini et al. (2018) prices market impact at 13.7 basis points on a trade of two percent of daily volume and 32.3 at ten percent, against 13.4 and 29.9 from (4.17) at an annualised volatility of 30%. Impact is also the only one of the four components that grows with the size of the book, which is what makes capacity a computable quantity rather than an assumption; Section 7.4 computes it.

Table 4.5 lists the parameters, which are fixed in advance and not calibrated to results. They also describe a deliberately plain execution: every fill crosses the quote, each rebalance completes within one day's volume, and leverage sits at a fixed ceiling. No rule is optimised against the costs it generates, so the net results price an unengineered implementation of the signal rather than a trading desk's best effort. Section 7.2 sets out which components are measured and which rest on assumption, and the bridge of Section 7.3 reports each component separately.

Table 4.5: Parameters of the realistic execution regime. Values are fixed before estimation and calibrated to the middle of the published evidence, so that net figures are read as a point estimate under a stated parameterisation.

Component	Parameter	Value
Entry and exit	Entry	Open of $t + 1$
	Exit	Close of the next rebalance
Spread	Charge per side	Half the name’s quoted spread, plus c^{SP}
	Commission c^{SP}	2 basis points per side
	Without a usable quote	Last known half-spread, else commission alone
Market impact	Prefactor κ	0.5, dimensionless, square-root law
	Volatility $\sigma_{i,t}$	Daily: annualised sizing estimate $/\sqrt{252}$
	ADV window	20 trading days
Stock loan	General collateral	50 basis points per year
	Hard-to-borrow premium	200 basis points per year
	Hard-to-borrow trigger	Average daily volume below \$20 million
Financing	Spread on the debit balance $L_t - 1$	100 basis points per year
Capital	Managed capital A	\$1 million in the base case

4.11 Reproducibility and audit trail

Each run writes a manifest containing the resolved configuration, a configuration hash, a feature-cache signature, content hashes of every source module, package versions, and the score and return series, the scores retained per seed as well as in ensemble form. Results in Parts II and III are produced from those stored artefacts and not from ad hoc recomputation, and each reported figure resolves through its run’s manifest to a configuration hash and a stored series. One set of figures has a different provenance and is marked where it appears: the activation diagnostic of Section B.6 is a property of the architecture rather than of any fit, and resolves to the resolved configuration and its generating script instead of to a stored return series. Nothing else in the thesis is obtained that way.

The pipeline, the configurations, the table and figure generators and the test suite are available at https://github.com/quantquant-max/Master_thesis. The licensed CRSP, Compustat and IBES inputs are not redistributable and are not included; the repository documents the fields drawn from each and the cache the pipeline expects.

Part II

Core Empirical Evidence

Chapter 5

Representation and Ranking-Objective Results

This chapter reports the first stage of the transmission chain. It asks whether the numerical mapping applied to within-date feature ranks changes what a neural ranking model learns, whether any such change survives into the traded book, and whether the gradient-boosted control behaves as the invariance argument of Section 4.5 predicts, and why Section 5.2 and Section 5.3 do not look alike at first sight (Section 5.4). Every headline figure is struck at one configuration, namely decile tails, the sizing rule of Table 4.4, and the gross return convention, so that the contrasts reported here are not confounded with the construction and cost decisions studied in Chapter 6 and Chapter 7. Several sections then ask whether a finding survives the rest of the grid or the cost stack, and name the configuration when they do.

5.1 Estimands and the three arms

Every characteristic enters the models as its within-date rank, and three encodings of that same rank are estimated: the centred percentile $u - \frac{1}{2}$, the inverse-normal $\Phi^{-1}(u)$, which spans roughly ± 3.09 at the realised universe size and has a standard deviation of ≈ 1 , and $\sqrt{12}(u - \frac{1}{2})$, which has a standard deviation of ≈ 1 and carries the percentile shape at inverse-normal scale.

That third arm is introduced here as an identification device, but it is not only that. Multiplying the centred percentile rank by $\sqrt{12}$ divides it by the standard deviation a uniform grid has, which at the realised universe size is its own to within a tenth of a percent, so the scale-matched arm is the field's default encoding standardised, which is the single most conventional preprocessing step there is outside this literature. Two features make the standardisation unusually clean here. Every column is a within-date rank, so all columns share one standard deviation and a single constant standardises the entire

surface; and the constant is analytic rather than estimated, so nothing is fitted on the training block and nothing can leak across the split. The arm is therefore both the device that identifies (5.1) and a specification a practitioner might deploy, and Section 5.3.1 reads it in the second role once the estimates are in.

Writing $\text{SR}_\ell(a)$ for the annualised gross long-short Sharpe ratio of objective ℓ under arm a , and s.m. for the scale-matched arm, the total representation contrast decomposes exactly:

$$\underbrace{\text{SR}_\ell(\text{int}) - \text{SR}_\ell(\text{pct})}_{\Delta_\ell^{\text{repr}}} = \underbrace{\text{SR}_\ell(\text{s.m.}) - \text{SR}_\ell(\text{pct})}_{\Delta_\ell^{\text{scale}}} + \underbrace{\text{SR}_\ell(\text{int}) - \text{SR}_\ell(\text{s.m.})}_{\Delta_\ell^{\text{shape}}}. \quad (5.1)$$

The first step changes the magnitude of the inputs and leaves their spacing alone; the second changes the spacing at matched magnitude. Δ^{repr} denotes the total throughout, and shape refers to the matched-scale step. Because the decomposition is an identity it cannot fail to allocate the effect, and it may allocate it somewhere the hypothesis would prefer it did not.

Interpretation follows the protocol of Section 3.7: the pooled Δ^{repr} is the primary family's single member, the five per-objective contrasts a gatekept secondary family, the two components of (5.1) a description of where the contrast sits rather than two further tests, and the gradient-boosted and linear cells members of no family (Section 5.5, Section 4.7). Direction and consistency across objectives carry more weight than the magnitude of any single cell: at 272 monthly observations the standard error on a single Sharpe difference is wide, and the five difference series are estimated on one history and are strongly correlated, so they are worth materially less than five independent replications, a dependence the pooled endpoint prices through the joint bootstrap and Section 5.8 examines directly. Ordering quality and portfolio quality are measured on the same score vectors and are not assumed to agree; where they diverge, the divergence is the object of study, and Section 5.4 shows that much of the apparent gap dissolves once statistics of the same kind are compared.

Reporting convention. Sharpe ratios are invariant to any constant rescaling of the book; returns and drawdowns are not, and the realised volatility of the raw book differs across cells. Two bases are therefore in use, and which one applies depends on whether realised risk is the object being measured or an obstacle to measuring something else. Where a return or a drawdown is set against another at a common risk target (the width grid of Table 6.1, the per-cell depths of Section 8.2), each series is first rescaled by a single scalar that sets its full-sample annualised volatility to 15%. Because that scalar is known only at the end of the sample, a rescaled figure is the one a manager would have recorded having chosen the right leverage in advance: a comparison device rather

than an achievable result. Where realised risk is itself one of the quantities reported, the raw book is given instead (Table 5.3, Table 5.8 and the leg decomposition of Table 6.3), because a common target would normalise away the very thing being measured, and because only on the raw series do the two leg contributions sum to the book’s own return. The return and drawdown columns of those tables are accordingly not comparable across cells on their own, which is why the volatility column sits beside them. Every table states which basis it is on, and every Sharpe ratio, contrast and standard error in this thesis is computed on the raw unscaled series and is untouched by the convention either way.

5.2 Ordering quality

This section measures the score vectors alone: does the encoding change the ordering the model produces, and if so, in what respect.

Two design points make the measurement clean. The universe holds 1000 names at every rebalance, so the traded fraction $k = 100$ is the top decile and a model with no skill scores a precision of 0.100; every precision figure below can be read directly as a lift over chance. And the rank information coefficient is reported both over the full cross-section and *within* each predicted tail. The distinction matters more than it may appear. The full-cross-section coefficient is dominated by the eight deciles the book never holds, so it can move for reasons the portfolio never experiences; the within-tail coefficient asks whether the model orders the hundred names it actually selects. For a fractile book the second is economically irrelevant by construction, since only membership matters, which is precisely why it is informative about *where* the model’s resolution sits. Table 5.1 reports these statistics cell by cell and Table 5.2 pools them with their inference; the paragraphs below read the pooled figures.

The level of ranking skill does not move. Pooled over the five neural objectives, the full-cross-section rank information coefficient is close to identical in all three arms, (Table 5.2), with the gradient-boosted control a little above them and the reference points of Section 5.6 below all four: the linear anchors at roughly three fifths of the neural level, the momentum rule at a fifteenth of it (Table 5.1, Table 5.8). The total contrast is +0.00196 with a HAC standard error of 0.00253, $t = 0.77$ and $p = 0.4390$, and neither step is sharply resolved on its own, at $p = 0.7505$ for the scale step and $p = 0.0781$ for the shape step. On the statistic the applied literature most often selects models by, the three arms are indistinguishable.

The coefficient is steadier, and only the first step makes it so. The standard deviation of the monthly coefficient falls from 0.1063 to 0.0899 and then to 0.0892. The pattern holds in every objective and is concentrated entirely in the first step, which

Table 5.1: Ordering quality, measured on the score vectors alone. IC is the mean per-date Spearman correlation over the full cross-section, sd its dispersion across dates and IR their ratio. IC_L and IC_S are the same coefficient computed *within* the $k = 100$ names selected on each side; P_L and P_S are the precision of each predicted tail against the realised decile. The universe is exactly 1000 names, so a no-skill model scores 0.100. Recall matches precision to within 0.0002 and is not printed. The scale-matched arm has no gradient-boosted counterpart.

Objective	Arm	Rank IC, full cross-section			Within tail		Precision	
		mean	sd	IR	IC_L	IC_S	P_L	P_S
RankNet	Percentile	0.0395	0.1057	0.373	+0.0086	+0.0250	0.1111	0.1515
	Scale-matched	0.0396	0.0897	0.442	+0.0070	+0.0280	0.1138	0.1548
	Inverse-normal	0.0406	0.0887	0.458	+0.0180	+0.0279	0.1146	0.1542
LambdaRankIC	Percentile	0.0389	0.1042	0.374	+0.0048	+0.0193	0.1040	0.1377
	Scale-matched	0.0396	0.0889	0.446	+0.0022	+0.0282	0.1040	0.1403
	Inverse-normal	0.0414	0.0888	0.466	+0.0163	+0.0123	0.1066	0.1406
Margin ranking	Percentile	0.0352	0.0992	0.355	+0.0145	+0.0198	0.1047	0.1508
	Scale-matched	0.0365	0.0928	0.393	+0.0195	+0.0230	0.1094	0.1495
	Inverse-normal	0.0384	0.0911	0.421	+0.0212	+0.0255	0.1093	0.1496
Soft-Spearman	Percentile	0.0391	0.1115	0.351	+0.0064	+0.0191	0.1129	0.1561
	Scale-matched	0.0396	0.0894	0.443	+0.0206	+0.0251	0.1144	0.1598
	Inverse-normal	0.0405	0.0888	0.456	+0.0129	+0.0176	0.1162	0.1619
RankCosine	Percentile	0.0389	0.1109	0.351	+0.0076	+0.0213	0.1131	0.1556
	Scale-matched	0.0400	0.0886	0.452	+0.0216	+0.0238	0.1142	0.1579
	Inverse-normal	0.0405	0.0884	0.458	+0.0128	+0.0190	0.1162	0.1603
XGBoost (control)	Percentile	0.0439	0.1034	0.424	+0.0204	+0.0354	0.1120	0.1588
	Inverse-normal	0.0439	0.1034	0.424	+0.0204	+0.0354	0.1120	0.1588

Table 5.2: Ordering statistics with inference, pooled across the five neural objectives. Levels are given for the percentile (pct.), scale-matched (s.m.) and inverse-normal (int.) arms of Table 4.1. The rank IC and precision statistics are formed per rebalance date, averaged across objectives, and tested with Newey–West HAC standard errors. The information ratio is not: it is a mean over a standard deviation and cannot be formed per date, so its levels are the mean of the five per-objective full-sample ratios and its contrasts come from the studentised circular block bootstrap the Sharpe contrasts use, on a Ledoit–Wolf HAC standard error. The three right-hand columns are the two steps of (5.1) and their sum, each a *paired* per-date difference, so common variation cancels before the test. Precision levels are tested against 0.100 and the rank IC rows against zero; the information ratio levels are not tested and carry no stars. Stars mark *unadjusted* values; of the seven ordering members of the exploratory family only the information ratio and the tail asymmetry survive Benjamini–Hochberg. ***, **, *: $p < 0.01, 0.05, 0.10$.

Statistic	Level by arm			Decomposition		
	pct.	s.m.	int.	Δ_{scale}	Δ_{shape}	Δ_{repr}
Rank IC, full cross-section	0.0383***	0.0391***	0.0403***	+0.0008	+0.0012*	+0.0020
Rank IC, within top decile	0.0084	0.0142***	0.0162***	+0.0058	+0.0020	+0.0079
Rank IC, within bottom decile	0.0209***	0.0256***	0.0205***	+0.0047	−0.0052	−0.0005
Precision, top decile	0.1092***	0.1112***	0.1126***	+0.0020	+0.0014**	+0.0034**
Precision, bottom decile	0.1503***	0.1524***	0.1533***	+0.0021	+0.0009	+0.0030
IC information ratio	0.3606	0.4351	0.4517	+0.0745***	+0.0166*	+0.0911***

compresses the dispersion by between 6.4% for Margin ranking and 20.1% for RankCosine, against between 0.2% and 1.8% for the second. Pooling the five per-objective log variance ratios, the scale step gives a ratio of standard deviations of 0.846. The estimator is the one Section 3.7 defers to this point: a circular block bootstrap on the pooled log variance ratio, resampling the same six-month blocks as the Sharpe bootstrap over 2000 replications, with the standard error taken as the dispersion of the replicates. It returns a t of -7.84 and no replicate of the two thousand as extreme as the point estimate ($p < 0.0005$, the resolution of that bootstrap), and all five objectives reject equality of variances individually, the Pitman–Morgan paired-variance test, which assumes away the serial dependence the block bootstrap prices in, agreeing cell by cell, so the finding does not rest on the block length (Section C.2). At the shape step the pooled ratio is 0.992 with $p = 0.3695$ and no objective rejects.

The information ratio of the coefficient. Mean divided by standard deviation is the ranking analogue of a Sharpe ratio, and it is the ordering statistic that moves. It is 0.361 in the percentile arm, 0.435 in the scale-matched arm and 0.452 in the inverse-normal arm. The scale step contributes $+0.0745$ ($t = 3.18$, $p = 0.0010$) and the shape step $+0.0166$ ($t = 1.86$, $p = 0.0635$), for a total of $+0.0911$ ($t = 3.77$, $p = 0.0005$). The total is one of only two ordering members of the exploratory family to survive the Benjamini–Hochberg correction of Section 3.7, at an adjusted p of 0.0057; it is, on that evidence, the one ordering statistic the representation genuinely moves. Four fifths of the improvement is delivered by a transformation that cannot reorder anything. Since the mean coefficient is flat across the arms, this is a statement about stability and not about accuracy, and the compression works on both tails of the monthly distribution: the worst months are less bad, and the best months are less good by nearly as much. What the inverse-normal arm buys is a steadier coefficient rather than a better one.

Membership at the traded fraction. Precision is significantly above the 0.100 no-skill benchmark in all three arms and markedly asymmetric between the tails, the bottom decile running roughly four percentage points above the top in every one (Table 5.2). The total contrast in the top decile is $+0.00341$ ($t = 2.04$, $p = 0.0411$) and, unlike everything above, it is the second step that is the better resolved: $+0.00142$ ($t = 1.96$, $p = 0.0499$) against $+0.00199$ ($p = 0.1908$) for the first. That is a statement about precision and not about size, the second step being the smaller of the two. The bottom decile moves $+0.00298$ ($p = 0.1519$) with neither step significant. Both totals are members of the ninety-seven-contrast exploratory family of Section 3.7, and neither survives the Benjamini–Hochberg correction at $q = 0.10$ within it: the adjusted values are 0.1813 for the top decile and 0.3684 for the bottom. The membership result is therefore recorded as suggestive. It is worth recording because it is the only ordering statistic on which

the shape of the mapping, rather than its magnitude, reaches conventional significance unadjusted. The shape step also moves the bottom-tail coefficient, as the next paragraph shows, though not by enough to separate it from zero.

Where the resolution sits. The within-tail coefficients are much smaller than the full-cross-section figure in both tails and all three arms (Table 5.2). Far less of the model's resolution lies in ordering the hundred names it selects than in separating them from the rest of the cross-section. For a book whose weights do not depend on the ordering inside a tail, which the inverse-volatility sizing of Table 4.4 is as much as equal weighting, that is the favourable configuration. The bottom tail is also the one place where the total conceals its parts: its contrast is -0.00045 , indistinguishable from nothing, and is the sum of two movements of similar size, $+0.00471$ at the scale step and -0.00516 at the shape step, that happen to cancel.

Asymmetry between the tails. Bottom-decile precision exceeds top-decile precision by $+0.0411$ pooled over the fifteen neural cells, at $t = 7.55$. No other contrast on a mean in this chapter comes close, and it clears the exploratory correction of Section 3.7 by the widest margin of any member. The models identify losers considerably more reliably than they identify winners, and the per-cell precisions of Table 5.1 run the same way in every neural cell, so it is not the property of one objective or one arm. The asymmetry sits awkwardly against the implementation. It is the short leg that bears the borrow charge of Chapter 7, and the short leg carries a negative standalone Sharpe ratio in every cell, between -0.292 and -0.414 across the ten neural specifications and at no cell reaching conventional significance (Section 6.4). The skill is concentrated where it is most expensive to monetise, and the finding is carried forward on that basis.

What the section establishes. Neither step changes the level of ranking skill by an amount the sample can resolve; what each does move is collected in Table 5.6.

5.3 Economic translation

The same score vectors are now run through the portfolio rules of Table 4.4. Table 5.3 reports the levels and Table 5.4 the paired contrasts with their inference.

Levels. The three arms occupy three non-overlapping bands of gross Sharpe ratio, though the margin that separates them is not uniform: the scale-matched arm clears the percentile arm by 0.111 of Sharpe ratio at the closest point, whereas the inverse-normal arm clears the scale-matched arm by 0.012 , which is nothing at this sample length. The non-overlap is a description of the fifteen cells and not a claim that any two of them are

Table 5.3: Gross long-short performance by objective and arm at decile width, over the 272 monthly out-of-sample rebalances from April 2003 to November 2025 on the top-1000 United States universe. Every figure is the *raw* book: return, volatility and drawdown carry no volatility normalisation, the realised risk being one of the quantities measured here. Turnover is one-side net per rebalance. Cells share the universe, label, folds, seeds and portfolio rules, and differ only in the encoding of the feature ranks.

Objective	Arm	Sharpe	Return	Volatility	Max DD	Turnover
RankNet	Percentile	1.009	17.77%	17.60%	-44.75%	2.65
	Scale-matched	1.289	19.93%	15.45%	-29.82%	2.76
	Inverse-normal	1.490	22.88%	15.35%	-34.53%	2.77
LambdaRankIC	Percentile	1.083	17.62%	16.27%	-42.69%	2.66
	Scale-matched	1.386	20.35%	14.68%	-28.75%	2.76
	Inverse-normal	1.586	22.80%	14.38%	-30.20%	2.77
Margin ranking	Percentile	1.123	17.39%	15.48%	-45.19%	2.63
	Scale-matched	1.285	20.01%	15.57%	-40.72%	2.70
	Inverse-normal	1.398	21.05%	15.06%	-45.46%	2.70
Soft-Spearman	Percentile	1.119	19.43%	17.36%	-42.43%	2.65
	Scale-matched	1.287	20.01%	15.54%	-34.56%	2.74
	Inverse-normal	1.452	21.99%	15.15%	-39.31%	2.73
RankCosine	Percentile	1.174	20.34%	17.33%	-42.27%	2.66
	Scale-matched	1.382	20.65%	14.95%	-32.02%	2.74
	Inverse-normal	1.431	21.97%	15.35%	-44.33%	2.74

Table 5.4: Paired Sharpe differences and their inference, computed on the 272 monthly rebalances of April 2003 to November 2025. Δ^{repr} is the inverse-normal minus percentile contrast of (5.1). Standard errors are Ledoit–Wolf HAC, intervals ± 1.96 of them, p_{HAC} the unadjusted two-sided value and p_{RW} the Romano–Wolf stepdown over the five objectives. The pooled row is the primary endpoint, carries a studentised bootstrap interval, and needs no multiplicity adjustment.

Objective	Contrast	ΔSR	95% CI	p_{HAC}	p_{RW}
RankNet	Total	+0.481	[+0.270, +0.692]	0.000	0.000
LambdaRankIC	Total	+0.503	[+0.299, +0.707]	0.000	0.000
Margin ranking	Total	+0.275	[+0.011, +0.540]	0.041	0.115
Soft-Spearman	Total	+0.333	[+0.084, +0.581]	0.009	0.036
RankCosine	Total	+0.257	[−0.018, +0.532]	0.067	0.115
Pooled	Total	+0.370	[+0.216, +0.523]	0.000	—

separated by the evidence. Raw annualised return rises from a pooled 18.51% to 20.19% to 22.14%, and raw annualised volatility falls from 16.81% to 15.24% to 15.06%, so the ordering by arm is reproduced in both components separately, though as the tests below show the two components are not moved by the same step.

The primary endpoint. The pooled total contrast is $D = +0.370$ gross with a Ledoit–Wolf HAC standard error of 0.083 at a Newey–West lag of 4, $t = 4.46$, studentised bootstrap $p < 0.00025$ and interval $[+0.216, +0.523]$. Net of costs it is $D = +0.307$, standard error 0.078, $t = 3.92$, $p = 0.00050$, interval $[+0.158, +0.457]$. This is a single pre-specified quantity and no multiplicity adjustment applies to it.

The secondary family tells a less uniform story, and it is reported for that reason. Under the Romano–Wolf stepdown three of the five per-objective contrasts clear the 5% level gross and only two net, RankNet and LambdaRankIC surviving in both regimes and Soft-Spearman gross alone (Table 5.4, Section C.2). Only two of the five objectives are individually significant in both cost regimes, and the pooled estimate is no licence to read the effect into the remaining three.

The decomposition, and what it costs the hypothesis. Gross, the scale term is $\Delta^{\text{scale}} = +0.224$ with a standard error of 0.097, $t = 2.32$, $p = 0.0197$ and interval $[+0.040, +0.409]$; the shape term at matched scale is $\Delta^{\text{shape}} = +0.146$, standard error 0.070, $t = 2.09$, $p = 0.0652$, interval $[-0.009, +0.301]$. Net the pattern repeats with both terms weaker: $+0.188$ ($t = 2.03$, $p = 0.0447$) and $+0.119$ ($t = 1.93$, $p = 0.0747$), the scale term keeping an interval clear of zero at $[+0.004, +0.373]$ and the shape term not, at $[-0.013, +0.251]$. Which of the two is the larger, gross, is the part of this result that depends on where the sample stops: the ordering printed here holds at terminal dates from April 2025 and reverses at every one from July 2013 to March 2025, for the reason Section 8.6 gives: the crossing sits inside the terminal drawdown, which the scale-matched arm weathered better than the percentile one.

A pure rescaling of the inputs, one that cannot alter any ordering and to which the gradient-boosted control is provably indifferent, moves the traded outcome by more than the change in spacing does. The shape term is only marginally significant in either regime and its confidence interval contains zero in both. The tail spacing is therefore not established as the carrier of the effect, and no claim that it is the sole or demonstrated mechanism is made anywhere in what follows.

What the larger step is instead is a question about the architecture rather than about the returns, and Section B.6 answers it from the network alone. The input scale governs how far the first block is driven into the curved region of its activation, and because the normalisation follows the activation rather than preceding it, nothing downstream undoes that. Measured at initialisation the scale step moves the first block’s departure from an

Table 5.5: Panel A splits the total contrast into its two steps through the scale-matched arm, which carries the percentile shape at inverse-normal scale and is abbreviated s.m. in the step labels. Δ^{scale} changes input magnitude alone, Δ^{shape} their spacing at matched magnitude, and the two sum to Δ^{repr} by construction, so the third row is an arithmetic check rather than a third estimate. Panel B re-estimates every step with each cell trained to its own validation optimum under the one-SE rule, beside the fixed-budget figure, so the question it answers is whether the budget determines how the total divides between the two steps and not merely whether the total survives. Standard errors are Ledoit–Wolf HAC; pooled rows carry studentised bootstrap intervals. Both panels are unadjusted: they partition or re-estimate a quantity already tested.

Panel A: the two steps (pooled over the five objectives)

Step	Regime	ΔSR	SE	95% CI	p
Δ^{scale} (pct $\rightarrow$ s.m.)	gross	+0.224	0.097	[+0.040, +0.409]	0.020
	net	+0.188	0.093	[+0.004, +0.373]	0.045
Δ^{shape} (s.m. $\rightarrow$ int)	gross	+0.146	0.070	[−0.009, +0.301]	0.065
	net	+0.119	0.062	[−0.013, +0.251]	0.075
Δ^{repr} (sum)	gross	+0.370	0.083	[+0.216, +0.523]	0.000
	net	+0.307	0.078	[+0.158, +0.457]	0.000

Panel B: each step by training schedule (pooled over five objectives)

Step	Regime	ΔSR fixed	ΔSR conv.-matched	95% CI	p
Δ^{repr}	gross	+0.370	+0.290	[−0.009, +0.590]	0.058
	net	+0.307	+0.191	[−0.068, +0.449]	0.145
Δ^{scale}	gross	+0.224	+0.176	[−0.032, +0.383]	0.091
	net	+0.188	+0.077	[−0.115, +0.270]	0.426
Δ^{shape}	gross	+0.146	+0.115	[−0.043, +0.272]	0.152
	net	+0.119	+0.113	[−0.020, +0.246]	0.095

Retained, gross: Δ^{repr} 78%, Δ^{scale} 78%, Δ^{shape} 79%. Descriptive: a ratio of two estimates, with no standard error of its own.

affine map and the shape step at matched scale leaves it where it found it (Section 4.2.1), so the two steps take the same order in the diagnostic that they take in the Sharpe contrast: the step that changes how much of the network’s nonlinearity is engaged is the larger and the better resolved, and the step that changes none of it is the smaller and the marginal one. Four cautions attach. The diagnostic is computed at initialisation and knows nothing about what these models went on to learn, so it is a mechanism consistent with the decomposition rather than a demonstration that the decomposition arises from it. The first-layer weights are free to adapt over the training budget and trained weights are not stored, so whether the gap survives ten epochs is open. The curve traced there is steep only between the two encodings a modeller would choose between and flattens on either side, so the channel is not monotone in input magnitude in any useful sense and multiplying the inputs by a larger constant again would not be expected to buy more. And the route it identifies is a property of this activation and this block order: it would be largely closed by a positively homogeneous activation, which is a prediction the design

does not test.

The split of the total into roughly three fifths and two fifths is a ratio of two estimated quantities and its sampling variability is much larger than the two point estimates suggest. What the chapter leans on is that the two steps are separately identified, and that they are identified as doing different things, which Section 5.3.2 establishes.

5.3.1 What this means for a practitioner

Read as a decomposition, (5.1) allocates an effect between two steps. Read as advice, it says something narrower and more useful, and the reading costs nothing because the arm that delivers it is already estimated.

The first step is standardisation. Going from the field's default encoding to the same encoding divided by its own standard deviation is worth three fifths of the total contrast in each regime at the full sample. Pooled over the five objectives it is positive in all ten of the width and weighting cells of Table C.1 and Table C.2, gross and net alike, running from +0.256 to +0.039 gross and from +0.188 to +0.038 net. It changes one constant. It fits nothing, so it cannot leak, and it leaves the ordinal content of the surface exactly where it found it.

Its share of the total is not constant across the grid, running gross from four fifths at the quintile to about a third at 2.5%, and net from about three quarters to the same third. At 1% the shape step turns negative and the total collapses to +0.040 gross and +0.051 net, a denominator too close to zero for the share to be identified. Nor is the sign constant cell by cell. Of the hundred individual objective-by-width-by-scheme cells across the two cost regimes, fourteen have the percentile arm ahead of the scale-matched one. They are not scattered: every one belongs to Soft-Spearman or RankCosine, the two listwise objectives, and ten of the fourteen sit at the 2.5% or 1% fractions. Both are also the objectives for which the fixed percentile target is itself an active channel, since their losses consume the encoded target numerically while the three pairwise losses are invariant to it (Section 4.3.1); in those cells a target-encoding interaction cannot be separated from a representation effect, and the design holds the target fixed rather than identifying it. The recommendation below is a pooled one and should be read as such.

None of this touches the advice, because the two steps are sequential rather than alternatives. Standardising comes first whichever of the two proves the larger, and its level, unlike its share, is stable: across the 153 stopping dates of Section 8.6 it runs from +0.111 to +0.237 gross and from +0.075 to +0.203 net at the traded decile, positive at every one. What the terminal date moves is how much the second step adds on top of it.

The second step is the inverse-normal transform, and it is the one the design cannot underwrite. At matched scale it adds +0.146 gross and +0.119 net, with an interval covering zero in both regimes (Table 5.5). It has the sign the tail-spacing claim of Hypothesis 1

predicts in every objective, and this sample cannot separate it from nothing.

Stated that way the finding is less exotic than the framing of the hypothesis and more directly usable. It is not that ranks should be mapped through Φ^{-1} . It is that percentile ranks presented on $[-\frac{1}{2}, \frac{1}{2}]$ are presented at a scale this class of network does not handle neutrally, that correcting the scale recovers between roughly 0.25 and 0.65 of what the inverse-normal encoding delivers depending on where the sample is stopped, and that whether the remaining distance is worth anything is an open question rather than a result. Section 9.3 carries the recommendation and Section B.6 the mechanism it points to.

Two limits belong with it. The scale-matched arm was pre-specified as an identification device and not as a candidate specification, so reading it as advice is a reading of the pre-specified decomposition rather than a hypothesis this study set out to test; nothing was selected on its outcome, since all three arms are reported at every operating point, but no confirmatory claim attaches to the recommendation either. And the standardising constant is exact only because every column is a rank. On a surface carrying raw characteristics alongside ranked ones the same step would require an estimated scale, at which point the leakage question the analytic constant avoids would have to be answered.

5.3.2 Which channel carries each step

A Sharpe ratio can rise because the numerator rises or because the denominator falls, and the two readings are not equivalent: the first is a claim about selection, the second a claim about the risk of the resulting book. (3.3) separates them exactly, and each is tested here in levels. The share each channel takes of a step is not identified at this sample length, its denominator being a Sharpe difference that passes close to zero in resampling. No proportions are quoted below.

The return channel moves at the second step and not at the first. The pooled annualised mean return rises by 1.68 percentage points at the scale step, with a HAC standard error of 1.51, $t = 1.11$ and $p = 0.2651$, and by 1.95 points at the shape step, standard error 0.80, $t = 2.44$ and $p = 0.0146$. Per objective the shape step is individually significant for RankNet (+2.95%, $p = 0.0107$) and LambdaRankIC (+2.45%, $p = 0.0157$) and no objective is individually significant at the scale step. The total is +3.63 points ($t = 2.80$, $p = 0.0050$).

The volatility channel moves at the first step and not at the second. Pooling the five per-objective log variance ratios, the scale step gives a volatility ratio of 0.907 with $t = -2.97$ and $p = 0.0020$, and four of the five objectives reject equality of variances individually under the Pitman–Morgan test, at p -values between 0.00001 and 0.0016. The shape step gives 0.988 with $p = 0.5865$, and no objective rejects. Margin ranking is the exception at the scale step, with a ratio of 1.006 and no reduction at all, as it is

in the corresponding test on the full contrast. Over the two steps together the ratio is 0.896 ($t = -4.23$, $p < 0.0005$). Net of costs the same pattern holds, with the shape step’s return effect at +1.72 points ($p = 0.0318$) and its volatility ratio at 0.993 ($p = 0.7530$).

Both movements rest on a test of their own, each absence measured against zero rather than inferred from the presence of the other. Separating the two steps from one another is a further question, and the last row of Table 5.6 answers it for dispersion and not for the means. One consequence cuts in the hypothesis’s favour and is easy to miss: the shape step’s return effect ($p = 0.0146$) is better resolved than the shape step’s Sharpe contrast ($p = 0.0652$). Combining a real movement in the numerator with a noisy absence in the denominator adds variance to the ratio, so the Sharpe understates how well that step is identified.

The volatility reduction is measured under inverse-volatility sizing, so the branch Hypothesis 4 distinguishes is live: a reduction appearing only under that sizing rule would be an interaction with the construction rather than a property of the selection. It is settled in Chapter 6 by the primary endpoint re-struct under equal weighting, which is what the hypothesis names as the test. The comparison between the two schemes is a different quantity and sits in the exploratory tier, where none of fifteen attains significance under the Benjamini–Hochberg correction.

5.3.3 The training budget

The estimates above are taken at one fixed training budget, and a fixed ten-epoch schedule may measure the arms at different distances from their own optima. Under the one-standard-error rule of Section 4.4.2 each cell trains to its own validation optimum, and the contrast is re-estimated on that basis.

The schedule is re-run for all five objectives in all three arms, so what is re-estimated is the decomposition and not only the total. Gross, the contrast keeps its sign in every objective, from +0.203 for margin ranking to +0.394 for RankNet, and pooled it stands at +0.290 against the fixed-budget +0.370, retaining 78%, with standard error 0.136 and interval $[-0.009, +0.590]$. At this operating point it does not clear conventional significance, $p = 0.0585$ gross and $p = 0.1447$ net, and the reason is set out below: it is a property of the treatment rather than of the effect.

The informative part is how that shortfall is distributed. Δ^{scale} falls from +0.224 to +0.176 and Δ^{shape} from +0.146 to +0.115 (Table 5.5, Panel B), retaining 78% and 79%. Little follows from the closeness of those two proportions on its own. Each is a ratio of two estimated quantities and neither carries a standard error: across the five objectives the scale retention runs from 0.38 to 1.09 and the shape retention from 0.41 to 3.79, so two pooled ratios landing a point apart is arithmetic rather than evidence. What the question is actually about is the gap between the two steps, and that is a quantity with a

distribution. Under the fixed budget the shape step sits 0.079 below the scale step; under convergence matching it sits 0.061 below. The change is +0.018 with standard error 0.139 ($p = 0.8988$), so gross the division of the total between the two steps does not move with the training budget. Net the change is larger, +0.104 against 0.136 ($p = 0.4501$), inside its own noise but not negligible as a point estimate; the closing paragraph of this subsection takes it up.

That matters for the reading of Section 5.3.2, because the most natural objection to the scale step is that the two arms sit at different points of their own optimisation paths after the same number of epochs, so the scale-matched arm might simply be the further advanced of the two at ten. Were that the mechanism, letting every cell train to its own optimum should cost the scale step far more than the shape step. Gross it costs the scale step 0.049 against the shape step's 0.031. That difference runs in the direction the objection predicts, and it is an eighth of its own standard error, which is not the margin the objection needs.

The wider standard errors are not a thinner sample. These cells average the same five seeds as every other neural cell, and the widening follows from the treatment itself. Re-tuning the epoch budget arm by arm makes the two books behave less alike: the paired correlation between them sits between 0.717 and 0.779 across the five objectives, against roughly 0.84 under the fixed budget. A paired Sharpe difference is estimated the more precisely the more correlated its two legs are, so standard errors rise from a range of 0.104 to 0.140 up to 0.145 to 0.246 on contrasts of much the same size, and only RankNet and LambdaRankIC clear conventional significance individually ($p = 0.029$ and $p = 0.015$). The same decorrelation reaches the pooled rows, and it reaches them unevenly. The standard error of the total rises from 0.083 to 0.136, while those of the two steps barely move, from 0.097 to 0.104 and from 0.070 to 0.076. A step spans one change of arm and the total spans two, so the total is where the lost pairing accumulates, and the total is accordingly the row whose interval widens, while each step is estimated about as sharply as it was before. The pooled sign is carried by all five; the individual intervals are simply wider.

Net, the two steps come apart, and the ordering between them reverses. Δ^{shape} is almost untouched, +0.113 against +0.119, while Δ^{scale} retains under half of its fixed-budget value, +0.077 against +0.188, so at this operating point the shape step is the larger of the two and the better resolved as well, at $p = 0.095$ against $p = 0.426$. The advantage the scale step buys is the one that costs absorb once each cell is trained to its own optimum, which is consistent with its being a volatility effect rather than a selection effect: Section 5.3.2 locates the scale step in the risk of the book, and a risk advantage is worth less after a cost stack charged on turnover. The difference between the two steps moves by +0.104 in the process, from -0.069 to $+0.036$ in favour of the shape step, three quarters of its own standard error, so the reversal belongs to the point estimates rather

than to what the sample resolves. It is nonetheless the second vantage point from which the division of the total does not hold, the first being the terminal date of Section 8.6; what survives both is the level of the scale step, which stays positive here as it does at every stopping date, and it is the level that Section 5.3.1 reads as advice.

The pooled shortfall of a fifth gross is not attributed here. It is small against standard errors of this size, and the design does not separate a genuine budget effect from the ordinary variation of a re-estimated optimum.

5.3.4 The path, turnover and the common drawdown

Figure 5.1 shows what these differences look like as a path instead of as a summary statistic. The question the figure answers is whether the arms are distinguished by a single episode or by a gap that accumulates gradually. A period-by-period attribution of the contrast is not reported, so the figure is read for shape. Two facts bear on the question from opposite sides. The gap is not the property of one decade: pooled over the five objectives it is +0.443 before 2010, +0.396 through the 2010s and +0.331 in the 2020s, so it narrows without closing. Against that, part of the separation is visible in the terminal episode, where on the volatility-scaled series the inverse-normal arm loses less over calendar 2025 for RankNet (-26.72% against -32.93%), LambdaRankIC (-21.52% against -33.32%) and Margin ranking (-28.51% against -34.82%), and more for Soft-Spearman (-27.79% against -22.88%) and RankCosine (-32.30% against -24.24%). That episode does not carry a common sign across objectives and cannot on its own be the source of the pooled contrast.

Figure 5.2 plots the per-objective contrasts and their two components on a common scale. Every objective places the inverse-normal arm ahead in both cost regimes. Gross, the total contrasts are +0.503 for LambdaRankIC, +0.481 for RankNet, +0.333 for Soft-Spearman, +0.275 for Margin ranking and +0.257 for RankCosine (Table 5.4); net, in that same order, +0.434, +0.404, +0.252, +0.265 and +0.183, so that Margin ranking and Soft-Spearman exchange places. The magnitudes do not pattern by family. The pairwise family supplies both the two largest contrasts and, in Margin ranking, one of the two smallest, and the listwise pair straddles it. Pairwise and listwise objectives interleave rather than separate.

The two components inherit the pooled ordering of Section 5.3 objective by objective. Gross, the scale step is the larger of the two in all five objectives; net, in four, Soft-Spearman being the one cell where the shape step is larger. The per-objective intervals overlap heavily and no cell separates its own two steps, so this is the pooled result seen unpooled rather than five further findings.

Turnover is higher in the inverse-normal arm for every objective, so the better-performing arm is also the more expensive one to trade, though the difference between

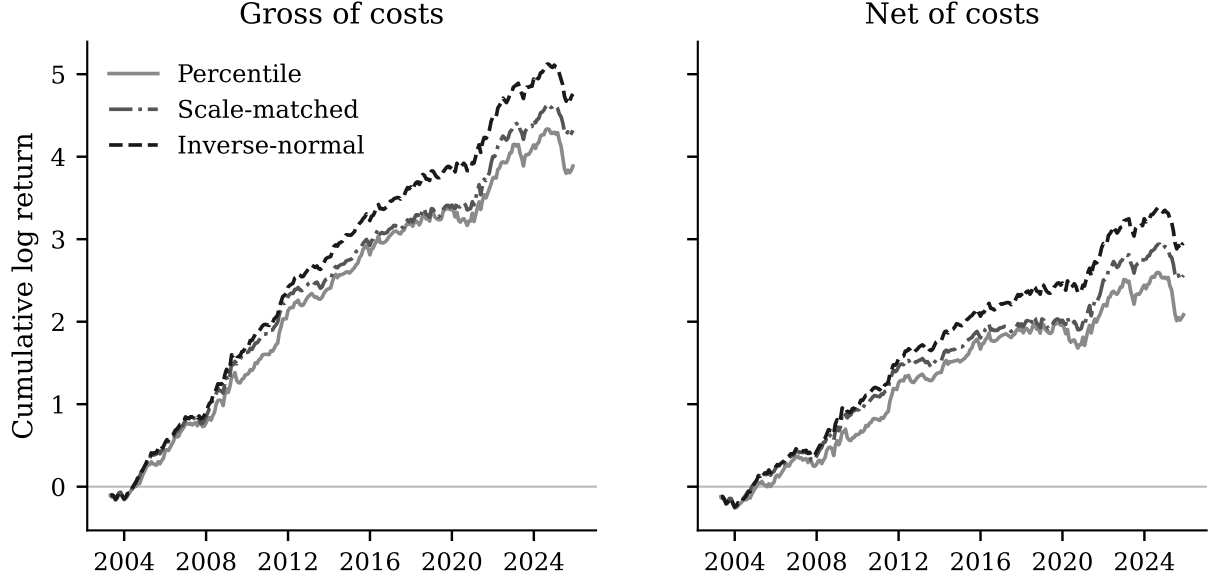


Figure 5.1: Cumulative log return of the headline long-short book by representation arm over April 2003 to November 2025, averaged across the five neural objectives, gross and net of costs. All three arms of Table 4.1 are drawn, the scale-matched one dot-dashed between them, and each is identified by the in-panel legend. The vertical scale is common to both panels, so the compression of the net curves relative to the gross curves is the cost of trading the strategy and not an artefact of scaling.

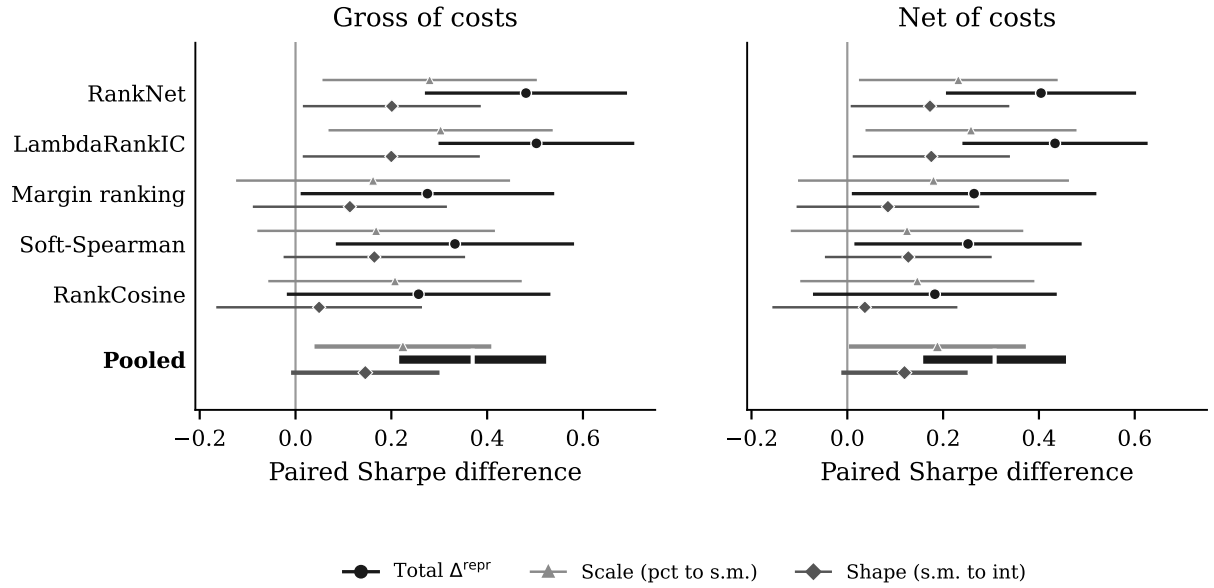


Figure 5.2: Paired Sharpe differences by ranking objective, gross and net, over the 272 rebalances of April 2003 to November 2025. The heavy dark marker on each row is the total contrast Δ^{repr} ; the two lighter markers offset above and below it are the steps of (5.1), the scale step from the percentile to the scale-matched arm and the shape step from the scale-matched to the inverse-normal arm. Bars are ± 1.96 Ledoit–Wolf HAC standard errors, the pooled rows carry studentised bootstrap intervals, and the vertical rule marks zero. The individual intervals are wide at this sample length, which is why the pooled contrast and not any single cell is the primary endpoint. The per-objective split of the two steps appears here and in no table; Table 5.5 reports it pooled.

the arms is small beside the turnover either of them runs. What that costs is settled in Chapter 7 and taken up again in Section 5.9.

The drawdown column moves with the Sharpe column only weakly, and for a reason that makes the comparison less informative than it appears: every one of the fifteen neural cells records its maximum drawdown in the same episode, in 2025, and none has recovered by the last observation. The gradient-boosted control, which is blind to the encoding, records it too. It is a short-leg event throughout, and Section 8.2 decomposes it across every specification.

5.4 Why the two do not look alike

Set the two preceding sections against one another and there is an apparent conflict. The level of ranking skill does not move at all, while the portfolio moves a great deal, and Figure 5.3 shows that the first of those two statements holds month by month and not only on average. Read literally, the inverse-normal arm orders the cross-section no better than the percentile arm and builds a substantially better book out of an equally good ordering, which is the non-equivalence (2.1) asserts. Most of the gap, however, is an artefact of which two statistics have been placed side by side.

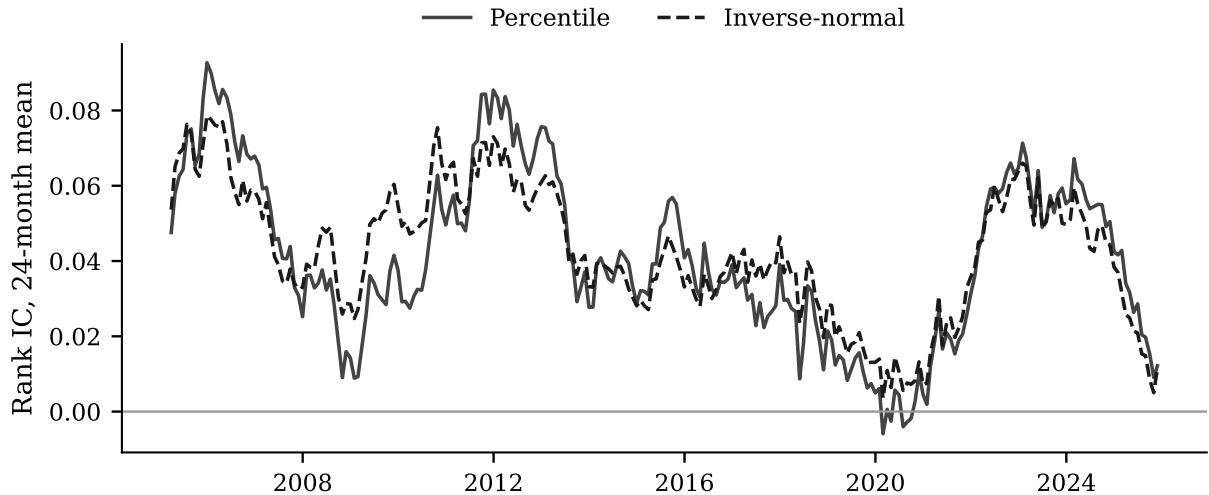


Figure 5.3: Trailing twenty-four-month mean of the monthly rank information coefficient, averaged across the five neural objectives, for the two arms a modeller would choose between. The two series track each other through the sample, the gap between them averaging 0.008 in absolute value against a level near 0.042, which is the sense in which the pooled difference of $+0.0020$ is small. Both decline over the sample and both reach their lowest reading in February 2020, the percentile arm at -0.006 and the inverse-normal arm at $+0.004$; the percentile arm is the only one that falls below zero, and it does so in five months of that year. Figure 5.5 tracks the dispersion of the same coefficient, which separates the arms far more cleanly than the level does.

Comparing like with like. A Sharpe ratio is a mean divided by a standard deviation. The mean rank information coefficient is a mean. The ordering statistic that corresponds to a Sharpe ratio is the information ratio of the coefficient, and it rises by 25%, from 0.361 to 0.452, against the Sharpe ratio's 34%. Measured the same way, the two criteria move by comparable amounts and in the same direction, and the divergence largely dissolves. What remains of it is the ordinary observation that a risk-adjusted statistic and a raw one need not agree.

What each step moves. Table 5.6 sets the four measurements against each other, and row by row the pattern is clean: the scale step moves the two dispersions and neither mean, the shape step moves the two means and neither dispersion.

That pattern is not by itself a dissociation. Each of those rows tests one step against zero, and a claim that the two steps move different things is a claim about the difference between them, which no row above tests. The distinction decides the matter here: on the return column the two estimates are +1.68% and +1.95%, and what separates them is that one clears 5% and the other does not. The last row of the table tests the difference itself, and it divides the claim in two.

On dispersion the claim holds, decisively on the score vectors and marginally on the book: the two steps differ in what they do to the steadiness of the coefficient by a ratio of 1.172 ($p < 0.001$), and in what they do to the volatility of the book by 1.089 ($p = 0.084$). On the means it does not hold. The gap between their return effects is +0.27 percentage points against a standard error of 2.03, and the gap between their effects on the mean coefficient is +0.00041 ($p = 0.869$); at a standard error seven times the estimate this sample is not close to resolving either. What the scale step acts on is established by its own rows, and Section 5.3.3 and Section 8.2 rest on that. Whether the shape step acts on the means instead of on the same channel is a question this evidence leaves open.

One fact seen twice, not two facts. The two levels of the chain are linked, and tightly enough that the agreement between them is worth less than it looks. Regressing the monthly book return on the monthly coefficient cell by cell gives a slope invariant across the arms to within 0.006 (Section C.2). Applying the square of the common slope to the measured fall in the variance of the coefficient accounts for 51.1 of the 56.4 squared percentage points by which the annualised variance of the book falls, or 91%, with a bootstrap 95% interval of [59%, 160%]. The improvement in the information ratio of the coefficient and the improvement in the Sharpe ratio are therefore substantially the same observation recorded at two points in the chain, and the chapter treats them as one piece of evidence rather than two.

Table 5.6: Each step of (5.1) against the ordering statistic and the Sharpe component it moves. Steps run percentile (pct) to scale-matched (s.m.) to inverse-normal (int). Every cell is a paired test on the monthly series, pooled as the mean of the five per-objective statistics. The rank IC columns are measured on the score vectors and the return and volatility columns on the resulting book at the headline decile under inverse-volatility sizing, gross, so the two halves are one intervention seen at two points in the chain. Mean columns carry Newey–West HAC standard errors; dispersion columns give the ratio of standard deviations and the p -value of a circular block bootstrap on the pooled log variance ratio, which has a block length rather than a bandwidth, and in parentheses how many of the five objectives reject individually under Pitman–Morgan. The last row tests the difference between the two steps directly: a dissociation is a claim about a difference, and reading one off which of the rows above carry stars would compare significances instead of testing one. It is not the difference of the two standard errors either – the steps share the scale-matched arm, so their errors are correlated and the gap is estimated jointly. ***, **, *: $p < 0.01, 0.05, 0.10$.

Step	Ordering (score vectors)		Economics (traded book)	
	mean rank IC	s.d. of rank IC	ann. return	ann. volatility
Scale (pct → s.m.)	+0.00077	0.846*** (5/5)	+1.68%	0.907*** (4/5)
Shape (s.m. → int)	+0.00119*	0.992 (0/5)	+1.95%**	0.988 (0/5)
Total (pct → int)	+0.00196	0.840*** (5/5)	+3.63%***	0.896*** (4/5)
Shape – scale	+0.00041	1.172***	+0.27%	1.089*

What survives as a genuine non-equivalence. One comparison in the study resists every reconciliation above, and it does not involve the treatment. The gradient-boosted control has the highest mean rank information coefficient of any specification reported, 0.0439, and, at the traded decile under inverse-volatility sizing, a gross Sharpe ratio of +1.335 that falls below every inverse-normal neural cell, which run from +1.398 to +1.586. The reason is visible in the same statistic this chapter has been tracking: the dispersion of its coefficient is 0.1034, which is a percentile-arm figure, so its information ratio is 0.424 and sits inside the inverse-normal range of 0.421 to 0.466 despite carrying the best mean coefficient in the study. A learner invariant to the encoding never receives the stability the encoding confers, and its ordering is accurate on average and unsteady month to month. That is the cleanest instance of (2.1) the study produces: the criterion the applied literature selects on ranks this specification first, and the portfolio ranks it below seven others. The same observation carries a second reading, taken up in Section 5.6.3: the specification the applied criterion ranks first is also the study’s strongest comparator, and what places it below the inverse-normal arm is precisely the stability the encoding supplies and invariance denies.

Figure 5.4 makes the same point without any inference. Realised volatility tracks the Sharpe column far more closely than the coefficient does: the inverse-normal arm runs at lower annualised volatility than the percentile arm for all five objectives, and it is the arm with the higher Sharpe ratio in every case. Neither statistic orders the cells on its own: at the traded decile the gradient-boosted specification runs a higher realised volatility

than any neural cell and still out-earns the entire percentile arm (Table C.1). Risk is the better predictor of the two, but it is a predictor within the comparison and not a general one, and neither ordering holds at every fraction of the grid.

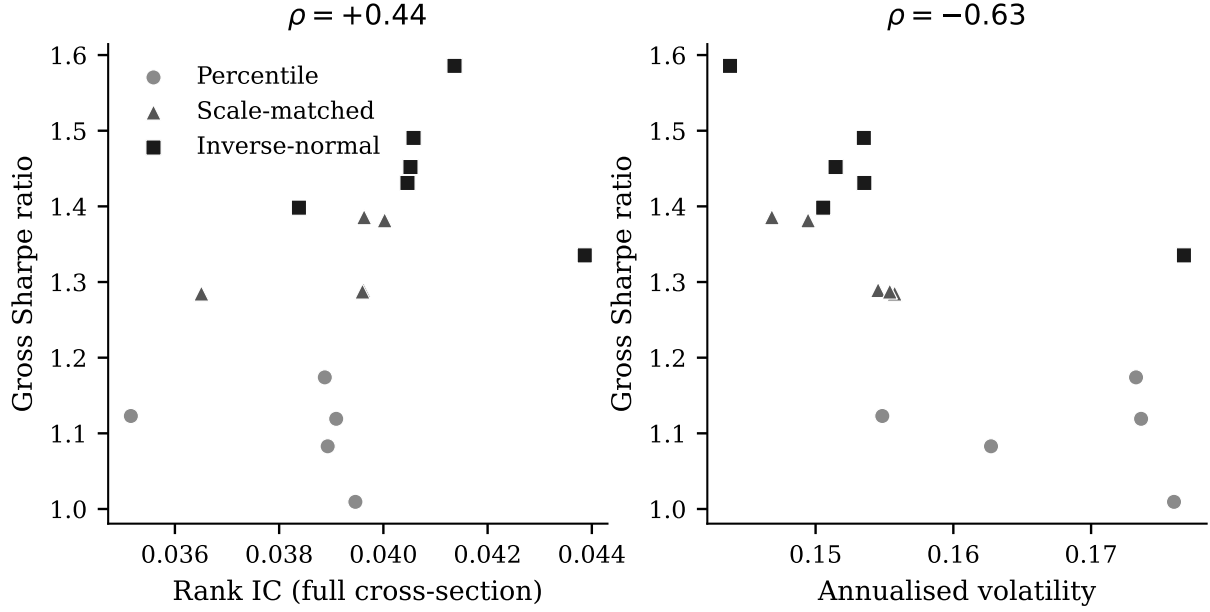


Figure 5.4: Each point is one sweep cell at the headline width and weighting, gross. Left: gross Sharpe against full-cross-section rank IC. Right: the same Sharpe ratios against realised annualised volatility. The seventeen cells are the fifteen neural ones and the two arms of the gradient-boosted control. The control carries the marker of its own arm rather than one of its own, and its two arms coincide to the last digit on all three coordinates, so it appears as the single rightmost point of each panel with the percentile marker hidden beneath the inverse-normal one. ρ is the correlation across the seventeen cells, and at that count neither ρ is sharply estimated. The contrast between the panels is (2.1): the portfolio outcome may track the risk the book ran more closely than the accuracy of the ordering behind it.

Figure 5.5 shows the mechanism directly, tracking the dispersion of the coefficient and the volatility of the book through the sample rather than summarising each in one number. The percentile arm runs above the midpoint of the other two in 99% of months on the left and 94% on the right, and the scale-matched and inverse-normal lines are close to indistinguishable in both, which is the dissociation of Table 5.6 at monthly frequency. The two gaps are not one series drawn twice. They correlate at 0.46 across the sample, and the episodes in which each is widest do not coincide: the dispersion gap is at its widest across 2016 and 2017, the volatility gap in the first months of 2021. Both consequences follow from the same step without arriving on the same schedule.

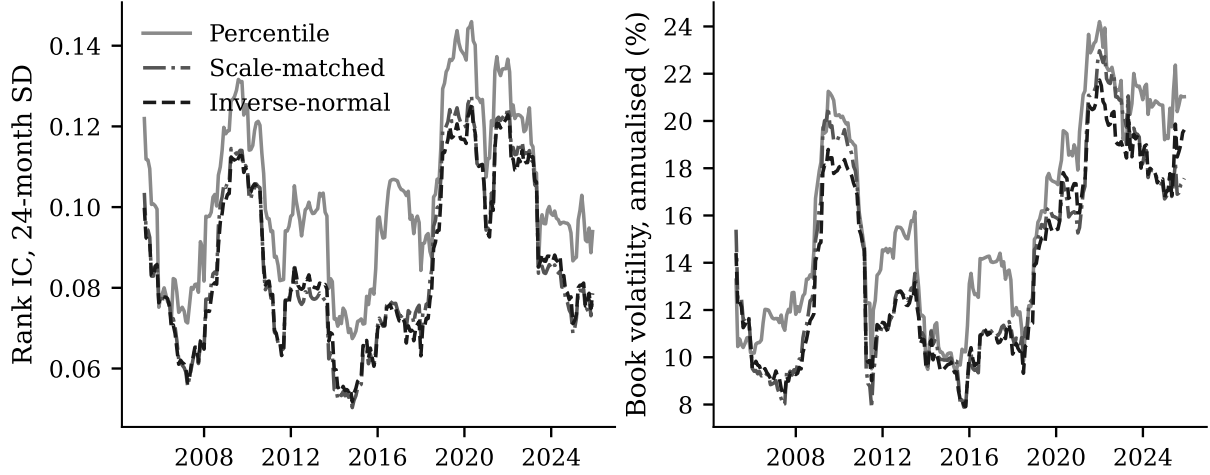


Figure 5.5: Left: trailing twenty-four-month standard deviation of the monthly rank information coefficient by representation arm, averaged across the five neural objectives. Right: trailing twenty-four-month annualised volatility of the corresponding long-short book. Each panel averages a per-objective rolling statistic across the five objectives, which is the pooling operator used throughout. The two panels track the same mechanism at two points in the chain, and the gap between the percentile arm and the other two is present in almost every month of both without the two gaps widening on the same schedule.

5.5 The gradient-boosted invariance control

The two gradient-boosted cells are read against Hypothesis 3: Section 4.5 predicts a contrast close to zero, with any residual attributable to histogram binning and floating-point representation, and Table 5.7 fixes how each outcome is read. The control is estimated in the two base arms, the scale-matched arm having no gradient-boosted counterpart, and it tests whether the neural result is a representation effect or an implementation artefact.

Table 5.7: Pre-specified interpretation of the gradient-boosted control. The mapping from outcome to inference is fixed before the estimates are inspected.

Observed pattern	Inference
Neural effect material, tree contrast near zero	Consistent with the optimisation-geometry mechanism of Section 2.3. The encoding changes how a smooth estimator fits, not what it was told.
Neural effect material, tree contrast also material	Points to an implementation artefact upstream of the estimator, since no information distinguishes the two arms for a threshold learner. The neural result would not be interpretable as a representation effect.

The control returns the first row of Table 5.7, and it returns it exactly. Its contrast is 0.000 in both cost regimes rather than merely small: the gradient-boosted cell records the same gross and net Sharpe ratio under both arms, and its mean rank information coefficient, its dispersion, its maximum drawdown and its calendar-2025 return are identical in both. The agreement reaches below the reported figures: Section C.6 compares the

two fitted arms field by field across all 214 trees of all five seeds in the retained fit, where everything that determines a prediction is identical and only the split thresholds differ, in encoding units. It is therefore a prediction that was tested rather than an identity that had to hold: Section 4.5 names two channels through which it could have failed, and both measure at zero here. On that evidence the neural estimates of Section 5.3 are not an implementation artefact upstream of the estimator, and the arm assignment reaches the models through the encoding alone: the tree shares the panel, the labels, the folds, the seeds and the portfolio construction, is blind to the encoding, and registers nothing.

The same cell has a second use, which Section 4.5 fixed in advance and which Section 5.6.3 takes up: being blind to the encoding is what makes it a control here, and it is also the strongest specification in the study against which the neural cells can be set. The two readings are kept apart because only the first belongs to a hypothesis.

5.6 Calibration against external reference points

Everything reported so far is an internal comparison, and a Sharpe ratio read on its own carries no scale. Table 5.8 sets the fitted models against three reference points of increasing strength: the twelve-month momentum rule of (4.9), the two linear anchors of Section 4.7, and the gradient-boosted ranker. All three are replayed through the identical universe, sector demeaning, selection rule, weighting and cost stack, so the only thing that differs is what produced the scores.

5.6.1 A naive rule

The naive rule is weak on this sample: the 12–1 momentum sort earns a gross Sharpe ratio of +0.130 and a net one of -0.085 , on a rank information coefficient, realised volatility and maximum drawdown far worse than any fitted specification (Table 5.8). Those last two carry the common gross exposure of Table 4.4, as every raw figure in the table does, so the -87.8% is a levered drawdown on a book running more than half again the volatility of the most volatile fitted one, and is read against the other raw drawdowns rather than as an unlevered loss. The book-level check is decisive for the fitted models and informative about the linear ones. The final column of Table 5.8 correlates each book’s gross monthly return series with the benchmark’s. For the neural composites and the gradient-boosted control it is slightly negative. Those models are not an expensive reconstruction of the momentum sort; whatever they are trading is close to orthogonal to it. The linear anchors do not behave that way: both correlate positively with the benchmark, the elastic net markedly so, so a penalised linear model on this surface holds a book that moves materially with the momentum sort while the fitted neural models do not. The orthogonality is a property of the estimator rather than of the characteristic

surface it reads.

Table 5.8: Three reference points against the fitted models, at the headline width and weighting over April 2003 to November 2025: a 12–1 momentum sort, the two linear anchors of Section 4.7, and the gradient-boosted control. All run through the identical universe, sector demeaning, selection rule, weighting and cost stack; the momentum sort alone is a fixed rule with no training, folds or seeds. Every row but the two neural ones is a single estimated specification. The neural rows average the five per-objective figures, the information ratio included, with the sole exception of the final column, which correlates the objective-averaged series; they are not otherwise the composite book of Table 8.2. Columns marked g are gross, and raw rather than rescaled, since the difference in realised risk is part of the comparison. Only the two Sharpe columns are stated in both regimes, so the drawdown beside a net Sharpe ratio is not that book’s — the neural inverse-normal book draws down -41.2% net against the -38.8% shown, and neither had recovered at the last observation (Table 8.4). The final column correlates each row’s gross monthly returns with the benchmark’s.

Specification	SR gross	SR net	Return ^{g}	Vol. ^{g}	Max DD ^{g}	Turnover	Rank IC	ICIR	Corr.
12–1 momentum	0.130	-0.085	4.1%	31.9%	-87.8%	1.54	+0.0026	+0.016	—
OLS, percentile	0.825	+0.289	12.2%	14.8%	-47.1%	2.74	+0.0260	+0.298	+0.033
OLS, inverse-normal	0.882	+0.360	13.3%	15.1%	-42.3%	2.73	+0.0252	+0.290	+0.123
Elastic net, percentile	0.466	+0.069	8.6%	18.5%	-47.4%	2.56	+0.0224	+0.220	+0.307
Elastic net, inverse-normal	0.280	-0.092	5.6%	20.1%	-59.5%	2.57	+0.0233	+0.224	+0.353
Neural, percentile	1.102	+0.625	18.5%	16.8%	-43.5%	2.65	+0.0383	+0.361	-0.066
Neural, inverse-normal	1.471	+0.932	22.1%	15.1%	-38.8%	2.74	+0.0403	+0.452	-0.090
Gradient-boosted control	1.335	+0.881	23.6%	17.7%	-40.9%	2.76	+0.0439	+0.424	-0.163

The correct claim from a weak benchmark is the narrow one. The sample opens in 2003, so it contains the 2009 momentum crash and the decade of regimes hostile to momentum that followed, and the sector demeaning imposed for comparability strips the industry component that carries much of the raw premium. The comparison establishes that the reported performance is not a momentum premium in disguise rather than that the models beat momentum by the printed margin. One piece of evidence speaks to the book-level question in the specified direction: in the six-factor attribution of the RankNet inverse-normal book, run on the volatility-rescaled series of Section 5.1, the momentum loading is -0.033 with $t = -0.30$, indistinguishable from no exposure at all, against an annualised alpha of $+18.17\%$ ($t = 5.41$) and a dominant profitability loading of $+0.895$ ($t = 7.21$). That exposure is large and it is not what the book earns: RMW returned $+2.41\%$ a year over the evaluation window, so the loading accounts for some $+2.2$ percentage points of annual return against the alpha above. The book is exposed to profitability without being a repackaging of it.

The benchmark’s turnover runs well below that of the fitted books, as a twelve-month sort must, and Section 8.1 takes up the comparison at the level of subperiods.

5.6.2 The linear anchors

The linear anchors are the harder reference point, and they are where Section 4.7 said the level would be read. Fitted on the identical surface under the folds, target encoding

and portfolio rules of the fixed surface, ordinary least squares is the stronger anchor, and against its two arms the pooled inverse-normal neural Sharpe is worth between roughly one and a half and two times the gross ratio and between two and a half and three and a quarter times the net one (Table 5.8). Two qualifications travel with that comparison. Each anchor is a single deterministic fit while every neural figure is a five-seed ensemble, and Section 8.5 shows that the ensemble is not neutral here, so part of the gap is an ensembling effect this design does not separate from the model class. And neither anchor is tuned beyond the penalty the elastic net selects on validation rank information coefficient, so what is compared is a linear model at conventional settings against a neural one at conventional settings, which is the fixed-surface rule of Section 4.4 applied across families rather than a contest between two tuned alternatives.

5.6.3 The gradient-boosted ranker as a comparator

The linear anchors set a floor and the momentum rule sets a lower one. The gradient-boosted ranker is the only specification in the study that competes. On the identical surface it earns a gross Sharpe ratio of +1.335 and a net ratio of +0.881, against a pooled +1.471 and +0.932 for the inverse-normal neural cells and a pooled +1.102 and +0.625 for the percentile ones. Gross, every one of the five inverse-normal cells is ahead of it, from +1.398 to +1.586, and every one of the five percentile cells is behind it, from +1.009 to +1.174; the scale-matched arm straddles it, three cells below and two above. Net the tree clears every percentile and every scale-matched cell, the latter a change from gross, and four of the five inverse-normal cells clear the tree, the fifth falling just short at +0.865 (Tables C.1 and C.2). Neither turnover nor return is what separates the two families: the tree runs 2.76 against 2.74 for the inverse-normal arm and earns more than any neural cell in either arm, and what places it below the inverse-normal one is realised volatility. Both orderings are struck under inverse-volatility sizing, and it is the inverse-normal half of each that does not survive the change of weighting scheme. Under equal weighting the control still clears every percentile cell gross and net, while the same three inverse-normal cells fall below it in each regime. The pooled margin of the arm over the control accordingly narrows from +0.136 to +0.054 gross and turns from +0.052 to -0.014 net.

The comparison is not tilted towards the network. As Section 4.5 records, the tree's capacity is selected fold by fold on the validation rank information coefficient under early stopping, while every neural cell trains for a fixed ten epochs and deploys the last of them. The family that wins gross is the family that was denied the selection.

Realised risk is the statistic this chapter has been tracking throughout, and Section 5.4 states it: a learner invariant to the encoding cannot receive what the encoding confers. The neural family's advantage over the tree is accordingly not a property of the architecture standing on its own: it is absent in the percentile arm, where the tree is ahead of

every cell at the traded decile, and it appears only once the encoding is changed.

The reading is bounded on four sides, and none of them is small. The gradient-boosted cell belongs to no inference family, by the pre-specification of Section 3.7, so no test is attached here and none is computed; admitting the contrast to a family after the estimates were read would move the thresholds carrying the other members. One tree configuration is not the tree family, and no search was run over it. The net margin, +0.052 pooled, is not a quantity this sample can resolve. The ordering itself is specific to the weighting scheme it is struck under, reversing in sign net under equal weighting. What the comparison licenses, once execution is priced, is therefore a claim about the gross ordering at one operating point and about the mechanism behind it, and not a claim about which family an investor should run.

5.7 Mechanism in the fitted function

Table 5.9 reports the diagnostics that distinguish a genuine change in what the model selects from a rescaling of what it already selected. Seed dispersion and score correlation are reported across all three arms, so that each can be attributed to one step of (5.1) rather than to the total alone; overlap and return correlation are struck on the total contrast.

Table 5.9: Agreement between the arms, by objective, across the three encodings of Table 4.1: percentile (pct), scale-matched (s.m.) and inverse-normal (int). Seed dispersion is the cross-sectional mean of the per-name standard deviation of rank percentile across the five seeds. Score correlation is the mean within-date Spearman correlation between two arms’ ensemble scores, reported for each step of (5.1) and for the total. Overlap is the share of names common to the two arms’ long deciles and return correlation is between the monthly series, both for the total contrast at the headline configuration, gross.

Objective	Seed dispersion			Score correlation			Total contrast	
	pct	s.m.	int	pct-s.m.	s.m.-int	pct-int	Overlap	Ret. corr.
RankNet	0.1254	0.1450	0.1449	0.887	0.952	0.864	0.641	0.847
LambdaRankIC	0.1270	0.1471	0.1463	0.887	0.948	0.862	0.643	0.839
Margin ranking	0.1235	0.1384	0.1396	0.892	0.932	0.873	0.646	0.848
Soft-Spearman	0.1243	0.1397	0.1398	0.857	0.931	0.832	0.605	0.857
RankCosine	0.1237	0.1396	0.1401	0.860	0.933	0.834	0.604	0.841

Book agreement between arms. Table 5.9 reports the within-date score correlation between arms, the share of names their long deciles hold in common, and the correlation of their monthly return series. Identical ordinal information does not produce an identical book. Read across the three arms, the score correlations place the scale-matched book much nearer the inverse-normal one than the percentile one: the correlation runs between 0.931 and 0.952 across the second step against between 0.857 and 0.892 across the first,

in every objective. Most of the distance the encoding travels in score space is therefore travelled by the change of scale, which is where Section 5.3 also locates the larger half of the Sharpe contrast. The evidence that the change is a genuine one rather than a rescaling is indirect, and it is quantified. Two specifications that receive identical ordinal information produce books of materially different realised risk, from 17.60% to 15.35% annualised volatility for RankNet and from 16.27% to 14.38% for LambdaRankIC, and materially different drawdowns; a rescaling of a common ordering could not do that. The gradient-boosted control shows what a genuinely common ordering yields, namely the same Sharpe ratio, the same drawdown and the same ordering statistics in both arms, and the linear anchors show that the feature surface alone does not produce the effect either, the contrast being $+0.056$ for ordinary least squares and -0.186 for the elastic net. Neither reproduces it: the first is nothing and the second, though larger in magnitude than the shape step under study, runs the other way, and nothing here explains why. Both are single deterministic fits carrying no test, and what they establish is a sign and not a magnitude. The representation effect lives in the neural fitted function, not in the data it is fitted to.

Seed dispersion and ensembling. Cross-seed dispersion is reported by arm and objective in Table 5.9, and it runs higher in the inverse-normal arm than in the percentile arm for every one of the five objectives. The third arm locates the increase precisely: the scale-matched dispersion is indistinguishable from the inverse-normal one in all five objectives, differing by at most 0.0012, so the whole of the rise arrives with the change of scale and the change of shape adds nothing to it. The comparison carries no test, so it is consistent with a rougher optimisation landscape at the larger input scale without establishing one. The seed dimension is characterised in Section 8.5 from the stored per-seed score vectors; the result that matters here is that the contrast at a single seed is roughly half its five-seed value, so what Section 5.3 reports is the effect of the representation on a five-member rank-averaged ensemble, not on a single network.

5.8 Pooling across objectives

Everything reported above is pooled over the five ranking objectives, and everything in the chapters that follow is too. A pooled figure carries a claim: that the objectives respond to the treatment closely enough for one number to stand for five. The claim is substantive, since the design spans five objectives, three arms, five traded widths, two weighting schemes and three cost regimes, and it is what allows a dozen numbers to speak for several hundred.

The machinery of Section 3.7 is turned on a different axis: within each arm, all ten pairwise differences between objectives are estimated on the paired monthly series

under the same bootstrap and stepdown. A second family asks whether the effect of the representation is itself the same across objectives, which for a Sharpe ratio is a difference of two Sharpe differences and is bootstrapped directly. Table 5.10 reports both families.

Table 5.10: Homogeneity of the five ranking objectives. Arms are the percentile (pct), scale-matched (s.m.) and inverse-normal (int) encodings of Table 4.1. Each entry counts, out of the ten pairwise objective comparisons, how many are significant at 5% after a family correction taken off one joint circular block bootstrap. The correction is a Romano–Wolf stepdown for the two Sharpe rows and the information ratio at the level stage; every other entry carries a single-step maximum-statistic correction, which is uniformly the more conservative of the two. That includes both precision rows, which hold the only non-zero counts in the two blocks above the equivalence tests. *Levels* compares the objectives inside an arm; *contrasts* compares the effect of the representation across them — a difference of two Sharpe differences on the Sharpe rows, of two mean differences on the rank IC and precision rows. A low count is what pooling requires: objectives may sit at different levels, as they do for precision, and still agree about the treatment. $|t|_{\max}$ is the largest $|t|$ in that family. The lower blocks add a two-one-sided-test count of pairs demonstrably closer than the pooled contrast, and the best-to-worst spread within an arm, to be read against a total of +0.370. The equivalence margin is the pooled contrast itself, estimated on the same months and carrying a standard error of 0.083 gross and 0.078 net that the test treats as zero; those counts are also the only ones in the table without a family correction, so both read as upper bounds.

Statistic	Levels, within arm				Contrasts, across objectives			
	pct	s.m.	int	$ t _{\max}$	scale	shape	total	$ t _{\max}$
Sharpe, gross	0/10	0/10	0/10	1.84	0/10	0/10	0/10	1.64
Sharpe, net	0/10	0/10	0/10	1.78	0/10	0/10	0/10	1.61
Mean rank IC	0/10	0/10	0/10	2.08	0/10	0/10	0/10	1.59
IC information ratio	0/10	0/10	0/10	2.25	—	—	—	—
Precision, top	6/10	3/10	5/10	8.33	0/10	0/10	0/10	2.34
Precision, bottom	4/10	5/10	5/10	11.17	0/10	0/10	0/10	2.37
<i>Pairs equivalent at the 5% level</i>								
gross, margin 0.370	10/10	10/10	7/10					
net, margin 0.307	8/10	9/10	5/10					
<i>Within-arm spread, gross Sharpe</i>								
best – worst objective	0.165	0.101	0.187					

On the Sharpe ratio the objectives are indistinguishable. Not one of the sixty pairwise level comparisons attains significance in any arm or either cost regime, and not one of the differences of differences does at any of the three steps, the largest $|t|$ anywhere being 1.84 (Table 5.10). The mean rank information coefficient and its information ratio repeat the pattern exactly.

That second family is also the evidence for Hypothesis 2. The hypothesis asks whether the effect is a property of the representation rather than of one loss surface, and the answer it wants is not that the five objectives perform alike but that the representation moves them alike. On the Sharpe ratio it does, in both cost regimes and at each step of (5.1), which is what Section 5.3 shows from the other direction when every objective places the inverse-normal arm ahead. The sign is common to all five and the size is not

distinguishable between them.

Precision does not. In levels the objectives genuinely differ in how accurately they populate a decile, with between three and six of the ten pairs separated inside each arm and t -statistics reaching 11.2, RankNet against LambdaRankIC the dominant pair throughout (Table 5.10). Yet the contrasts pool cleanly: none of the ten differences of differences is significant at any step for either tail. The objectives disagree about the level of membership accuracy they achieve and agree about how much the representation changes it. Pooling is therefore licensed for contrasts throughout and for levels everywhere except precision, which Table 5.1 reports cell by cell for that reason.

Two further checks bound what the non-rejections are worth. The first is a matter of magnitude: the spread between the best and worst objective within an arm is 0.165, 0.101 and 0.187 in gross Sharpe, against a total representation contrast of +0.370. The axis being pooled over is smaller than the axis being tested, but only against the total. It is not smaller than the shape step at +0.146 in the two base arms, and that is the one place where objective heterogeneity and the effect are of the same order.

The second check is that failing to reject equality is not the same as demonstrating it. A two-one-sided-test against a margin equal to the pooled contrast asks whether each pair is demonstrably closer together than the effect the study is trying to measure. Gross, all ten pairs are equivalent at the 5% level in the percentile arm and all ten in the scale-matched arm, but only seven of ten in the inverse-normal arm. Net, where the margin is smaller, the counts fall to eight, nine and five.

Two things keep those counts from being read as a positive result. The margin is not a constant: it is the primary endpoint, estimated on the same 272 months as the pairs it bounds and carrying a standard error of 0.083 gross and 0.078 net, roughly a quarter of its own size, which the test treats as zero. And the ten per-pair tests carry no family correction, unlike the level and contrast rows above them in the same table. The nominal 5% is therefore not the attained level, and since margin and pairs are estimated from the same months on overlapping books the error does not even have a signed direction. What survives is the conservative reading, which is the one this chapter takes: equivalence is harder to demonstrate in the stronger arm, and pooling in the inverse-normal arm rests on a non-rejection rather than on a positive result.

A pooled figure in this thesis is always the mean of the five per-objective statistics, and the definition carries weight. Averaging the per-objective Sharpe ratios gives 1.102, 1.326 and 1.471 across the three arms and a total contrast of +0.370; the Sharpe ratio of a single book formed from all five score vectors gives 1.169, 1.430 and 1.592 and a total of +0.423. The extra 14% is the diversification of five imperfectly correlated books, which is not something the representation produced.

The full per-objective cross is carried in Appendix C at every width.

5.9 What is carried forward

Chapter 6 and Chapter 7 vary construction and cost at a fixed set of model specifications. All five objectives go forward, pooled as Section 5.8 licenses, and both base arms go forward with them. The construction and cost analyses therefore run on both sides of the contrast: a representation advantage that exists gross is worth reporting only if it survives construction and execution. Whether it does is a live question, because the inverse-normal arm is also the more expensive one to trade, and Section 7.3 prices the answer.

The scale-matched arm stops here as a treatment. Its work in the design is finished once the decomposition of (5.1) has been estimated, and the construction and cost chapters vary their axes against the two arms a modeller would choose between rather than against three. That is a decision about what the following chapters test and not a judgement on the arm: as Section 5.3.1 reads it, standardised percentile ranks are a specification anyone might deploy, and their record at every width and under both weighting schemes is in Table C.1 and Table C.2 rather than in the chapters below. The one thing not available for it is the component-wise cost bridge of Section 7.3, which the run computes for the two base arms alone.

Chapter 6

Portfolio Construction and Tail Width

Chapter 5 measured a modelling choice at a fixed portfolio configuration. This chapter varies the configuration, taking the specifications carried forward from Section 5.9 and asking whether the representation contrast survives the change.

Costs are modelled and stress-tested in Chapter 7. Every net figure quoted in this chapter is struck at the base case of Table 4.5, among them the bridge of Table 7.1 that the width rule of Section 6.5 needs: a rule stated in gross terms would select a book that execution then reorders.

6.1 Width and weighting as treatments

Narrowing the traded fraction is usually described as concentrating the signal, and if the model is informative the average quality of the selected names does rise as α falls. The remaining effects run the other way: concentration raises idiosyncratic risk and turnover, since membership of a narrow tail is less stable across rebalances than membership of a wide one, reduces the number of positions across which estimation error can diversify, and raises execution cost more than proportionately, since the same book must be traded through fewer names. The width that maximises gross risk-adjusted performance is therefore not necessarily the width that maximises net performance, which is the content of Hypothesis 5 and the reason widths and costs are reported jointly and not sequentially.

The weighting scheme is a treatment for the reason set out in Section 2.4. Both schemes are crossed with all five widths at the common gross exposure of Table 4.4.

6.2 Results across widths

Table 6.1 reports the width grid of (4.16) under both weighting schemes; the realised names per side, the $k_{\min}$ floor and the turnover definition are in Section C.5. One row needs a word before the rest can be read.

The 1% row is a concentration stress test rather than a candidate operating point, since a book of ten positions per side is dominated by name-level idiosyncratic risk and by frictions the cost model of Section 4.10 calibrates on averages. No conclusion in this thesis rests on it.

Table 6.1: Gross performance across the traded-fraction grid under both weighting schemes, averaged across the five neural objectives in the inverse-normal arm. **Two bases share this table.** Return and maximum drawdown are stated on the book rescaled to 15% annualised volatility (Section 5.1), so that widths are comparable at a common risk target; volatility is the *realised* figure, which is the quantity that differs across widths and would be 15% in every row if it were normalised too. Return divided by volatility is therefore not the Sharpe ratio in the column beside it: the raw return is Sharpe $\times$ realised volatility. Sharpe and turnover are invariant to the normalisation. Turnover is one-side net per rebalance. Names is the realised count per side, not an effective count: under inverse-volatility sizing the concentration-adjusted figure is strictly smaller.

Weighting	α	Names	Sharpe	at 15% vol	realised	at 15% vol	Turnover
				Return	Vol.	Max DD	
Equal	20%	200	1.388	20.8%	13.2%	-35.2%	2.30
	10%	100	1.361	20.4%	17.5%	-39.5%	2.65
	5%	50	1.251	18.8%	22.7%	-35.2%	2.91
	2.5%	25	1.153	17.3%	30.1%	-33.0%	3.10
	1%	10	0.903	13.5%	42.3%	-29.2%	3.33
Inverse-vol.	20%	200	1.448	21.7%	11.8%	-32.0%	2.39
	10%	100	1.471	22.1%	15.1%	-38.6%	2.74
	5%	50	1.312	19.7%	19.6%	-35.3%	2.98
	2.5%	25	1.199	18.0%	26.3%	-32.0%	3.17
	1%	10	0.939	14.1%	37.9%	-28.6%	3.38

Gross, concentration does not pay. Performance improves as the book widens over most of the grid, and the cause is a variance effect rather than a return effect. Raw return and raw volatility both fall as the tails are widened, volatility faster, so thinning them raises the risk of the book more quickly than the improving average quality of the selected names raises what it earns. Under equal weighting the raw pair runs from 38.2% and 42.3% at the one-percent fraction to 18.3% and 13.2% at the quintile. The return column of Table 6.1 is stated at a common volatility and therefore moves in the opposite direction.

The two schemes part company at the wide end. Under equal weighting gross performance is still rising at the quintile; under inverse-volatility sizing it turns at the decile, 1.471 against 1.448, and nothing separates the schemes statistically (Section 6.3).

Net settles what gross leaves open. Net Sharpe in Table 7.1 is highest at the decile under both schemes and falls away on both sides of it, giving up 0.09 under inverse-volatility sizing and 0.04 under equal weighting as the book widens to the quintile. The headline fraction is therefore the interior peak of the net profile once costs are charged,

bracketed on both sides rather than sitting at the edge of the grid, which is a stronger statement than the same fraction winning a grid that stopped there. It is a statement about where the point estimates peak: the two contrasts that bracket the decile carry t -ratios of 0.80 and 0.89 under the paired test of Section 3.7, so the location of the maximum is not itself an established result.

The percentile arm fixes the scope of that claim. Its gross profile rises all the way to the quintile under both schemes, but its net profile is non-monotone and peaks at the 2.5% fraction, so there costs reorder the grid outright: the quintile leads that arm on gross and the 2.5% fraction leads it on net. Whether frictions preserve the gross ordering is a property of the arm and the cost stack rather than a general feature of the grid, and nothing gross would have revealed it.

Turnover falls monotonically as the fraction widens under both schemes, so the widest book is also the cheapest to keep in place. The mechanism is membership stability: a thin tail is a narrow slice of the score distribution, so a given month-to-month movement in the scores displaces a larger share of its members than the same movement displaces from a decile tail. Turnover does not settle the width question on its own, because only two of the four charged components follow it. Table 7.1 decomposes the stack at every width, and Section 7.3 reads it at the headline one.

Maximum drawdown does not follow the Sharpe ratio down the grid. The column carries the volatility normalisation of Section 5.1, so the widths are comparable on it, and the deepest decline of the whole grid sits at the traded decile, -39.5% under equal weighting and -38.6% under inverse-volatility sizing, while the narrowest fractions are the shallowest, -29.2% and -28.6% at one percent. At a common risk target the concentrated book therefore suffered the milder worst case, which runs against the direction of every other column in the table. A maximum drawdown is one order statistic of a single realised path and carries no test here, so it qualifies the width recommendation rather than competing with it; Section 8.2 takes the episode itself apart.

Figure 6.1 puts the three profiles on one axis: turnover falls as the book widens in both arms, gross Sharpe turns only at the quintile, and neither net profile is monotone once costs are charged.

What the grid does not do is separate one width from another. Those comparisons sit in the exploratory tier, they are made on gross Sharpe against the decile, and one of the sixty clears Benjamini–Hochberg at $q = 0.10$, and it does so on the composition of the family rather than on its own evidence, since the width family taken alone yields nothing from sixty comparisons (Section C.2). What is settled is the gross-to-net step itself: fifteen of the fifteen netting contrasts are significant, and Section 7.3 prices them. At the headline width the inverse-normal arm nets between $+0.865$ and $+1.022$ across the five objectives, against $+0.551$ to $+0.719$ in the percentile arm. The headline specification is also the net-optimal one, but it was not chosen for that reason (Section 6.5).

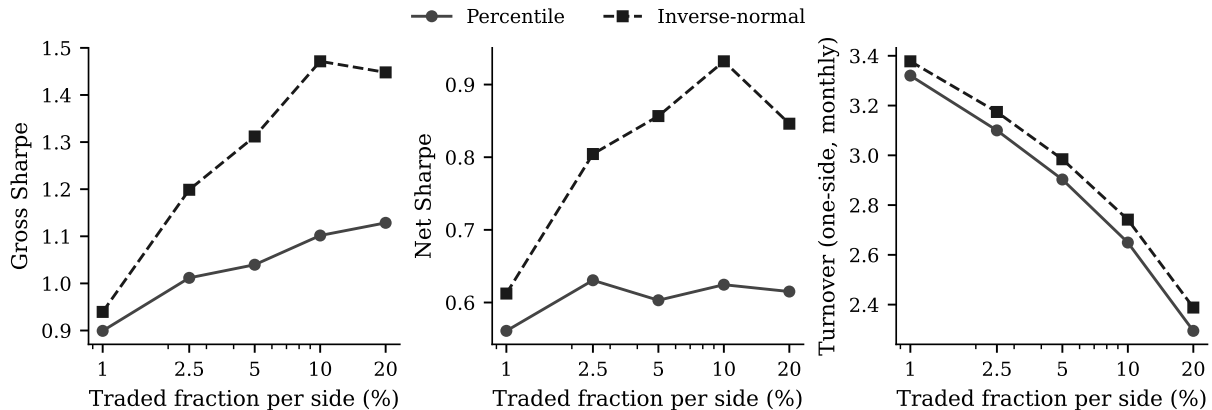


Figure 6.1: Gross Sharpe ratio, net Sharpe ratio and one-side net turnover against the traded fraction per side, by representation arm, averaged across the five neural objectives under inverse-volatility sizing. The horizontal axis is logarithmic. Turnover falls as the book widens in both arms. Gross Sharpe rises, turning only at the quintile and only in the inverse-normal arm. Net Sharpe peaks at the decile in the inverse-normal arm and at the 2.5% fraction in the percentile arm, so neither profile is monotone once costs are charged.

The representation contrast is not an artefact of the width at which it is measured, though it is not uniform across the grid either. Table 6.2 reports it at all five fractions under both schemes and it is positive in all ten cells, largest at the decile and falling in both directions, with the bootstrap intervals ceasing to exclude zero at the two narrowest; the per-width net contrasts of Table 7.2 follow the same profile. Point estimates and resolution deteriorate together at the concentrated end, so the grid cannot separate a genuine attenuation of the effect in a twenty-five-name book from the wider standard errors such a book carries, and no claim is made for those widths in either direction. The table pools the five objectives, so the number holding the sign is stated only at the headline width: there all five are positive gross and net, though only RankNet and LambdaRankIC clear the Romano–Wolf stepdown in both regimes. Six of the fifty objective-by-width-by-scheme cells turn slightly negative at the two narrowest fractions, gross and net, the largest -0.089 (Tables C.1 and C.2).

6.3 Selection against sizing

Hypothesis 4 is a statement about mechanism, and the width grid is what makes it testable. An equally weighted book is a function of membership alone, so the branch the hypothesis distinguishes turns on whether the effect of Chapter 5 survives equal weighting, which would make it selection, or appears only under inverse-volatility sizing.

Table 6.2 reports the representation contrast separately under each scheme at every width. The quantity of interest is whether the contrast between representation arms is of the same sign and comparable size under both schemes, not which scheme performs better in levels. That second question is in any case not settled here: inverse-volatility sizing

leads gross at every width, equal weighting leads net at every width in the percentile arm and at three of the five in the inverse-normal arm, and none of the fifteen weighting comparisons survives Benjamini–Hochberg.

Table 6.2: Paired Sharpe difference between the arms under each weighting scheme at each traded fraction, gross, averaged across the five neural objectives. Formed per objective and then averaged. Point estimate and interval both come from the pooled estimator of Section 3.7 – one circular-block bootstrap across the five objectives, studentised pooled pivot – so the inverse-volatility row at 10% reproduces the primary endpoint of Table 5.4 by construction rather than by a second method agreeing with the first. A contrast surviving equal weighting is evidence about selection; one confined to inverse-volatility sizing is evidence about the sizing rule.

α	$\Delta\text{SR, equal}$	95% CI	$\Delta\text{SR, inv. vol.}$	95% CI
20%	+0.282	[+0.146, +0.419]	+0.319	[+0.170, +0.468]
10%	+0.316	[+0.156, +0.476]	+0.370	[+0.216, +0.523]
5%	+0.251	[+0.054, +0.448]	+0.272	[+0.075, +0.470]
2.5%	+0.153	[−0.080, +0.385]	+0.187	[−0.033, +0.407]
1%	+0.008	[−0.238, +0.255]	+0.040	[−0.214, +0.294]

The evidence selects the selection branch of Hypothesis 4. The decile column under equal weighting is the primary endpoint re-struck under the second construction rule, in the sense Section 3.7 fixes, and not a member of the exploratory weighting family, which sets one scheme against the other rather than one arm against the other. The contrast is positive under equal weighting at all five widths, largest at the decile with +0.316 and falling away on both sides, to +0.282 at the quintile and +0.008 at one percent, and the bootstrap interval excludes zero at the three widest levels while covering it at the two narrowest. It is therefore not an artefact of the sizing rule: a contrast that survives equal weighting is a statement about which names the model selects.

That has to be squared with Section 5.2, where membership accuracy at the traded fraction moves by too little to survive the multiplicity correction. The two statements are compatible because precision counts hits and is blind to what is hit: a book can hold a materially different set of names, and a set with a different risk profile, while placing almost the same number of them in the realised top decile. What the equal-weight contrast establishes is that the effect is carried by composition rather than by the sizing rule. It does not establish that the selected names are better ordered, and Section 5.2 finds that they are not.

The inverse-volatility column is uniformly larger, by +0.054 at the decile and by between +0.021 and +0.037 elsewhere, so the gap is small and roughly flat rather than concentrated at either end. The representation effect and the risk-based sizing rule therefore reinforce one another without the first depending on the second, which is the weaker of the two readings Hypothesis 4 distinguishes and the one consistent with the volatility channel of Section 5.3.2. The ordering of the two columns is stable across the grid, so the answer does not depend on the width at which it is asked.

6.4 The two legs

A combined long-short figure can conceal two very different books, and the legs of a long-short equity strategy have no reason to behave symmetrically (Daniel and Moskowitz, 2016).

Two decompositions answer different questions: the additive one of Section 4.8 attributes what each leg contributed to the book actually held, and the standalone one asks whether a leg would have been viable alone. Only the first is additive, and attribution uses it.

Table 6.3 reports both for every sweep specification at the headline width, together with the correlation between the leg return series, which is what determines whether the short leg is contributing return or diversification. The run computes the leg split at that configuration only, so the decomposition is read across objectives and arms rather than down the width grid.

The classification rule was fixed before the estimates were read. If the short leg has a standalone expected return that is not distinguishable from zero, but the combined book has a materially higher Sharpe than the long leg alone, then the short leg is functioning as a hedge and not as a source of alpha. That is a legitimate and common outcome, and the rule commits to reporting it as one. The distinction matters for practice, because a hedge and an alpha source justify different capital allocations and respond differently to borrow cost, which Chapter 7 charges only to the short leg.

Table 6.3: Long/short decomposition of the gross book at the headline width and weighting. Additive contributions split the combined annualised return by the sign of the weight and sum to it exactly; standalone Sharpe ratios renormalise each leg as its own portfolio and are *not* additive. Correlation is between the two monthly leg series. Computed at this configuration only.

Objective	Arm	Additive		Standalone Sharpe			Corr.
		Long	Short	Long	Short	Combined	
RankNet	Percentile	+35.33%	−17.56%	+0.896	−0.414	+1.009	−0.874
	Inverse-normal	+37.55%	−14.68%	+0.950	−0.333	+1.490	−0.905
LambdaRankIC	Percentile	+34.12%	−16.50%	+0.886	−0.395	+1.083	−0.886
	Inverse-normal	+37.37%	−14.57%	+0.972	−0.350	+1.586	−0.912
Margin ranking	Percentile	+34.26%	−16.87%	+0.874	−0.399	+1.123	−0.901
	Inverse-normal	+37.59%	−16.54%	+0.965	−0.397	+1.398	−0.906
Soft-Spearman	Percentile	+34.75%	−15.32%	+0.877	−0.322	+1.119	−0.888
	Inverse-normal	+36.14%	−14.15%	+0.922	−0.292	+1.452	−0.914
RankCosine	Percentile	+35.52%	−15.18%	+0.903	−0.321	+1.174	−0.887
	Inverse-normal	+36.33%	−14.36%	+0.927	−0.301	+1.431	−0.910
Gradient boosting	Percentile	+37.79%	−14.19%	+0.981	−0.320	+1.335	−0.870
	Inverse-normal	+37.79%	−14.19%	+0.981	−0.320	+1.335	−0.870

The short leg meets the return half of the rule, though its point estimates all sit

on one side of zero. The rule fixed above asks for a standalone expected return *not distinguishable from zero*. Run as a standalone portfolio at the headline width, and struck on the excess basis a one-sided book requires (Section 4.10), the leg returns a Sharpe ratio that is negative in every cell, between -0.292 and -0.414 across the ten neural specifications, at Newey–West t -statistics between -1.30 and -1.84 . No cell reaches conventional significance, so the antecedent the classification depends on is satisfied and the hedge reading is licensed rather than weakened. What the consistency of the sign carries is the negative half of the same statement: across all ten cells the leg is never a return source, and the classification rests on that regularity rather than on any single cell’s test. Its additive contribution is negative as well: in the inverse-normal RankNet cell the decomposition assigns $+37.6\%$ a year to the long side and -14.7% to the short.

The variance argument does hold, and gross it holds everywhere. The combined book reaches a gross Sharpe ratio of $+1.490$ in that cell against $+0.950$ for the long leg standing alone, and the correlation column is the reason: the correlation between the two leg return series runs between -0.874 and -0.914 across the neural cells, so the short leg buys variance reduction and pays for it in return. Replaying every specification’s long leg on its own confirms it is not a property of one cell. The long book is therefore struck at a gross exposure of one, where it draws no debit balance and pays no financing. Financing is kinked there and impact grows faster than the notional traded, so a figure struck at one exposure cannot be converted into a figure at another and each end is run separately. That is also the exposure at which the leg’s volatility, 17.7% a year, sits nearest the combined book’s 16.2% , so the two are compared at matched risk and not merely at a matched Sharpe ratio. Gross, the combined book wins in all twelve specifications, by between $+0.11$ and $+0.61$.

Net of costs the picture divides along the representation axis. Charged by the same spread and impact engine and entered on the same next-open convention, the long leg delevered to the book’s risk pays about 246 basis points a year against roughly 805 for the long-short book: it turns over one side rather than two, borrows nothing, and draws no margin loan. The composition differs as much as the level. Of the long leg’s 246, 138 is the entry convention, which costs the combined book only 73 of its 805; the majority of what the long book gives up is the overnight gap between the signal and the fill, and is therefore a consequence of direction rather than of trading.

On the same excess basis the long leg’s net Sharpe ratio runs between $+0.725$ and $+0.826$ against between $+0.551$ and $+1.022$ for the corresponding long-short book. Pooled, the two are a wash, the combined book ahead by 0.016 on average. That average conceals a clean split. The combined book wins in all five inverse-normal cells, by between $+0.049$ and $+0.197$, and in both tree cells; the long leg wins in exactly the five percentile cells, by between 0.038 and 0.197 . The cheaper book is therefore not the better one: the long leg surrenders between $+0.11$ and $+0.61$ of gross Sharpe ratio to buy back some 560 basis

points of cost, and whether that trade pays is not a property of long-short construction in general but of the representation the book is built on.

Three qualifications travel with the comparison. The delevered long-only book is a different product: it carries a market beta of essentially one against roughly 0.1 for the long-short book, so it is a long-biased holding whose Sharpe ratio embeds the equity premium, and a mandate that requires dollar or beta neutrality cannot hold it at all. That beta is also what makes the entry convention cost it, per unit of net asset value, roughly twice what the same convention costs the combined book, so the largest single term in the long book's cost and the largest caveat against holding it are one fact seen twice. The comparison is also a full-sample average of realised quantities with no standard error attached, of the same descriptive kind as the leg statistics themselves. And the split depends on where the long book is struck against the gross ceiling of Table 4.4: financing is 100 of the roughly 805 basis points the long-short book pays and none of the 246 the long leg pays, because a long book held at a gross exposure of one draws no debit balance. Struck instead at the book's own ceiling of four, the leg runs at four times the volatility, pays the financing spread on three units of margin loan and roughly eight times the impact by (4.17), and loses one of the five cells it wins here. What the comparison settles is worth stating exactly: the short leg is a variance hedge gross in all twelve specifications, it pays for itself net in seven of them at matched volatility, and the five where it does not are exactly the five percentile cells.

That reading is carried into Section 7.5, where borrow is charged to the short leg alone and is therefore a fee levied on a diversification benefit that, on the evidence above, covers it in seven of the twelve cells, and into Section 8.3.

It also locates the worst episode of the sample, the drawdown of Section 8.2 that every sweep specification shares. In 2025 the same inverse-normal RankNet book lost 27.3% at the headline width, gross, and the additive decomposition of Section 4.8 places the whole of that loss on one side: compounded over the year the long leg contributed +10.3% and the short leg -36.6%, the two struck separately and so not summing exactly to the book. The leg that exists to reduce variance is the leg that produced the drawdown. Outside the sweep, the elastic-net anchors and the momentum benchmark trough instead in 2021, 2010 and 2010 respectively, so the episode spans the neural cells, the tree control and the least-squares anchors without being a property of everything the study prices; Section 8.2 decomposes it.

6.5 Choosing a width

The headline width itself was fixed in Section 4.9 before any estimate existed and is not revisited here. What this section fixes is a checking rule, stated in advance for the same reason: it names the width the net evidence would select, so that the pre-fixed headline

is confronted with that selection rather than quietly assumed optimal. A width chosen after inspecting the grid would be a selected parameter, and the inference of Section 3.7 would no longer apply to it at face value; had the rule selected a different width, the headline would have stood and the divergence would have been reported beside it.

The rule selects the width that maximises net Sharpe under the base-case cost parameterisation of Table 4.5, not gross Sharpe, so the selected width need not be the one with the best gross figures. Where two widths are tied on that criterion to the printed precision, the wider is preferred, since it holds more names, turns over less, and is less exposed to the calibration of the cost model. And the 1% column is excluded from selection, for the reason given in Section 6.2.

Because the criterion is net, the rule cannot be applied until the cost evidence is on the table: Section 7.5 applies it to Table 7.1. The widths are largely not separated even before costs are charged (Section 6.2), and the two contrasts that bracket the decile do not separate it from its neighbours either, so the selection rests on the point estimates and the tie-break clause is a stated convention rather than an evidenced one. Section 7.5 should be read with that in mind.

Chapter 7

Trading Frictions and Implementability

This chapter prices the books of Chapter 6, and closes by separating four distinct claims that the phrase “the strategy works” routinely conflates.

7.1 Replay design

Costs are not applied during estimation. Every model in Chapter 5 and Chapter 6 is trained and scored under the gross convention, the resulting per-date score vectors are frozen, and the execution layer is applied afterwards by replaying those vectors through the entry convention and the cost stack of Section 4.10.

Freezing the scores is admissible because nothing in the cost stack enters the training objective or the model-selection criterion: the losses of Section 4.3 are functions of scores and labels alone, the only quantities selected anywhere in the design are selected on validation rank information coefficient alone, and no gradient reaches a cost term. It therefore changes only when the arithmetic is performed, not what is computed, and it means a single trained model can be priced under many cost parameterisations without retraining, which is what makes the capacity analysis of Section 7.4 feasible at all.

Because costs never influence estimation, no model here is cost-aware: none has been trained to prefer a cheaper ordering, and none penalises turnover in its objective. The net figures in this chapter are therefore the net performance of a gross-trained model, not the net performance attainable by a model that had been optimised for net performance. DeMiguel et al. (2020) show the two differ, and the gap is a matter of design and not of estimation error. Where this chapter reports that costs erode a difference, the correct reading is that the difference does not survive execution *under a gross-trained model*.

7.2 Cost assumptions

The parameters are those of Section 4.10, fixed before estimation and not calibrated to results, the impact prefactor being the one the live execution record of Frazzini et al. (2018) supports. The net figures are therefore read as a point estimate under a stated parameterisation, not as a bound in either direction. A reader who holds a different view of any single parameter can substitute it directly, because the bridge of Section 7.3 reports each component separately.

One of the four components is measured and one rests on an assumption the panel cannot verify. The spread is measured: the panel carries a closing bid and ask on almost every name-date, so each name is charged half its own quoted spread on the rebalance date rather than a common rate. That matters because the quoted spread varies widely across the cross-section and declines over the sample, so a flat charge would systematically overstate cost for the most liquid names and understate it for the least liquid, and would misstate the level in the early years. The configured basis-point rate is added on top as a commission. A name without a quote at the rebalance is charged its own last observed half-spread, and only a name never yet quoted falls back to the commission alone. The three tiers carry 97.83%, 2.11% and 0.06% of traded notional, the last confined to the opening months of the sample, before the carried history has filled. Even so, a quoted spread is not the effective spread actually paid, since an order need not execute at or inside the quote, so the charge is a lower bound on the true cost of crossing.

Borrow is the component that remains an assumption. It is keyed on a liquidity proxy rather than on measured lending-market utilisation, so it captures the direction of borrow scarcity but not its episodic spikes, and it will understate cost precisely in the periods when a crowded short book is most expensive to hold. That is not defensible as a measurement; it is defensible only as a transparent convention, stated rather than asserted. Its sensitivity is not traced here: no borrow curve other than the base case of Table 4.5 is estimated anywhere in this thesis, so every net figure below is conditional on that convention holding.

The financing term is charged on the debit balance of (4.18), the long market value less the unit of equity. The gross ceiling of Table 4.4 binds at every rebalance, fixing that balance at one unit of net asset value. Financing is therefore a standing charge on a borrowing the book never lets go of rather than an occasional one, and it is close to invariant to the traded fraction. It enters every width's return identically for that reason, and reaches the width decision of Section 6.5 only through the differing volatilities of the books rather than through the charge itself.

The ceiling is not a flattering choice, and the direction can be shown rather than asserted: the net Sharpe ratio is strictly decreasing in a common rescaling of the book, so Section 7.3 prices it at the most punitive point of the range from a gross ceiling of one

to the four used here. Section B.3 gives the component scaling this rests on, and the two things it sets aside.

7.3 The gross-to-net bridge

Table 7.1 decomposes the distance between gross and net performance into its four components at every width and under both weighting schemes, for the reason given in Section 2.5: a single net number would conceal which constraint binds and therefore which assumption the conclusion depends on.

Spread and impact scale with turnover, and turnover rises as the tail narrows, while borrow and financing are levied on exposure and are close to flat across widths. The expected pattern is therefore that the net-of-cost width profile is flatter than the gross profile, and that if a narrow width wins gross it wins by less net, or loses.

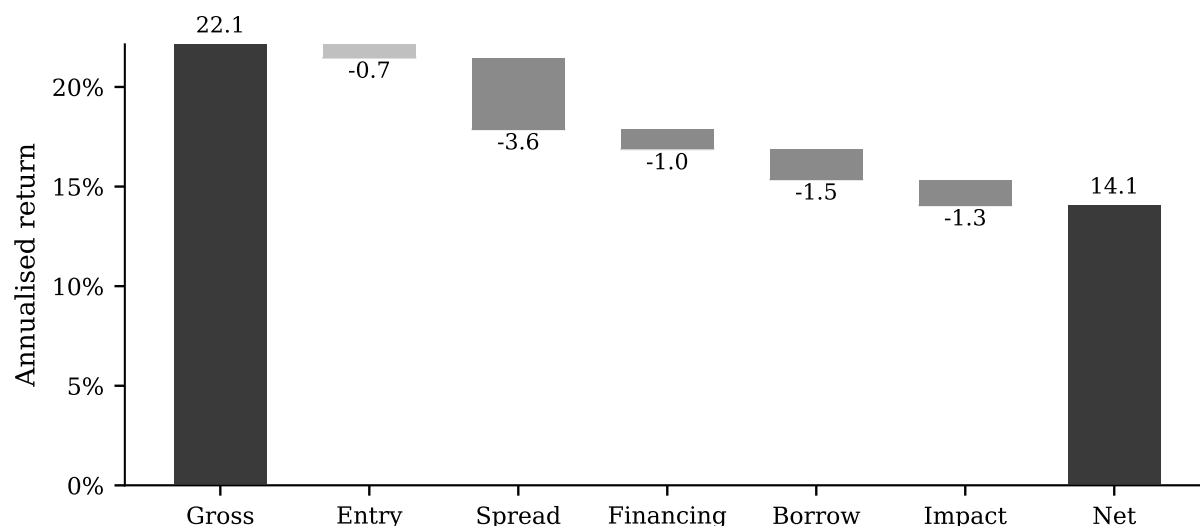


Figure 7.1: Annualised return from gross to net at the headline width and weighting, averaged across the five inverse-normal neural objectives, which is the book Table 7.1 prices. The first step is the executable entry convention, priced against an intermediate book that enters at the next open and pays nothing; the four charged components are levied on that book. The steps reconstruct the distance exactly, so the bars close on the net bar. The labels are rounded to a tenth of a point and for that reason do not quite add to it on the page; the closure is checked on the unrounded returns.

Figure 7.1 shows the arithmetic before the table breaks it out by width.

At the headline decile width under inverse-volatility sizing, the inverse-normal book earns 2214 basis points per year gross and 1405 net, so execution takes 809 basis points a year, a little over a third of what the book earns before costs, of which the executable entry convention accounts for 67 and the four charged components for 742. The spread dominates that stack at base-case capital and at this width, at 360 basis points, and impact is not the binding charge: at 130 it sits below borrow at 152. The two charges a turnover-based haircut omits altogether, borrow and financing, come to 252 between

them, a third of the charged stack and nearly twice the impact term, which is why a decomposition and not a blended rate is reported. These are annual drags on portfolio return rather than rates per traded dollar; no single per-traded-dollar spread rate exists, since each name is charged half its own quoted spread with the commission added on top, and Table 7.1 carries the same decomposition at every width and under both weighting schemes.

Impact is the component practitioners most often cite as the binding constraint on a strategy of this kind. Scaling it by the name's own volatility keeps it a visible share of the drag even for a book far too small to move the market, which a flat dollar coefficient would report as a rounding error, and it is the term whose weight grows with capital in the capacity analysis of Section 7.4, where the assumed book size is the quantity being varied.

The entry convention accounts for the remainder, and it is measured rather than inferred. The bridge runs through three books, not two: the gross book entering at the signal close, an intermediate book entering at the next open with every cost switched off, and the net book carrying the full stack. The step from the first to the second is the executable entry priced on its own, and at the headline configuration it is worth 0.67 percentage points of the 8.1-point distance between gross and net, or 8% of it. The remainder is the charged stack above. Reporting the intermediate book is what turns that term from a plug into an estimate: under the two-regime bridge it could only be obtained by subtraction, so any cost escaping the component isolation would have been silently relabelled as an entry effect. The term is not signed across the grid. It is positive at every fraction but the narrowest under both schemes, and turns slightly negative there, -44 basis points under equal weighting and -5 under inverse-volatility sizing, and nothing should be read into that: it is the 1% column of Section 6.2, on which no conclusion in this thesis rests.

The design gives the entry convention more prominence than the number supports: for the reason given in Section 3.4, the move that survives in a dollar-neutral book is more than an order of magnitude smaller than the price of trading the deciles. The overnight gain on each leg separately, and the t -statistic on their difference, are not among the frozen bridge outputs and are not reported here, so the statement rests on the aggregate entry term rather than on a leg-level decomposition of it. That term is enough for the question at issue: implementability does not bind at the entry. The close-to-close convention flatters the book, and it must be removed for the net figures to describe a portfolio anyone could have held, but it is a correction of the right sign and the wrong order of magnitude to be where the strategy is lost.

No specification anywhere in the grid retains three quarters of its gross Sharpe ratio once the entry convention is made executable and the stack is charged, and at the headline width and weighting none retains two thirds, the highest being the gradient-boosted

control at 0.659 (Tables C.1 and C.2), so any performance figure quoted for this strategy is uninterpretable unless the convention it refers to is stated.

Table 7.1: Gross-to-net annualised return in basis points per year, under the base case of Table 4.5, averaged across the five neural objectives in the inverse-normal arm. *Entry* is the next-open convention priced on its own, against an intermediate book that enters at the next open and pays nothing; the four charged components are levied on that book, and the six columns reconstruct the net return. Each entry is rounded to the nearest basis point on its own, so a row can miss its printed net by one; the closure is checked on the unrounded returns. The sample is the 272 monthly rebalances of April 2003 to November 2025 on the top-1000 United States universe. These are annual drags on portfolio return, not rates per traded dollar.

Weighting	α	Gross ret.	Entry	Spread	Impact	Borrow	Financing	Net ret.	Net SR
Equal	20%	1831	34	310	86	152	100	1148	0.878
	10%	2373	31	359	139	148	100	1596	0.918
	5%	2828	21	395	217	143	100	1952	0.865
	2.5%	3445	13	425	333	140	100	2435	0.815
	1%	3818	-44	459	573	135	100	2595	0.609
Inverse-vol.	20%	1708	66	312	79	156	100	996	0.846
	10%	2214	67	360	130	152	100	1405	0.932
	5%	2559	48	395	204	147	100	1665	0.857
	2.5%	3130	50	424	314	143	100	2098	0.804
	1%	3557	-5	454	542	138	100	2328	0.612

The representation contrast is then re-estimated on net returns. Table 7.2 reports the paired Sharpe difference between arms before and after costs at each width. A contrast can narrow under costs for either of two reasons, and they are not the same finding. The better-ordering arm may hold the more expensive book, which is the reordering Jensen et al. (2026) describes and that this thesis is built to detect. Or the two arms may part with the same money against different realised volatilities, in which case the narrowing is arithmetic on the denominator of a Sharpe ratio and carries no implication about implementability at all. Differencing two Sharpe ratios cannot separate them; Table 7.1 can, because it prices each arm’s stack rather than netting the two.

Table 7.2: Paired Sharpe difference between the arms, gross and net, at each traded fraction under inverse-volatility sizing, averaged across the five neural objectives. The last column is the share of the gross contrast surviving execution. Computed from the reported Sharpe levels; the headline row matches Table 5.4 to within 0.001.

α	$\Delta\text{SR gross}$	$\Delta\text{SR net}$	Share surviving
20%	+0.319	+0.231	72%
10%	+0.370	+0.307	83%
5%	+0.272	+0.254	93%
2.5%	+0.187	+0.174	93%
1%	+0.040	+0.051	128%

The contrast narrows under costs but neither reverses nor loses significance. The pooled contrast is the primary endpoint, struck in Section 3.7 under the gross convention and re-struck here under the realistic one: the paired Sharpe difference falls from

+0.370 gross ($t = 4.46$, $p < 0.00025$) to +0.307 net ($t = 3.92$, $p = 0.00050$), with a net confidence interval of $[+0.158, +0.457]$ that excludes zero. The net figure prices the same pre-specified quantity through execution; it is not a second endpoint, and it carries no multiplicity claim of its own. The advantage therefore survives execution without surviving it intact. Every per-objective contrast keeps its sign, but in the secondary family the Romano–Wolf stepdown leaves only two of the five individually significant net (Section C.2). The pooled result must not be read as five separate findings.

On the bridge it is the second reason and not the first. The inverse-normal arm is the busier book, turning over 2.75 per rebalance against 2.66, and it does pay twelve basis points more at the spread; but it gives sixteen back at the entry convention, so the two arms part with almost exactly the same money, 809 basis points a year against 804. What separates them is what that charge is divided by. The inverse-normal book runs a realised volatility of 15.1% against the percentile book’s 16.8%, so an identical deduction costs the better book more of its Sharpe ratio: had the two paid precisely the same money, the attenuation would have been 0.056 of the 0.062 observed. Nine tenths of it is the denominator.

The width grid settles the mechanism, because the two candidate explanations disagree across it. The volatility gap between the arms falls from 15.8% at the quintile to -2.3% at one percent, and the attenuation falls with it monotonically, from +0.088 to -0.011 , changing sign exactly where the gap does. The cost gap predicts none of that: it is erratic and twice negative, the inverse-normal arm being the *cheaper* book at both the five percent and the one percent fractions while the contrast still narrows at the first and gains at the second. So the share of the gross advantage surviving execution in Table 7.2 rises as the book narrows, from 72% at the quintile to 128% at one percent, and it does so although the narrow book is by far the more expensive one to trade, at 1229 basis points against 713.

The reordering that Jensen et al. (2026) describes is therefore not what this contrast exhibits. Costs compress the distance between the arms at the wider fractions without inverting it at any of them, but they do not compress it by charging the better-ordering book more: at the headline width that book is dearer by five basis points on a stack of eight hundred. What the net re-strike establishes is that the advantage survives execution, not that it was ever a trading-cost artefact.

That compression is nonetheless the smaller of the two effects at work in this chapter. Netting is the only group within the exploratory family in which every member is significant under Benjamini–Hochberg at $q = 0.10$, all fifteen of them, with differences ranging from -0.564 to -0.455 , which is more than the representation choice adds gross. The representation effect is real and it survives; it is not large beside the cost of trading.

7.4 Capacity

The base case sets managed capital at \$1 million, and at that size each position trades a small share of the average daily volume of a top-1000 name, the resulting charge of about two basis points per traded dollar sitting well inside the spread-and-commission rate. The base case therefore reports a book that is small enough not to move the market, and a net figure obtained at that size says nothing about the capacity of the strategy.

That is why capacity is analysed. Because (4.17) makes impact depend on the traded fraction of average daily volume, the curve of net performance against managed capital is the economically informative object: it identifies the capital at which net performance reaches zero. Table 7.3 reports that curve. The narrow widths are expected to saturate first, because the same notional is pushed through fewer names, which is the mechanism by which concentration and capacity trade against one another.

Table 7.3: Net annualised Sharpe ratio by managed capital and traded fraction, holding the frozen scores and Table 4.5 fixed, averaged across the five inverse-normal neural objectives under inverse-volatility weighting. Only impact varies with capital, scaling as $\sqrt{A}$ at fixed weights, so the break-even row follows in closed form from the base case. Participation is priced as if each rebalance trades in a single day.

Managed capital A	$\alpha = 20\%$	10%	5%	2.5%	1%
\$1 million	0.85	0.93	0.86	0.80	0.61
\$10 million	0.70	0.75	0.63	0.54	0.30
\$100 million	0.24	0.16	-0.09	-0.27	-0.64
\$500 million	-0.60	-0.90	-1.33	-1.64	-2.12
\$1 billion	-1.23	-1.66	-2.18	-2.54	-3.00
Break-even A	\$183 million	\$140 million	\$84 million	\$59 million	\$28 million

The curve confirms both expectations. Net Sharpe at the decile falls monotonically as managed capital rises, and the break-even capital in the final row of Table 7.3 falls monotonically as the traded fraction narrows, so the narrow widths saturate first and concentration is bought at the price of capacity: the same notional pushed through fewer names exhausts the market sooner. The base-case ordering of the components does not generalise upward. Impact enters the bridge third of the four at the decile, above financing and below borrow, and is the only one that grows with the size of the book: financing is a standing charge on a debit balance the gross ceiling fixes at every rebalance, borrow is levied on short exposure and holding period, and the spread is charged per traded dollar rather than per dollar managed. Because impact alone scales as $\sqrt{A}$ at fixed weights, and already exceeds financing at the base case, it must overtake borrow first and then the spread, and each of those crossings arrives at a smaller capital the narrower the book. The net Sharpe levels and the break-even capital are read from Table 7.3; the capital at which impact overtakes each of the other charges is not among the frozen outputs and is

not quoted.

Break-even capital therefore scales as $1/\kappa^2$ in the impact prefactor, so a reader who holds a different view of κ can rescale the whole table without re-estimating anything (Section B.3). The value used here reproduces the costs Frazzini et al. (2018) record at the participation rates they report, so the capacities are priced at the execution a large institutional manager achieved. The operative assumption is instead the single-day participation convention stated in the caption. A desk that works each rebalance over several days pays less impact than the convention charges, though less than the arithmetic of the law alone implies: the costs the prefactor is calibrated to were themselves realised over trade horizons that run past a single day, so part of that scheduling is already in the figure. What is not priced anywhere is the tracking error such a schedule buys against the paper book.

7.5 What survives

Four claims are commonly conflated under the phrase “the strategy works”, which compresses them into one and thereby overstates the weakest.

- C1.** *Statistically predictive.* The scores order the cross-section better than chance, measured by the rank information coefficient and its information ratio over the evaluated dates.
- C2.** *Gross profitable.* The traded book earns a positive risk-adjusted return before costs, under an entry convention no investor could have achieved.
- C3.** *Net profitable.* The same book earns a positive risk-adjusted return after the entry convention is made executable and the cost stack is charged.
- C4.** *Capacity bearing.* The net result persists at a capital base large enough to be of interest to an institutional allocator, rather than only at the base case.

A strategy can satisfy C1 and C2 while failing C3. The width-checking rule of Section 6.5, applied to Table 7.1, returns the 10% fraction under both weighting schemes, by a margin in net Sharpe of 0.075 over the 5% level under inverse-volatility sizing and 0.040 over the quintile under equal weighting.

The step from C2 to C3 bundles two changes of very unequal size, and the bridge of Section 7.3 separates them. Making the entry executable is what the claim turns on logically: a close-to-close figure is not a portfolio anyone could have held, so C2 is an accounting benchmark and not a performance claim. The entry convention accounts for at most 67 basis points a year anywhere in the grid, against a gross-to-net distance of between 682 and 1229, so the step from C2 to C3 is almost entirely the price of trading and not the cost of waiting a day to enter.

The classification also determines what may be said about the short leg. Since Section 6.4 finds the short leg to be a variance hedge gross and not a return source, borrow is a charge levied on a diversification benefit rather than on an alpha stream, and the correct question is whether the reduction in volatility is worth the fee. On the net comparison of Section 6.4 the answer depends on the encoding. The reduction covers its fee in seven of the twelve cells, and the five where it does not are exactly the five percentile ones, which makes the verdict a statement about the representation the book is built on as much as about the leg or this cost parameterisation. The alternative it is measured against holds only for an investor free to run the resulting market exposure.

C1 and C2 are supported, C3 is supported at every traded fraction and under both weighting schemes at the base-case parameterisation of Table 4.5, and C4 is not. The scores order the cross-section better than chance in both neural arms (Table 5.2), and C3 survives the entry convention and the full charged stack across the whole width grid, the narrowest fraction still netting a Sharpe ratio of 0.61 (Table 7.1). The decile is where the margin is widest rather than where the sign turns.

Two qualifications travel with that verdict, and the classification exists to keep them visible. The first is that C1 does not carry C3. At the traded decile under inverse-volatility sizing the XGBoost control has the highest rank information coefficient of any cell, yet a gross Sharpe ratio below every inverse-normal neural cell and a net Sharpe ratio that four of the five exceed; neither ordering holds at every fraction of the grid. The pooled net margin is +0.052, which this sample cannot resolve, and it turns to -0.014 under equal weighting, so ordering quality and portfolio quality are ranked differently by the same evidence while the two families are not separated by it net of costs (Section 5.6.3). The second is that C4 is answered at \$1 million and only through the replayed grid: the break-even capital in the final row of Table 7.3 marks where the net result of C3 expires, and at the decile it arrives at \$140 million, which is short of institutional scale. That figure carries the impact prefactor with it, and Section 7.4 gives the rescaling. The prefactor is not itself a conservative choice: it reproduces the costs Frazzini et al. (2018) record at the participation rates they report. What the break-even does inherit is the single-day participation convention of Table 7.3, and Section 7.4 states how far that carries. It also inherits the terminal date: struck at the family book's last peak before the episode it is \$193 million rather than \$140 million (Table 8.6), so the figure carried here is the low end of the range the sample supports rather than a point estimate. Capacity is also the one verdict in this chapter whose *shape* turns on a single component: borrow and financing do not vary with managed capital, so they do not bend the curve and impact alone does. They do enter its level, and quadratically, since the break-even is struck where the base-case net return is exhausted and that return is net of all four charges. A reader who prices borrow or financing differently therefore moves the break-even without moving the shape of the curve, which is why the level results of Section 7.3 are insensitive to the

parameter that decides this one.

To these must be added a caveat that no full-sample average conveys: every sweep specification, the XGBoost control included, reaches its maximum drawdown in 2025, and none has recovered by the last observation, so the deepest episode in the sample is still open at the point where the net figures above are struck.

Part III

Robustness, Failure, and Interpretation

Chapter 8

Robustness, Subperiods, and Failure

The preceding chapters report averages over the full evaluation sample. This chapter asks when those averages held and when they did not, and confronts the alternative explanations under which the results of Chapter 5 would not mean what they appear to mean.

No period is removed from any headline estimate on the basis of its outcome; Section 3.6 fixed this before the sample was examined. The subperiod analysis below is a decomposition of the reported result, not a selection from it.

8.1 Subperiods

The sample is partitioned into six economically defined regimes instead of into equal calendar blocks. Equal blocks are simpler to defend against a charge of ex-post construction, but they cut across the events that plausibly govern a cross-sectional equity strategy, and a partition that separates the global financial crisis from the decade that followed is more informative than one that splits the decade in half. The boundaries are drawn on macroeconomic and market events that are not specific to this strategy, and they are fixed for every specification, so no partition is tuned to a result. Table 8.1 states the resulting partition.

Two of the six regimes contain fewer than thirty monthly observations, and a Sharpe ratio estimated on twenty-four months has a standard error wide enough to accommodate almost any hypothesis. The subperiod figures are read for the pattern they trace across regimes and for the identification of which environments were adverse, never as a set of independently testable estimates.

The gross record deteriorates across the coarse three-block split behind Figure 8.1: for the RankNet inverse-normal cell the gross Sharpe ratio is +1.882 before 2010 (an annualised return of +27.71%), +1.884 over the 2010s (+22.60%), and +0.832 over the 2020s (+16.01%). No measure of cross-sectional dispersion is reported alongside these

Table 8.1: Subperiod partition of the evaluation sample. Boundaries are drawn on market-wide events and are identical for every specification. A rebalance is stamped on the date its book is formed while its return is earned over the following month, so a rebalance is assigned to the regime containing the month it earned in; the counts below therefore run from May 2003 to December 2025 and sum to the 272 evaluated rebalances.

Regime	Span	Rebalances
Pre-crisis expansion	2003–2007	56
Global financial crisis and aftermath	2008–2012	60
Low-rate, large-cap growth	2013–2019	84
COVID shock and recovery	2020–2021	24
Inflation and rate shock	2022	12
Concentration	2023–2025	36

Table 8.2: Performance by regime for the equally weighted composite of the five neural inverse-normal books, at the headline width and weighting. Compositing first diversifies objective-specific noise, so the composite’s full-sample Sharpe (1.59) exceeds the average of the five (1.471, Table 5.8); they are different objects. Return is compounded while volatility and both Sharpe ratios use arithmetic monthly moments, so Return over Vol. does not reproduce SR gross. None of these figures is tested.

Regime	Window	Mths	Return	Vol.	SR gross	SR net	% pos.
Pre-crisis expansion	2003–2007	56	20.3%	11.1%	1.73	0.79	75%
Global financial crisis and aftermath	2008–2012	60	41.8%	12.6%	2.88	2.11	80%
Low-rate, large-cap growth	2013–2019	84	19.6%	10.8%	1.72	1.10	71%
COVID shock and recovery	2020–2021	24	32.6%	20.7%	1.47	1.19	62%
Inflation and rate shock	2022	12	40.4%	14.1%	2.50	1.82	75%
Concentration	2023–2025	36	-0.2%	18.4%	0.08	-0.27	53%
Full sample	2003–2025	272	23.4%	13.9%	1.59	1.01	71%

Table 8.3: Gross and net Sharpe ratio of the momentum benchmark and of the neural family, by the regime partition of Table 8.1. The neural columns are the equally weighted composite of the five inverse-normal books, the same object and the same figures as Table 8.2, and not the mean of the five per-objective Sharpe ratios that Table 5.8 reports and that Section 5.8 fixes as this thesis’s pooled figure. Compositing first diversifies objective-specific noise, worth +0.12 of full-sample gross Sharpe ratio (1.59 against 1.471), so the neural side carries a diversification the single fixed rule beside it does not; the two operators differ by less than 0.03 at the Concentration regime, which is the row Section 8.1 reads. Neither set is tested; both are read for the pattern across regimes.

Regime	Window	Mom. gross	Mom. net	Neural gross	Neural net
Pre-crisis expansion	2003–2007	0.25	-0.08	1.73	+0.79
Global financial crisis and aftermath	2008–2012	-0.28	-0.46	2.88	+2.11
Low-rate, large-cap growth	2013–2019	0.18	-0.03	1.72	+1.10
COVID shock and recovery	2020–2021	-0.12	-0.63	1.47	+1.19
Inflation and rate shock	2022	1.84	+1.87	2.50	+1.82
Concentration	2023–2025	0.51	+0.47	0.08	-0.27
Full sample	2003–2025	0.13	-0.09	1.59	+1.01

figures, so the conjecture that the strongest regimes are the widest-dispersion regimes is not testable here and is not asserted. The regime-level compression from gross to net can be read off Table 8.2, but none of those figures is tested and no regime is identified as the most compressive; the full-sample wedge is priced in Section 7.3.

The terminal regime is the one that bears on any forward-looking claim, and it is much the weakest in the record: the concentration regime of Table 8.2 contains the episode of Section 8.2, which is still open when the sample ends. Two readings are available. The first is an adverse draw, although the coarse split contains no earlier block of comparable weakness against which to calibrate one. The second is the post-publication decay of McLean and Pontiff (2016), under which the terminal record is the beginning of the effect's erosion rather than a passage through it. A three-year window cannot separate them, and no attempt is made here to adjudicate.

The benchmark removes the most comfortable explanation. It is a weak competitor throughout, running a raw volatility more than half again the highest any fitted book runs (Table 5.8) and a geometric mean of -1.90% a year; but in the final regime of Table 8.3 it is the fixed rule that pays and the neural family that does not, gross and net alike. Neither set of regime figures is tested, and the comparison is read for its pattern only. Subject to that, a general withdrawal of cross-sectional dispersion is not what happened in that window, and the stronger and less comfortable statement stands: in the terminal regime the models failed relative to an available alternative, not in an environment where nothing worked.

The representation contrast is not decomposed by regime. It is estimated pooled, by objective, by width, by weighting, and separately for each step of (5.1), but not within regimes: Figure 8.1 reports the two base arms separately over three coarse blocks, without a paired contrast and its own standard error inside each. No across-time version of the stability is therefore evidenced here and none is claimed. The stability Hypothesis 2 does ask about runs across objectives, and it holds in sign for all five contrasts gross and net; individual significance is partial, with only RankNet and LambdaRankIC surviving the Romano–Wolf stepdown in both cost regimes (Section 5.3).

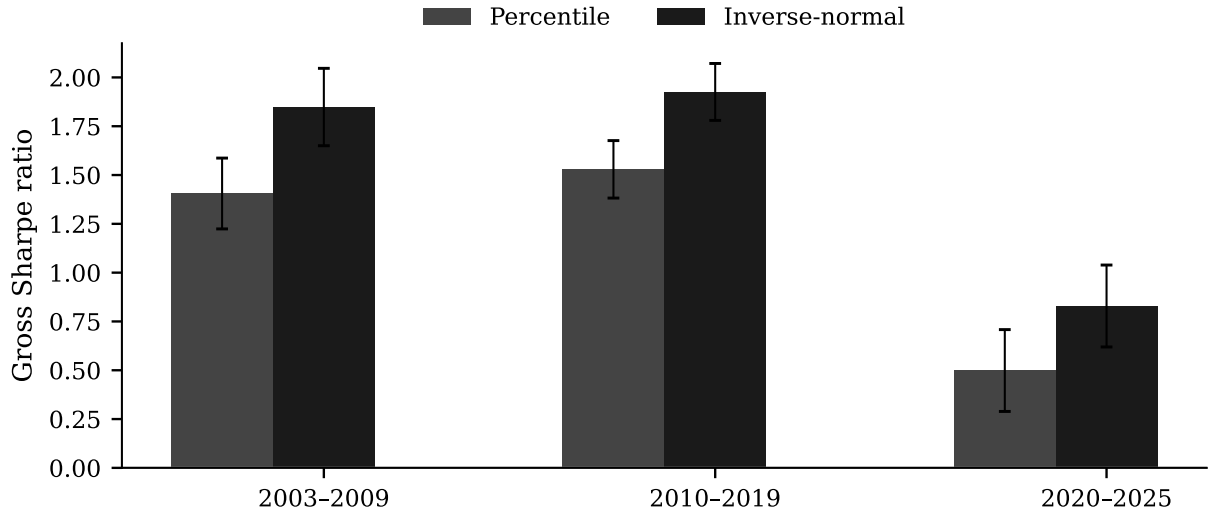


Figure 8.1: Gross Sharpe ratio for the percentile and inverse-normal arms over the three coarse subperiods for which HAC standard errors were computed, averaged across the five neural objectives. The scale-matched arm does not appear: the subperiod statistics behind this figure were computed for the two base arms only. Bars are the mean per-cell Newey–West standard error divided by $\sqrt{5}$, the standard error the mean would carry were the five objectives independent; since their returns are strongly correlated (Section 5.8), the bars understate the uncertainty of the mean, and the figure is read for the pattern across blocks rather than for the separation of the arms within one. The finer six-regime split of Table 8.2 is more informative about which environments were adverse; this figure is included because it carries the uncertainty, which the table does not.

8.2 The worst common drawdown

Every specification estimated in this thesis experienced substantial drawdowns, and the deepest of them are shared across specifications rather than idiosyncratic to any one. A failure common to otherwise independent specifications is evidence about the model family, which is why it is treated here at length.

The episode is located from the data and not asserted in advance. Forming the equally weighted average of all sweep specifications gives a family book, and the deepest peak-to-trough decline of that book defines the episode. Doing this separately for the gross and the net series turns out to identify two different events, and the distinction is itself informative, so both are carried through instead of reconciled into one window.

The episode admits two readings, and they are distinguished by measurement rather than by assertion: the drawdown may be specific to a representation, an objective or a width, and so belong in the assessment of the choices this thesis studies, or it may be common to the family, and so be evidence about cross-sectional machine-learning strategies in that environment. A diagnostic that would settle a third question is missing. The share of names held in common across specifications would separate books that merely experienced the same environment from books that were in fact the same book; it requires per-rebalance weight vectors that are not among the stored artefacts, and the

section closes on what its absence costs.

Figure 8.2 shows the episode against every specification at once. The lower panels make the second reading the obvious one: the decline arrives at the same dates and to comparable depths across cells, and the gradient-boosted control, which is blind to the treatment, sits inside the same band as the neural cells.

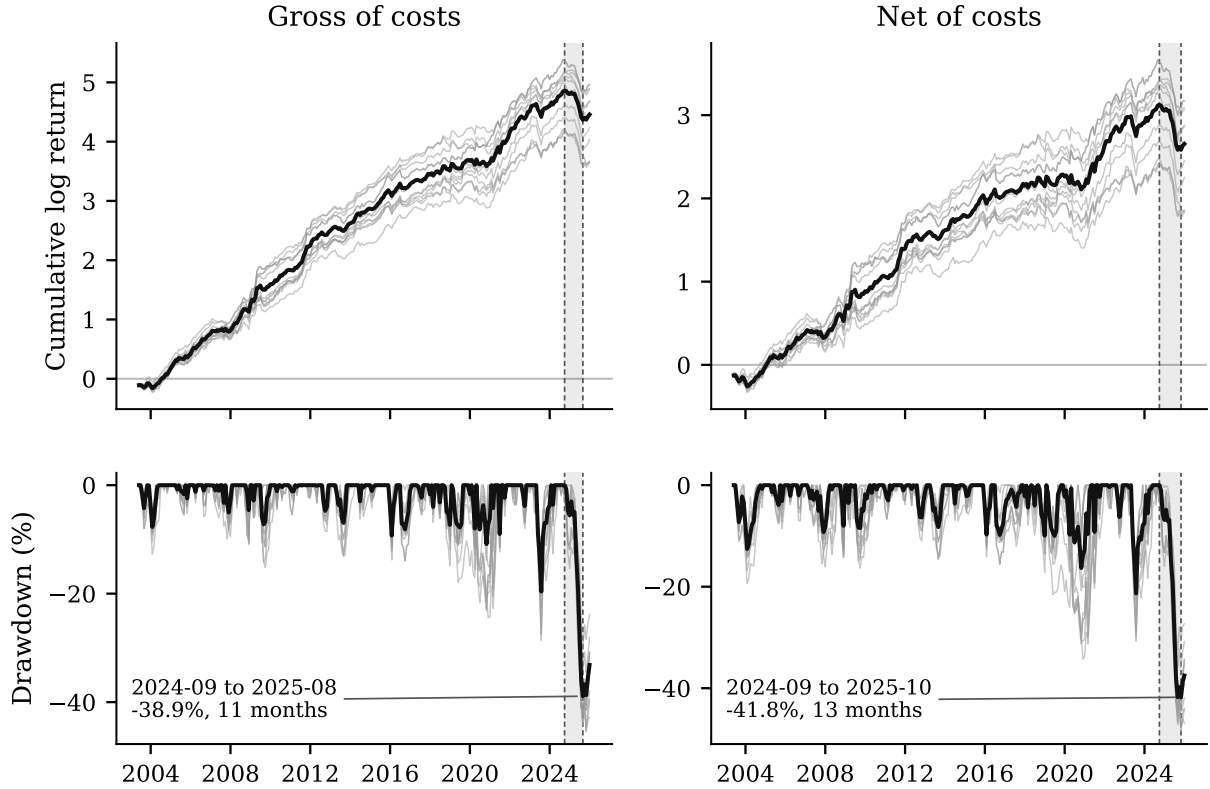


Figure 8.2: Cumulative log return (upper) and drawdown path (lower) for every specification, gross and net of costs. Grey lines are the individual specifications; the heavy line is the equally weighted family book. The shaded band and the dashed rules mark the deepest drawdown of the family book, which need not be the same episode gross and net. Dispersion in onset and depth across specifications is the quantity of interest.

The episode closes the sample, and it belongs to the family rather than to the treatment. Every one of the twelve sweep specifications, the XGBoost control included, reaches its maximum drawdown in 2025, with troughs falling between July and October, and none recovers before the sample ends: the deepest decline of the study is still open at the last observation. Measured on the volatility-scaled series the depths run from -31.32% for LambdaRankIC inverse-normal to -45.33% for margin ranking inverse-normal, and calendar-2025 cumulative returns from -21.52% to -34.82% . That the tree control reaches its own maximum drawdown of -35.81% in the same year carries most of the attribution, since its two arms are indistinguishable in every reported respect, with identical Sharpe ratios gross and net, identical rank IC, and a representation contrast of exactly 0.000: the treatments studied here did not protect against the episode, which is a weaker and more accurate statement than that they caused it or were immune to it. Mean pairwise

Table 8.4: The deepest drawdown of the equally weighted family book, then measured on each arm. Depth is peak-to-trough on the *raw* series, so the arms are compared at the risk each actually ran; duration counts months from peak to trough. Correlations are the mean pairwise figure across the twelve sweep specifications. The gross and net episodes are different events and are not forced onto a common window.

	Gross		Net	
	Depth	Duration	Depth	Duration
Percentile	-41.7%	11	-44.4%	13
Inverse-normal	-36.4%	11	-39.6%	13
Family book	-38.9%	11	-41.8%	13
Episode	2024-09 to 2025-08		2024-09 to 2025-10	
Recovered by	not in sample		not in sample	
Mean pairwise corr., episode		0.791		0.772
Mean pairwise corr., full sample		0.790		0.786

correlation across the twelve specifications during the episode is essentially unchanged from its full-sample level in both regimes (Table 8.4), these books having been close substitutes throughout, so the commonality registers in the timing rather than in a rise in comovement. The gross and net episodes are dated separately in that table rather than forced onto a common window, and neither is recovered within the sample.

It is a short-leg event in every specification, and Section 6.4 gives the split for the RankNet inverse-normal cell in a year when the market factor returned +12.81%. Applying the additive decomposition of Section 4.8 to the months of each specification's own decline puts the same sign on the short leg in all twelve: summed over that window it contributes between -33.3 and -59.0 percentage points, more than the whole of that cell's drawdown in every case, while the long leg's contribution is small and of mixed sign, between -8.6 and $+5.6$. The long side neither caused the decline nor offset it, and the whole of the drawdown, and more, is attributable to the short book. Nor is the pattern particular to this episode: the three cells outside the sweep whose deepest decline falls elsewhere in the sample (Section 6.4) decompose the same way, so across every cell with a stored leg series there is no exception. The leg classified in Section 6.4 as a hedge against alpha is the leg that produces the worst episode of every specification this thesis prices. The obvious factor explanation covers part of the episode and not the whole of it. The book carries a large profitability loading (RMW +0.895, $t = 7.21$) and RMW fell 11.13% in 2025, so the exposure accounts for close to ten points of the family book's calendar-2025 return of -29.98%: roughly a third of it, and the largest single attribution the factor model supplies. The window carries part of that share. Over the eleven months of the gross episode of Table 8.4 RMW fell 7.24% instead, against a decline deeper than the calendar year's, so the same loading is worth nearer a sixth of the episode than a third of it: the calendar year drops the close of 2024 from the fall while retaining the worst of

the profitability reversal. On neither reading is the exposure the mechanism behind the remainder. Nor is the episode an artefact of the panel: the universe is a full thousand names in every month of 2025 and the median fill touches no more of the feature surface in the terminal year than it does across the sample, so coverage in the terminal year is no worse than the historical norm.

What the episode *was* can be established further than that, and with the artefacts already stored. Replaying each specification's frozen scores through the production selection and weighting rules, and retaining the per-name contributions the backtest sums away, separates three readings of the same twelve months. The window used throughout is the twelve returns earned from October 2024 to September 2025: the eleven-month gross episode of Table 8.4 together with the month that follows its trough, which is carried so the forensics can see whether the mechanism it identifies persists past the low.

The first reading is that the ordering itself stopped working. It is measured by comparing the realised return of the names each side selects against the equally weighted mean of the whole eligible universe on the same date, which strips out leverage, the sizing rule and the cost stack and leaves the selection. Before the episode the short side of every specification picks names that underperform the average name, by between 0.42 and 0.61 percentage points a month; over the episode all twelve pick names that *outperform* it, by between 1.07 and 1.69. The shift is large in every specification, with Welch statistics between 2.13 and 2.91 quoted as a gauge of magnitude rather than as tests, under the subperiod convention of Section 3.7. They could not be read as tests even descriptively: the window is the deepest realised peak-to-trough of the family book, so a statistic computed on it is selected as well as estimated, and its reference distribution is not the one a Welch statistic assumes. What carries the finding is that all twelve specifications shift the same way. The long side moves the same way but less decisively, from between +0.46 and +0.63 to between -0.08 and -0.66, at statistics between -1.54 and -3.12. Formed without any portfolio rule at all, the raw long-minus-short selection spread runs between +0.9 and +1.2 percentage points a month before the episode and is positive in roughly seven months in ten; across the episode it runs between -1.2 and -2.2 and is positive in between a quarter and two fifths of them. The ordering inverted, and it inverted upstream of every construction choice this thesis studies.

The second reading is a short squeeze, and the composition of the book argues against it as the principal mechanism. Its median constituent traded about \$70–82 million a day before the episode and between \$184 and \$219 million during it, so the names it held grew more liquid rather than harder to borrow, and the typical bad month was no more concentrated in a few names than the typical bad month before it; the one squeeze-consistent trace is a thickening right tail of shorted-name returns (Section C.4). Crowded positions were being unwound; they were not what produced the drawdown.

The third reading is a factor rotation, and the regression argues against it. Refitted

on the twelve months of the episode window alone, and read descriptively under the subperiod convention of Section 3.7, since no test estimated on a window of that length would carry weight, the six-factor specification of Table 8.5 returns an intercept that is negative in every one of the twelve cells, running from -49.7% to -66.9% annualised on the unrescaled book, against the $+12.29\%$ to $+19.95\%$ Table 8.5 reports at matched 15% risk. The unanimity of the sign across the twelve cells, its reversal against the full-sample fit, and the size of the residual against the drawdown itself, are what the reading rests on. Realised factor returns over the same window are $+13.5\%$ for the market, -9.3% for RMW, $+10.3\%$ for momentum and -8.1% for size, so the profitability unwind does cost the book something through the loading documented above, and rather more than the correlation test earlier in this section suggests; but market and momentum exposure both worked in its favour over the episode, and what is left once all six are charged is a residual larger than the drawdown itself. The loss is in the part of the book the factor model does not span.

Taken together these say what the episode was without saying why it happened. A signal that discriminated for twenty-two years stopped discriminating, in the same months, in twelve specifications that differ in loss function and in representation, among the largest and most liquid names in the universe, and not through any exposure the standard factor model prices. Crowding is the reading most consistent with that description, and it is precisely the one this design cannot test: every specification here is fitted on one public characteristic surface, so the overlap that would matter is overlap with books this thesis never observes. Short interest, holdings disclosures or a measured borrow curve would discriminate; none is in the panel. The competing reading, that the signal decayed as the literature documents for published anomalies, is equally consistent with these measurements and equally untested by them, and Section 9.5 treats both as open. What the evidence does close off is the reading that the drawdown was mechanical in origin, a liquidity accident or a factor tilt large enough to account for it: the profitability exposure carries a minority of the fall, the residual is where the ordering failure sits, and the ordering is what this thesis claims to have improved.

Depth does not order with the representation arm once the difference in risk is removed. On the volatility-scaled series the inverse-normal cell is shallower than its percentile twin for RankNet (-33.86% against -39.36%) and for LambdaRankIC (-31.32% against -40.02%), and deeper for the other three objectives. The arm-level depths of Table 8.4, measured at the risk each book actually ran, do rank the inverse-normal arm shallower gross and net, but that is the volatility channel of Section 5.3.2 showing through: a book that runs lower volatility throughout, as the inverse-normal arm does in every objective, will have shallower drawdowns for that reason alone, so the agreement corroborates the volatility channel rather than providing independent evidence for it. The second qualification stated in advance also holds: a drawdown depth is one realised extremum,

no standard error attaches to it, and no test is run on it.

The missing overlap diagnostic matters here. Correlated returns can in principle arise from books that are simply the same book, and overlap is what separates that from a shared environment; the correlation evidence is reported without it. Given that the specifications differ only in the representation and the loss, and that they select from a common thousand-name universe at a hundred names per side, high overlap should be expected, and the commonality finding should be read as consistent with rather than proof of a shared shock.

8.3 The legs under stress

Section 6.4 classified the short leg on full-sample evidence. This section asks whether that classification is stable, since a leg that diversifies on average and fails in drawdowns is not providing the protection the average suggests.

Three questions bear on it: whether the long leg's contribution is concentrated in a few regimes, whether the short leg's contribution is counter-cyclical or pro-cyclical, and whether the combined result is attributable to market exposure rather than to selection. A dollar-neutral book need not be beta-neutral, and if the market loading is materially non-zero then part of what Chapter 6 reports as selection is directional exposure.

All three are answerable from the frozen run, though not to the same standard. The factor regression is estimated on the combined series for every specification and is reported in Table 8.5. The first two are read from the monthly leg contributions, which the run stores per rebalance date, so Figure 8.3 traces each leg's cumulative path and the counts below describe their co-movement. No standard error attaches to either: both are descriptions of one realised history, and no test is run on them.

Over the full sample the legs behave as Section 6.4 reports, the short leg a drag in level and a diversifier in covariance. The combined series is not principally market exposure: regressed on the five Fama–French factors and momentum it earns an intercept of +18.17% a year ($t = 5.41$) with a market loading of +0.174 ($t = 3.13$) and an R^2 of 0.178, the dominant exposure being profitability (Table 8.5). The pattern is general rather than particular to that cell: across the twelve specifications of Table 8.5 the annualised intercept runs from +12.29% to +19.95% and is significant at better than the 0.1% level in every one, while the market loading never exceeds +0.178 and reaches conventional significance in only three, all of them in the inverse-normal arm. None of the remaining four loadings attains significance in any cell. The momentum loading is the one worth stating, and it is read from the same regressions rather than from the printed columns of Table 8.5, which carry only the intercept, the market and profitability loadings and the fit: it runs between -0.177 and $+0.087$ with a largest $|t|$ of 1.27, on a feature surface where momentum and trend account for thirty-nine of the 144 columns.

Table 8.5: Fama–French five-factor plus momentum regressions of the gross long-short monthly series, by objective and arm. Each series is rescaled to 15% annualised volatility before fitting, so the annualised intercept and the loadings are stated on that book. The regression carries all six factors; the table shows the two the chapter argues from, the market because a market-exposure reading turns on it and profitability because it is the largest. t -statistics and R^2 are invariant to the rescaling. Market exposure in disguise would show a large market loading and a null intercept; the pattern is the opposite. t -statistics in parentheses; these alphas are unadjusted and are not members of the exploratory family.

Objective	Arm	α p.a.	Mkt-RF	RMW	R^2
RankNet	Percentile	+12.29% (3.77)	+0.101 (1.55)	+0.708 (5.49)	0.127
	Inverse-normal	+18.17% (5.41)	+0.174 (3.13)	+0.895 (7.21)	0.178
LambdaRankIC	Percentile	+13.56% (4.01)	+0.091 (1.42)	+0.670 (5.18)	0.110
	Inverse-normal	+19.95% (6.01)	+0.178 (3.00)	+0.782 (6.84)	0.168
Margin ranking	Percentile	+13.93% (4.26)	+0.088 (1.37)	+0.805 (5.65)	0.168
	Inverse-normal	+17.45% (5.14)	+0.124 (2.18)	+0.889 (7.01)	0.194
Soft-Spearman	Percentile	+15.01% (4.54)	+0.012 (0.17)	+0.600 (4.46)	0.088
	Inverse-normal	+18.51% (6.12)	+0.050 (0.89)	+0.981 (7.19)	0.180
RankCosine	Percentile	+15.64% (4.65)	+0.020 (0.28)	+0.638 (4.77)	0.087
	Inverse-normal	+18.36% (5.80)	+0.050 (0.88)	+0.927 (7.01)	0.162
Gradient boosting	Percentile	+17.39% (5.43)	+0.084 (1.28)	+0.763 (6.29)	0.167
	Inverse-normal	+17.39% (5.43)	+0.084 (1.28)	+0.763 (6.29)	0.167

One column of Table 8.5 moves with the treatment. The profitability loading is larger in the inverse-normal arm than in the percentile arm in all five objectives, from +0.708 to +0.895 for RankNet and from +0.600 to +0.981 for Soft-Spearman, and the regression R^2 rises alongside it in all five; the gradient-boosted control, blind to the encoding, carries the same loading and the same R^2 in both of its rows. The arm reported here as the better one is therefore the more factor-spanned one. Each series is rescaled to a common 15% volatility before the regression is fitted, so this is exposure at matched risk and not a residue of the volatility difference of Table 5.3; the inverse-normal loading remains the larger in all five objectives once the rescaling is undone, even though undoing it favours the percentile book, which is the more volatile of the two.

The consequence for Section 5.3.2 runs against the obvious reading. Systematic and residual variance follow from the printed R^2 and the realised volatility of Table 5.3 as $R^2\hat{\sigma}^2$ and $(1 - R^2)\hat{\sigma}^2$, and taken that way the systematic part is the larger of the two arms in the inverse-normal cell in all five objectives while the residual part is the smaller in all five. The fall in book volatility that carries the scale step is a fall in the part of the book the factor model does not span, obtained alongside a slight increase in the part it does, which is not what a comparison of total volatilities on its own would suggest. Two limits apply. Table 8.5 is estimated for the two base arms only, so the exposure contrast cannot be divided between the steps of (5.1), and its attribution to the scale step rests

on the volatility channel and not on the factor evidence; and no test attaches to any of these differences, which are read as a pattern across cells in the manner of the rest of this section. The direction nonetheless qualifies Section 8.2 in a way worth stating: the arm this thesis prefers is the arm carrying more of the exposure to which the largest single factor attribution of the terminal episode is charged.

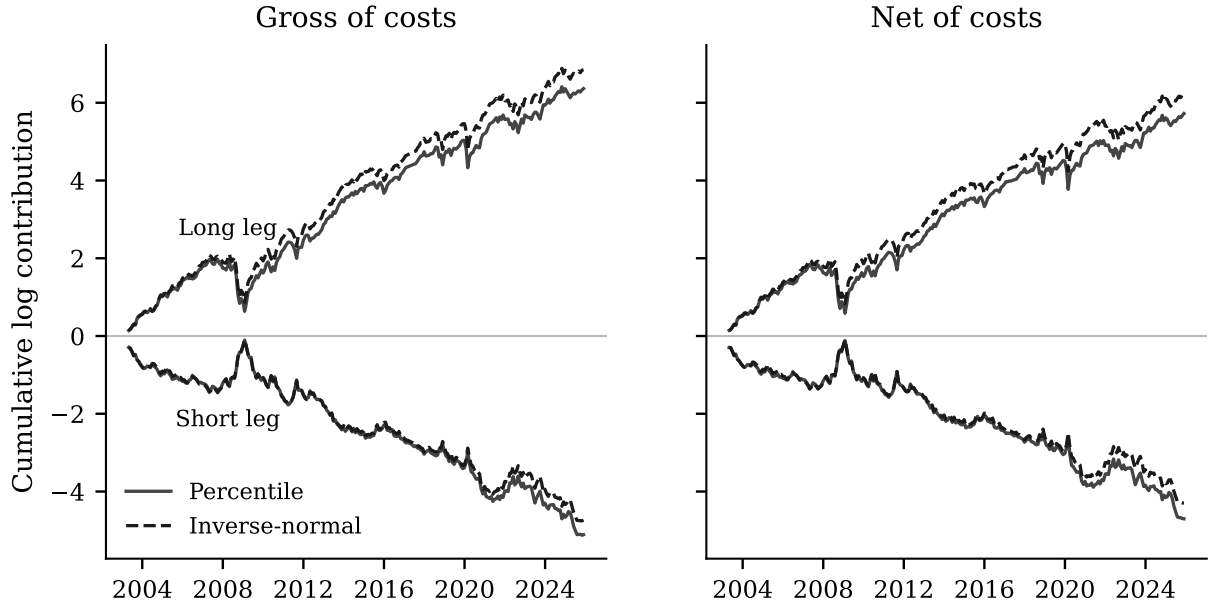


Figure 8.3: Cumulative log contribution of the long and short legs by arm, averaged across the five neural objectives, gross and net. Each series is the leg’s *additive* contribution, the form Table 6.3 reports; the standalone leg backtests there carry the full gross notional and are twice as large. The monthly contributions sum to the book of Figure 5.1, but their cumulative logs do not, so the two curves are not meant to be added. Panels share a vertical scale.

The two paths are close to mirror images. Pooled over the ten neural cells the long leg contributes +35.9% a year and the short leg −15.6%, summing to the +20.3% the combined book earns. That mirroring is what the diversification argument of Section 6.4 rests on, and Figure 8.3 shows it holding across the whole sample rather than only on average.

The second question resolves in favour of the counter-cyclical reading almost everywhere. In 79 of the 86 months in which the long leg loses, the short leg gains, so the offset operates in 92% of the months that call for it. Both legs lose together in 7 months out of 272, and those months are severe: across them the combined book loses at an annualised rate of −80.5%. The 2008 and 2009 window is the configuration the classification predicts, the long leg giving up 5.7% while the short leg earns 21.9%.

The legs are the two extreme bins of a sort, and Figure 8.4 shows the eight bins between them. The ordering is monotone in both arms and leaves no room for interpretation: the annualised mean rises at every one of the nine steps from D_1 to D_{10} , in the percentile arm from 7.75% to 17.62% and in the inverse-normal arm from 6.95% to 18.82%, and the rank correlation between bin number and bin mean is +1.000 in each. The signal is therefore

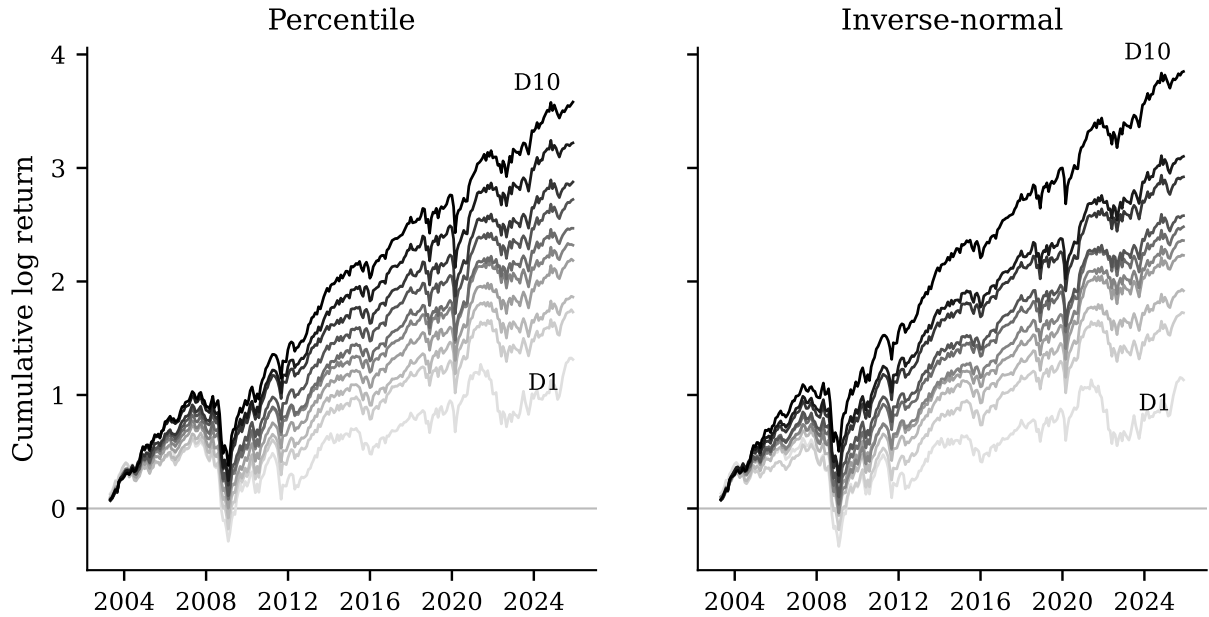


Figure 8.4: Cumulative log return of all ten bins of the sorted cross-section, by arm, averaged across the five neural objectives, gross. Scores are demeaned within sector before binning, as they are before selection, so the sort is the one the book trades; shading runs light to dark from D_1 to D_{10} . Each line is a long-only, equally weighted, unlevered bin, so all ten rise with the market and the levels are not comparable with Figure 5.1 or Figure 8.3. The construction reproduces the stored equally weighted book to within 0.008 in any month.

a gradient across the whole cross-section rather than an ability to isolate two unusual buckets, which is the reading under which the leg classification above makes sense. The spread from top bin to bottom is 9.87 percentage points a year under percentile ranks and 11.87 under inverse-normal ones, so the representation contrast of Chapter 5 is visible in the raw sort before any portfolio rule is applied to it.

The exception is the terminal episode, and it is the sharpest one. Over calendar 2025 the pooled short leg gave up 39.2% of contribution while the long leg gained 10.3%. The offset that holds in nine months out of ten failed in the one episode where it was most needed, so the classification of Section 6.4 holds on average and through the 2008 crisis, and does not hold at the end of the sample.

8.4 Alternative explanations

Six explanations would make the results of Chapter 5 mean something other than what they appear to mean. Three are addressed by the design; three are genuinely unresolved.

Shape confounded with scale. Measured rather than answered. (5.1) decomposes the contrast into its two components exactly, and the allocation Section 5.3 reports runs against the shape reading, the larger and better identified of the two steps at the full sample being the one that cannot reorder anything. That ordering is a property of the

terminal date rather than of the decomposition: Section 8.6 restrikes the contrast at 153 stopping dates and the scale step is the larger at twelve of them. The design identifies the split, and the split it identifies is the one struck at matched second moments, in the sense of Section 4.2.1; it does not dispose of the objection. It does now name the route, which Section 4.2.1 states and Section B.6 measures from the architecture alone. The scale channel is therefore not an unexplained residue, and it is also not the tail spacing the hypothesis was written about.

Selection among objectives and widths. Answered by construction. The concern would hold had the winner been reported, but Table 5.4 carries every objective, Table 6.1 every width, and Section 5.9 fixes the carry-forward rule before the estimates are inspected.

The hyperparameters favour one arm. The fixed-surface rule of Section 4.4 holds the architecture and optimiser settings identical across every cell and performs no per-cell tuning. This controls the contrast but does not establish that a tuned configuration would preserve the gap, since the comparison is conducted at settings optimised for neither arm. Consistency across five objectives mitigates the concern without removing it.

The setting with the sharpest version of this concern is the training length, where a fixed budget is a controlled operating point and not a neutral one (Section 4.4.2). The convergence-matched re-estimate of Section 5.3.3 is the answer: it repeats the contrast with each cell trained to its own validation optimum, so the conclusion does not rest on either operating point alone.

Data snooping. The architecture and its capacity were not searched. The settings of Table 4.3 were fixed once and are common to every cell. The representation contrast and its decomposition were estimated once, on the three arms of Section 4.2, and are not the surviving members of a wider comparison, and the primary endpoint, the inference tiers of Section 3.7, the width-checking rule of Section 6.5 and the interpretation of the tree control in Table 5.7 were all fixed before the estimates were read. None of this converts the study into a genuine out-of-sample test of a pre-registered hypothesis, and it is not presented as one.

The book is a profitability premium in disguise. Answered by the attribution itself. The exposure is real and it is the largest the factor model finds, but it accounts for some +2.2 percentage points of annual return against the intercept of Section 5.6. What the exposure does reach is the terminal drawdown, where Section 8.2 attributes about a third of the calendar-2025 fall, and nearer a sixth of the episode itself, to the unwind; the standing exposure and the episode are different questions and the answers differ. The

exposure also moves with the treatment, which Section 8.3 measures and locates in the systematic rather than the residual half of the variance.

Costs are assumptions, not observations. They are, and Section 7.2 states which components rest on assumptions the panel cannot verify. The per-component bridge of Table 7.1 states how much of the conclusion each assumption carries, and a reader who rejects the parameterisation should read the gross results and Table 7.3 rather than the net headline.

8.5 Initialisation noise and the seed dimension

The standard errors of Parts II and III carry no term for initialisation noise (Section 3.7). This section characterises that missing dimension from the retained per-seed score vectors; it stops short of the variance component that would enter the standard error.

Two readings of the contrast are available: that it is a property of the representation and would appear in any single fit, or that it is produced by averaging an ensemble the representation makes more diverse. The ensemble-size curve discriminates them, and it favours the second without eliminating the first. Recomputed over the $\binom{5}{m}$ subsets of m seeds within each objective and pooled across objectives, the mean Δ^{repr} rises monotonically with the number of members, from +0.192 at $m = 1$ to +0.370 at the reported $m = 5$ (Section C.3). The contrast is already positive at a single seed, so part of it is a property of the representation; but a single-seed study would have recovered +0.192 against the +0.370 reported, with a dispersion of 0.124 across seeds within objective, and would in all likelihood have missed the effect. Ensemble diversity is therefore part of the mechanism and not merely noise averaging: part of what the encoding delivers is a more diverse ensemble whose idiosyncratic errors cancel. Neither reading isolates the geometry of the transform, which is what the decomposition of Section 5.3 finds from a different direction, the mechanism there being at least partly the conditioning of the optimisation.

8.6 What this evidence does not establish

Four claims are outside what this design can support.

The study does not establish that the inverse-normal encoding is intrinsically superior as a representation for tabular learning. It measures one contrast, on one panel, under one architecture, with one target encoding, for one class of ranking objectives. A result that is consistent across five objectives on two decades of months is evidence; it is not a general property of the transform.

It does not establish that the selected specification will perform out of sample. Every estimate here is out of sample with respect to its own fold, but a fold structure cannot

price the choice of what to run, and the 2025 drawdown is a direct demonstration that the environment can change in ways the sample did not contain.

It does not establish a cost figure. It establishes net performance under a stated parameterisation. A different spread convention, a measured borrow curve, or a different impact calibration would produce different numbers, and the per-component bridge of Table 7.1 is what lets a reader substitute their own parameters and recompute the consequence.

And it does not establish that the allocation between the two steps of (5.1) is independent of where the sample stops. The evaluation ends fifteen months after the family book's last peak with the episode of Section 8.2 unrecovered, so the contrast is restruct on every terminal date from March 2013 forward, of which Table 8.6 reports three. The total barely moves: across the 153 stopping dates it runs from +0.367 to +0.467 gross and from +0.306 to +0.379 net, and its ratio to its own standard error never falls below 2.49 gross or 2.32 net, so the headline is not an artefact of the ending. The split is another matter. The scale step's share of the total runs from 0.27 to 0.63 gross and from 0.23 to 0.64 net, and gross it is the larger of the two at twelve of those dates: the eight from April 2025 onward, and the four earliest, where the sweep has barely ten years behind it and neither step is separated from the other. Its level moves far less than its share: across the same dates the scale step runs from +0.111 to +0.237 gross and from +0.075 to +0.203 net at the traded decile, so it never changes sign, which is the stability Section 5.3.1 reads as advice. At the 141 dates in between, from July 2013 to March 2025, the shape step is the larger gross and also the better resolved, +0.212 against +0.190 at the pre-episode peak with ratios of 3.56 and 1.83. The crossing sits inside the drawdown, and the episode says why: over the fifteen months from September 2024 the percentile arm lost 36.3%, the scale-matched arm 26.5% and the inverse-normal arm 30.5%, so the scale step earned its full-sample lead by falling less and the shape step gave part of its own back. Protection in a fall is what Section 5.3.2 would expect of a step that moves the volatility of the book rather than its mean, so the reading is consistent; but which half of the gap is the larger is a statement about one drawdown rather than a property of the decomposition.

Table 8.6: The primary contrast and the two steps of (5.1) restruck on the first T rebalances, pooled over the five neural objectives at the traded decile under inverse-volatility sizing. August 2024 is the family book’s last peak before the episode of Section 8.2 and is shown as a landmark, not as an alternative headline: it is located from the outcome, and the statement the section makes rests on the whole sweep. The figure in parentheses is the ratio of the estimate to its pooled HAC delta-method standard error, carried as a gauge of how sharply each step is resolved; following Section 3.7 no test is attached to a truncation. The break-even row restrikes the base-case net return of Table 7.3 and holds the impact charge at its full-sample value.

		Aug 2024 $T = 257$	Mar 2025 $T = 264$	Full sample $T = 272$
Gross	Δ^{repr}	+0.402 (4.27)	+0.394 (4.39)	+0.370 (4.46)
	Δ^{scale}	+0.190 (1.83)	+0.193 (1.94)	+0.224 (2.32)
	Δ^{shape}	+0.212 (3.56)	+0.201 (3.49)	+0.146 (2.09)
Net	Δ^{repr}	+0.334 (3.81)	+0.326 (3.87)	+0.307 (3.92)
	Δ^{scale}	+0.162 (1.64)	+0.162 (1.70)	+0.188 (2.03)
	Δ^{shape}	+0.171 (3.13)	+0.163 (3.08)	+0.119 (1.93)
Capital	Break-even	\$193 million	\$171 million	\$140 million

Chapter 9

Conclusion

9.1 Answers to the research questions

Section 1.3 posed six questions, ordered so that each presupposes an answer to the one before it. They are answered here in the same order and with the same scope restrictions, so that the boundaries of Section 1.5 apply to the answers as stated.

Q1. Does the numerical mapping of within-date ranks change portfolio performance? Yes, and materially: Hypothesis 1 is supported at the primary tier, in the direction it specified. The inverse-normal arm is ahead of the percentile arm in every one of the five objectives, gross (Table 5.3) and net (Table C.2). The primary endpoint, the pooled contrast of (5.1), is +0.370 gross and +0.307 net, both clearing the studentised bootstrap on intervals excluding zero (Tables 5.4 and 5.5, Section 5.3). As the single pre-specified contrast it carries no multiplicity adjustment. The secondary family is weaker than the pooled figure alone would suggest: under the Romano–Wolf stepdown only RankNet and LambdaRankIC are individually significant in both regimes (Section C.2).

Table 5.5 qualifies the answer, and this is where it is least comfortable. The decomposition of (5.1) allocates +0.224 of the gross contrast to the pure scale term, which cannot reorder anything, against +0.146 to the shape term at matched scale. The scale term is significant in both regimes; the shape term is marginal in both. Net, the scale interval clears zero and the shape interval does not. The two steps are separated from each other on dispersion but not on the means (Section 9.2, Table 5.6). That the mapping changes performance is established. That the carrier is the geometry of the transform rather than the conditioning of the optimisation is not, and the architecture, measured at initialisation, points the other way (Section 4.2.1, Section B.6): the step the returns identify as the larger one is the step that changes how much of the network’s nonlinearity is engaged.

That has a practical form, and it is the more useful statement of the result. The scale-

matched arm is the field’s default encoding divided by its own standard deviation, so the first step of (5.1) is standardisation and nothing more exotic. The defensible conclusion is therefore that the convention of presenting ranks on $[-\frac{1}{2}, \frac{1}{2}]$ is a mis-scaling rather than that ranks should be mapped through Φ^{-1} ; Section 5.3.1 sets out the evidence and Section 9.3 the recommendation. What the convergence-matched re-estimate settles is narrower, and it is worth stating precisely: whether the division between the two steps is an artefact of the fixed budget. Gross it is not. The gap between the two steps changes by +0.018 against a standard error of 0.139, and letting each cell train to its own optimum costs the scale term 0.049 against the shape term’s 0.031, which is not the margin the conditioning objection needs (Section 5.3.3). The pooled total does not clear 5% gross, at $p = 0.0585$, while retaining four fifths of its fixed-budget value. Net it falls further short (+0.191 against +0.307, $p = 0.1447$, on an interval covering zero), the estimate itself giving up nearly two fifths rather than only its precision, and the ordering of the two steps reverses; at three quarters of a standard error that is a reading this sample does not resolve, and Section 9.4 carries it.

Q2. Is the effect a property of one objective or of the representation? Of the representation, so far as five correlated objectives can establish it. The sign of Δ^{repr} favours the inverse-normal arm in all five objectives in both panels of Figure 5.2, which is what Hypothesis 2 predicts. Individual resolution is a second question and the two answers differ. Under the Romano–Wolf stepdown that adjusts the secondary family, three of the five contrasts clear 5% gross and two clear it in both cost regimes (Section C.2), so the hypothesis is supported on the stability of the sign and only partly on the resolution of its parts. That the first is the more informative of the two here is a judgement about what 272 months of one noisy cell can separate, and it was fixed with the hypotheses (Section 2.6, Section 3.7): a sign held by all five and a stepdown that only two survive are the same evidence read at two resolutions rather than a favourable and an unfavourable verdict. The magnitudes do not pattern by family: the pairwise family supplies both the largest contrast and one of the smallest, so the families interleave rather than separate and the five cells are closer to five readings of one quantity than to five replications.

The gradient-boosted control behaves as Hypothesis 3 and the argument of Section 4.5 require. Its representation contrast is 0.000 in both cost regimes, and since that cell shares the panel, the labels, the folds, the seeds and the portfolio rules and differs only in being blind to the encoding, the neural result cannot be an artefact of the data handling: the first row of Table 5.7 is the one returned. The linear anchors bound the reading from a third direction, and it is narrower than it appears. Both are invariant to a common rescaling of their inputs, so the scale channel reaches neither and a linear contrast between the arms isolates the shape channel alone (Section 4.7). On that channel the two anchors disagree, +0.056 for ordinary least squares against −0.186 for the elastic net, untested

by design: neither reproduces the neural shape step and the second runs against it, which tightens the limit Section 5.3.3 already places on the shape reading rather than corroborating the total. On the scale step the anchors are silent by construction, and nothing empirical is needed there; whatever that step does, it does to a fit and not to the features.

One further reading of the same cell belongs here rather than under Hypothesis 3, which concerns only the size of the tree’s own contrast. Fitted on the identical surface and selected fold by fold where the neural cells select nothing, the gradient-boosted ranker still sits above every percentile cell and below every inverse-normal one gross at the traded decile under inverse-volatility sizing (Section 5.6.3). The neural family is ahead of the strongest available comparator only in the arm this thesis argues for, and behind it in the arm the literature supplies by default. That is the model-class statement of the representation result, not a second finding. Four limits travel with it. The comparison carries no test; the tree is one configuration and no search was run over it; the net margin of +0.052 is not resolved at this sample length; and under equal weighting the control clears three of the five inverse-normal cells and the pooled margin changes sign.

Q3. Selection or sizing? Selection, not sizing. The representation contrast is positive under equal weighting at every traded fraction of Table 6.2, and at the headline decile it is +0.316 on a studentised bootstrap interval of $[+0.156, +0.476]$, with $p = 0.0005$ from the same joint bootstrap, which that table reports as an interval rather than as a p -value. That cell is the primary endpoint re-struck under the second weighting rule rather than a further test (Section 3.7), and an equally weighted book is a function of tail membership alone, so the effect cannot be an interaction with the inverse-volatility rule: it is a statement about which names the model selects. The contrast is larger under inverse-volatility sizing at every width, by +0.054 at the decile and by less elsewhere (Table 6.2, Section 6.3), which is consistent with the volatility channel of Section 5.3.2, but the weighting axis sits in the exploratory tier, the intervals at the two narrowest fractions include zero under both schemes, and none of its fifteen comparisons is significant under Benjamini–Hochberg at $q = 0.10$, so the two schemes are not separated by the evidence. The branch of Hypothesis 4 that would have made the finding a weaker one is closed by the re-struck endpoint; what the exploratory tier does not do is separate the two schemes from each other.

Q4. How does tail width affect the book? Concentration does not pay. Turnover rises monotonically as the fraction narrows and gross performance improves as the book widens over most of the range (Table 6.1); the two cost regimes disagree only at the wide end, gross favouring 10% under the inverse-volatility scheme while net the two schemes agree on 10% (Section 6.2). The headline fraction is therefore the interior peak

of the net profile once costs are charged, with the grid bracketing it on both sides rather than stopping at it, which is the stronger form of the claim, though the location of the maximum rests on the point estimates, the two contrasts bracketing the decile carrying t -ratios below 0.9. Drawdown is the exception. On the volatility-matched basis of Table 6.1 the deepest decline of the grid is at the decile itself and the narrowest books are the shallowest, so at a common risk target concentration bought a milder worst case while costing risk-adjusted return. No test attaches to a maximum drawdown here, and the observation qualifies the width recommendation without displacing it. The width axis nonetheless remains the one the evidence separates least: of the sixty width comparisons in the exploratory family, only one is significant under Benjamini–Hochberg at $q = 0.10$. The 1% column is the weakest cell of the gross grid under both schemes and is read as a concentration stress test rather than as a candidate operating point; no conclusion rests on it.

Q5. What retains economically meaningful performance after execution? C1 and C2 are supported, C3 is supported at a reduced level, and C4 is not. The scores order the cross-section better than chance in both arms (Table 5.2). Execution then removes a little over a third of what the gross book earns, the charged stack and the executable entry convention together (Table 7.1), a reduction larger than the treatment this thesis studies: all fifteen netting comparisons are significant under Benjamini–Hochberg at $q = 0.10$, against a representation effect of +0.370. Hypothesis 5 is an existence claim about the two maxima rather than a point prediction, and the grid bears it out on one scheme and not the other (Section 6.2). What the evidence supports is that the two maxima can part company, not a measurement of how far apart they sit. C4 is where the design stops: Table 7.3 traces the net ratio downwards as managed capital rises and locates a break-even that is modest by institutional standards, with the narrow widths saturating first. That degradation is what an unengineered implementation would have paid; Section 9.4 weighs how far a desk could improve on it and what the answer is not allowed to change.

Q6. What do the legs, the subperiods, and the worst drawdowns reveal? The record is weakest at the end, and by a wide margin: the two blocks before 2020 are all but identical and the 2020s fall away from them. Figure 8.1 carries that coarse split with its standard errors, and the finer regime partition of Table 8.2 places its weakest reading, gross and net, in the closing regime.

The deepest drawdown is a family-level event and it is unresolved. Every sweep specification reaches its maximum drawdown in 2025, including the gradient-boosted control, which is invariant to the treatment, and none has recovered by the last scored rebalance (Table 8.4); the mean pairwise correlation across specifications during the episode is essentially its full-sample level, so the commonality is the family’s ordinary condition

rather than a correlation spike. It is a short-leg episode, and the obvious explanation accounts for part of it and not the larger part: the profitability exposure that dominates the book's factor loadings is the largest single attribution the factor model supplies, close to a third of the calendar-2025 fall but nearer a sixth of the episode itself, and what the six factors leave is a residual the ordering failure sits in rather than an accounted remainder (Section 8.2, Section 8.3). Coverage is not the cause either. The treatments studied here neither caused the episode nor protected against it. Whether the terminal weakness is general to cross-sectional selection or specific to these models is not settled: the fixed momentum rule was not the weaker book over the closing regime of Table 8.3, which removes the most comfortable reading without supplying an explanation.

The short leg is a hedge and not a return source, and gross it earns the label: its standalone Sharpe ratio and its additive contribution to return are negative in every cell, and its correlation with the long leg runs between about -0.87 and -0.91 (Table 6.3), while the combined book gross outperforms the long leg alone in all twelve specifications. Net it does so in seven, and the split is not a wash even though the pooled figure is: the combined book wins every inverse-normal cell and both tree cells, the long leg wins exactly the five percentile cells (Section 6.4). The diversification is real, and whether it covers its own price depends on the encoding the book is built on; that the leg supplying it is also the leg that produced 2025 is the second half of the same finding.

9.2 What the criteria chain shows

The answers above rest on the primary endpoint and its decomposition; four findings sit across them rather than under any single question.

The criteria of (2.1) are not equivalent, and the shape of the non-equivalence is the study's most informative result. It is not that the ordering fails to improve. The *level* of ranking skill does not move, the paired coefficient contrast being indistinguishable from zero at $p = 0.4390$; but the coefficient becomes markedly steadier, its standard deviation falling by 16.0% with all five objectives rejecting equality of variances individually, and its information ratio rises from 0.361 to 0.452 ($+0.0911$, $t = 3.77$, $p = 0.0005$ unadjusted and 0.0057 after the correction applied to the exploratory tier it belongs to, one of only two ordering members that correction leaves standing and the one ordering statistic the representation itself moves). Measured the same way the portfolio is measured, as a mean over a standard deviation, the ordering improves by 25% against the Sharpe ratio's 34%. The apparent divergence between a strong portfolio result and a null ranking result was largely an artefact of comparing a risk-adjusted statistic with a raw one.

What the two steps of (5.1) do has the shape of a double dissociation, though its two halves are not equally well resolved. Rescaling the inputs steadies the coefficient and lowers the volatility of the book without moving either mean; restretching the tails at

matched scale does the reverse and leaves both dispersions untouched, but only its return leg is firmly resolved, and the gain in membership accuracy at the traded fraction does not survive the exploratory correction and is recorded in Chapter 5 as suggestive (Table 5.6, Section 5.3.2). Only the dispersion half is established as a dissociation. Tested directly, the two steps differ in what they do to the steadiness of the coefficient ($p < 0.001$) and, marginally, to the volatility of the book ($p = 0.084$); on the two means they are not distinguished, the gap between their return effects standing at +0.27 percentage points against a standard error of 2.03. Measuring each absence against zero is not the same as separating the two steps from one another, which only the last row attempts. These are in any case the attribution tests of Section 3.7, and the pattern describes where the effect went rather than adding a confirmed result to the ones above.

One comparison resists every reconciliation, and it does not involve the treatment. The gradient-boosted control has the highest mean rank information coefficient of any cell and, at the traded decile under inverse-volatility sizing, a gross Sharpe ratio below every inverse-normal neural cell, because the dispersion of its coefficient is a percentile-arm figure: a learner invariant to the encoding never receives the stability the encoding confers. That is the cleanest instance of (2.1) the study produces, and it is the one a reader should take away: the criterion the applied literature selects on ranks that specification first, and the portfolio ranks it below seven others at that operating point.

The consequence for model choice is the one this thesis is in a position to state. On this surface a neural ranker is worth its complexity over a gradient-boosted one under the inverse-normal encoding and is not worth it under the percentile encoding, so a study run on the field's default spacing would have reached the opposite conclusion about the architecture while changing nothing about the data, the folds or the portfolio rules. The statement is descriptive, it rests on one tree configuration, it is struck at the traded decile under inverse-volatility sizing and does not hold across the grid, and net of costs the two families are not separated; what it establishes is that a comparison between model classes on this kind of panel is not identified until the encoding, the traded fraction and the weighting rule are held fixed and stated.

One juxtaposition is uncomfortable and worth stating last. Bottom-decile precision exceeds top-decile precision by +0.0411 with $t = 7.55$, so the models' sharpest ordering skill sits in the leg that bears borrow, that contributes negative return, and that produced the drawdown of Section 8.2.

9.3 Implications for practice

Selection metrics. A model selected on the rank information coefficient alone is selected on a criterion weighting the eight deciles never traded as heavily as the two that are, so the tail diagnostics of Chapter 5 belong beside it rather than in place of it.

Standardise the ranks. A centred percentile rank on $[-\frac{1}{2}, \frac{1}{2}]$ has standard deviation $1/\sqrt{12}$, and at that scale the first hidden block of a network of this kind operates, at initialisation, within a few percent of an affine map, so the layer that reads the features does almost no nonlinear work there (Section B.6). Dividing the ranks by their own standard deviation, which for a surface of ranks is one analytic constant with nothing estimated and nothing to leak, is worth three fifths of the representation contrast at the traded decile (Section 5.3.1). Its share of the gap is not constant across the grid or across stopping dates, running from roughly 0.25 to 0.65 as the terminal date moves, though its level stays positive at every one of them (Section 8.6), and fourteen of the hundred individual cells run the other way, all of them in the two listwise objectives and mostly at the narrow fractions (Section 5.3.1). The further step to an inverse-normal encoding adds +0.146 gross and +0.119 net with an interval covering zero, so it is the part of the recommendation this study cannot underwrite. A practitioner who takes one thing from this thesis should take the scale and treat the shape as an open question.

The representation axis. The mapping applied to within-date ranks is a modelling decision with a testable consequence and not a convention to be inherited unexamined, and varying it along an ordered axis instead of bundling it with other changes is the difference between a treatment and a confound. Reporting the interval is part of that: a study that says it used centred ranks without saying which interval it means has left its network’s operating point unstated, and on this evidence that is not a detail.

Comparisons between model classes. The same point has a consequence for how a neural model is set against a tree (Section 9.2). A study that reports one model class as preferable without stating the spacing its features were mapped onto has not, on the evidence here, reported enough for the comparison to be reproduced, and the encoding belongs in the specification of such a study alongside the architecture and the feature set. This study’s own comparison shows why the warning is not academic: the neural margin over the gradient-boosted control changes sign between the two weighting schemes, so a single quoted margin would have reported either verdict depending on a choice unrelated to the model class. Whether the sign moves that way on other panels is untested; that it moved on this one is enough to make the omission worth avoiding.

Two weighting schemes. Running equal weighting alongside inverse-volatility weighting at a common gross exposure is the only way to separate a claim about which stocks a model picks from a claim about how they are sized; a comparison conducted under a risk-based sizing rule alone cannot support the first.

Implementation reporting. A single net figure at a single capital base conceals which cost component binds and at what size the strategy saturates; the gross-to-net decomposition of Table 7.1 and the capacity curve of Table 7.3 cost one replay each once scores are frozen. The ordering of the components matters more than their total. The largest is the cost of crossing, on a book that trades a multiple of its own net asset value every year, and the second is the borrow charged on the short leg. Impact, the component practitioners cite first, ranks third of the four at base-case capital at the traded decile, above only the financing spread, which is a standing charge on the debit balance that no improvement in modelling can touch; and impact is the only one of the four that grows with the size of the book, already the largest of them at the narrowest fraction of the grid. Execution engineering is therefore not a downstream detail of this design but a first-order determinant of whether it delivers anything. One warning stands regardless of the numbers: the crossing charge combines the measured quoted spread with a configured 2 basis-point commission, and a quoted spread is a lower bound on the effective cost of crossing rather than a measurement of it, so the gap between the two remains an assumption about execution quality and not a property of the panel.

Short exposure. The short leg is a variance hedge rather than a return source (Section 6.4). Borrow, at 152 basis points a year, is therefore a fee paid for variance reduction and not a charge against an alpha stream, and the question a desk should ask is whether the reduction is worth the fee. The answer turns on the encoding, and Section 6.4 gives the split. The long leg is in any case a long-biased holding carrying a market beta of essentially one that a neutral mandate could not run. The episode of Section 8.2 sharpens the question: the drawdown common to every sweep specification is a short-leg event, so the leg supplying the diversification is also the leg that can lose more in a year than the book earns in several, and neither the representation nor the objective offered protection, since the invariant gradient-boosted control participated on the same scale. Short-side position caps, borrow-aware sizing, and an explicit limit on short exposure are the obvious candidates; this thesis supplies the motivation for them and no evidence at all on their effect.

9.4 Limitations

Section 8.6 states what the evidence does not establish. This section states the design features that constrain it, in the order in which they bind.

One market and one panel. The study covers the liquid large-capitalisation segment of one national equity market over one history. The liquidity screens of Section 3.3 are what make the cost model credible, and they simultaneously remove the segment in which many documented anomalies concentrate. The results are statements about the segment

in which a long-short book of the size considered can actually be traded, and they do not transfer to the unfiltered cross-section.

One walk-forward history. All 272 out-of-sample observations come from a single expanding schedule on a single panel. Adjacent Sharpe ratios cannot be treated as sharply distinguished population parameters at that sample size, which is why Section 3.7 reads the descriptive and exploratory tiers for direction, economic size, and consistency rather than for significance alone, and why a single favourable comparison is not presented as evidence of superiority.

The headline is an ensemble quantity, and no interval prices the ensemble. Every neural figure averages five seeds, and roughly half the contrast is delivered by that averaging: a single seed recovers +0.192 against the reported +0.370, with a standard deviation of 0.124 across seeds within an objective and two of the twenty-five individual seeds negative (Section 8.5, Section C.3). No standard error in this thesis carries an initialisation-noise term (Section 3.7), so every interval above understates total uncertainty by whatever that component is worth. The curve has not flattened by the fifth seed and both arms gain under averaging, so the direction of the omission is toward understating the contrast rather than manufacturing it. What the design does not supply is a per-step seed dispersion, so the screen of Section 5.7, that an effect smaller than the spread across seeds is not a finding, is applied to the total and never to its two halves, which is the same limit the split reaches from a different direction below.

The cross-cell return and drawdown comparisons use a look-ahead scalar. Wherever returns or drawdowns are set against one another across cells they are first rescaled to a common 15% annualised volatility, and the scalar that does it is computed on the full sample (Section 5.1). Those figures are comparison devices rather than results anyone could have recorded, the width-grid drawdown ordering read under Q4 included. The leg decomposition of Q6 is not among them: it is struck on the raw series, which is what lets its two contributions sum to the book's own return exactly. Every Sharpe ratio, contrast and standard error in this thesis is computed on the raw unscaled series and is untouched by the convention.

Common conservative settings, not tuned ones. The architecture and optimiser settings are standard values held fixed across every cell, so no cell is reported at its best attainable configuration; Section 8.4 weighs what that control does and does not establish.

One target encoding. The label is presented to every neural cell on the centred percentile encoding, so the treatment runs through the feature channel alone. For the three pairwise objectives that holds by construction, two of them reading the target only through the direction of each comparison and the third recomputing an integer rank inside its own loss; for Soft-Spearman and RankCosine it does not, since both consume the encoded target numerically (Section 4.3.1). Those are also the two objectives in which the scale step reverses sign at the narrow fractions (Section 5.3.1), so in those cells the

design cannot separate a target-encoding interaction from a representation effect.

The choice of what to study is not priced. The three arms, the five objectives and the panel were set at the outset rather than selected on results, but no walk-forward structure prices the act of choosing them; the pre-specified endpoints and the fixed carry-forward and width-checking rules are what constrain the inference that follows (Section 8.4).

Costs are partly measured and partly a parameterisation. Section 7.2 states which components are measured and which are assumed, and the per-component bridge of Table 7.1 states exactly how much of the result rests on each.

Execution is deliberately unengineered. The realistic regime prices a plain set of rules. Every share crosses the quote at its full half-spread plus commission, each rebalance completes within one day's volume, and gross leverage sits at a fixed ceiling that binds at every rebalance, so the financing spread is a standing charge on the debit balance it fixes: about one hundred basis points a year, an eighth of the total drag in Table 7.1, and Section 7.2 shows that a lower ceiling would report a higher net Sharpe ratio at every point of the range, so this rule too errs against the results reported here. None of these rules was chosen to minimise cost. A desk would work part of its orders passively, schedule trades against the signal's decay, and set leverage against its own funding terms rather than a ceiling. The impact prefactor is the exception: $\kappa = 0.5$ reproduces the costs Frazzini et al. (2018) record at the participation rates they report, so that component is calibrated to live institutional execution and carries no margin of the same kind, which is what Section 7.4 turns on. The assumptions leaning the other way are fewer, and they are stated where they arise: borrow is charged at a liquidity proxy that misses episodic spikes, every short is assumed borrowable, and slower execution buys its lower charges with decay and tracking error that none of these figures price. On balance the net levels are a conservative reading of a stated stack rather than a forecast of an engineered implementation, and the capacity figures are conservative through the trading schedule alone. The conservatism attaches to the levels alone: both arms pay the same rules, so the representation contrast is measured under the stack rather than assumed away by it, and a cheaper execution would narrow the gross-to-net reductions reported above while leaving the effect standing.

The result does not reach institutional scale. The net figures above are struck at \$1 million of managed capital. Break-even arrives at \$140 million at the traded decile (Table 7.3), and at \$193 million struck at the family book's last peak before the episode, so C4 of Section 7.5 is the one criterion in the chain the evidence does not satisfy. Break-even scales as $1/\kappa^2$, so a reader who prices impact differently can rescale it; nothing in the design moves it into the range an institutional mandate would require.

Short-sale feasibility is assumed. Every short position is assumed borrowable at the modelled rate. The universe is restricted to liquid large capitalisations partly to make that assumption defensible, but the panel contains no lending-market data with which to

verify it, and episodic recall or borrow spikes are outside what can be measured here.

Dependence across observations. The label windows partition the sample, so consecutive observations do not overlap, but the return series are serially dependent and the specifications compared share dates, universe, and environment. The paired inference and the block bootstrap of Section 3.7 address the dependence; they do not create independent observations.

A common failure the sample cannot explain. Every sweep specification reaches its maximum drawdown in 2025 and none has recovered by the end of the sample, including the gradient-boosted control, which is invariant to the treatment and therefore cannot be failing for any reason varied here. The family shares a vulnerability the rest of the sample did not reveal. Section 8.2 establishes that the ordering itself inverted, upstream of every construction choice studied here, and rejects the leading mechanical readings (a coverage artefact, a short squeeze and a factor rotation), while the profitability unwind, the one exposure that does reach the episode, is priced rather than rejected and accounts for a minority of it; Section 8.1 rules out a general withdrawal of cross-sectional dispersion. None of this supplies a cause. No result reported above should be read as evidence that the vulnerability has been understood.

One terminal date, one training budget, and a split that is sensitive to both. The total contrast is close to invariant to where the sample is stopped, holding between +0.367 and +0.467 gross across every terminal date from March 2013 forward. Its division between the two steps of (5.1) is not: at no terminal date between July 2013 and March 2025 is the scale step the larger, and the lead it holds at the full sample is earned inside the drawdown, where the scale-matched arm fell almost ten points less than the percentile one (Section 8.6). The magnitudes reported above are therefore conditional on an ending that sits in the middle of an unrecovered episode, and the ordering of the two halves more so than the total. The same ordering fails a second time under the convergence-matched schedule of Section 5.3.3, where net the scale step gives up its whole lead and the shape step is the larger. What survives both cuts is the level of the first step, positive at every stopping date and in every width-and-weighting cell under both cost conventions; what neither supports is the rank of the two.

The convergence-matched control may be measuring its own stopping rule. Letting each cell train to its own validation optimum costs the contrast 0.080 gross and 0.116 net (Table 5.5), and the loss is not shared between the arms. The percentile arm is unmoved gross, 1.102 against 1.101, and slightly better net, 0.625 against 0.643; the inverse-normal arm gives up 1.471 to 1.391 gross and 0.932 to 0.833 net. The whole of the shrinkage is one arm losing rather than the other catching up, which is what a schedule effect would not look like and a measurement artefact might. The suspect is the stopping statistic. Each cell stops at the epoch maximising the validation rank information coefficient over the full cross-section, and Section 5.2 establishes that this is precisely the statistic the

representation does *not* move: the mean coefficient is indistinguishable between the arms at $p = 0.4390$, while what the encoding changes is the steadiness of the coefficient and the precision of the traded tails. A rule that stops on a quantity the treatment leaves alone has no reason to stop either arm where its traded book is best, and it will misplace the arm whose advantage is concentrated where the rule is not looking. On that reading the convergence-matched figures bound the contrast under one particular stopping rule rather than under free training, and the re-estimate is weaker evidence against the fixed-budget result than its face value suggests. It is also a testable claim this thesis does not test: stopping on the information ratio, or on tail precision, would separate the two readings, and neither was run.

One benchmark is a floor and one is a comparator, and neither is a contest. The momentum rule of Section 4.6 establishes that the reported performance is not a momentum premium in disguise, which the momentum loading of -0.033 ($t = -0.30$) and the -0.090 correlation between the benchmark book and the neural composite in Table 5.8 both confirm; its level is a floor for the reasons given in Section 5.6. The gradient-boosted ranker is a genuine comparator on the identical surface and it is the one the neural family is measured against, but the comparison is descriptive by pre-specification, it rests on a single configuration that received no search, and its margin after costs is inside the resolution of this sample. What it supports is that the neural advantage over that comparator is delivered by the encoding and not by the architecture; what it does not support is a general preference between the two families.

9.5 Future research

The extensions below follow from the limitations above.

The most valuable is also the cheapest: extending the sample forward. The terminal deterioration sits at the end of the history, the least informative position for an event of that kind, since there is little post-event data with which to distinguish a regime change from an ordinary adverse stretch. Re-estimating on data that accumulates after it would separate the two readings that Section 8.1 cannot.

The representation axis can then be extended beyond the three points studied here. The design compares a uniform spacing against a normal one and separates their scale from their shape, but none of the three arms identifies where within the tail the effect is generated. A family of transforms indexed by a continuous tail-emphasis parameter, of which the arms studied here are interior points, would convert the contrast into a response curve and locate the mechanism more precisely than a difference can.

The scale channel supports a sharper test than that, and a cheap one. The account Section 4.2.1 gives, and Section B.6 measures, makes two predictions the present design does not test. Re-running the arms under a positively homogeneous activation, for which

the normalisation absorbs a rescaling up to the first-layer bias, should collapse the scale step while leaving the shape step standing. Re-running them with the normalisation moved ahead of the activation should do the same. Either one costs a single sweep, neither touches the panel or the portfolio rules, and a result in either direction is informative: a scale step that survives both would mean the account in Section 4.2.1 is wrong, and one that disappears under both would establish that the field’s default encoding is mis-scaled for this class of network rather than mis-shaped, which is a more useful thing to be able to tell a practitioner than the contrast on its own.

Cost-aware estimation is the natural sequel to Chapter 7, none of the models here having been given the opportunity to prefer a cheaper ordering (Section 7.1). Two routes are open and they are not the same thing. The loss can carry a turnover penalty, which acts on how much the book churns without altering what it finds attractive; or the label of (3.1) can be struck net, charging each name the spread and, on the short side, the borrow it would incur, which changes the ordering itself by demoting names that are expensive to hold rather than merely expensive to reach. The second is the more direct expression of the goal and the more awkward to specify, since only part of the stack is a property of a name: spread and borrow attach to the security, while impact depends on participation and turnover on the transition between two books, neither of which is defined for a single forward return. Whether either route recovers the performance that execution removes is a question this design poses and does not answer.

Execution modelling is its counterpart, and on the evidence of Section 9.3 the larger of the two prizes. Cost-aware estimation asks the model to want a cheaper book; execution modelling asks what the same book would actually cost to trade. The present cost stack prices a deliberately plain execution: every share crosses a quote at the rebalance close, with no passive fill and no scheduling against the signal’s decay (Section 4.10). That is conservative in what it forgoes and optimistic in what it charges, since the quoted half-spread it applies is itself a lower bound on the effective spread (Section 7.2), so the net figures are a mid-range estimate rather than a bound in either direction. Execution is also the one layer of the problem with a mature learning literature of its own, reaching back to the reinforcement-learning order-placement policies of Nevmyvaka et al. (2006); that machinery optimises the trading of a given book and is deliberately outside the scope of this design. A replay in which the crossing charge is a function of participation, urgency and order type rather than a constant would not change a single fitted model, since the scores are frozen, and it would establish how much of the drag attributed to spread is a property of the strategy rather than of the execution assumed for it.

Borrow can be measured instead of proxied. The current specification keys borrow on liquidity, which captures the direction of scarcity but not its episodic behaviour, and episodic behaviour is what matters for a crowded short book during a drawdown.

The commonality documented in Section 8.2 raises the last and least tractable ques-

tion. Specifications that differ in loss, in representation, and in fitted scores may nonetheless hold overlapping books and fail together, so diversification across ranking objectives may be worth considerably less than the count of objectives suggests, and the overlap plausibly does not stop at the boundary of this study, since the characteristic surface used above is the public one, assembled from definitions the literature has standardised (Green et al., 2017; Gu et al., 2020). Crowding of that kind is invisible to every diagnostic in this thesis and would present exactly as Section 8.2 documents. Testing it needs data this study does not have; reducing the exposure needs a feature surface that is not the standard one, and an extension built on such a surface would separate compensation for a priced risk from queue position in a trade that many others hold.

The architecture is the least explored axis. Every cell applies one fitted function to every date, so a single ordering must serve environments as different as the 2008 dislocation and the terminal concentration regime, across which Section 8.1 shows performance differing by more than any treatment studied here. A mixture-of-experts formulation is the natural response inside the learning-to-rank framework: several ranking experts fitted on the same surface, combined by a gate that routes on observable market state. Whether regime-conditional routing recovers any of the terminal deterioration, or merely relabels it once the regime is already visible, is what such a design would settle.

Bibliography

- Almgren, R. and Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2):5–39.
- Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56.
- Ang, A., Chen, J., and Xing, Y. (2006). Downside risk. *Review of Financial Studies*, 19(4):1191–1239.
- Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.
- Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, 99(2):427–446.
- Barroso, P. and Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics*, 116(1):111–120.
- Baz, J., Granger, N., Harvey, C. R., Le Roux, N., and Rattray, S. (2015). Dissecting investment strategies in the cross section and time series. Working Paper 2695101, SSRN.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.
- Blitz, D., Huij, J., and Martens, M. (2011). Residual momentum. *Journal of Empirical Finance*, 18(3):506–521.
- Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. (2020). Fast differentiable sorting and ranking. In *Proceedings of the 37th International Conference on Machine Learning*, pages 950–959.
- Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M. (2018). *Trades, Quotes and Prices: Financial Markets Under the Microscope*. Cambridge University Press, Cambridge.

- Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96.
- Burges, C. J. C. (2010). From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research.
- Burges, C. J. C., Ragno, R., and Le, Q. V. (2006). Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19, pages 193–200.
- Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82.
- Chen, A. Y. and Velikov, M. (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, 58(3):968–1004.
- Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Coqueret, G. and Guida, T. (2020). *Machine Learning for Factor Investing*. Chapman and Hall/CRC, Boca Raton, FL.
- Daniel, K. and Moskowitz, T. J. (2016). Momentum crashes. *Journal of Financial Economics*, 122(2):221–247.
- DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953.
- DeMiguel, V., Martin-Utrera, A., Nogales, F. J., and Uppal, R. (2020). A transaction-cost perspective on the multitude of firm characteristics. *Review of Financial Studies*, 33(5):2180–2222.
- Detzel, A., Novy-Marx, R., and Velikov, M. (2023). Model comparison with transaction costs. *Journal of Finance*, 78(3):1743–1775.
- Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

- Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.
- Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.
- Frazzini, A., Israel, R., and Moskowitz, T. J. (2018). Trading costs. Technical report, SSRN Working Paper No. 3229719.
- Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969.
- Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943.
- Green, J., Hand, J. R. M., and Zhang, X. F. (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *Review of Financial Studies*, 30(12):4389–4436.
- Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, New York, 2nd edition.
- Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.
- Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.
- Guijarro-Ordóñez, J., Pelger, M., and Zanotti, G. (2025). Deep learning statistical arbitrage. *Management Science*. Forthcoming (Articles in Advance).
- Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034.

- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.
- Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133.
- Hribar, P. and Collins, D. W. (2002). Errors in estimating accruals: Implications for empirical research. *Journal of Accounting Research*, 40(1):105–134.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456.
- Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.
- Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91.
- Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518.
- Jensen, T. I., Kelly, B. T., Malamud, S., and Pedersen, L. H. (2026). Machine learning and the implementable efficient frontier. *Review of Financial Studies*. Advance article, published online 15 March 2026.
- Jobson, J. D. and Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36(4):889–908.
- Kelly, B., Malamud, S., and Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, 79(1):459–503.
- Kelly, B. and Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3–4):205–363.
- Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.
- Korajczyk, R. A. and Sadka, R. (2004). Are momentum profits robust to trading costs? *Journal of Finance*, 59(3):1039–1082.
- Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.
- LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–48. Springer, 2nd edition.

- Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.
- Lin, Y., Su, Y., and Yang, Y. (2026a). LambdaRankIC: Directly optimizing rank IC for financial prediction. arXiv:2605.00501 [cs.LG].
- Lin, Y., Su, Y., and Zhu, Z. (2026b). Empirical asset pricing via learning-to-rank. Technical report, SSRN Working Paper No. 6348379. Available at SSRN: <https://ssrn.com/abstract=6348379>.
- Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.
- Lo, A. W. (2002). The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52.
- López de Prado, M. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, NJ.
- Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.
- McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32.
- Memmel, C. (2003). Performance hypothesis testing with the Sharpe ratio. *Finance Letters*, 1:21–23. Global EcoFinance, Edinburgh.
- Mertens, E. (2002). Comments on the correct variance of estimated Sharpe ratios in Lo (2002, FAJ) when returns are IID. Technical report, University of Basel. Unpublished note.
- Moreira, A. and Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance*, 72(4):1611–1644.
- Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, 31(1/2):13–19.
- Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653.

- Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, 103(3):429–453.
- Novy-Marx, R. and Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147.
- Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31(1/2):9–12.
- Poh, D., Lim, B., Zohren, S., and Roberts, S. (2021). Building cross-sectional systematic strategies by learning to rank. *Journal of Financial Data Science*, 3(2):70–86.
- Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 263–270. Wiley, New York.
- Qin, T., Liu, T.-Y., and Li, H. (2010). A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*, 13(4):375–397.
- Qin, T., Zhang, X.-D., Tsai, M.-F., Wang, D.-S., Liu, T.-Y., and Li, H. (2008). Query-level loss functions for information retrieval. *Information Processing and Management*, 44(2):838–855.
- Romano, J. P. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.
- Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.
- Romano, J. P. and Wolf, M. (2016). Efficient computation of adjusted p -values for resampling-based stepdown multiple testing. *Statistics and Probability Letters*, 113:38–40.
- Sato, Y. and Kanazawa, K. (2025). Strict universality of the square-root law in price impact across stocks: A complete survey of the Tokyo stock exchange. *Physical Review Letters*, 135(25):257401.
- Shumway, T. (1997). The delisting bias in CRSP data. *Journal of Finance*, 52(1):327–340.
- Simon, F., Weibels, S., and Zimmermann, T. (2025). Deep parametric portfolio policies. CFR Working Paper 23-01, Centre for Financial Research (CFR), University of Cologne. Revised February 2025; also SSRN Working Paper No. 4150292.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

- Tóth, B., Lempérière, Y., Deremble, C., de Lataillade, J., Kockelkoren, J., and Bouchaud, J.-P. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006.
- Wang, X., Li, C., Golbandi, N., Bendersky, M., and Najork, M. (2018). The LambdaLoss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1313–1322.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Appendix A

Feature Surface

Section 3.5 describes the families of the characteristic surface and the rules that govern its construction. This appendix carries the inventory, the construction of the derived columns, and the measured completeness of both.

Table A.1 lists every one of the 144 characteristics with its definition, grouped by the families of Table 3.2. The listing is generated directly from the column manifest of a completed run rather than transcribed, and the generating script fails if the manifest and the inventory disagree in either direction, so a characteristic cannot enter the model without appearing here and no characteristic is described here that the model does not receive.

Table A.1: Complete inventory of the 144 admitted characteristics, grouped by family. Windows are in trading days unless stated otherwise. Every characteristic is mapped to a within-date rank before the model sees it, under the encoding of Section 4.2; the two seasonality columns are the sole exception and are noted as such.

Characteristic	Definition
Total momentum (8 columns)	
ret_63	Cumulative log return from 63 trading days before the rebalance to 21 days before it, $\log(P_{\tau-21}/P_{\tau-63})$: the strict 3-1 window of Jegadeesh and Titman (1993), skipping the reversal month.
ret_126	Cumulative log return from 126 trading days before the rebalance to 21 days before it, $\log(P_{\tau-21}/P_{\tau-126})$: the strict 6-1 window of Jegadeesh and Titman (1993), skipping the reversal month.

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>ret_252</code>	Cumulative log return from 252 trading days before the rebalance to 21 days before it, $\log(P_{\tau-21}/P_{\tau-252})$: the strict 12-1 window of Jegadeesh and Titman (1993), skipping the reversal month.
<code>nret_63</code>	<code>ret_63</code> divided by $\hat{\sigma}\sqrt{42}$, where $\hat{\sigma}$ is the 63-day exponentially weighted daily return volatility and 42 is the formation length.
<code>nret_126</code>	<code>ret_126</code> divided by $\hat{\sigma}\sqrt{105}$, where $\hat{\sigma}$ is the 63-day exponentially weighted daily return volatility and 105 is the formation length.
<code>nret_252</code>	<code>ret_252</code> divided by $\hat{\sigma}\sqrt{231}$, where $\hat{\sigma}$ is the 63-day exponentially weighted daily return volatility and 231 is the formation length.
<code>ret_63_252_spread</code>	<code>ret_63</code> minus <code>ret_252</code> : the slope of the momentum term structure, positive when recent momentum exceeds the annual measure. In logs this telescopes to minus the return over months 3–12.
<code>ret_126_252_spread</code>	<code>ret_126</code> minus <code>ret_252</code> , the same slope measured at the six-month point; in logs, minus the intermediate-horizon return over months 6–12.
Residual momentum (6 columns)	
<code>resid_ret_63</code>	Compounded daily residuals from 63 trading days before the rebalance to 21 days before it (the same strict 3-1 window as <code>ret_63</code> , so short-term reversal cannot contaminate the window), from a rolling 252-day regression of daily excess returns on the five Fama-French factors, the momentum factor, and the equal-weighted return of the firm's industry peers.
<code>resid_ret_126</code>	Compounded daily residuals from 126 trading days before the rebalance to 21 days before it (the same strict 6-1 window as <code>ret_126</code> , so short-term reversal cannot contaminate the window), from a rolling 252-day regression of daily excess returns on the five Fama-French factors, the momentum factor, and the equal-weighted return of the firm's industry peers.

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>resid_ret_252</code>	Compounded daily residuals from 252 trading days before the rebalance to 21 days before it (the same strict 12-1 window as <code>ret_252</code> , so short-term reversal cannot contaminate the window), from a rolling 252-day regression of daily excess returns on the five Fama-French factors, the momentum factor, and the equal-weighted return of the firm's industry peers.
<code>resid_nret_63</code>	<code>resid_ret_63</code> scaled by the trailing residual volatility times $\sqrt{42}$, the formation length.
<code>resid_nret_126</code>	<code>resid_ret_126</code> scaled by the trailing residual volatility times $\sqrt{105}$, the formation length.
<code>resid_nret_252</code>	<code>resid_ret_252</code> scaled by the trailing residual volatility times $\sqrt{231}$, the formation length.
Trend (MACD and moving averages) (25 columns)	
<code>macd_S8L24_lag0</code>	Moving-average convergence-divergence signal for the span pair (8, 24), normalised, observed at the rebalance date.
<code>macd_S8L24_lag21</code>	Moving-average convergence-divergence signal for the span pair (8, 24), normalised and observed 21 trading days before the rebalance.
<code>macd_S8L24_lag63</code>	Moving-average convergence-divergence signal for the span pair (8, 24), normalised and observed 63 trading days before the rebalance.
<code>macd_S8L24_lag126</code>	Moving-average convergence-divergence signal for the span pair (8, 24), normalised and observed 126 trading days before the rebalance.
<code>macd_S8L24_lag252</code>	Moving-average convergence-divergence signal for the span pair (8, 24), normalised and observed 252 trading days before the rebalance.
<code>macd_S16L48_lag0</code>	Moving-average convergence-divergence signal for the span pair (16, 48), normalised, observed at the rebalance date.
<code>macd_S16L48_lag21</code>	Moving-average convergence-divergence signal for the span pair (16, 48), normalised and observed 21 trading days before the rebalance.
<code>macd_S16L48_lag63</code>	Moving-average convergence-divergence signal for the span pair (16, 48), normalised and observed 63 trading days before the rebalance.

continued overleaf

Table A.1 continued

Characteristic	Definition
macd_S16L48_lag126	Moving-average convergence-divergence signal for the span pair (16, 48), normalised and observed 126 trading days before the rebalance.
macd_S16L48_lag252	Moving-average convergence-divergence signal for the span pair (16, 48), normalised and observed 252 trading days before the rebalance.
macd_S32L96_lag0	Moving-average convergence-divergence signal for the span pair (32, 96), normalised, observed at the rebalance date.
macd_S32L96_lag21	Moving-average convergence-divergence signal for the span pair (32, 96), normalised and observed 21 trading days before the rebalance.
macd_S32L96_lag63	Moving-average convergence-divergence signal for the span pair (32, 96), normalised and observed 63 trading days before the rebalance.
macd_S32L96_lag126	Moving-average convergence-divergence signal for the span pair (32, 96), normalised and observed 126 trading days before the rebalance.
macd_S32L96_lag252	Moving-average convergence-divergence signal for the span pair (32, 96), normalised and observed 252 trading days before the rebalance.
macd_composite	Mean across the three span pairs of the response-shaped MACD signal $\phi(y) = y e^{-y^2/4}/0.89$, which damps extreme readings before aggregation (Baz et al., 2015).
macd_S8L24_accel_0_21	macd_S8L24_lag0 minus macd_S8L24_lag21: one-month acceleration of the fast trend signal.
macd_S8L24_accel_21_63	macd_S8L24_lag21 minus macd_S8L24_lag63: the same acceleration measured one month earlier, giving trend curvature.
macd_S16L48_accel_0_21	One-month acceleration of the medium-speed trend signal.
macd_S32L96_accel_0_21	One-month acceleration of the slow trend signal.
dist_to_ma_200	P_τ divided by its 200-day moving average, minus one.
ma_50_to_200	The 50-day moving average divided by the 200-day moving average, minus one: a trend-regime anchor.

continued overleaf

Table A.1 continued

Characteristic	Definition
ker_63	Kaufman efficiency ratio over 63 trading days: the absolute net price change $ P_\tau - P_{\tau-63} $ divided by the summed absolute daily changes over the same window. Bounded in $[0, 1]$, it is near one when a move is achieved in a straight line and near zero when the same net move is achieved through a noisy path, so it separates trend quality from trend size.
ker_126	Kaufman efficiency ratio over 126 trading days: the absolute net price change $ P_\tau - P_{\tau-126} $ divided by the summed absolute daily changes over the same window. Bounded in $[0, 1]$, it is near one when a move is achieved in a straight line and near zero when the same net move is achieved through a noisy path, so it separates trend quality from trend size.
ker_252	Kaufman efficiency ratio over 252 trading days: the absolute net price change $ P_\tau - P_{\tau-252} $ divided by the summed absolute daily changes over the same window. Bounded in $[0, 1]$, it is near one when a move is achieved in a straight line and near zero when the same net move is achieved through a noisy path, so it separates trend quality from trend size.
Mean reversion: residual shocks (6 columns)	
resid_cum_5	Sum of the last 5 daily residuals, using betas from the trailing 252-day window described above.
resid_cum_10	Sum of the last 10 daily residuals, using betas from the trailing 252-day window described above.
resid_cum_21	Sum of the last 21 daily residuals, using betas from the trailing 252-day window described above.
resid_z_5_252	resid_cum_5 standardised by the trailing 252-day residual volatility scaled to a 5-day horizon.
resid_z_10_252	resid_cum_10 standardised by the trailing 252-day residual volatility scaled to a 10-day horizon.
resid_z_21_252	resid_cum_21 standardised by the trailing 252-day residual volatility scaled to a 21-day horizon.
Mean reversion: path shape (15 columns)	
resid_worst_1d_21	Minimum single-day residual over the trailing 21 days.
resid_best_1d_21	Maximum single-day residual over the trailing 21 days.

continued overleaf

Table A.1 continued

Characteristic	Definition
resid_downside_vol_21	Root mean square of the negative parts of the 21 daily residuals, $\sqrt{\min(e_t, 0)^2}$: downside dispersion of the idiosyncratic shock, which separates a name that fell steadily from one that fell in a single jump.
resid_autocorr1_21	First-order autocorrelation of the 21 daily residuals. Negative values indicate day-to-day reversal in the idiosyncratic component, positive values indicate continuation.
resid_sign_flip_rate_21	Fraction of consecutive residual pairs within the 21-day window whose signs differ, a non-parametric counterpart to the autocorrelation.
overnight_ret_5	Compounded overnight return $\prod(1 + O_t/C_{t-1} - 1) - 1$ over five days. Days whose open is imputed from the close are excluded rather than booked as overnight moves.
overnight_ret_21	The same compounded overnight return over 21 days.
intraday_ret_5	Compounded intraday return $\prod(C_t/O_t) - 1$ over five days. Together with the overnight leg this decomposes the total return into the two segments, which reverse at different speeds.
intraday_ret_21	The same compounded intraday return over 21 days.
close_location_21	Position of the closing price within the trailing 21-day intraday range, $(P_\tau - \min_{21} L_t)/(\max_{21} H_t - \min_{21} L_t)$ with L_t/H_t the adjusted daily low and high, equal to zero at the low and one at the high.
drawdown_from_21d_high	P_τ divided by the 21-day maximum price, minus one.
bounce_from_21d_low	P_τ divided by the 21-day minimum price, minus one.
days_since_21d_low	Number of trading days since the 21-day minimum price was set. Resolves into roughly twenty levels; retained and discussed in Section 3.5.
resid_cum_21_chg_1m	Change in resid_cum_21 from the previous rebalance: whether the residual shock is building or decaying.
bounce_from_21d_low_chg_1m	Change in bounce_from_21d_low from the previous rebalance, capturing the transition of the name's position relative to its recent low.
Risk and style (14 columns)	

continued overleaf

Table A.1 continued

Characteristic	Definition
beta_mkt_63	Rolling 63-day beta of daily excess returns on the value-weighted Ken French market factor, computed as cov/var .
beta_mkt_252	Rolling 252-day beta of daily excess returns on the value-weighted Ken French market factor, computed as cov/var .
downside_beta_63	The same beta re-estimated over the subset of the trailing 63 days on which the market factor is below its own window mean (Ang et al., 2006).
downside_beta_252	The same beta re-estimated over the subset of the trailing 252 days on which the market factor is below its own window mean (Ang et al., 2006).
idio_vol_63	Annualised standard deviation over the trailing 63 days of the residuals from the factor model used throughout (five Fama-French factors, momentum, and the industry peer return), so idiosyncratic volatility and the residual momentum/reversal features share one definition of idiosyncratic.
idio_vol_252	Annualised standard deviation over the trailing 252 days of the residuals from the factor model used throughout (five Fama-French factors, momentum, and the industry peer return), so idiosyncratic volatility and the residual momentum/reversal features share one definition of idiosyncratic.
ret_skew_63	Skewness of daily returns over the trailing 63 days.
ret_skew_252	Skewness of daily returns over the trailing 252 days.
crash_min_63	Minimum single-day return over the trailing 63 days.
crash_min_252	Minimum single-day return over the trailing 252 days.
drawdown_from_63d_high	P_τ divided by the 63-day maximum price, minus one.
drawdown_from_252d_high	P_τ divided by the 252-day maximum price, minus one.
max_21	Maximum single-day return over the trailing 21 days: the lottery or MAX characteristic of Bali et al. (2011), included so that the surface carries the demand-for-skewness channel alongside the volatility measures.
sigma_an_63	Annualised volatility from the 63-day exponentially weighted daily return variance, $\hat{\sigma}\sqrt{252}$. This is the same estimate used for inverse-volatility sizing in Chapter 6.

continued overleaf

Table A.1 continued

Characteristic	Definition
Liquidity and size (12 columns)	
log_mcap	$\log(1 + \text{market capitalisation})$ at the rebalance date.
log_adv20	$\log(1 + 20\text{-day average dollar volume})$.
adv_shock_5_63	Five-day average dollar volume divided by its 63-day average: a short-horizon volume surprise, distinct from the turnover shock below in that it is not normalised by shares outstanding.
amihud_63	63-day mean of the daily Amihud ratio $ r_t /\text{dollar volume}_t$ (Amihud, 2002): the quarterly counterpart of <code>amihud_21</code> .
amihud_21	21-day mean of the daily Amihud ratio.
amihud_z_21_252	Time-series z -score of $\log(\text{amihud_21})$ over a trailing 252-day window: illiquidity relative to the name's own recent norm rather than the cross-section.
turnover_21	21-day mean of daily share turnover (shares traded divided by shares outstanding).
turnover_63	63-day mean of daily share turnover: the quarterly counterpart of <code>turnover_21</code> .
turnover_shock_5_63	Five-day mean turnover divided by its 63-day mean.
listing_age_years	Years between the start of the security's CRSP record (<code>SecurityBegDt</code>) and the rebalance date, floored at zero. This is an age within CRSP coverage rather than an age since flotation, and is left-censored for securities already listed when coverage begins.
net_share_issuance_63	Split-adjusted shares outstanding divided by their value 63 trading days earlier, minus one.
net_share_issuance_252	The same issuance measure over 252 trading days.
Seasonality (2 columns)	
month_sin	$\sin(2\pi m/12)$ for calendar month m of the rebalance date.
month_cos	$\cos(2\pi m/12)$. The pair encodes the month on a circle so that December and January are adjacent. These two columns are the sole exception to the rank mapping of Section 4.2: they are already bounded and equally spaced by construction, and they enter the model on their native $[-1, 1]$ scale in every arm, so the representation treatment does not apply to them.
Fundamentals (44 columns)	

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>fund_bm</code>	Book-to-market equity, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_ep_exi</code>	Earnings-to-price excluding extraordinary items, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_ep_inc</code>	Earnings-to-price including extraordinary items, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_cf_to_price</code>	Cash flow to price, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_ebitda_to_ev</code>	EBITDA to enterprise value, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_roa</code>	Return on assets, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_roe</code>	Return on equity, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_gp_to_assets</code>	Gross profitability: gross profit over total assets, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_asset_turnover</code>	Asset turnover, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_capex_to_assets</code>	Capital expenditure to total assets, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_asset_growth_yoy</code>	Year-on-year growth in total assets, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_debt_to_equity</code>	Debt to equity (ratio file), as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>fund_stmt_accruals</code>	Accruals computed from statement-level data, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_stmt_debt_to_equity</code>	Debt to equity computed from statement-level data, as of the most recent point-in-time observation admissible under the availability rules of Section 3.5.
<code>fund_bm_ind_centered</code>	Within-date rank of <code>fund_bm</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_ep_exi_ind_centered</code>	Within-date rank of <code>fund_ep_exi</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_ep_inc_ind_centered</code>	Within-date rank of <code>fund_ep_inc</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_cf_to_price_ind_centered</code>	Within-date rank of <code>fund_cf_to_price</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_ebitda_to_ev_ind_centered</code>	Within-date rank of <code>fund_ebitda_to_ev</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_roa_ind_centered</code>	Within-date rank of <code>fund_roa</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_roe_ind_centered</code>	Within-date rank of <code>fund_roe</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_gp_to_assets_ind_centered</code>	Within-date rank of <code>fund_gp_to_assets</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_asset_turnover_ind_centered</code>	Within-date rank of <code>fund_asset_turnover</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_capex_to_assets_ind_centered</code>	Within-date rank of <code>fund_capex_to_assets</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>fund_asset_growth_yoy_ind_centered</code>	Within-date rank of <code>fund_asset_growth_yoy</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_debt_to_equity_ind_centered</code>	Within-date rank of <code>fund_debt_to_equity</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_stmt_accruals_ind_centered</code>	Within-date rank of <code>fund_stmt_accruals</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_stmt_debt_to_equity_ind_centered</code>	Within-date rank of <code>fund_stmt_debt_to_equity</code> centred within the firm's industry, formed only where at least twenty contemporaneous peers are available.
<code>fund_roe_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_roa_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_gp_to_assets_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_asset_turnover_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_debt_to_equity_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_stmt_accruals_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_bm_chg_6m</code>	Change in the corresponding level characteristic over the trailing 6 monthly rebalances.
<code>fund_ep_exi_chg_6m</code>	Change in the corresponding level characteristic over the trailing 6 monthly rebalances.
<code>fund_cf_to_price_chg_6m</code>	Change in the corresponding level characteristic over the trailing 6 monthly rebalances.
<code>fund_ep_inc_chg_6m</code>	Change in the corresponding level characteristic over the trailing 6 monthly rebalances.
<code>fund_ebitda_to_ev_chg_6m</code>	Change in the corresponding level characteristic over the trailing 6 monthly rebalances.
<code>fund_ep_exi_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_cf_to_price_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.

continued overleaf

Table A.1 continued

Characteristic	Definition
<code>fund_asset_growth_yoy_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_capex_to_assets_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
<code>fund_stmt_debt_to_equity_chg_12m</code>	Change in the corresponding level characteristic over the trailing 12 monthly rebalances.
Analyst surprise (12 columns)	
<code>analyst_sue_score</code>	Standardised unexpected earnings at the most recent admissible announcement.
<code>analyst_sue_abs_score</code>	Absolute value of the surprise score, a magnitude measure that does not distinguish direction.
<code>analyst_sue_surprise_to_estimate</code>	Earnings surprise scaled by the consensus estimate.
<code>analyst_sue_score_avg_4y</code>	Mean surprise score over the four most recent consecutive annual announcements (the delivered file contains annual periods only, hence the <code>_4y</code> suffix).
<code>analyst_sue_score_std_4y</code>	Standard deviation of the surprise score over the same four announcements: the consistency of the surprise series.
<code>analyst_sue_score_chg_4y</code>	Change in the surprise score across four annual announcements.
<code>analyst_sue_score_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_score</code> .
<code>analyst_sue_abs_score_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_abs_score</code> .
<code>analyst_sue_surprise_to_estimate_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_surprise_to_estimate</code> .
<code>analyst_sue_score_avg_4y_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_score_avg_4y</code> .
<code>analyst_sue_score_std_4y_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_score_std_4y</code> .
<code>analyst_sue_score_chg_4y_ind_centered</code>	Industry-centred counterpart of <code>analyst_sue_score_chg_4y</code> .

Twenty-six of the 144 columns are built on the monthly rebalance grid rather than from the daily panel. They are first differences of another characteristic, or spreads between the same characteristic measured at two horizons: sixteen six- and twelve-month changes in the fundamental ratios, a change in the surprise score across four annual announcements with its industry-relative twin, four accelerations of the trend signals, two

spreads between cumulative returns at different horizons, and two mean-reversion state transitions. These carry transition information that a level-only surface cannot express, since a firm whose profitability has just improved to a given level and one that has been at that level throughout are indistinguishable in the levels alone. With one exception, the differences are taken on the common rebalance calendar and not on each firm’s own history of appearances: a name absent from the universe at the lagged date yields a missing difference rather than one silently spanning a longer window. The exception is the surprise change, which, together with its industry-relative twin, is defined across the firm’s four most recent annual announcements and therefore moves on the announcement calendar. They are also computed on the pre-imputation values of their parents, so a difference can never reflect the movement of the imputed cross-sectional median rather than of the firm. They are formed after the level characteristics and are subject to the same availability rules, because a difference of two point-in-time observations is itself point-in-time.

Two per-characteristic diagnostics are computed on every feature build and are not typeset in Table A.1: the source archive and staleness rule behind each characteristic, and the missing rate and exact-zero mass over the evaluation sample, both measured before the median substitution and recorded with the build (`feature_quality_prefill.csv` and its by-year companion). The missing rate is the diagnostic referred to in Section 3.5, where a high rate is noted to produce a tied block at the cross-sectional median whose spacing differs between representation arms.

Measured over the evaluation sample, the surface is close to complete: the median characteristic carries 0.8% missing values, and one of the 144 exceeds 15%. Among the derived columns above, the fundamental and surprise differences carry the highest rates, by construction of their lagged-membership requirement, while the accelerations and return spreads are formed from the complete price panel and carry none. No characteristic carries a zero mass above 14%, and the largest masses are economically meaningful states rather than artefacts.

A.1 Point-in-time availability

A characteristic may enter the feature vector for a rebalance only if the record behind it was observable on that date. Section 3.5 states the principle; Table A.2 states it per source, because the policy is only as good as the weakest date field it relies on and those differ between CRSP, Compustat and IBES.

Three rules govern when a record becomes visible, and each is a deliberate choice of the conservative option over the convenient one. A fourth governs which security a record attaches to.

An announcement date is trusted only when it is not impossible. Compustat carries three candidate announcement fields, and the build takes the first available of `rdq`, `pdateq` and `fdateq`. Any of them is accepted only when it falls on or after the fiscal period end `datadate`. A field that predates the period it reports is not a publication date, whatever the vendor stores in it, and accepting one would date a statement to before the quarter it describes had finished.

A missing announcement date makes a record later, never earlier. Where no candidate survives that test, the observation is dated by a fallback of ninety days after the fiscal period end, extended to a hundred and twenty days for fourth quarters, which report on a longer statutory clock. The fallback is always weakly later than a genuine announcement date would have been, so a gap in the vendor’s coverage cannot make a statement usable sooner. This is the rule that matters most: the alternative convention, dating a statement to its fiscal period end when the announcement is missing, would grant the model roughly a quarter of hindsight on precisely the observations whose provenance is least certain.

Nothing is carried forward indefinitely, and nothing arrives from the future. Every availability date is moved forward to the first trading day strictly after it: one session of publication lag for a record already dated on a trading day, and the next session for one dated on a weekend or holiday. Records are then joined to rebalances by a backward `merge_asof` subject to two conditions: the age of the record must be non-negative, and it must not exceed the family’s staleness cap. The first condition is what makes lookahead structurally impossible rather than merely discouraged, since a record whose availability date falls after the rebalance is discarded by the join itself. The second stops a single stale statement populating a feature for years: a quarterly statement expires after 548 days and a monthly ratio after 120.

Industry-relative variants are formed only where at least twenty peers share the sector on the date, so a thin industry cannot produce a relative measure defined against two or three names.

A reused ticker is discarded, not resolved. I/B/E/S identifies firms by exchange ticker, and tickers are reused across companies over time, so surprises are joined to securities through a ticker-to-company concordance, built once from I/B/E/S ticker–CUSIP records and anchored to the company identifier through CUSIP. Any ticker that ever points to more than one company is discarded rather than resolved. The concordance resolves 91.2% of surprise announcements on its own; those it does not reach are matched on the exchange ticker against the identifier bridge, under the same uniqueness rule, which takes the total to 93.5%. That last step tests uniqueness against the bridge rather than

against the CUSIP records, so the 2.2 points it adds carry a weaker guarantee than the rest.

Table A.2: When each upstream record becomes usable. A characteristic may enter the feature vector for a rebalance only if its availability date falls on or before that date. Announcement dates are trusted only when they fall on or after the fiscal period end, so an erroneous `rdq` cannot make a statement usable early; otherwise the conservative fallback applies, which is never earlier. The date is then moved forward to the first trading day strictly after it. A record older than the staleness cap is dropped rather than carried forward. All joins are backward `merge_asof` requiring a non-negative age, so a record dated after the rebalance is discarded by construction.

Source	Date used	Fallback when absent	Lag	Staleness cap
CRSP daily prices and volume	Trade date	None: the record is the observation	None	None
Compustat quarterly statements	<code>rdq</code> , else <code>pdateq</code> , else <code>fdateq</code>	<code>datadate</code> + 90 days (Q1–Q3), + 120 days (Q4)	1 day	548 days
Monthly financial ratios	<code>public_date</code>	Not applicable	1 day	120 days
IBES earnings surprise	<code>anndats</code>	Not applicable	1 day	548 days

A.1.1 Where the invariance control puts its weight

The rules above are a claim about the data. Table A.3 offers a check from the other direction, on the gradient-boosted control of Section 4.5 fitted to the same surface rather than on the inputs. Its gain profile is identical in both representation arms (Section C.6), so what it measures is a property of the surface rather than of one encoding.

The reasoning is narrow, and it is worth stating precisely because the temptation is to claim more. Feature importance cannot detect lookahead: a leaked value and a legitimate one are indistinguishable to any attribution method, since the question is whether a number could have been known and not how much it contributed. What importance can rule out is the specific failure that would matter here. A leak large enough to generate the performance of Chapter 5 would have to sit in a characteristic that control leans on heavily, and no characteristic is leaned on heavily. The largest single gain share is 1.41%; the Herfindahl index over the 144 characteristics is 0.0073, an effective count of 137 of them; and every characteristic is used.

The sharper cut is by source. Of the 144 characteristics, 56 come from Compustat or IBES, the two sources whose dating depends on the rules above rather than on a trade. Those 56 are 39% of the surface and carry 37.6% of the gain, which is 0.671% each against 0.709% for the characteristics dated by a trade. The families most exposed to a dating error are therefore slightly less influential per characteristic than the families that carry

no dating risk at all, which is the opposite of the pattern a lookahead problem would produce.

Table A.3: Total gain of the gradient-boosted control, by characteristic family. Taken from the fitted boosters of the inverse-normal cell, averaged over the five seeds of the final fold, not from a probe refitted for the purpose. Averaging the twenty-three folds instead moves no characteristic’s share by more than a third of a percentage point. Gain per feature divides the family share by the number of characteristics in it, so a merely large family does not look influential. The last column marks how firmly the upstream record is dated: trades stamp their own date, statements and analyst records can be revised. This describes where a tree spends its splits; it is not a test for lookahead, which no attribution method can perform.

Family	Features	Gain share	Per feature	Dating
Fundamentals	44	31.0%	0.70%	Restatement risk
Trend (MACD and moving averages)	25	17.3%	0.69%	Trade date
Mean reversion: path shape	15	11.0%	0.73%	Trade date
Risk and style	14	11.0%	0.79%	Trade date
Liquidity and size	12	7.8%	0.65%	Trade date
Analyst surprise	12	6.6%	0.55%	Revision risk
Total momentum	8	6.0%	0.75%	Trade date
Mean reversion: residual shocks	6	3.7%	0.62%	Trade date
Residual momentum	6	3.5%	0.58%	Trade date
Seasonality	2	2.1%	1.05%	Trade date
All	144	100.0%	0.69%	

The most important single characteristic carries 1.41% of total gain and the Herfindahl index over the 144 characteristics is 0.0073, an effective count of 137. No characteristic is unused. The 56 characteristics whose availability is not fixed by a trade are 39% of the surface and carry 37.6% of gain, 0.671% each against 0.709% for the rest.

Appendix B

Protocol and Configuration

B.1 Walk-forward calendar

The schedule of Section 3.6 yields 23 folds, the first testing on 30 April 2003 to 31 March 2004 and the last covering the eight months to 28 November 2025, and the 23 test blocks tile the 272 evaluated rebalances without overlap.

The fold-by-fold calendar, the training, validation and test spans with their purge boundaries and the number of rebalances each test block contributes, is determined entirely by the constants of Table 3.3 and the purge rule of Section 3.6, including the truncation of the final fold at the end of the panel. It is checked by the splitter at run time on the realised dates and written to the run log, and is not typeset here.

B.2 Resolved configuration

Three properties of the stored configuration would mislead a reader who inspected a run manifest without also reading the evaluation module, and all three are recorded here.

The cost stack is applied at replay, not at training. The manifest of every training run carries the estimation-time configuration, in which costs are disabled and the impact, hard-to-borrow and financing coefficients are therefore zero. That is the correct setting for the gross convention under which every model is fitted (Section 7.1), and it is not the parameterisation any net figure is computed under. The net regime is a single override applied when the frozen scores are replayed, and its values are exactly those of Table 4.5: a 2 basis-point commission, a volatility-scaled square-root impact prefactor $\kappa = 0.5$ on a 20-day average daily volume window, 50 basis points of general-collateral borrow, a 200 basis-point hard-to-borrow premium below \$20 million of average daily volume, \$1 million of managed capital, and 100 basis points of financing on the debit balance of (4.18), the long market value less the unit of equity, which is one unit at $G = 4$.

The feature encodings carry legacy mode names. The percentile arm is stored un-

der the mode name `gaussian`, the inverse-normal arm under `normal_score`, and the scale-matched arm under `centered_scaled`. The first name is historical and no longer describes what it does: the transform it selects is the centred percentile of (4.2), and the normal score of (4.3) is reached only through the second. The cells are assigned as Table 4.1 states; the name is the only thing that is misleading, and it is recorded here so that the manifests can be read without inverting the treatment.

The volatility target is not a risk constraint. The manifest records `vol_target: 0.15`, which reads as though the book were held at 15% annualised volatility. It is not. The value enters position sizing as the constant v^* of (4.12), where the leg-level rescaling of (4.13) divides it out, so it reaches the fitted book only through the per-name ceiling and is inert at these values, binding only for a name whose annualised volatility falls below about 3%, which does not occur in this universe; the scale of the book is set by the gross exposure G alone. The volatility estimate feeding $1/\hat{\sigma}_i$ is guarded at both ends by the volatility floor and fill of Table 4.4, and both guards shrink the position rather than enlarge it: an understated volatility enters the denominator and produces an outsized position, whereas an overstated one merely shrinks the position towards zero, and a name whose risk cannot be measured takes one of the smallest positions in its leg. The same field is read a second time, and for an unrelated purpose, when a run is summarised: it is the target of the volatility rescaling that Section 5.1 applies so that returns and drawdowns can be compared across cells. A reader who took the field for a risk constraint would conclude that the realised volatility differences of Table 5.3, and with them the volatility channel of Section 5.3.2, could not be what they are.

The resolved configuration of each cell, its configuration hash, the feature-cache signature, the content hashes of every source module, and the package versions are stored in the run manifest, as Section 4.11 describes, and are not reproduced here. The feature-cache signature is stored per arm, since a signature shared across arms would indicate that the treatment had not been applied.

Two considerations fix the gross exposure at $G = 4$. The raw book then sits at the 15% at which every cross-cell comparison of Section 5.1 is struck, so the rescaling that convention applies is a factor between 0.85 and 1.04 rather than a large adjustment to the quantities the tables report. And two units a side on a dollar-neutral book of large, liquid names is the conventional operating point for the mandate the screens of Section 3.3 already describe. The choice is not a flattering one, and Section 7.2 shows in which direction it errs.

B.3 Gross ceiling and cost scaling

The ceiling is not a flattering choice, and the direction can be shown rather than asserted. Under a common rescaling of the book to sG the components of (4.19) do not move alike:

spread, borrow and the entry term scale with gross exposure, market impact scales as its three-halves power by (4.17), and financing, being charged on the debit balance $sG/2 - 1$, is affine in s rather than proportional to it, and vanishes entirely below a ceiling of two, where the long side is covered by equity and no margin loan is drawn. Dividing through by the volatility, which scales with s as well, the net Sharpe ratio is strictly decreasing in s over the whole range from a gross ceiling of one to the four used here, wherever impact is positive and wherever the financing spread is charged at all. Every lower ceiling therefore reports a higher net Sharpe ratio than the one carried below, so Section 7.3 prices the book at the most punitive point of that range and not at its most favourable. No grid over G is replayed: the statement is algebra on how the four components scale, it holds at fixed managed capital, and it sets aside the share rounding that would begin to bite on a book this small at a low ceiling.

The same scaling fixes the capacity table. Because impact scales exactly as $\sqrt{A}$ at fixed weights, the break-even figures follow in closed form from the base-case bridge, and the replayed grid reproduces them to numerical precision. The same closed form makes the figures inherit the prefactor: break-even capital scales as $1/\kappa^2$, so a reader who holds a different view of κ can rescale the whole table without re-estimating anything.

B.4 Soft-rank temperature

The soft-rank temperature of Table 4.2 is inherited rather than tuned, and Section 4.3 states the reason in one sentence. The measurement behind it is recorded here. The soft rank (4.7) is monotone in the score at every temperature, so the induced ordering is exact in all cases and the surrogate tracks the realised Spearman deficit essentially perfectly across two orders of magnitude in τ . What the temperature governs is where gradient is spent. A sharp sigmoid saturates for distant pairs, so only locally adjacent names contribute, and the extreme names (which have neighbours on one side only) are starved. Measured shares of total gradient falling in the top and bottom deciles are 4.5% and 3.2% at $\tau = 0.02$, against 6.8% and 7.0% at $\tau = 1.0$. Since the traded book is formed from precisely those deciles, the flatter profile is the appropriate operating point.

B.5 Hyperparameters

The settings not reported in Table 4.3, Table 4.2 and Table 4.4 are identical across cells except where an objective requires otherwise, and they are recorded in the resolved configuration each run manifest stores rather than typeset here.

A cross-section of a thousand names carries roughly half a million unordered pairs, so any feasible budget is a sample; 32,768 pairs cover about 6.6% of them. The budget's size

affects the sampling variance of the gradient but not the target it estimates: the RankNet and margin objectives average over the sample, and (4.5) rescales the sampled sum by $N/|\mathcal{P}|$, so all three are unbiased for their full-enumeration counterparts at any budget. The variance the sample leaves behind is not innocuous under a ten-epoch schedule, which is one more reason the budget is held identical across every cell; the listwise objectives read the entire cross-section and never consume it.

Gradient clipping is disabled uniformly, as a condition of the cross-objective comparison. AdamW is invariant to a constant rescaling of the loss up to the small stabilising constant in its denominator, which sits orders of magnitude below the gradient scales here, so a clip norm is the only material channel through which the absolute magnitude of an objective can affect the fitted model, and the objectives do not share a magnitude: (4.5) sums over sampled pairs and is orders of magnitude larger than the averaged objectives (4.4) and (4.6). A common clip norm would bind at almost every step for the LambdaRankIC cells and at almost none for the others, confounding a comparison across objectives with a difference in effective step size.

On tabular problems, a simple well-tuned baseline remains competitive with specialised architectures (Gorishniy et al., 2021), and a parsimonious backbone keeps the interpretation anchored on the treatments and not on the network.

B.6 Where each arm sits on the activation curve

Section 4.2.1 argues that a common rescaling of the inputs is not neutral for this architecture, and that the route through which it acts is the curvature of the activation rather than the size of the optimiser’s step. The argument is about the network and not about the panel, so it is settled here without a fitted model, without returns and without a training run: the scorer is built at the configured settings, a tie-free cross-section of 1000 names and 144 independently ordered columns is pushed through its first block under each encoding, and the block is compared against the affine map it would be if the activation did not bend. Table B.1 reports the result. The generating script is `scripts/build_activation_diagnostic.py` and it reads the architecture from the same resolved configuration the runs use, so the diagnostic cannot drift from the model it describes.

The statistic is the share of a hidden unit’s post-activation variance that the best affine function of its own pre-activation leaves unexplained, averaged across units. It is zero when the block is affine and rises as the activation is driven into its curved region. Three readings follow.

Panel A traces it against the input scale and shows the two encodings sitting on different parts of one curve. The departure is 0.4% at an input standard deviation of 0.1, 3.2% at the $1/\sqrt{12}$ of the centred percentile, 19.7% at unit scale, and flattens near 29%

beyond that. The curve is steepest between the two encodings a modeller would choose between, which is the interval this thesis measures across, and it flattens beyond them. That has a consequence worth stating because it bounds the finding as much as it supports it: the returns to input magnitude diminish above that interval, so multiplying the inputs by a larger constant again would be expected to buy less than the first multiplication did.

Panel B gives the three arms and, for each, every hidden block rather than the first alone. At the first block the percentile arm reads 3.2%, the scale-matched arm 19.7% and the inverse-normal arm 19.6%. The change of scale accounts for the whole of the movement and the change of shape at matched scale accounts for none of it, which is the same division Section 5.3.2 finds in the volatility of the book, arrived at from the architecture instead of from the returns.

The remaining columns localise the channel. Blocks two, three and four read a normalised output rather than the encoding, so they sit at a common scale whatever encoding was applied at the input, and the three arms are indistinguishable across all of them, from 17.5% to 20.4% with no ordering by arm. The encoding therefore does not make one network globally more nonlinear than another. It costs the percentile arm the use of exactly one of its four nonlinear layers, the one that reads the features, and leaves the rest untouched. That is a narrower and more checkable statement than a claim about the loss surface as a whole, and it is the reason a diagnostic taken at the first block is the right one.

Two limits attach to the whole of this section and are not small. The diagnostic is computed at initialisation, and the first-layer weights are free to adapt over the ten epochs the deployed models train for, so nothing here establishes that the gap at the first block survives training. Trained weights are not among the stored artefacts (Section 4.11), so answering that would require a re-fit rather than a reading, and it is left open. The correspondence with the empirical decomposition is accordingly a consistency check and not a second piece of evidence: this diagnostic knows nothing about what any of these models went on to learn.

Panel C asks how much of a rescaling survives the whole block once the normalisation has acted. Under the configured GELU the two arms' post-block activations differ by up to 1.999 and correlate at 0.9412. Under a rectified linear activation, which is positively homogeneous and for which the normalisation would absorb a rescaling exactly were the first-layer bias zero, the difference falls to 0.980 and the correlation rises to 0.9718. It does not fall to zero, and the residue is the bias route named in Section 4.2.1. The activation is therefore the larger part of the channel and the bias the smaller, and neither is closed by the normalisation.

Panel D closes the route that is most often assumed to be the operative one. A LambdaRankIC loss is backpropagated through the initialised network under each encoding

and one AdamW step is taken. The norm of the gradient with respect to the first-layer weights differs between the arms by 2.6%, not by the factor $\sqrt{12}$ a reader would expect from the input scaling alone, because the normalisation’s backward pass removes almost all of it before it reaches the first layer. The realised update on those weights is then identical to the five decimals printed, since AdamW divides by a running estimate of the gradient’s own magnitude and is invariant to a constant rescaling of it. The decoupled weight-decay pull is the same in both arms at initialisation, the weights being identical there, and is smaller than the gradient step by a factor of about forty thousand. Whatever the scale channel is doing, it is not letting one arm travel further per epoch than the other, which is why the objection Section 4.4.2 answers is about the number of epochs and not about the size of the step.

Table B.1: The first block’s departure from an affine map, by input scale and by representation arm. Computed at initialisation on a tie-free cross-section of 1000 names and 144 columns, at the architecture of Table 4.3, with no fitted model and no returns involved. The statistic is the share of a hidden unit’s post-activation variance that the best affine function of its own pre-activation leaves unexplained, averaged over every fourth unit of the block: zero when the block is an affine map. Only the first block reads the encoding itself; every later block reads a normalised output and therefore sits at the same scale in all three arms, which is what confines the channel to one layer. Panel C reports how much of a common rescaling survives the whole block, once the normalisation has acted, under the configured activation and under a positively homogeneous one. Panel A is evaluated at target input scales, two of which (0.289 and 0.577) are the centred-rank conventions of Section 2.3; Panel B reports the scale each arm realises on the simulated cross-section, which differs from the target by at most 0.006. Panel D backpropagates a LambdaRankIC loss through the initialised network and takes one optimiser step, so that the gradient reaching the first layer, the movement it actually produces, and the decoupled weight-decay pull can be read against one another.

Panel A. Departure from affinity against input scale

Input s.d.	Pre-activation s.d.	Departure
0.100	0.057	0.4%
0.289	0.166	3.2%
0.500	0.287	8.2%
0.577	0.332	10.2%
1.000	0.575	19.7%
1.732	0.996	27.6%
3.000	1.725	29.8%
6.000	3.449	28.4%

Panel B. The three arms, block by block

Arm	Input s.d.	Block-1 pre-act. s.d.	Departure, by hidden block			
			1	2	3	4
Percentile	0.288	0.166	3.2%	18.6%	17.5%	20.4%
Scale-matched	0.999	0.574	19.7%	19.5%	18.2%	19.9%
Inverse-normal	0.994	0.571	19.6%	19.3%	18.2%	19.7%

Panel C. Percentile against scale-matched, after the first block

Activation	Max abs. difference	Correlation
GELU (configured)	1.999	0.9412
ReLU (positively homogeneous)	0.980	0.9718

Panel D. The step-size route, on the first-layer weights

Arm	$\ \nabla_{W_1}\ $	Realised AdamW step	Decay pull
Percentile	643.6	0.19200	4.61×10^{-6}
Scale-matched	660.3	0.19200	4.61×10^{-6}
Ratio, scale-matched to percentile	1.026	1.000	1.000

Appendix C

Full Results

C.1 Representation by objective

Chapter 5 through Chapter 8 report figures pooled over the five ranking objectives, on the licence Section 5.8 establishes. This section is the unpooled record those figures average, given so that the licence can be checked rather than taken on trust, and so that no combination is absent from the record on the basis of its outcome. Table C.1 and Table C.2 carry the annualised Sharpe ratio of every estimated specification at every traded fraction under both weighting schemes, gross and net respectively. The primary endpoint of Section 3.7 is a paired difference between two of these rows, and the secondary family is the same difference taken objective by objective.

Three features of the tables are worth pointing out, because each bears on a claim made in the main text and each is visible without any inference.

The ordering of the arms at the headline decile holds in every objective and under both weighting schemes: the percentile row sits below the scale-matched row, which sits below the inverse-normal row, in all five families. That is the pattern Section 5.3 reports pooled, shown cell by cell.

The gradient-boosted control is identical across its two arms in every column of both tables, not merely at the configuration Section 5.5 inspects. Invariance to a strictly monotone re-encoding is therefore verified across the whole grid rather than at one operating point, which is a stronger statement than that section makes.

The ordering of the arms is not preserved at the narrow fractions. At the 1% column several objectives reverse, and the reversals do not agree between the two weighting schemes. Nothing in the thesis rests on that column, which Section 6.5 excludes from selection by construction and reads as a concentration stress test, but the reversals are part of the record and are shown rather than suppressed.

Of the fifty objective-by-width-by-scheme cells, the six that turn slightly negative at the two narrowest fractions are three in Soft-Spearman, two in margin ranking and one

Table C.1: Annualised gross Sharpe ratio for every estimated specification, at every traded fraction and under both weighting schemes. The unpooled record behind the main text, and the table Section 5.8 defers to. Every cell shares the universe, label, folds, seeds and portfolio rules. The gradient-boosted control is invariant to the encoding, so its two rows agree to the printed precision; it and the linear anchors are estimated in the two base arms only.

Specification	Arm	Inverse-volatility					Equal				
		20%	10%	5%	2.5%	1%	20%	10%	5%	2.5%	1%
RankNet	Percentile	1.079	1.009	1.027	0.998	0.808	1.116	1.000	1.009	0.994	0.774
	Scale-matched	1.358	1.289	1.298	1.070	0.952	1.328	1.241	1.258	1.033	0.895
	Inverse-normal	1.388	1.490	1.354	1.326	0.938	1.401	1.411	1.304	1.292	0.959
LambdaRankIC	Percentile	1.106	1.083	1.068	0.975	0.875	1.132	1.062	1.017	0.984	0.850
	Scale-matched	1.388	1.386	1.362	1.193	1.036	1.371	1.341	1.294	1.134	0.968
	Inverse-normal	1.425	1.586	1.532	1.485	0.983	1.433	1.536	1.467	1.408	0.860
Margin ranking	Percentile	1.170	1.123	1.023	0.999	0.963	1.093	1.016	0.970	1.025	0.969
	Scale-matched	1.247	1.285	1.213	1.121	1.039	1.209	1.193	1.197	1.134	1.050
	Inverse-normal	1.335	1.398	1.251	1.088	1.000	1.269	1.269	1.207	1.017	0.917
Soft-Spearman	Percentile	1.133	1.119	1.018	1.083	0.897	1.075	1.043	0.984	1.029	0.918
	Scale-matched	1.413	1.287	1.034	0.978	0.950	1.313	1.183	0.940	0.937	0.932
	Inverse-normal	1.549	1.452	1.205	1.077	0.808	1.422	1.296	1.124	1.070	0.842
RankCosine	Percentile	1.154	1.174	1.062	1.003	0.954	1.114	1.101	1.020	0.971	0.961
	Scale-matched	1.518	1.382	1.114	1.028	0.942	1.414	1.247	1.015	0.959	0.904
	Inverse-normal	1.543	1.431	1.218	1.018	0.968	1.416	1.292	1.151	0.978	0.935
XGBoost (control)	Percentile	1.368	1.335	1.225	1.011	0.942	1.375	1.307	1.237	1.017	0.949
	Inverse-normal	1.368	1.335	1.225	1.011	0.942	1.375	1.307	1.237	1.017	0.949
OLS	Percentile	0.950	0.825	0.833	0.645	0.538	0.829	0.750	0.772	0.559	0.483
	Inverse-normal	0.870	0.882	0.768	0.786	0.389	0.748	0.775	0.672	0.689	0.410
Elastic net	Percentile	0.515	0.466	0.446	0.445	0.687	0.440	0.415	0.438	0.380	0.713
	Inverse-normal	0.390	0.280	0.366	0.423	0.358	0.377	0.321	0.387	0.393	0.375

in RankCosine, the same six gross and net.

Table C.2: Annualised net Sharpe ratio for every estimated specification, at every traded fraction and under both weighting schemes. The unpooled record behind the main text, and the table Section 5.8 defers to. Every cell shares the universe, label, folds, seeds and portfolio rules. The gradient-boosted control is invariant to the encoding, so its two rows agree to the printed precision; it and the linear anchors are estimated in the two base arms only.

Specification	Arm	Inverse-volatility					Equal				
		20%	10%	5%	2.5%	1%	20%	10%	5%	2.5%	1%
RankNet	Percentile	0.568	0.551	0.593	0.626	0.470	0.664	0.600	0.622	0.668	0.475
	Scale-matched	0.748	0.783	0.851	0.711	0.614	0.802	0.816	0.872	0.714	0.592
	Inverse-normal	0.783	0.956	0.895	0.923	0.606	0.878	0.969	0.904	0.938	0.649
LambdaRankIC	Percentile	0.582	0.588	0.608	0.560	0.530	0.672	0.634	0.620	0.626	0.550
	Scale-matched	0.769	0.846	0.889	0.801	0.659	0.836	0.889	0.893	0.795	0.629
	Inverse-normal	0.814	1.022	1.044	1.031	0.648	0.901	1.065	1.061	1.021	0.561
Margin ranking	Percentile	0.602	0.600	0.571	0.615	0.617	0.623	0.589	0.604	0.703	0.673
	Scale-matched	0.693	0.780	0.747	0.721	0.702	0.736	0.782	0.821	0.803	0.769
	Inverse-normal	0.767	0.865	0.788	0.706	0.654	0.793	0.845	0.835	0.702	0.625
Soft-Spearman	Percentile	0.652	0.664	0.601	0.719	0.562	0.661	0.657	0.624	0.707	0.606
	Scale-matched	0.811	0.789	0.606	0.595	0.624	0.816	0.778	0.574	0.603	0.639
	Inverse-normal	0.942	0.916	0.782	0.710	0.511	0.919	0.855	0.760	0.751	0.566
RankCosine	Percentile	0.671	0.719	0.642	0.632	0.627	0.698	0.717	0.656	0.645	0.664
	Scale-matched	0.899	0.865	0.676	0.642	0.613	0.902	0.824	0.644	0.623	0.607
	Inverse-normal	0.925	0.902	0.775	0.653	0.643	0.900	0.855	0.766	0.662	0.645
XGBoost (control)	Percentile	0.852	0.881	0.820	0.629	0.639	0.934	0.932	0.901	0.699	0.685
	Inverse-normal	0.852	0.881	0.820	0.629	0.639	0.934	0.932	0.901	0.699	0.685
OLS	Percentile	0.334	0.289	0.367	0.276	0.236	0.335	0.329	0.402	0.254	0.215
	Inverse-normal	0.253	0.360	0.334	0.435	0.146	0.264	0.366	0.326	0.400	0.196
Elastic net	Percentile	0.049	0.069	0.083	0.139	0.428	0.064	0.100	0.141	0.120	0.483
	Inverse-normal	-0.054	-0.092	0.056	0.163	0.129	0.019	0.011	0.127	0.176	0.189

C.2 Multiplicity-adjusted inference

Two per-objective diagnostics that Chapter 5 reads pooled are recorded here cell by cell.

At the scale step of (5.1) all five objectives reject equality of the rank information coefficient’s variances individually, four of them at $p < 0.0001$ and Margin ranking at $p = 0.0163$; the traded book’s volatility is the column where one objective does not (Table 5.6). The Pitman–Morgan paired-variance test, which assumes away the serial dependence the block bootstrap prices in, returns the same verdict cell by cell, so the finding does not rest on the block length.

Regressing the monthly book return on the monthly coefficient cell by cell and averaging gives a slope of 0.3535 in the percentile arm, 0.3595 in the scale-matched arm and 0.3543 in the inverse-normal arm, with HAC standard errors near 0.026 and t -statistics between 11.4 and 15.3. The slope is invariant across the arms to within 0.006, and the regression explains between 51% and 63% of the monthly variance of the book.

Section 3.7 fixes three inference tiers before any result is read, and assigns each comparison to one of them. Table C.3 is the resulting bookkeeping: which comparisons belong to which family, and how many survive before and after the correction that family carries.

The primary endpoint is a single contrast specified in advance, so there is no family

to correct and none is applied. It clears the threshold gross and net.

The secondary family is the same contrast taken objective by objective, and Panel B gives it under both cost regimes. Four of the five objectives are significant on unadjusted values in each regime; Romano–Wolf leaves three gross and two net. RankNet and LambdaRankIC survive everywhere, Soft-Spearman survives gross only, and Margin ranking and RankCosine survive neither. That spread is the reason the pooled contrast is the primary endpoint and the per-objective pattern is secondary: the pooled figure is estimated on all five, while any single objective is estimated on 272 months of one noisy cell.

The exploratory family needs a caveat that the counts alone do not convey. All fifteen netting contrasts are significant, which is not a discovery, and because their p -values are the smallest in the family they take the lowest ranks. The Benjamini–Hochberg critical value at rank k grows with k , so every remaining test is compared against a larger threshold than it would face in a family without them.

Removing them is therefore the sharper reading of Panel C, and it separates the axes cleanly. Of the tests that remain once the netting contrasts are set aside, the only two that clear $q = 0.10$ are ordering statistics: the information ratio of the rank information coefficient and the asymmetry between the two tails. The width family taken on its own yields nothing from sixty comparisons, and the weighting family nothing from fifteen. The single width contrast that appears to survive in the full family does so on the composition of the family it sits in rather than on its own evidence, and it should not be read as a separable result.

That pattern is the thesis in miniature. What survives multiplicity correction is the evidence on how well the scores order the cross-section, which is where Chapter 5 locates the effect. The two construction axes do not separate, which is why Section 6.2 treats the width grid as descriptive and Chapter 9 reports it as the axis the evidence separates least.

The netting contrasts stay in the family even so. Removing comparisons because their outcome is not in doubt is selection on the outcome, and the alternative to keeping them is a family whose membership was decided after the p -values were seen. Keeping them is the right bookkeeping and it does make the correction more permissive for everything else, which is why both counts are reported here rather than only the favourable one.

Table C.3: Family membership for each inference tier, and what the multiplicity correction removes. The tiers are fixed in Section 3.7 before any result is read. Survivors are counted at the conventional threshold for the tier: $p < 0.05$ for the primary and secondary families and $q < 0.10$ for the exploratory one. The two right-hand columns of Panel A differ only through the correction, so the gap between them is what multiplicity control costs. Panel B gives the secondary family under both cost regimes, the gross half of which is Table 5.4. Panel C breaks the exploratory family into the questions it answers.

Panel A. Tiers						
Tier	Family	m	Correction	Raw	Adjusted	
Primary (gross)	Pooled Δ^{repr}	1	None	1	1	
Primary (net)	Pooled Δ^{repr}	1	None	1	1	
Secondary (gross)	Per-objective Δ^{repr}	5	Romano–Wolf	4	3	
Secondary (net)	Per-objective Δ^{repr}	5	Romano–Wolf	4	2	
Exploratory	4 question groups, Panel C	97	Benjamini–Hochberg	24	18	

Panel B. Secondary family, both regimes						
Objective	Gross			Net		
	ΔSR	t	p_{RW}	ΔSR	t	p_{RW}
RankNet	+0.481	+4.47	0.0005	+0.404	+3.99	0.0012
LambdaRankIC	+0.503	+4.82	0.0002	+0.434	+4.40	0.0005
Margin ranking	+0.275	+2.04	0.1155	+0.265	+2.03	0.1215
Soft-Spearman	+0.333	+2.62	0.0362	+0.252	+2.08	0.1215
RankCosine	+0.257	+1.83	0.1155	+0.183	+1.41	0.1830

Panel C. Exploratory family			
Question	Tests	Raw $p < 0.05$	BH $q < 0.10$
Traded fraction against the decile	60	4	1
Net against gross	15	15	15
Equal against inverse-volatility	15	2	0
Ordering quality, arm against arm	7	3	2
All	97	24	18

C.3 Per-seed results

Every neural figure in this thesis comes from an ensemble of five seeds, and the comparison of Section 5.7 sets the standard that a representation effect smaller than the spread across seeds is not a finding. Table C.4 is the evidence for that comparison, computed from the retained per-seed score vectors and replayed through the same portfolio rules as everything else.

Panel A shows that the dispersion is not small. Taken one at a time, the twenty-five individual seeds put Δ^{repr} anywhere between -0.044 and $+0.500$, and the standard deviation across seeds within an objective averages 0.124 . Two of the twenty-five are negative. A reader who ran a single seed and stopped would have found $+0.192$ on average with that much noise around it, which is not a result anyone should report.

Panel B is the part that answers the objection rather than restating it. As seeds are added the contrast does not decay toward zero, which is what a draw would do. It rises at every step, from $+0.192$ at one seed to $+0.370$ at five. Both arms improve under averaging, so the growth is not an artefact of ensembling one of them: the percentile arm gains 0.111 of Sharpe ratio between one seed and five and the inverse-normal arm gains 0.290 . The reported contrast is a property of the converged ensemble, and since the curve has not flattened by the fifth seed, it is more likely to understate the limit than to overstate it.

The asymmetry has a reading, and it is worth separating from the obvious one. It is not that the inverse-normal arm trains more stably. Seed by seed the two arms are almost equally noisy, with mean within-objective standard deviations of 0.161 and 0.176 for the percentile and inverse-normal levels respectively. What differs is how much of a single seed’s ordering is peculiar to its initialisation and therefore cancels when seeds are averaged. More of it cancels in the inverse-normal arm. That is a statement about the representation the network is given, which is the subject of Chapter 5, and not about the optimisation.

One consequence reaches past this thesis. A learning-to-rank study run on one seed, or on a small ensemble, will recover roughly half of a representation effect of this kind and may well conclude there is nothing there. The ensemble size is not a free parameter tuned on the outcome here: five seeds were fixed in advance, the curve is monotone over its whole range, and every value of m on it gives the same sign.

Table C.4: The representation contrast Δ^{repr} across individual seeds, and across ensemble size. Panel A gives the contrast computed from each seed on its own; the last column is the five-seed ensemble actually reported. Panel B averages over every subset of seeds of a given size, so the row at m pools $\binom{5}{m}$ subsets in each of the five objectives. Both arms gain from averaging and the inverse-normal arm gains more, so the contrast widens monotonically in m rather than converging to zero. A study run on one seed would have recovered roughly half of it. Dispersion is summarised by the mean of the five per-objective standard deviations, matching the pooling operator used throughout Chapter 5.

Panel A. Δ^{repr} by seed							
Objective	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	SD	Five-seed
RankNet	+0.268	+0.142	+0.329	+0.213	+0.221	0.069	+0.481
LambdaRankIC	+0.097	+0.135	+0.285	+0.262	+0.286	0.090	+0.503
Margin ranking	+0.010	−0.003	+0.306	+0.197	+0.052	0.134	+0.275
Soft-Spearman	−0.044	+0.171	+0.394	+0.016	+0.222	0.174	+0.333
RankCosine	+0.082	+0.244	+0.500	+0.189	+0.213	0.155	+0.257
Pooled						0.124	+0.370

Panel B. Ensemble size					
Seeds	m	Percentile	Inverse-normal	Δ^{repr}	Subsets
1		0.990	1.182	+0.192	25
2		1.050	1.329	+0.279	50
3		1.076	1.389	+0.313	50
4		1.081	1.419	+0.339	25
5		1.102	1.471	+0.370	5

Going from one seed to $m = 5$ raises the percentile arm by +0.111 of Sharpe ratio and the inverse-normal arm by +0.290. The gap is +0.192 at $m = 1$ against +0.370 at $m = 5$, and it rises at every step in between. The two arms are equally noisy seed by seed: the mean within-objective standard deviation of the level is 0.161 for the percentile arm and 0.176 for the inverse-normal arm, so the widening is not a difference in seed stability.

C.4 The short-squeeze reading of the terminal episode

Section 8.2 rejects a short squeeze as the principal mechanism behind the terminal episode. The measurements behind that verdict are given here. The window is the twelve returns earned from October 2024 to September 2025: the eleven-month gross episode of Table 8.4 together with the month that follows its trough.

A squeeze concentrates in names that are small, illiquid and expensive to borrow. The short book did the opposite: its median constituent traded about \$70–82 million a day before the episode and between \$184 and \$219 million during it, and the share of short positions below the \$20 million threshold that Table 4.5 treats as hard to borrow falls from between 12.5% and 13.7% to between 1.3% and 1.9%. The book was less exposed to borrow scarcity during the episode than at any earlier point, not more. Nor is the loss the work of a few positions: measured as the share of a losing month contributed by its five worst names, concentration over the episode runs between 26.2% and 33.8% for the ten neural cells against between 31.5% and 34.4% historically, so the typical bad month of the

episode is no more concentrated than the typical bad month before it, and the median shorted name rose in each of the four worst months. The gradient-boosted control is the one exception, its worst five names carrying 43.0% against 34.7% historically. One squeeze-consistent trace does survive: the right tail of shorted-name returns thickens, the share rising above 25% in a month going from between 1.9% and 2.6% to between 4.5% and 6.3%, and above 50% from between 0.2% and 0.3% to between 0.7% and 1.8%.

C.5 Tail-width tables

Section 6.2 reads the width grid on figures averaged over the five objectives in the inverse-normal arm. Table C.5 is the unpooled turnover behind that average.

Only one of the three columns of Table 6.1 is worth showing cell by cell. Sharpe ratios across the same grid are already the complete cross of Table C.1 and Table C.2. The realised name count is fixed by the fraction on a universe of 1000 names, so it is two hundred, one hundred, fifty, twenty-five and ten down every column, the two rebalances of 999 names rounding to the same counts, and a table of it would carry no information. The rounding of (4.10) is unchanged by the missing name, and the floor $k_{\min}$ is met but nowhere binds. Turnover is one-side net per rebalance, measured by (4.14) against the drifted book of (4.15). Turnover is the column that varies by specification, and it is the one the width decision is carried on, so it is the one given here.

What the table mostly shows is an absence. Turnover is a property of the traded fraction and very nearly nothing else: within an inverse-volatility column the standard deviation across the twelve specifications runs between 0.045 and 0.057, against a range of 1.16 across that half of the grid. The width moves turnover roughly twenty times as far as the choice of objective and encoding does. That is worth stating because it licenses the treatment of turnover in Chapter 7 as a function of α , which the cost bridge assumes when it decomposes the stack by width rather than by cell.

Two regularities survive that flatness, and both bear on claims made elsewhere.

The gradient-boosted control records identical turnover in its two arms at every fraction under both schemes. Table C.6 establishes that its two arms fit the same trees; this extends the same statement to the traded book across the whole grid, which is a stronger check than the headline configuration alone provides, and one the main text does not make.

The inverse-normal arm turns over more than the percentile arm at every fraction, by between 0.057 and 0.094 per rebalance across the ten neural cells under inverse-volatility sizing; the control is left out of that average, its two arms being identical. The direction matters for how the net results of Section 7.3 are read. The arm that nets more does so while paying more to trade, not less, so the net advantage is earned against a turnover penalty rather than assisted by a turnover saving. A reader who suspected the represen-

tation contrast of surviving costs because the favoured arm happened to trade less can see from this table that the opposite holds.

The leg decomposition of Table 6.3 is not extended across widths. The run computes it at the headline configuration only.

Table C.5: One-side net turnover per rebalance, for every specification at every traded fraction under both weighting schemes. The unpooled version of the turnover column of Table 6.1. Sharpe ratios across the same grid are in Table C.1 and Table C.2 and are not repeated; the realised name count is fixed by the fraction on a universe held at 1000 names, so it is constant down every column and is not shown. Turnover is measured against the drifted book, so a position whose target weight is unchanged still contributes the trade that restores it.

Objective	Arm	Inverse-volatility					Equal				
		20%	10%	5%	2.5%	1%	20%	10%	5%	2.5%	1%
RankNet	Percentile	2.292	2.648	2.905	3.108	3.338	2.202	2.560	2.826	3.044	3.296
	Inverse-normal	2.406	2.768	3.004	3.195	3.411	2.320	2.678	2.928	3.120	3.359
LambdaRankIC	Percentile	2.293	2.656	2.925	3.140	3.379	2.217	2.583	2.858	3.082	3.343
	Inverse-normal	2.403	2.769	3.026	3.217	3.426	2.333	2.696	2.954	3.145	3.381
Margin ranking	Percentile	2.269	2.631	2.892	3.076	3.290	2.174	2.543	2.807	2.999	3.239
	Inverse-normal	2.338	2.702	2.963	3.167	3.365	2.250	2.617	2.878	3.094	3.307
Soft-Spearman	Percentile	2.310	2.654	2.887	3.079	3.296	2.216	2.559	2.805	3.010	3.242
	Inverse-normal	2.397	2.726	2.953	3.139	3.332	2.301	2.631	2.871	3.067	3.289
RankCosine	Percentile	2.309	2.658	2.905	3.098	3.299	2.218	2.567	2.821	3.024	3.247
	Inverse-normal	2.398	2.742	2.975	3.154	3.356	2.306	2.650	2.896	3.086	3.320
XGBoost (control)	Percentile	2.418	2.758	2.996	3.171	3.400	2.322	2.663	2.903	3.088	3.336
	Inverse-normal	2.418	2.758	2.996	3.171	3.400	2.322	2.663	2.903	3.088	3.336

Turnover is a property of the fraction rather than of the model. Across the twelve specifications the standard deviation within an inverse-volatility column runs from 0.045 to 0.057, against a range of 1.16 across that half of the grid, so the width moves turnover roughly twenty times as far as the choice of objective and arm does. Two regularities survive that flatness. The gradient-boosted control records identical turnover in its two arms at every fraction under both schemes, which extends the invariance of Table C.6 from the fitted trees to the traded book. And the inverse-normal arm turns over more than the percentile arm at every fraction, by between 0.057 and 0.094 averaged over the five objectives under inverse-volatility sizing, so the arm that nets more in Table 7.2 does so while paying more to trade, not less.

C.6 Gradient-boosted benchmark

Section 4.5 predicts that a threshold-splitting learner is invariant to a strictly monotone re-encoding of its features, and qualifies the prediction with two residual channels through which the encoding could still reach it. This section reports the configuration those channels depend on, and then reports whether either one moved anything.

The first channel is the one worth taking seriously. With `tree_method` set to `hist`, candidate split points are not the observed values but the edges of a bucketed sketch of them, and Table C.6 records that `max_bin` is left at the library default of 256. A monotone re-encoding redistributes observations within the range of a feature without reordering them, so a bucket edge that fell between one adjacent pair under the percentile encoding can fall between a different pair under the inverse-normal one. Where that happens the two arms are offered different candidate splits and may take them. The second channel, floating-point representation, is smaller and works the same way.

Panel B is the audit. Across all 214 trees of all five seeds, the two arms agree on the feature chosen at every node, on the topology of every tree, on the gain realised at every split, on the value at every leaf, on the hessian mass at every node and on the direction taken for missing values. Gain, weight and cover agree over all 144 characteristics. One field differs, and it is the one that has to: the split thresholds, which are recorded in the units of whichever encoding produced them.

The two arms therefore did not fit similar models. They fit the same model, and the exact agreement Section 5.5 reports on the Sharpe ratio, the rank information coefficient and the drawdown profile is a consequence of that rather than an artefact of the printed precision. Neither residual channel was active on this data, so the bound they were introduced to provide is zero here, and the invariance the control was built to test holds exactly rather than approximately.

Table C.6: Resolved configuration of the gradient-boosted control, and how exactly its two arms agree. Panel A is the configuration as resolved at run time. `max_bin` is not set, so the library default of 256 applies; with `tree_method` at `hist` that is the histogram resolution Section 4.5 identifies as one of the two channels through which a monotone re-encoding could still reach a threshold-splitting learner. Panel B compares the two fitted arms field by field across all 214 trees of each of the five seeds, taken from the final fold, which is the fit the run retains. Split thresholds are expressed in the units of whichever encoding produced them and so must differ; everything that determines a prediction does not.

Panel A. Resolved configuration	
Setting	Value
<code>objective</code>	<code>rank:pairwise</code>
<code>tree_method</code>	<code>hist</code>
<code>max_depth</code>	4
<code>eta</code>	0.03
<code>min_child_weight</code>	20
<code>subsample</code>	0.8
<code>colsample_bytree</code>	0.7
<code>gamma</code>	0.0
<code>reg_alpha</code>	0.1
<code>reg_lambda</code>	10.0
<code>lamdarank_pair_method</code>	<code>mean</code>
<code>lamdarank_num_pair_per_sample</code>	4
<code>validation_metric</code>	<code>rankic</code>
<code>max_bin</code>	256 (library default, not set)
<code>num_boost_round</code>	800 with early stopping at 75
Trees retained after early stopping	214 in every seed
Library version	xgboost 3.2.0
Panel B. Percentile arm against inverse-normal arm	
Quantity compared	Result
Feature chosen at every node	Identical
Tree topology (left children)	Identical
Tree topology (right children)	Identical
Gain at every split	Identical
Leaf and node values	Identical
Hessian mass at every node	Identical
Missing-value direction	Identical
Split thresholds	Differs, in encoding units
Gain, weight and cover over 144 features	Identical

Use of generative artificial intelligence

Generative AI was used in the preparation of this thesis. Anthropic’s Claude, accessed through the Claude Code command-line interface, assisted between February and August 2026 with the implementation of the research pipeline, the generation of tables and figures, the drafting and revision of prose, and the auditing of numerical claims against the stored run artefacts. The research questions, the experimental design, the pre-specified protocol and the interpretation of the results are my own. Every reported number is produced by the pipeline of Chapters 3 and 4 and is reproducible from the stored artefacts, and every reference has been verified against its original source. I remain solely responsible for the content of this thesis, including any remaining errors.