BEYOND NOMINAL WIDTH: COMPARISON-NOISE GEOMETRY OF MODEL VALIDATION UNDER CROSS-SECTIONAL DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated model discovery makes candidate generation cheap, but comparing candidates requires more than a large validation set. Counting evaluation units ignores both their dependence and how shared fluctuations affect differences between candidates. We introduce comparison effective width to characterize this comparison-noise geometry. The measure pairs candidate-centered score geometry with dependence among evaluation units. A projection-sufficiency result establishes that comparison rules invariant to common additive shifts depend only on centered candidate evaluation scores. Exact identities link comparison effective width to a comparison design effect and a second-order noise descriptor, all computable before reading evaluation outcomes. In label-free measurements across 2,258 matched cells from CSI300, a large-cap Chinese equity universe, as nominal width grows from 30 to 120, comparison effective width relative to nominal width falls from 0.726 to 0.389. The cell-level per-added-asset slope declines in 2,258/2,258 matched cells. CSI500, a disjoint universe in the same market, shows the same descriptive pattern with weaker attenuation. A same-candidate decomposition locates the CSI500–CSI300 gap in log comparison effective width mainly in return-side dimensionality and score–return alignment. Controlled synthetic experiments further show that the eigenvalue spectrum of the full covariance matrix alone does not determine split-half ranking agreement. An empirical relation combining cross-half candidate-IC covariance with zero-signal split-half IC-difference dispersion predicts this agreement on held-out synthetic worlds with cross-fitted $R^2 \approx 0.88$ and systematic calibration deviations. Together, these results show why nominal evaluation counts should be complemented by a measure that accounts for evaluation-unit dependence and candidate-score geometry.

1 INTRODUCTION

How much does adding evaluation units help distinguish competing models? Nominal evaluation width is easy to report, but it does not specify the noise faced by a ranking rule. The relevant structure involves multiple candidates, a shared validation target, and dependence among evaluation units: common fluctuations can affect candidates together, while other fluctuations alter their ordering. This distinction motivates measuring the noise that remains relevant to comparison. We study one instance where evaluation-unit dependence can be estimated before evaluation, rather than claiming empirical coverage across application domains.

Marginal participation ratio and effective rank describe spectral diversity within an evaluation population (Roy & Vetterli, 2007), but model comparison also depends on how evaluation-unit dependence interacts with differences between candidates. Two joint evaluation designs can therefore share the full covariance spectrum yet induce different comparison-noise levels for the same candidate family. Classical dependence-adjusted inference already pairs covariance structures (Clifford et al., 1989; Dutilleul, 1993), but our comparison setting introduces an additional question: which part of the candidate-score geometry should be paired with evaluation-unit dependence? If a comparison rule is unchanged when the same additive shift is applied to every candidate's comparison statistic, then candidate-common fluctuations do not affect the decision. Proposition 1 formalizes this invariance and shows that the rule depends only on the candidate-centered comparison vector. This reduction

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

motivates pairing candidate-centered score geometry with evaluation-unit dependence. The resulting comparison effective width n_{eff}^{c} depends on both the marginal dimensionality of these two structures and their relative alignment, with exact algebraic links to a comparison design effect and a second-order noise descriptor, D_{geom}.

To examine what marginal spectra leave unresolved, we use controlled synthetic interventions that preserve the eigenvalue spectrum of the full joint evaluation-unit covariance while changing its orientation relative to candidate-score geometry. These interventions vary geometric information that is invisible to the full-covariance spectrum and examine its consequences for comparison noise and split-half ranking reliability. Geometry, however, characterizes only the noise side of model comparison; reliability also depends on signal. This distinction is consistent with work separating signal and noise in evaluation (Heineman et al., 2025). We examine split-half reliability through a separate empirical decomposition into a cross-half agreement channel and a noise channel estimated from zero-signal replicates, and evaluate cross-fitted predictions on held-out synthetic worlds. Geometry alone does not establish which candidate is better.

We instantiate the framework in cross-sectional factor evaluation, where large candidate families make the comparison problem practically relevant (Yu et al., 2023; Shi et al., 2025). Here, each candidate is a predefined signal that assigns a score to every asset, each asset is an evaluation unit, and candidates are compared by the Spearman rank correlation (RankIC) between their scores and subsequent returns. This setting also lets us estimate evaluation-unit dependence before evaluation: asset dependence is estimated from prior return history, while point-in-time membership and pre-evaluation data availability define the geometry-eligible support without filtering on future-label availability. Nested subsets of these eligible assets then let us vary evaluation width under a matched design. Our main measurements use two disjoint Chinese A-share universes, the large-cap CSI300 and the smaller-cap CSI500, evaluated with the same candidate roster and market calendar. We additionally examine a study-specific exchange-traded fund (ETF) panel to probe a structurally different dependence regime, treating it as boundary evidence rather than as a third replication benchmark. The reported real-data geometry measurements do not use future evaluation labels and do not test realized candidate-selection reliability. Whether the estimated dependence adequately characterizes actual comparison errors remains a separate empirical question.

Our contributions are threefold:

- **Orientation beyond marginal spectra.** Covariance-spectrum-preserving interventions show that orientation relative to candidate scores changes split-half ranking reliability while the full covariance spectrum remains fixed. A relation between a cross-half agreement channel and a zero-signal noise channel predicts this reliability with cross-fitted $R^2 \approx 0.88$ on held-out synthetic worlds, alongside a structured calibration bias.
- **Comparison-specific geometry and effective width.** Candidate-shift invariance motivates projecting out candidate-common variation before pairing score geometry with evaluation-unit dependence. The resulting comparison effective width n_{eff}^{c} has exact algebraic links to a comparison design effect and the second-order noise descriptor D_{geom}.
- **Matched-width structural measurements.** Label-free matched-width measurements on two equity universes show a declining per-added-asset gain in comparison effective width over $m = 30$–120; a same-roster decomposition and an ETF boundary locate the differences in return-side dimensionality and alignment.

2 Related work

Evaluation validity under repeated and dependent evaluation. Repeated interaction with a fixed evaluation signal can bias model comparison; reusable holdouts, leaderboard mechanisms, and information-theoretic analyses formalize aspects of this adaptive-validity problem (Dwork et al., 2015a;b; Blum & Hardt, 2015; Russo & Zou, 2016). Empirical work likewise documents model-selection overfitting and its prevalence across machine-learning studies (Cawley & Talbot, 2010; Roelofs et al., 2019; Mania et al., 2019). More recent benchmark analyses show that evaluation composition, item dependence, underspecification, and aggregation rules can materially change uncertainty or rankings (Siska et al., 2024; Wein et al., 2023; Jayawardana et al., 2022; Colombo et al., 2022; Reuel et al., 2024). We ask a complementary pre-outcome question: before repeated

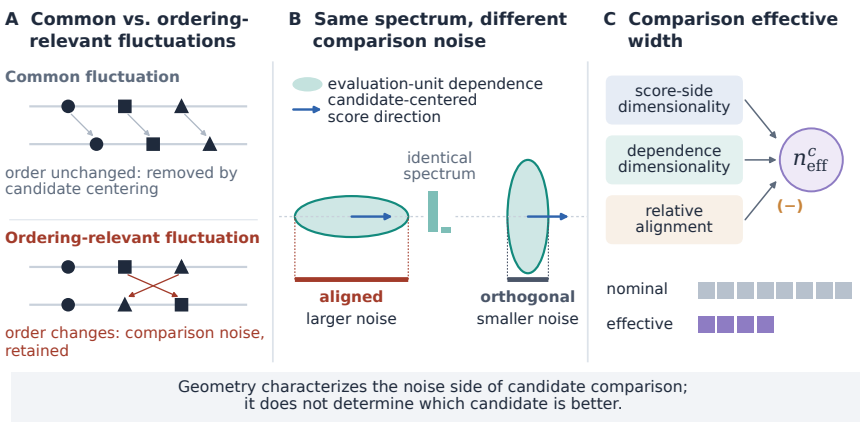

Figure 1: Comparison-noise geometry. (A) Fluctuations shared by all candidates leave their order unchanged and are removed by candidate centering; ordering-relevant fluctuations remain as comparison noise. (B) Two joint dependence structures with the same full-covariance spectrum produce different comparison noise, depending on whether the dependence is aligned with or orthogonal to candidate-centered score directions. (C) The comparison effective width n_{eff}^{c} combines score-side dimensionality, dependence dimensionality, and their relative alignment, which enters with a negative sign. Geometry characterizes the noise side of candidate comparison and does not determine which candidate is better.

reuse or realized-ranking correction, how does cross-sectional dependence shape the noise side of candidate comparison?

Effective dimensionality and comparison-relevant geometry. Classical dependence-adjusted inference already uses paired covariance corrections (Clifford et al., 1989; Dutilleul, 1993); these trace-overlap formulations provide the algebraic starting point, but they do not specify which score geometry is relevant when a decision rule ranks many candidates and ignores their common shifts. Effective degrees of freedom (Bretherton et al., 1999) and marginal eigenvalue summaries (Li & Ji, 2005; Roy & Vetterli, 2007) provide related foundations. ENSD combines participation ratios and alignment to measure shared dimensions (Giaffar et al., 2024); alignment has the opposite directional role in our width. Kernel alignment, centered alignment, and CKA are established ingredients (Cristianini et al., 2001; Cortes et al., 2012; Kornblith et al., 2019). Our contribution concerns the comparison-specific projection and interpretation, rather than alignment algebra itself. Evaluation signal/noise decomposition (Heineman et al., 2025) further clarifies why noise geometry alone does not establish predictive quality.

Automated search and financial evaluation. Automated alpha discovery provides a concrete setting in which large candidate families are generated, combined, and repeatedly screened (Yu et al., 2023; Shi et al., 2025; Luo et al., 2026; Chen et al., 2026). Related cross-sectional forecasting systems emphasize strong predictive architectures and evaluation pipelines (Fan & Shen, 2024; Li et al., 2024; Song et al., 2026), while finance-wide replication evidence highlights the broader need to distinguish robust evidence from large candidate inventories (Jensen et al., 2023). Our contribution is neither a new search algorithm nor a forecasting architecture: it is a measurement and mechanism framework for comparison-relevant validation geometry under dependence.

3 Comparison-Noise Geometry

3.1 Projection Sufficiency and Geometric Objects

Consider one evaluation date with n evaluation-time geometry-eligible assets and K locally estimable candidates. The support U_t intersects the fixed eligible universe, point-in-time membership, and finite scores for all candidates; future-label availability does not filter this support. Scores are rank-transformed, centered, and standardized (sample variance, denominator $|U_t| - 1$) on full U_t, then restricted to form $Z \in \mathbb{R}^{n \times K}$ without re-ranking or re-centering. The strictly pre-evaluation matrix

C_r is the sample correlation of cross-sectionally demeaned close-to-close returns over the 250 returns before the origin, without shrinkage. Assets with missing required histories are excluded; restriction uses principal submatrices.

Let $P_n = I_n - \mathbf{1}\mathbf{1}^\top/n$ and $P_K = I_K - \mathbf{1}\mathbf{1}^\top/K$. On the asset axis, Spearman RankIC is Pearson correlation after midrank transformation and therefore depends only on centered midrank vectors. On the candidate axis, comparison rules such as $\arg\max$ or top-m selection are invariant to common additive shifts.

Proposition 1 (Projection sufficiency)**.** *(i) For midrank vectors u, w, their centered correlation is unchanged by positive rescaling and common shifts. (ii) If $\delta(x + c\mathbf{1}) = \delta(x)$ for every c, then $\delta(x) = \delta(P_K x)$, so additive perturbations affect the decision only through $P_K\varepsilon$. The candidate-centered width inherits this invariance. (iii) If ε is Gaussian with covariance Σ, or elliptical with dispersion Σ and fixed generator, the projected noise law is determined by $P_K\Sigma P_K$ within that family.*

Proof. Since $P_n\mathbf{1} = 0$, common shifts vanish and positive scales cancel in correlation. For candidate comparison, write $x = P_K x + \bar{x}\mathbf{1}$ and apply translation invariance. Gaussian and elliptical families are closed under linear projection with covariance or dispersion transformed by congruence. □

Four distinct objects. For nonzero positive semidefinite C, let $\mathrm{PR}(C) = \mathrm{tr}(C)^2/\mathrm{tr}(C^2)$. Define

$$d_{\mathrm{ret}} = \mathrm{PR}(C_r), \qquad C_s^c = K^{-1} Z P_K Z^\top, \tag{1}$$

$$n_{\mathrm{eff}}^c = \frac{\mathrm{tr}(C_s^c)\,\mathrm{tr}(C_r)}{\mathrm{tr}(C_s^c C_r)}, \tag{2}$$

$$\Sigma_{\mathrm{pred}} = \frac{Z^\top C_r Z}{(n-1)^2}, \qquad v_{\mathrm{cand}} = K^{-1}\,\mathrm{tr}(\Sigma_{\mathrm{pred}}), \tag{3}$$

$$\Sigma_c = P_K \Sigma_{\mathrm{pred}} P_K, \quad d_{\mathrm{search}} = \mathrm{PR}(\Sigma_c), \quad D_{\mathrm{geom}} = \frac{\mathrm{tr}(\Sigma_c)}{K-1}. \tag{4}$$

These describe return-only structure, comparison width, marginal candidate noise, and centered search directions, respectively. Nonzero trace overlap and $K, n > 1$ are required for the displayed ratios. Equation 3 is a second-order descriptor; how closely it approximates actual Spearman RankIC comparison noise is an empirical question.

Writing $C_s = ZZ^\top/K$ and $\bar{z} = Z\mathbf{1}_K/K$ gives $\mathrm{tr}(C_s) = \mathrm{tr}(C_s^c) + \|\bar{z}\|^2$. The centered width is invariant to $Z \mapsto Z + b\mathbf{1}_K^\top$; the original uncentered width can change (by up to 33.5 in the numerical check). The supplement gives a lower bound and an example exceeding ambient dimension.

Exact comparison identities. Taking traces after candidate projection yields

$$D_{\mathrm{geom}} = \frac{K\,\mathrm{tr}(C_s^c)\,\mathrm{tr}(C_r)}{(K-1)(n-1)^2 n_{\mathrm{eff}}^c}. \tag{5}$$

Using the term design effect in its classical sense (Kish, 1965), define the comparison design effect

$$\mathrm{DEFF}_{\mathrm{cmp}} = \frac{n}{n_{\mathrm{eff}}^c} = \frac{\mathrm{tr}(\Sigma_c)}{\mathrm{tr}(\Sigma_c^{\mathrm{iso}})}, \tag{6}$$

where Σ_c^{iso} substitutes $(\mathrm{tr}(C_r)/n)I$ for C_r. These are algebraic identities, not outcome-validity results. In the real-data nested-width protocol, only candidate-axis projection is applied to restricted Z: full-support standardization does not commute with restriction, so $P_n Z$ need not equal Z.

Observed outcome-side target. For a frozen replicate, the realized support is split into fixed halves A and B. Spearman ranks are formed locally within each actual half for the observed statistic, while Z_h and $C_{r,h}$ in the descriptor are obtained only by restriction of their frozen pre-label objects. With a common locally estimable candidate mask,

$$D_{\mathrm{obs}} = \tfrac{1}{2}\,\mathrm{Var}_k[\mathrm{RankIC}_A(k) - \mathrm{RankIC}_B(k)]\,. \tag{7}$$

This is an observed split-half dispersion, not pure sampling noise by definition: differencing removes components common to the halves but does not eliminate support-specific signal heterogeneity.

3.2 PROTOCOL AND EVIDENCE UNITS

Empirical setting. We study CSI300 and CSI500, two disjoint Chinese A-share equity universes traded in Shanghai and Shenzhen. CSI300 contains large, liquid securities; CSI500 selects from the eligible population outside CSI300 under size and liquidity rules (China Securities Index Co., Ltd., 2023; 2022). Detailed index construction is given in the supplement. The pairing holds the national market, calendar, and institutional environment largely fixed, so CSI500 is a same-market cross-universe replication, not an independent-market replication.

Study-specific ETF panel. The ETF testbed is different by design: it is a study-specific point-in-time panel assembled for this work from historical ETF market and instrument records, not an off-the-shelf public benchmark dataset. Universe support, return-continuity segments, feature-scale continuity, and corporate-action boundaries were frozen before paper-facing geometry was computed. The direct ETF–CSI300 comparison is restricted to the mechanically maximal 137-date common calendar at $m = 60$ and ETF origins O01/O02.

Candidate family. Both universes use the same frozen Qlib Alpha158 candidate family, a public hand-crafted daily equity feature set implemented in Qlib (Yang et al., 2020). The label column is excluded, and the ordered 158-candidate roster is fixed across universes. Because those candidates are evaluated on different asset populations, identical candidate identity does not imply identical score-side geometry.

The empirical design uses this common 158-dimensional candidate roster at three nominal widths $m \in \{30, 60, 120\}$. For each origin-date cell, a deterministic random priority defines nested asset subsets. The same priority is reused across widths so that the width response is a within-cell intervention on asset support rather than a comparison across calendar periods. Score geometry is computed on the full finite point-in-time support before width restriction; return geometry is origin-fixed and then restricted by principal submatrix. Width therefore changes the selected support but does not trigger score refitting, reranking, covariance re-estimation, replacement assets, or outcome-conditioned filtering.

The primary statistical unit for the width-response pattern is an origin-date cell. The 32 deterministic randomizations quantify composition sensitivity; they are not treated as 32 independent observations. For a cell i, let $\bar{n}_i(m)$ denote n_{eff}^c averaged across its 32 replicates. We summarize local marginal gain by

$$s_i^{30\to60} = \frac{\bar{n}_i(60) - \bar{n}_i(30)}{30}, \qquad s_i^{60\to120} = \frac{\bar{n}_i(120) - \bar{n}_i(60)}{60}, \tag{8}$$

and local log–log elasticity by the corresponding finite differences in $\log \bar{n}_i$ over $\log m$. The primary descriptive statement concerns these cell-level aggregates. Replicate-level directional counts are reported separately because randomized compositions are dependent within a cell.

All real-data structural experiments in Section 4 are label-free. Section 5 uses controlled synthetic outcomes and synthetic measurement stress with real geometry only. No real held-out validation is reported.

4 REAL-WORLD STRUCTURAL EVIDENCE

4.1 CSI300 AND CSI500 MATCHED-WIDTH RESPONSE

The nested-support design compares nominal widths within the same origin-date cell. Each universe has 2,258 matched cells and 32 outcome-independent randomizations per cell, giving 216,768 width-specific rows. Cells, rather than dependent replicates, are the reporting unit.

CSI300 mean n_{eff}^c is 21.77, 33.69, 46.63, with n_{eff}^c/m of 0.726, 0.562, 0.389. Per-added-asset gain falls from 0.397 to 0.216. The second-interval slope is smaller in 2,258/2,258 cells, as is local elasticity; replicate-level elasticity declines in 55,749/72,256 comparisons (77.2%).

CSI500 mean n_{eff}^c is 26.87, 48.34, 80.94, with n_{eff}^c/m of 0.896, 0.806, 0.675. Per-added-asset gain falls from 0.716 to 0.543. The second-interval slope is smaller in 2,258/2,258 cells, as is local elasticity; replicate-level elasticity declines in 55,893/72,256 comparisons (77.4%).

For CSI300, the absolute increment is larger over the second interval (12.94 versus 11.92) because twice as many assets are added. CSI500 shows weaker attenuation. These are descriptive regularities over $m = 30$–120, not saturation or a scaling law. Candidate centering changes CSI300 cell values by -0.9% to $+4.1\%$; the median common-component share is 0.009. Uncentered results are reported in the supplement.

4.2 CSI500–CSI300 DESCRIPTIVE DECOMPOSITION

Define

$$\mathrm{ALIGN}^c(C_s^c, C_r) = \frac{\mathrm{tr}(C_s^c C_r)}{\sqrt{\mathrm{tr}((C_s^c)^2)\,\mathrm{tr}(C_r^2)}}, \qquad n_{\mathrm{eff}}^c = \frac{\sqrt{\mathrm{PR}(C_s^c)\mathrm{PR}(C_r)}}{\mathrm{ALIGN}^c(C_s^c, C_r)}. \tag{9}$$

For matched rows from the common 158-candidate roster,

$$\Delta \log n_{\mathrm{eff}}^c = \tfrac{1}{2}\Delta \log \mathrm{PR}(C_s^c) + \tfrac{1}{2}\Delta \log \mathrm{PR}(C_r) - \Delta \log \mathrm{ALIGN}^c. \tag{10}$$

The cell-level decomposition has a stable hierarchy at all widths: return-side dimensionality is the dominant descriptive term, alignment is a substantial secondary term, and the score-side term is small and directionally mixed. At $m = 120$, the mean log gap is 0.5606, comprising $+0.3413$ from return-side dimensionality, $+0.2020$ from lower CSI500 alignment, and $+0.0173$ from the score side. Lower ALIGN^c mechanically raises n_{eff}^c because it is in the denominator of equation 9; this is comparative noise geometry, not predictive signal quality.

Return-side separation persists after dimension normalization: across 21 paired origins, median $\mathrm{PR}(C_r)/n$ is 0.1401 for CSI300 and 0.2424 for CSI500, and leading-eigenvalue and top-10 spectral shares are lower for CSI500 in all 21 (Figure 3).

4.3 ETF CROSS-DOMAIN BOUNDARY

The ETF panel is a study-specific point-in-time fund panel built under separate support and continuity rules, and it serves only as cross-domain structural boundary evidence relative to the same-market CSI500 replication. On the mechanically maximal 137-date common calendar (2020-03-31 to 2020-12-07, ETF origins O01/O02, $m = 60$), score-side participation is nearly matched: $\mathrm{PR}(C_s)/m = 0.1046$ for ETF versus 0.1031 for CSI300. Return-side participation differs sharply, $\mathrm{PR}(C_r)/m = 0.0744$ versus 0.4125, and the original (uncentered) paired effective dimension n_{eff}/m is 0.1618 versus 0.5818. ETF also has stronger return concentration and higher alignment position (Table 1, panel C). A frozen near-duplicate-reduced sensitivity preserves 98.0% of the $\mathrm{PR}(C_r)/m$ gap, 94.9% of the paired-n_{eff}/m gap, and 78.5% of the alignment-position gap, computed on the same 68 retained O02 dates for both arms. This weakens but does not eliminate the near-duplicate explanation; O01–O12 persistence is an ETF-only extension. Alignment position is a spectral-bound-normalized trace overlap (defined in the supplement); it differs from ALIGN^c.

5 CONTROLLED RELIABILITY MECHANISM

Because return-spectrum shape and its orientation relative to score geometry vary together in real data, we use controlled synthetic interventions that hold the full covariance spectrum fixed while changing that orientation.

5.1 COVARIANCE-SPECTRUM-PRESERVING INTERVENTIONS

Controlled synthetic setting. A synthetic *world* is one independently seeded realization under the frozen generator and analysis pipeline. Nominal width selects the number of evaluation units, SNR indexes the frozen signal-strength condition, and θ indexes a continuous rotation of return-covariance orientation relative to fixed score geometry. We study two predeclared intervention families: a categorical three-arm return intervention and a continuous orientation-sweep intervention. For each condition, repeated synthetic outcomes are passed through the same split-half comparison operator, producing split-half rank reliability ρ. The cross-half agreement channel τ^2_{emp} is the mean cross-half candidate-IC covariance over positive-signal replicates, and D_{noise} is the split-half comparison noise estimated from zero-signal replicates. Because the two halves share one joint covariance, τ^2_{emp} is

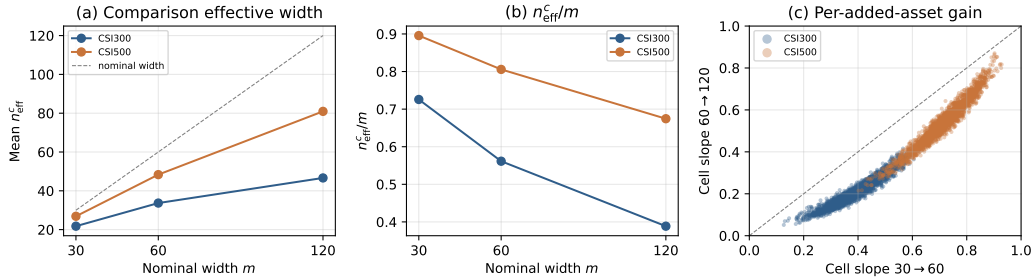

Figure 2: Descriptive matched-width measurements. Comparison effective width rises while n_{eff}^c/m and per-added-asset gain decline. The cell-level slope and elasticity decline counts are 2,258/2,258 in each universe; replicates are dependent within a cell.

Table 1: Real-data evidence summary. Panel A reports matched-width measurements. Panel B reports the exact same-candidate-roster CSI500–CSI300 mean log-gap decomposition. Decomposition terms are descriptive, not causal shares.

A. Matched-width response

	Mean n_{eff}^c			n_{eff}^c/m		
Universe	30	60	120	30	60	120
CSI300	21.7699	33.6875	46.6314	0.7257	0.5615	0.3886
CSI500	26.8699	48.3408	80.9413	0.8957	0.8057	0.6745

B. CSI500–CSI300 mean $\Delta \log n_{\text{eff}}^c$

m	score-side	return-side	alignment	total
30	0.0149	0.1684	0.0341	0.2173
60	0.0163	0.2559	0.0980	0.3703
120	0.0173	0.3413	0.2020	0.5606

C. ETF–CSI300 boundary ($m = 60$, uncentered)

Metric	ETF	CSI300
$\text{PR}(C_s)/m$	.1046	.1031
$\text{PR}(C_r)/m$	.0744	.4125
Leading-eigenvalue share	.3809	.1314
Spectral entropy	.5252	.8931
Alignment position	.5789	.3801
Paired n_{eff}/m	.1618	.5818

not a pure signal quantity: under zero signal it stays near zero in the three-arm family but reaches 54–72% of its positive-signal value, by layer, in the orientation family. On replayed zero-signal streams, D_{geom} and its cross-half counterpart correctly order single-half noise variance and cross-half noise covariance across orientations in all 192 development and held-out world–strata, with about half the predicted variation (median within-world slopes 0.55–0.59); split-half difference noise is ordered correctly in 86–89% of them, whereas an isotropic comparator is constant within a world. In the three-arm family only the cross-half ordering is reliable (supplement).

We construct controlled return-covariance families that preserve the eigenvalue spectrum, trace, $\text{PR}(C_r)$, effective rank, and mean off-diagonal covariance, but not the diagonal, while rotating covariance orientation relative to fixed score geometry. The object is a $2m \times 2m$ covariance, with $m \times m$ half-specific principal submatrices, not a unit-diagonal correlation matrix. At width 30, $\text{PR}(C_r) = 7.2926$ and the maximum arm difference in $\text{PR}(C_r)$ is approximately 1.7×10^{-14}. Full-population spectral diagnostics therefore cannot distinguish these conditions, whereas paired geometry can; the spectra of the two half-sample submatrices are not held fixed and change with orientation (supplement). In the synthetic geometry pipeline, score columns are centered separately within each evaluation half before computing D_{geom}; this differs from the real-data nested-width protocol, which retains full-support standardization after restriction.

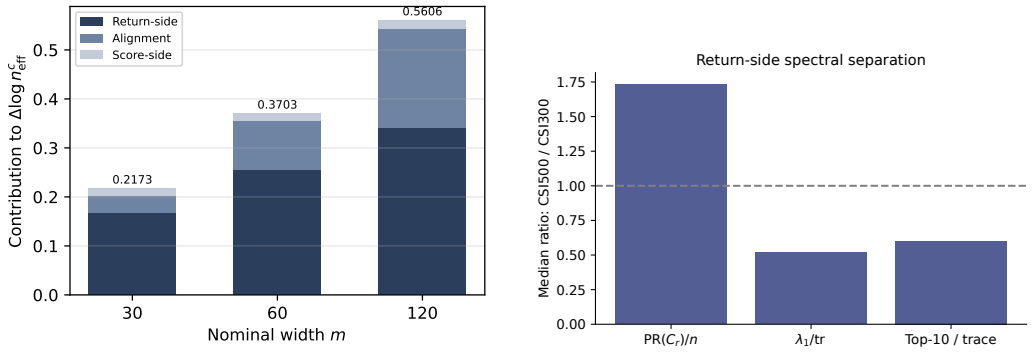

Figure 3: Real-universe descriptive mechanism. Left: CSI500–CSI300 mean cell-level log-gap components; return-side dimensionality is largest, alignment is second, and the score-side term is small. Right: dimension-normalized return-spectrum diagnostics; all 21 paired origins agree in direction.

5.2 Shared reliability relation

Raw geometry alone is not a family-independent reliability coordinate: a continuous orientation sweep does not collapse under the one-parameter form $\rho = \tau^2/(\tau^2 + aD_{\text{geom}})$ (frozen $R^2 = -1.316$; RMSE 0.1849), and the raw association reverses direction across the two controlled intervention families. This motivates separating a cross-half agreement channel from a zero-signal noise channel rather than fitting a post-hoc nonlinear function of raw geometry. The relation

$$\rho = \frac{\tau^2_{\text{emp}}}{\tau^2_{\text{emp}} + aD_{\text{noise}}} \tag{11}$$

is fitted on development worlds and describes both intervention families at the pooled-family level. These fits are in-sample. The three-arm return intervention gives $a = 1.058$ ($R^2 = 0.986$), the orientation-sweep intervention gives $a = 1.113$ ($R^2 = 0.943$), and the joint shared-a fit gives $a = 1.089$ ($R^2 = 0.953$). The shared empirical relation describes both families despite their opposite raw geometry associations. Both intervention families are well described at the pooled-family level by the shared form, while family- and stratum-level calibration remains heterogeneous.

5.3 Rank attenuation and held-out cross-fitting

Using the prespecified local Gaussian-copula factor $k = 3/\pi$, the rank-corrected development coefficient is $a = 0.993$ (95% CI $[0.975, 1.011]$); fixing $a = 1$ gives in-sample $R^2 = 0.955$.

On 32 held-out synthetic worlds, each cell's 100 replicates are divided into disjoint halves: one estimates τ^2_{emp} and D_{noise}, the other supplies target reliability, and the roles are swapped. The resulting 1,920 predictions from 960 cells give $R^2 = 0.8814$ (95% world-cluster bootstrap interval $[0.859, 0.900]$; slope 0.836 $[0.817, 0.854]$) with a fixed from development; the prespecified zero-parameter form gives $R^2 = 0.8845$. Same-half and full-sample comparisons give $R^2 = 0.9324$ and 0.9313: removing coupling, rather than halving replicate count, explains the decrease. Calibration is weakest in width-120/SNR-30 ($R^2 = 0.770$ $[0.725, 0.805]$, slope 0.783). The same-sample zero-parameter comparator ($R^2 = 0.9345$) has mean residual -0.0131 $[-0.0153, -0.0108]$.

A separate same-sample 3×2 factorial diagnostic shows structured residuals: within strata the residual becomes more negative as predicted reliability increases, and a between-stratum width-associated offset remains that a reliability-level local approximation does not explain, so no post-hoc correction is introduced. Twenty-two of 24 prespecified zero-signal checks passed; the two deviations were conservative (supplement).

6 Measurement–validation boundary

A planned real-outcome validation on a held-out CSI300 population was gated by a prespecified synthetic estimator-recovery check. At zero injected effect, all 256 worlds in each of two scenarios

Table 2: Development fits and held-out synthetic cross-fitting. The latter covers only three orientation-family strata. The zero-parameter form uses $k = 3/\pi$.

Setting	a	R^2	Evaluation
Three-arm return	1.058	.986	Development, in-sample
Orientation sweep	1.113	.943	Development, in-sample
Joint shared-a	1.089	.953	Development, in-sample
Rank-corrected	.993	–	95% CI [.975,1.011]
Rank-corrected, fixed	1	.955	Development, in-sample
Development-fixed coefficient	1.089	0.8814	Held-out, cross-fitted
Zero-parameter form	1	0.8845	Held-out, cross-fitted

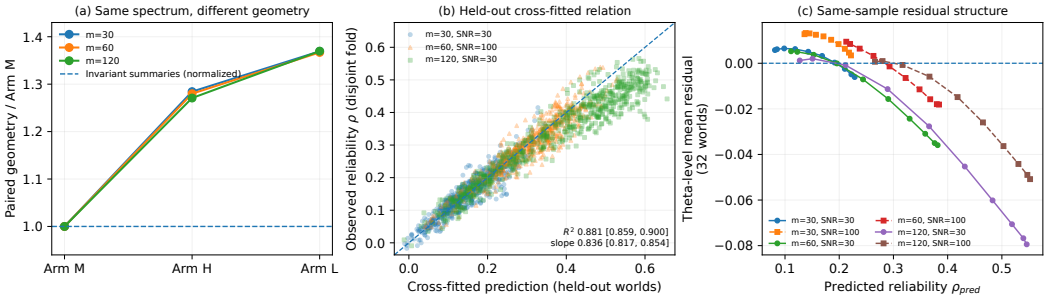

Figure 4: Controlled mechanism under covariance-spectrum-preserving interventions. (a) Within each nominal width, three full-covariance spectral summaries (trace, participation ratio, effective rank) and mean off-diagonal covariance overlap at the invariant baseline while paired geometry changes across the frozen categorical intervention arms; the arm labels are identifiers, not a monotone dose ordering. (b) Held-out cross-fitted relation on 32 synthetic worlds: predictor channels from one replicate fold, observed reliability from the disjoint fold, with a fixed from development; the dashed line is the identity, and bracketed values are 95% world-cluster bootstrap intervals. The in-sample development fit is shown in the supplement. (c) Same-sample residual diagnostics on held-out synthetic worlds: the complete 3×2 width$\times$SNR factorial is shown using 60 theta-level means, each averaging 32 frozen worlds.

returned a positive lower confidence bound (estimated slope approximately 0.72), exceeding the false-positive criterion. The real-outcome stage was not authorized; the held-out population was read in none of the audited execution lineages. Later revisions remained synthetic, including a change of estimand, and established no outcome bridge. This check is distinct from Section 5's residual calibration. We therefore make no claims about realized ranking error, candidate-selection regret, portfolio performance, or forecasting alpha.

7 Discussion and Limitations

The CSI decomposition is descriptive: lower ALIGN^c raises n_{eff}^c algebraically without implying more signal. Width measurements are local to $m = 30$–120, with no saturation claim. Covariance sufficiency in Proposition 1 holds within Gaussian or fixed-generator elliptical families. The reliability relation is fitted on two synthetic families, but held-out cross-fitting covers only three orientation-family strata; calibration remains heterogeneous. CSI300 and CSI500 share a market and calendar; ETF provides only a 137-date direct boundary.

8 Conclusion

Nominal width alone does not determine comparison-noise geometry: candidate-axis projection makes score geometry comparison-specific, and covariance orientation matters beyond full-population spectra. Real-data measurements describe local regularities, and a pooled synthetic relation predicts held-out split-half reliability with cross-fitted $R^2 \approx 0.88$ and structured bias. These findings motivate reporting comparison-noise geometry alongside nominal evaluation counts.

AI USE STATEMENT

This work was directed by a single author, who formulated the research question and principal comparisons, chose the estimands and evaluation populations, and set the experimental, claim, and data-access boundaries. The author retained the full CSI500 comparison and the ETF boundary analysis despite AI suggestions to narrow their scope.

Generative AI systems assisted with literature retrieval and organization; critique of research framing, methods, and experimental protocols; review of mathematical arguments and proof drafts; implementation and review of code; discussion of results; and drafting and editing of text and figures. AI agents also executed prespecified computations under written work orders that specified configurations, permitted data reads, and required outputs. Changes in scope required the author's authorization. Survey, interview, transcription, and qualitative-analysis tasks were not applicable to this work.

The author checked AI-assisted material used in the submission against source literature and executed artifacts, including independent recomputation of numerical results where appropriate. For example, the author required an absolute tolerance of 10^{-15} and a completed robustness scan after an AI audit used relative differences near zero and prematurely reported verification; rejected a proposed multivariate-t stress test whose common radial scale cancels under correlation and ranks; and rejected a shape-drift comparison that paired a 120-observation statistic with a 60-observation noise reference. A prespecified synthetic estimator-recovery check prevented the planned held-out real-outcome stage from proceeding; the held-out population remained unread in all audited execution lineages.

The author reviewed the final submission and takes responsibility for its text, claims, code, figures, and other artifacts, including material produced with AI assistance.

REPRODUCIBILITY STATEMENT

The supplement documents the comparison-width identities, preprocessing, aggregation levels, synthetic interventions, and held-out evaluation. The data-construction pipeline and synthetic generator will be released with any camera-ready version. Licensed raw market data, including the study-specific ETF panel inputs, are not redistributed. Real-data results use label-free geometry only; Section 5 uses controlled synthetic outcomes only.

REFERENCES

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blum15.html.

Christopher S. Bretherton, Martin Widmann, Valentin P. Dymnikov, John M. Wallace, and Ileana Bladé. The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, 12(7):1990–2009, 1999. doi: 10.1175/1520-0442(1999)012⟨1990:TENOSD⟩2.0.CO;2.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL http://jmlr.org/papers/v11/cawley10a.html.

Binqi Chen, Hongjun Ding, Ning Shen, Taian Guo, Jinsheng Huang, Luchen Liu, and Ming Zhang. AlphaSAGE: Structure-aware alpha mining via GFlowNets for robust exploration. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zRKF4ln2VE.

China Securities Index Co., Ltd. Methodology for the CSI 200, CSI 500, CSI 700 and CSI 800 Indices. Technical report, China Securities Index Co., Ltd., May 2022. URL https://oss-ch.csindex.com.cn/static/html/csindex/public/uploads/indices/detail/files/zh_CN/000907_Index_Methodology_cn.pdf. Version V1.2.

China Securities Index Co., Ltd. CSI 300 Index Methodology. Technical report, China Securities Index Co., Ltd., September 2023. URL https://oss-ch.csindex.com.cn/static/html/csindex/public/uploads/indices/detail/files/en/000300_Index_Methodology_en.pdf.

Peter Clifford, Sylvia Richardson, and Denis Hémon. Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45(1):123–134, 1989. doi: 10.2307/2532039.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan CLEMENCON. What are the best systems? new perspectives on NLP benchmarking. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=kvtVrzQPvgb.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Pierre Dutilleul. Modifying the t test for assessing the correlation between two spatial processes. *Biometrics*, 49(1):305–314, 1993. doi: 10.2307/2532625.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a. doi: 10.1126/science.aaa9375.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 117–126, New York, NY, USA, 2015b. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746580. URL https://doi.org/10.1145/2746539.2746580.

Jinyong Fan and Yanyan Shen. StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i8.28681. URL https://doi.org/10.1609/aaai.v38i8.28681.

Hamza Giaffar, Camille Rullán Buxó, and Mikio Aoi. The effective number of shared dimensions between paired datasets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4249–4257, 2024. URL https://proceedings.mlr.press/v238/giaffar24a.html.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hanna Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2508.13144.

Vindula Jayawardana, Catherine Tang, Sirui Li, Dajiang Suo, and Cathy Wu. The impact of task underspecification in evaluating deep reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023. doi: https://doi.org/10.1111/jofi.13249. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13249.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, 1965.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Jian Li and Lei Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i1.27767. URL https://doi.org/10.1609/aaai.v38i1.27767.

Haochen Luo, Ho Tin Ko, Jiandong Chen, David Sun, Yuan Zhang, and Chen Liu. AlphaBench: Benchmarking large language models in formulaic alpha factor mining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d97Q8r7ZKZ.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9993–10002, 2019.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=hcOq2buakM.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:207979247.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/russo16.html.

Hao Shi, Weili Song, Xinting Zhang, Jiahe Shi, Cuicui Luo, Xiang Ao, Hamid Arian, and Luis Angel Seco. AlphaForge: A framework to mine and dynamically combine formulaic alpha factors. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i12.33365. URL https://doi.org/10.1609/aaai.v39i12.33365.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560/.

Chen-Hui Song, Shuoling Liu, and Liyuan Chen. The label horizon paradox: Rethinking supervision targets in financial forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=G43CIfmmxh.

Shira Wein, Christopher M. Homan, Lora Aroyo, and Chris Welty. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3138–3161, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.196. URL https://aclanthology.org/2023.findings-acl.196/.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An AI-oriented quantitative investment platform, 2020. URL https://arxiv.org/abs/2009.11189.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5476–5486, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599831. URL https://doi.org/10.1145/3580305.3599831.