FACET: GEOMETRY AND GUARANTEES AT THE CONFORMAL PREDICTION-BOX INTERFACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate conformal methods are calibrated and compared as prediction *regions*, while many downstream interfaces expose only coordinatewise lower and upper bounds: the axis-aligned box of that region. These are different prediction tasks. We characterize the resulting interface. For any nested box family, exact box membership depends only on its coordinate-envelope path, so two arbitrarily different native regions with the same envelopes become the identical box predictor after exact recalibration; for homothetic families this reduces the geometry to its axial support vector, and for Mahalanobis ellipsoids to $\sqrt{\operatorname{diag}\Sigma}$. Containing conversions preserve coverage but can be asymptotically vacuous: the box of a calibrated Gaussian ellipsoid has coverage tending to one and is $\Omega(\sqrt{N/\log N})$ wider than a valid centered box. Finite non-containing conversions admit no distribution-free transfer and can instead lose nearly all coverage. We test these predictions in a locked, external-only benchmark on high-dimensional forecasting panels. At the native threshold, containing conversions add 9.8–11.6 percentage points of coverage, whereas a rank-truncated non-containing conversion loses 86.3 points in median. The box-task winner differs from the native-region winner in 69.3% of eligible cells. Exact online box calibration reduces median absolute coverage error from 2.47 to 0.10 percentage points at a median width ratio of 0.980, and equal envelope paths yield equal boxes to 4.25×10^{-10}. The strongest external geometry remains 28% wider in median than a validation-selected coordinatewise-support box. On four untouched panels, time-varying coordinate supports reduce width by 30.2% over static supports, while pooled support models win 1 of 140 comparisons. The result is a task-faithful evaluation principle and a simple prescription: predict coordinate supports, then calibrate exact membership in the box that is actually delivered.

1 INTRODUCTION

A high-dimensional forecast often reaches its user as a list of lower and upper limits. A forecaster may internally construct a 90% ellipsoid, a union of sampled neighborhoods, or a learned nonconvex region, but a downstream interface may accept only one interval per coordinate. The natural conversion is the region's axis-aligned coordinate envelope. This raises a basic question: *what guarantee, and what geometric information, survives that conversion?* Native-region coverage does not answer it, and native-region volume is not the size of the box the user receives.

Conformal prediction attaches coverage to a particular set, the sublevel set of a calibrated score (Vovk et al., 2005; Lei et al., 2018). A later representation change need not inherit it. A containing box preserves coverage but may become needlessly conservative; a finite non-containing box has no distribution-free transfer guarantee at all. We show that both cases become extreme in high dimension: the first can approach coverage one while growing far wider than necessary, and the second can approach coverage zero while the native region remains exactly calibrated.

This distinction cuts across multivariate conformal prediction. Existing methods construct ellipsoids, copula-coupled rectangles, sampled unions, optimal-transport regions, and learned flow-based sets (Wang et al., 2023; Sun & Yu, 2024; Xu et al., 2024; Thurin et al., 2025; Lee et al., 2026); recent work compares native regions in unified frameworks (Dheur et al., 2025; Johnstone & Ndiaye, 2025) or optimizes boxes directly (Schlembach et al., 2025). These are legitimate but different targets. Our

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

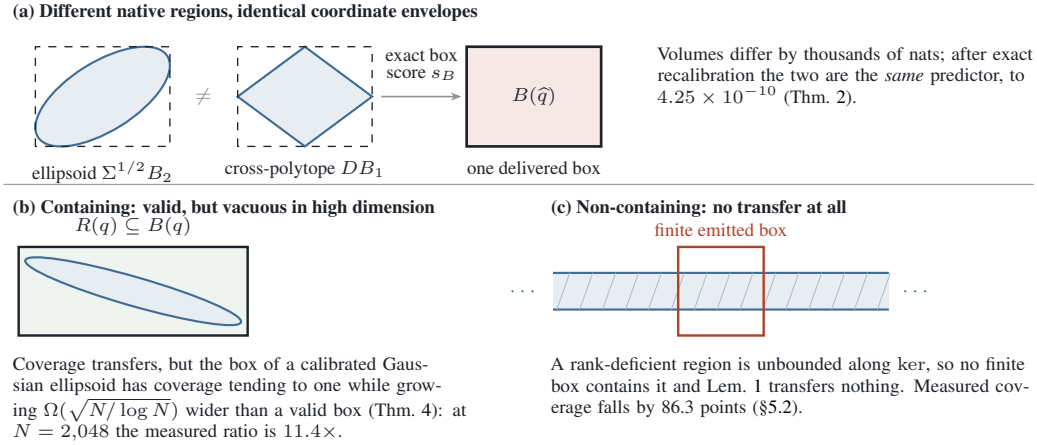


Figure 1: **The prediction-box interface.** Native methods remain valid for their stated task throughout; the mismatch appears only when the required output changes to an axis-aligned box.

goal is not to relabel valid region methods as invalid, but to separate the *native-region task* from the *prediction-box task* and to compare methods only after they solve the same one.

Our main result is an interface theorem. Let $R_t(q)$ be any nested native-region path and $B_t(q) = \prod_{i=1}^{N}[\ell_{t,i}(q), u_{t,i}(q)]$ the nested box delivered at level q. Exact entrance into $B_t(q)$ is determined by the first level at which all coordinate constraints hold, so exact box calibration can see the native region only through the path $q \mapsto (\ell_t(q), u_t(q))$. Two regions may differ arbitrarily in topology, volume, or dependence geometry; if their coordinate envelopes agree, their exact box scores, calibrated thresholds, and delivered boxes agree for every outcome sequence. For a homothetic region $\mu_t + qK_t$ only the axial support values of K_t survive, and for a Mahalanobis ellipsoid these are $\sqrt{\operatorname{diag}\Sigma_t}$, so fitted off-diagonal covariance disappears once the envelope is fixed. True residual dependence is not ignored: it still determines the distribution of the resulting maximum score and is learned by calibration.

The theorem has two sharp consequences. If $R_t(q) \subseteq B_t(q)$, then calibrating exact membership in B_t is always weakly sharper than calibrating R_t and boxing afterward. For Gaussian ellipsoids the latter box has coverage tending to one while being at least order $\sqrt{N/\log N}$ wider than a valid centered box. Conversely, a rank-deficient pseudo-Mahalanobis region is an unbounded cylinder in its nullspace, so any finite coordinate summary fails to contain it and inherits no universal coverage transfer. Figure 1 summarizes both failure modes.

We evaluate the interface in four explicit stages: the native region (A), its coordinate envelope at the same threshold (B), exact split recalibration of that box (C), and a common exact online recalibration (D). The primary benchmark contains no method of our own design and spans published ellipsoidal, sample-based, copula, and box-native families on 9 panels with 128–2,327 coordinates, multiple forecast origins, two backbones, and three coverage levels; methods, hypotheses and selection rules were fixed before the confirmation experiments were run.

The results follow the theorem. At nominal 90%, containing conversions add 9.8–11.6 coverage points, while the rank-truncated non-containing conversion loses 86.3 points in median. The box-task winner differs from the native-region winner in 69.3% of eligible cells. After exact online recalibration, equal coordinate supports produce equal boxes to 4.25×10^{-10}, and median absolute coverage error falls from 2.47 to 0.10 percentage points at a median width ratio of 0.980. The strongest published geometry on the common benchmark remains 28% wider in median than a validation-selected coordinatewise-support box.

This suggests a simple constructive division of labor. Predict any positive support vector v_t before observing Y_t, and calibrate

$$S_t = \max_i \frac{|Y_{t,i} - \widehat{\mu}_{t,i}|}{v_{t,i}}. \tag{1}$$

One online scalar controls overall size; v_t controls efficiency. On 4 held-out panels, validation-selected coordinatewise conditioning reduces width by 30.2% over static supports, while pooled

factor and shared supports win 1 of 140 eligible comparisons: one under-scaled coordinate inflates the maximum score and hence every side of the box. The scalar update is standard in spirit (Gibbs & Candès, 2021; Zaffran et al., 2022; Angelopoulos et al., 2023; Areces et al., 2025); our contribution is to attach it to exact delivered-box membership and to identify the coordinate-support path as the only geometry it needs.

Contributions.

- **Box-interface theory.** We characterize coverage transfer by containment, prove coordinate-envelope collapse for arbitrary nested paths, and establish the dual high-dimensional failures of containing and non-containing conversions.
- **Task-faithful evidence.** A pre-registered benchmark of published methods separately evaluates native regions, their coordinate envelopes without recalibration, and exactly recalibrated boxes, revealing large distortion, frequent ranking reversals, and exact equivalence when envelopes agree.
- **A constructive prescription.** Exact online box feedback restores reliable realized coverage under drift, while a held-out confirmation shows that validation-selected coordinatewise support prediction supplies the dominant efficiency gain.

Our claims are interface-specific: native-region geometry can remain valuable when the region itself is the output, and we stop before the separate problem of turning a delivered box into a downstream action.

2 From Native Regions to Prediction Boxes

Native and operational sets. Condition on everything available before the response is observed. Let $\mathcal{Q} \subseteq [0, \infty]$ be totally ordered; for each q let $R(q) \subseteq \mathbb{R}^N$ be the *native region* and

$$B(q) = \prod_{i=1}^{N} \big[\ell_i(q), u_i(q)\big] \tag{2}$$

the *box* exposed by the coordinate interface. We assume only that both paths are nested and that entrance levels are attained, so that the infimum below is a minimum; this holds on any computational grid and, in general, whenever the path is right-continuous (Appendix A.1). Convexity, central symmetry, homothety, exchangeability, and any particular score construction are *not* required. Define the exact entrance scores

$$s_R(y) = \inf\{q : y \in R(q)\}, \qquad s_B(y) = \inf\{q : y \in B(q)\}, \tag{3}$$

so that $y \in B(q) \iff s_B(y) \le q$. That equivalence is the entire exact-set principle: calibrating s_B calibrates the set the user actually receives, and nothing else.

Four evaluation stages. Writing $\widehat{q}_R$ and $\widehat{q}_B$ for thresholds calibrated on s_R and s_B, we use throughout

A: $R(\widehat{q}_R)$ native; **B:** $B(\widehat{q}_R)$ same threshold; **C:** $B(\widehat{q}_B)$ split; **D:** $B_t(q_t)$ online.

The distinction is what makes the comparison fair. A and B differ only in which set coverage is measured on, so a method can succeed at A and fail at B without failing the task it was designed for; we call such a gap *compilation distortion* and never a defect of the published method.

Metrics and provenance. Stage A uses native coverage and, where the native region has finite volume, log-volume per coordinate; stages B–D use joint box coverage with mean, geometric-mean and maximum coordinate width and the unbounded-step fraction. Width is $+\infty$ at an unbounded step and headline widths are medians over *all* evaluated steps: a calibrator that falls behind can cover trivially by emitting the whole space, and averaging only bounded steps would let it report a small width for doing so. Each family is labeled *native box*, *documented projection*, or *derived box* according to how its envelope is obtained; the per-family algebra behind that label is Appendix B.

3 What Survives the Box Interface?

3.1 Universal transfer and coordinate-envelope collapse

Containment is not merely a convenient sufficient condition for a guarantee to carry over; distribution-free, it is the exact geometric condition.

Lemma 1 (Universal transfer). *For measurable $R, E \subseteq \mathbb{R}^N$, the containment $R \subseteq E$ holds if and only if $\mathbb{P}(Y \in R) \le \mathbb{P}(Y \in E)$ for every law of Y. Consequently, if $R \not\subseteq E$ then $\inf\{\mathbb{P}(Y \in E) : \mathbb{P}(Y \in R) \ge 1-\alpha\} = 0$, so no nontrivial distribution-free lower bound on the coverage of E follows from the coverage of R.*

One direction is pointwise domination of indicators; for the other, pick $y_0 \in R \setminus E$ and let $Y \equiv y_0$. For fitted regions the statement is applied conditionally, so what is needed is $R(q) \subseteq E(q)$ almost surely for every q.

The main result identifies how much of a native region an axis-aligned interface can see. For each coordinate define the entrance times

$$\tau_{i,-}(z) = \inf\{q : \ell_i(q) \le z\}, \qquad \tau_{i,+}(z) = \inf\{q : z \le u_i(q)\}. \tag{4}$$

Theorem 2 (Box-interface collapse). *Let $B(q)$ be a nested box path* (2) *whose entrance levels are attained. Then for every $y \in \mathbb{R}^N$,*

$$s_B(y) = \max_{1 \le i \le N} \big\{\tau_{i,-}(y_i),\, \tau_{i,+}(y_i)\big\}. \tag{5}$$

Hence the coordinate-envelope path $q \mapsto (\ell_1(q), \dots, \ell_N(q), u_1(q), \dots, u_N(q))$ is a complete invariant of the exact box interface: any two native region paths inducing the same envelope path have identical exact box scores, and are therefore identical under every split, rolling, or online calibration procedure driven only by those scores. If moreover $R(q) \subseteq B(q)$ for every q, then $s_B(y) \le s_R(y)$ for every y, so for every coordinatewise nondecreasing calibration functional Γ the thresholds $\widehat{q}_B = \Gamma(s_B(Y_1), \dots, s_B(Y_n))$ and $\widehat{q}_R = \Gamma(s_R(Y_1), \dots, s_R(Y_n))$ satisfy $\widehat{q}_B \le \widehat{q}_R$, hence $B(\widehat{q}_B) \subseteq B(\widehat{q}_R)$.

Proof sketch. Nesting makes ℓ_i nonincreasing and u_i nondecreasing, so $y \in B(q)$ exactly when $\tau_{i,-}(y_i) \le q$ and $\tau_{i,+}(y_i) \le q$ for every i; the smallest such q is their maximum, which is (5). If $s_R(y) = q$ then $y \in R(q) \subseteq B(q)$, so $s_B(y) \le q$; monotonicity of Γ and nesting of B give the rest (Appendix A.3). □

That last inequality is the formal reason calibrating a native region and then projecting is never sharper than exactly recalibrating the same box path—deterministically, on every calibration sample, with no distributional assumption at all.

3.2 Homothetic regions and ellipsoids

The specialization that does most of the work is the scale family. Let $R(q) = \mu + qK$ with K bounded and star-shaped about the origin, with positive one-sided axial supports $b_i^+ = \sup_{z \in K} z_i$ and $b_i^- = -\inf_{z \in K} z_i$. The exact enclosing-box score is

$$s_B(y) = \max_i \Big\{ \tfrac{(y_i - \mu_i)_+}{b_i^+},\ \tfrac{(\mu_i - y_i)_+}{b_i^-} \Big\}, \qquad \text{and} \qquad s_B(y) = \max_i \frac{|y_i - \mu_i|}{b_i} \tag{6}$$

when K is centrally symmetric. For a Mahalanobis ellipsoid $K_\Sigma = \{z : z^\top \Sigma^{-1} z \le 1\}$ the axial support is $b_i = \|\Sigma^{1/2} e_i\|_2 = \sqrt{\Sigma_{ii}}$, so an $O(N^2)$ fitted covariance geometry reduces to N coordinate scales.

This deserves a precise reading. Interior geometry disappears once the coordinate-envelope path is fixed: all fitted *off-diagonal* covariance information leaves the exact box score. It does not follow that the coordinates are independent or that dependence is irrelevant: the true joint law still governs the distribution of $\max_i |Y_i - \mu_i|/\sqrt{\Sigma_{ii}}$, and calibration learns exactly that. What becomes unnecessary is encoding dependence in a fitted region's interior.

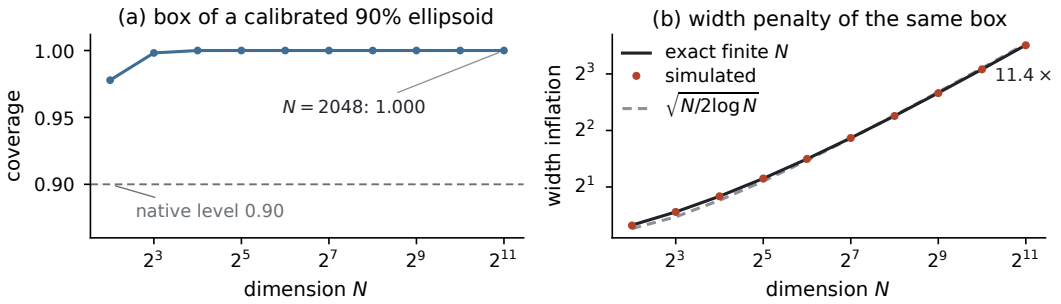


Figure 2: **Both halves of Theorem 4, measured.** Gaussian simulation, $\alpha = 0.10$, 20,000 draws per dimension. The simulated ratio in (b) tracks the *exact* finite-N value $q_{N,\alpha}/c^{\mathrm{iid}}_{N,\alpha}$ to 4.3×10^{-3} relative error; the asymptotic rate is drawn separately because it is visibly a different number at these dimensions.

3.3 The two high-dimensional extremes

Rank deficiency gives the first extreme, immediately.

Corollary 3 (No finite hull for a pseudo-Mahalanobis cylinder)**.** *Let $A \succeq 0$ have rank $r < N$ and put $R(q) = \{z : z^\top A^+ z \le q^2\}$. For every $h \in \ker(A)$ and $z \in R(q)$ we have $z + th \in R(q)$ for all $t \in \mathbb{R}$. Hence $R(q)$ admits no finite axis-aligned containing box, every finite coordinate compilation of it is non-containing, and by Lemma 1 such a compilation receives no universal coverage transfer from the native region.*

The second extreme is the opposite failure: containment is preserved, and the box becomes vacuous anyway.

Theorem 4 (Gaussian ellipsoid-to-box blow-up)**.** *Let $Y_N \sim \mathcal{N}(\mu_N, \Sigma_N)$ with $\Sigma_N \succ 0$, let $q^2_{N,\alpha}$ be the $(1-\alpha)$-quantile of χ^2_N, and let E_N be the coordinate bounding box of the calibrated ellipsoid, with half-widths $q_{N,\alpha}\sigma_{N,i}$ and $\sigma_{N,i} = \sqrt{\Sigma_{N,ii}}$. Then, uniformly over correlation structures,*

$$\mathbb{P}(Y_N \in E_N) \;\ge\; 1 - 2N\overline{\Phi}(q_{N,\alpha}) \;\longrightarrow\; 1. \tag{7}$$

If $W^\star_N$ is the smallest mean half-width of any valid centered box and $c^{\mathrm{Bon}}_{N,\alpha} = \Phi^{-1}(1-\alpha/2N)$, then

$$\frac{W(E_N)}{W^\star_N} \;\ge\; \frac{q_{N,\alpha}}{c^{\mathrm{Bon}}_{N,\alpha}} \;=\; (1+o(1))\sqrt{\frac{N}{2\log(2N/\alpha)}}. \tag{8}$$

Under $\Sigma_N = I_N$ the optimal centered box has exact radius $c^{\mathrm{iid}}_{N,\alpha} = \Phi^{-1}\big((1+(1-\alpha)^{1/N})/2\big)$, and the ratio is asymptotic to $\sqrt{N/(2\log N)}$.

The intuition is a mismatch of tail geometry. A χ^2_N radius grows like $\sqrt{N}$ because an ellipsoid must absorb deviations spread across all N coordinates at once, whereas a box needs each coordinate only within about $\sqrt{2\log N}$ standard deviations; the bounding box inherits the ellipsoid's radius in every direction and buys coverage it was never asked for. Comparing against an explicitly valid Bonferroni box is what makes (8) hold for *every* correlation structure (Appendix A.6).

4 Exact Online Calibration of a Predicted Box

Theorem 2 says the only geometry an axis-aligned interface can use is a coordinate-support path. That is a constructive instruction, not just a negative result: predict the supports, and calibrate exact membership in the box they define. We call the construction *Online ExactBox*; it is deliberately plain, because the content is in what it calibrates rather than in its machinery.

Predicted support, exact score. Let $\widehat{\mu}_t$ and $v_t \in \mathbb{R}^N_+$ be predictable, that is, computed before Y_t is revealed. Emit

$$C_t(q_t) = \prod_{i=1}^{N} \big[\widehat{\mu}_{t,i} - q_t v_{t,i},\ \widehat{\mu}_{t,i} + q_t v_{t,i}\big], \qquad S_t = \max_i \frac{|Y_{t,i} - \widehat{\mu}_{t,i}|}{v_{t,i}}. \tag{9}$$

By (6) this is the exact membership score of the emitted box: $Y_t \in C_t(q_t) \iff S_t \le q_t$. It is the only score the constructive method uses. Note what v_t is *not* asked to be. It need not be a covariance, a quantile, or a correctly specified volatility model, and need not be accurate at all; it only has to be predictable and positive.

Split and online calibration. Stage C takes the usual split-conformal quantile of S_t, which is standard validity under exchangeability applied to the delivered set rather than a new theorem (Appendix A.7).

Exchangeability is exactly what a drifting panel violates, so stage D replaces the fixed quantile by an online update. Let $q_t = \widehat{Q}_t(1-\lambda_t)$ be the empirical $(1-\lambda_t)$-quantile of a trailing window of scores, let $M_t = \mathbb{1}\{Y_t \notin C_t(q_t)\}$ be the *realized miss of the emitted box*, and update

$$\lambda_{t+1} = \Pi_{[\lambda_-,\lambda_+]}\big[\lambda_t + \eta(\alpha - M_t)\big], \qquad r_t = \lambda_{t+1} - \lambda_t - \eta(\alpha - M_t), \tag{10}$$

where r_t is the projection correction. Summing (10) and rearranging gives the pathwise identity

$$\frac{1}{T}\sum_{t=1}^{T} M_t - \alpha = \frac{\lambda_1 - \lambda_{T+1}}{\eta T} + \frac{1}{\eta T}\sum_{t=1}^{T} r_t, \tag{11}$$

so the realized miss rate is controlled by $D/(\eta T)$ with $D = \lambda_+ - \lambda_-$ whenever the projection is inactive.

Three qualifications apply to (11). It assumes no exchangeability, independence, stationarity, or correctness of the support model; it concerns the realized *average* miscoverage over the horizon; and it is not conditional coverage at each time, which we do not claim.

Support selection. The efficiency question is now entirely about v_t. Scalar calibration removes global scale error, since $S_{cv} = S_v/c$ while the emitted widths scale by c; what matters is the *relative* coordinate profile, and a projective perturbation of v costs at most a bounded width factor (Appendix A.9). The predictor therefore selects on validation among a small factor-free library,

$$\big\{\text{Static},\ \text{EWMA}_{12,48,192},\ \text{Seasonal-EWMA}_{8,32}\big\}, \tag{12}$$

each a coordinatewise estimate of scale from that coordinate's own history. Factor and pooled-shared predictors are deliberately *not* in the algorithm; they appear in §5.4 as ablations. Appendix C gives the pseudocode and the initialization and boundary conventions.

5 Experiments

5.1 Does converting a native region into a box change the task?

Protocol. The primary benchmark comprises 9 forecasting panels with 128–2,327 coordinates, all public except one synthetic control (Appendix D), each evaluated at several non-overlapping chronological origins, with ridge autoregression as the primary backbone and persistence as a sensitivity check, at $\alpha \in \{0.05, 0.10, 0.20\}$. Each origin is cut into five disjoint blocks: design, two validation blocks, final calibration, and test. Every structural and calibration choice—rank, window, shrinkage, learning rate, and which configuration represents a family—is made on validation only; test is scored once. Uncertainty is a hierarchical bootstrap over panels and then origins. The benchmark contains only published methods: no variant designed by us appears in it. The protocol, including the criterion each claim had to meet, was fixed before the confirmation experiments and is summarized in Appendix H; code that reruns every experiment from the public data is available online.[1]

Table 1 reports the benchmark at all four stages. Stage A is reported throughout, so each native method remains evaluated on its own task. Two families need a word on eligibility. FCP (Lee et al., 2026) appears at the native stage only: its exact coordinate extrema require globally optimizing over a flow pushforward, and the pre-registered protocol does not allow a sampled hull in their place, since that would be an *inner* approximation biased in favor of our conclusions. This is an ineligibility under our criterion, not a failure of FCP. SPACE (Li et al., 2026) enters as an auxiliary sample-cloud

[1] `https://anonymous.4open.science/r/facet-anonymous`

Table 1: **The benchmark at four evaluation stages** ($\alpha = 0.10$, ridge-AR, medians over 9 panels and origins). Stage A evaluates the *native* object on its own task; stage B reads the same threshold through the coordinate box without recalibration; C and D recalibrate exact box membership, offline and online. A gap between A and B is compilation distortion, not evidence that a method fails its stated task. D/C w. is the ratio of median stage-D to median stage-C width. The last column is the median ratio of stage-D mean width to that of the validation-selected coordinate-support box of §4, pooled over all levels and backbones; it is the quantity plotted in Fig. 3(c) and quoted in the text.

family	interface	A cov.	B cov.	C cov.	D cov.	D/C w.	D w./ref
BONFERRONI	native box	0.869	0.869	0.885	0.900	0.99	1.40
COPULACPTS	native box	0.689	0.689	0.896	0.900	1.00	1.47
DCP-MT	native box	0.816	0.816	0.884	0.900	1.01	1.37
MAXNORM	native box	0.895	0.895	0.895	0.900	0.87	1.33
L2-BALL	derived, containing	0.895	1.000	0.899	0.900	0.92	1.36
LEDOIT-WOLF	derived, containing	0.884	1.000	0.895	0.900	0.96	1.28
MAHALANOBIS	derived, containing	0.883	1.000	0.892	0.900	0.92	1.28
POET	derived, containing	0.886	1.000	0.890	0.900	0.98	1.34
PCP	derived, containing	0.895	1.000	0.898	0.900	0.94	1.29
MULTIDIMSPCI	documented, non-cont.	0.909	0.000	0.892	0.900	0.89	1.34
SPACE	sample cloud	–	–	–	0.901	–	1.70
ONLINE EXACTBOX	predicted support	–	–	–	0.900	–	1.00

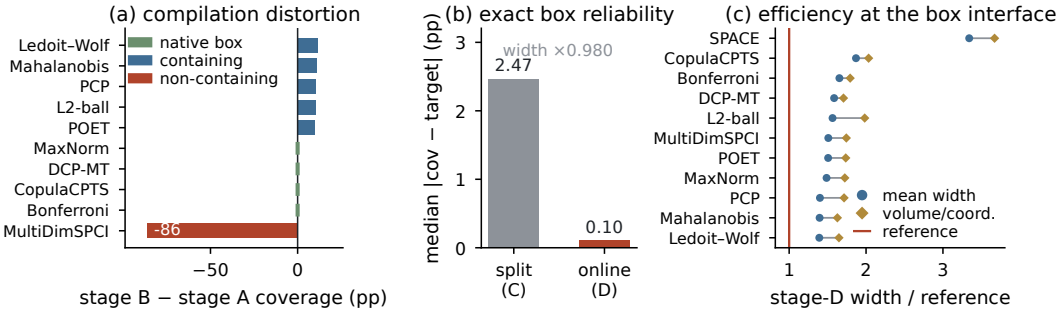


Figure 3: **The benchmark at the box interface.** Compilation distortion separates the three regimes cleanly (a); exact online recalibration cuts median absolute coverage error from 2.47 to 0.10 points while *narrowing* boxes slightly (b); and every published geometry sits to the right of the reference line under both efficiency metrics (c).

comparison: our clouds come from a pre-specified causal residual bootstrap rather than the original method's deep forecasters (Appendix B).

5.2 WHAT FAILS UNDER COMPILATION?

Reading the same calibrated threshold through the coordinate box splits the methods into exactly the three regimes Lemma 1 predicts, and Figure 3(a) shows the split is clean rather than graded. Methods that already emit coordinate intervals have $B - A = 0$ by construction. Containing conversions gain 9.8–11.6 coverage points: valid, but superfluous, since the native regions were already calibrated. The rank-truncated documented compilation loses 86.3 points in median and is non-containing in 302 of 306 cells.

This last result needs careful interpretation. A pseudo-Mahalanobis region remains well defined on its native score—median stage-A coverage 0.909, on target—but by Corollary 3 its finite documented compilation cannot contain the unbounded nullspace cylinder. The native and box guarantees refer to different events. The method is not invalid; its compilation simply carries no guarantee.

5.3 DOES EXACT BOX RECALIBRATION CHANGE CONCLUSIONS?

It changes both reliability and rankings. Across all methods and cells, the median absolute stage-D coverage error is 0.11 points $[0.07, 0.17]$. Of the 3,510 method–cell pairs with at least 500 test steps, 3,503 are within one point of target; none undercovers by more than three points, and

only 1 overcovers by that much.[2] Paired against split calibration on the same envelope, online calibration improves median absolute error from 2.47 to 0.10 points at a median width ratio of 0.980 $[0.964, 1.001]$: the boxes get slightly *narrower* as they become far more reliable, because a static threshold on a drifting panel must be conservative on average to be adequate at the worst moment.

The ranking result is the one with methodological consequences. Comparing each family's native-task efficiency with its exact-online-box efficiency in the same cell gives $1{,}069$ reversals among $2{,}950$ ordered pairs (36.2%), and the top-ranked method changes in 138 of 199 eligible cells. Selecting a method by native-region efficiency therefore picks a different box-task winner in roughly seven of every ten eligible cells.

Efficiency at that interface is the last column of Table 1 and Figure 3(c): every published geometry is wider than the validation-selected coordinate-support box under both metrics, the closest being LEDOIT-WOLF at 28% in median.

Finally, Theorem 2 makes a testable point prediction: whenever two families' coordinate supports agree, their delivered boxes must agree. Among $1{,}568$ matched-support pairs, $1{,}562$ are bitwise identical in width and coverage, and the largest relative discrepancy anywhere is 4.25×10^{-10}: round-off from the different linear algebra by which each family reaches the same mathematical support, and within the 10^{-9} tolerance to which the supports themselves agree.

5.4 WHAT DETERMINES SHARP BOXES?

Because both the seasonal support model and the learning rate were chosen after seeing exploratory results, we test them on 4 held-out panels that played no part in developing the method, with the library (12) and the update fixed in advance and no further configuration allowed. All numbers in this subsection are from those panels.

Validation-selected coordinatewise conditioning reduces width by 30.2% $[20.6, 96.7]$ over static support at comparable coverage, against a pre-registered threshold of 5%. The mechanism is seasonality, but a strong seasonal cycle does *not* imply that conditioning on it helps—which is why the selector is part of the method rather than a convenience. On LargeST-SD (seasonal share 0.91) seasonal EWMA gains 20.4% against ordinary EWMA's 14.4%, and the selector chooses it in 33 of 36 cells. On LargeST-GBA the seasonal share is a comparable 0.83, yet seasonal conditioning is harmful (11.9% against 29.2%); the selector takes it in 1 of 36 cells and keeps the full 29.2%. The conclusion is therefore not "use seasonal EWMA on seasonal panels" but that time-varying coordinate supports matter and validation selection among plausible temporal support models is necessary.

Pooled cross-coordinate supports do not help: they win 1 of 140 eligible comparisons, at panelwise pooled-to-selected width ratios of $1.4\times$, $3.1\times$, $14.9\times$ and $130.2\times$ (Figure 4(c)). The mechanism is the interface itself: $S_t = \max_i |e_{t,i}|/v_{t,i}$ is set by the most under-scaled coordinate, so a pooled model that badly underpredicts one series inflates the scalar radius and hence all N sides at once. The ratios measure width, not a coverage trade-off: both sides cover, both with negligible unbounded fraction. On these panels, tested pooled support predictors do not improve the exact-box objective; we do not generalize this to all pooled models.

5.5 THEORY CHECKS AND ROBUSTNESS

A synthetic suite tests the theory against analytic objects up to $N = 2{,}048$ under 4 covariance structures. The independent-Gaussian width ratios reproduce the exact finite-N calculation to at most 4.3×10^{-3} relative error (Figure 2), and every arbitrary-correlation case satisfies (7) and (8). Matched-support ellipsoid, box and cross-polytope procedures agree at conditioning-appropriate precision—worst discrepancy 7% of its own tolerance—while differing in log-volume by up to 8,679 nats: Theorem 2 in its starkest form. Rank truncation undercovers under all 4 structures and collapses below 50% under 4 of them, with the native region exactly calibrated in every case. A multivariate-t arm is reported in Appendix F as an illustration and is not used to validate the Gaussian constants. On

[2] PCP on solar137 at $\alpha = 0.20$, where the sampled hull already covers 86.5% at $q = 0$, an irreducible coverage floor that no downward calibration can cross. This is a property of that envelope, not a failure of the calibrator (Appendix D).

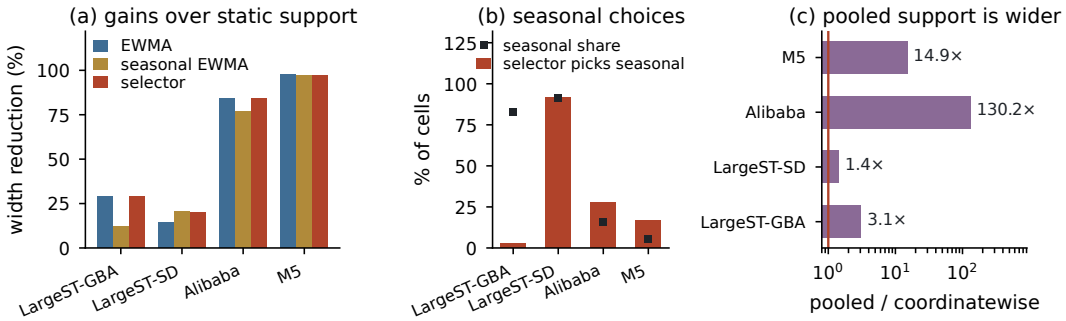


Figure 4: **Held-out confirmation** on 4 panels not used during development, with the support library and the update fixed in advance. In (b) the selector's seasonal choice does not track the panel's seasonal share, which is the point: LargeST-GBA is strongly periodic and the selector correctly avoids seasonal conditioning there.

the benchmark itself, compilation distortion, online reliability, and the width disadvantage of every published geometry are the same at each coverage level and under the persistence backbone.

6 Related Work and Limitations

Related work. We organize prior work by the object it targets. One group constructs *native multivariate regions*—sampled unions (Wang et al., 2023), Mahalanobis ellipsoids (Xu et al., 2024; Limmer, 2026), optimal-transport regions (Thurin et al., 2025), flow pushforwards (Lee et al., 2026), and time-local sample-cloud ellipsoids (Li et al., 2026). A second targets *boxes* directly, coupling per-coordinate intervals through a copula (Sun & Yu, 2024) or optimizing a coordinate error budget under joint validity (Schlembach et al., 2025). A third calibrates *a scalar level online* under distribution shift (Gibbs & Candès, 2021; Zaffran et al., 2022; Angelopoulos et al., 2023; Areces et al., 2025), and a fourth compares native regions in unified frameworks (Dheur et al., 2025; Johnstone & Ndiaye, 2025). Each asks how to build an efficient native region, an efficient native box, or how to calibrate a level online. We study the representation map *between* those tasks and what survives it, which is why our benchmark re-evaluates published methods at a stage they were not designed for while reporting their native-task performance unchanged alongside.

Limitations. Our results concern an axis-aligned output contract; nothing here argues against native geometry when the region itself is consumed, and Table 1 reports stage A so this stays visible. The stage-D guarantee is the realized-average-miss identity (11), not conditional coverage at each time. Exact boxification can be computationally unavailable for flexible learned regions, as for FCP here, so fewer published methods can be compared at the box interface than at the native one. That coordinatewise support forecasting dominates is benchmark-specific, and not claimed for pooled models we did not test; the interface theorem, by contrast, is distribution-free. Converting a box into a downstream action is a separate interface.

7 Conclusion

A conformal guarantee is attached to a particular prediction set, not to every representation derived from it. At an axis-aligned interface, exact recalibration reduces a native region to its coordinate-envelope path: containing conversions may become arbitrarily conservative, while finite non-containing conversions may lose coverage entirely. Across the evaluated high-dimensional forecasting methods, this representation change frequently reverses rankings, whereas exact online box calibration restores reliable realized coverage at no width cost. The remaining efficiency problem is therefore direct: predict useful coordinate supports and calibrate the box that is actually delivered.

Reproducibility Statement

The repository `https://anonymous.4open.science/r/facet-anonymous` contains code for every experiment. It needs no credentials, GPU or job scheduler: readers for the twelve public panels download the data on demand, and every input matrix, including that of the seeded synthetic control, is checked against a recorded SHA-256 checksum. One command runs the main

design—the 306 benchmark cells, the 144 confirmation cells and the Gaussian simulations—and a second regenerates from its outputs every number, table entry and figure in this paper, using the same selection and aggregation code that evaluates the pre-registered criteria; the auxiliary SPACE comparison and the heavy-tailed simulation are run separately (Appendices D and F). An independent second implementation, which does not import that code, then recomputes the headline aggregates. For a quicker check, the repository includes reference outputs for five runs, and recomputing a full-size benchmark cell and a Gaussian simulation study against them takes a few minutes on a laptop CPU, with no downloads. Appendix D lists the panels, their sources, the chronological split and the backbones; Appendix B gives, for each published family, its native set, native score, envelope formula and whether that envelope is exact—the information needed to reproduce or contest any row of Table 1; Appendix C gives the pseudocode and numerical conventions of Online ExactBox; Appendix A contains complete proofs with all assumptions stated; and Appendix H summarizes the pre-registered protocol, lists every deviation from it, and describes how the reported numbers were verified.

Ethics Statement

This work is a methodological contribution to uncertainty quantification and uses only public sensor, energy, cloud-telemetry, and retail benchmarks that contain no personal data. Its intended effect is to make multivariate forecast uncertainty harder to overstate: we show that a coverage guarantee established for one prediction set can be silently lost when the deployed system exposes a different set, and that in one documented case the deployed object retains almost none of it. The main risk we see is the converse of the usual one—a reader who takes our stage-D empirical coverage as a guarantee under arbitrary distribution shift. We state in §4 exactly what the online update controls, namely the realized average miss rate over the horizon, and say explicitly that it is not conditional coverage at each time.

AI use statement

Generative AI tools were used to assist with drafting and polishing the manuscript; to support the verification and stress-testing of mathematical arguments; and as a supporting tool for implementing, debugging, and testing the numerical experiments. All mathematical claims and proofs were independently derived and verified by the authors. All AI-assisted text and code were reviewed and validated by the authors. The authors take full responsibility for the final manuscript, its mathematical claims, the experimental implementation, and the reported results.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23047–23074. Curran Associates, Inc., 2023. doi: 10.52202/075280-1000. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/47f2fad8c1111d07f83c91be7870f8db-Abstract-Conference.html.

Felipe Areces, Christopher Mohri, Tatsunori Hashimoto, and John Duchi. Online conformal prediction via online optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1604–1649. PMLR, 2025. URL https://proceedings.mlr.press/v267/areces25a.html.

Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 153–167, 2017.

Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4883–4894, 2020.

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13444–13485. PMLR, 2025. URL https://proceedings.mlr.press/v267/dheur25a.html.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4): 603–680, 2013.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2021.

Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. Who limits the resource efficiency of my datacenter: An analysis of Alibaba datacenter traces. In *Proceedings of the IEEE/ACM International Symposium on Quality of Service*, pp. 1–10, 2019.

Chancellor Johnstone and Eugene Ndiaye. Exact and approximate conformal inference for multi-output regression. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 153–172. PMLR, 2025. URL https://proceedings.mlr.press/v266/johnstone25a.html.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Junghwan Lee, Chen Xu, and Yao Xie. Flow-based conformal prediction for multi-dimensional time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Uv3efQiPBZ.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Baishi Li, Kelvin J. L. Koa, and Ke-Wei Huang. SPACE: Sample-cloud predictive adaptive conformal ellipsoids for multivariate time-series forecasting. *arXiv preprint arXiv:2608.17333*, 2026. doi: 10.48550/arXiv.2608.17333. URL https://arxiv.org/abs/2608.17333.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Yannick Limmer. Filtered conformal ellipsoids for graph-native time series. *arXiv preprint arXiv:2606.17014*, 2026. doi: 10.48550/arXiv.2606.17014. URL https://arxiv.org/abs/2606.17014.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. LargeST: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Filip Schlembach, Evgueni Smirnov, and Mark H. M. Winands. Dynamic conformal prediction for multi-target regression: Optimising informational efficiency under joint validity. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 193–213. PMLR, 2025. URL https://proceedings.mlr.press/v266/schlembach25a.html.

Sophia Huiwen Sun and Rose Yu. Copula conformal prediction for multi-step time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ojIJZDNIBj.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527. PMLR, 2025. URL https://proceedings.mlr.press/v267/thurin25a.html.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8814–8836. PMLR, 2023. URL https://proceedings.mlr.press/v206/wang23n.html.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55076–55099. PMLR, 2024. URL https://proceedings.mlr.press/v235/xu24m.html.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25834–25866. PMLR, 2022.

A Complete Theory

A.1 Setup, attainment, and measurability

Fix a probability space and let $\mathcal{F}_t$ be the σ-field generated by everything available before Y_t is observed. All regions below are $\mathcal{F}_t$-measurable set-valued maps; every statement is made conditionally on $\mathcal{F}_t$, so each proof is a deterministic argument about one realization and the conditional versions follow by taking expectations. We suppress t where it is not needed.

Let $\mathcal{Q} \subseteq [0, \infty]$ be totally ordered and $q \mapsto A(q)$ nested. The entrance score is $s_A(y) = \inf\{q \in \mathcal{Q} : y \in A(q)\}$ with $\inf \varnothing = +\infty$. We assume *attainment*: if $Q_y := \{q : y \in A(q)\} \neq \varnothing$ then $\inf Q_y \in Q_y$.

Lemma 5 (Attainment is exactly the sublevel property)**.** *Nesting alone gives $s_A(y) \leq q \Rightarrow y \in A(q')$ for all $q' > q$, and $y \in A(q) \Rightarrow s_A(y) \leq q$. Attainment holds if and only if $y \in A(q) \iff s_A(y) \leq q$ for all y, q.*

Proof. If attainment holds and $s_A(y) \leq q$ then $y \in A(s_A(y)) \subseteq A(q)$ by nesting; the converse implication is definitional. Conversely, if the equivalence holds then $s_A(y) \leq s_A(y)$ gives $y \in A(s_A(y))$, which is attainment. □

Attainment is automatic when $\mathcal{Q}$ is finite or well-ordered, in particular on any computational grid, and holds in general whenever $A(q) = \bigcap_{q'>q} A(q')$, i.e. the path is right-continuous. It is not a technicality that can be dropped: with $\mathcal{Q} = \mathbb{R}$ and $A(q) = (-\infty, q)$ one has $s_A(q) = q$ but $q \notin A(q)$, and calibrating at s_A would emit a set missing its own boundary. Every path we construct is right-continuous by taking closed boxes.

A.2 Universal transfer

Proof of Lemma 1. If $R \subseteq E$ then $\mathbb{1}\{y \in R\} \le \mathbb{1}\{y \in E\}$ pointwise, so the inequality holds in expectation under every law. Conversely suppose $R \not\subseteq E$ and pick $y_0 \in R \setminus E$. Under $Y \equiv y_0$ we get $\mathbb{P}(Y \in R) = 1 > 0 = \mathbb{P}(Y \in E)$, contradicting the hypothesis. For the second claim, fix α and let $y_0 \in R \setminus E$ as above; then $Y \equiv y_0$ satisfies $\mathbb{P}(Y \in R) = 1 \ge 1 - \alpha$ while $\mathbb{P}(Y \in E) = 0$, so the infimum is 0 and attained. □

Two consequences are worth recording. First, the equivalence is with respect to *all* laws; if the law is restricted—for instance to those with a density bounded on a compact set—non-containing compilations can retain partial guarantees, and our claim is only that none is available distribution-free. Second, for a fitted region the required statement is $\mathbb{P}(R(q) \subseteq E(q)\ \forall q) = 1$; a compilation that contains the region on most fits and not all inherits a guarantee degraded by exactly the failure probability.

A.3 Box-interface collapse

Proof of Theorem 2. (i) The score formula. Nesting of B forces each ℓ_i to be nonincreasing and each u_i nondecreasing: if $q \le q'$ and some $\ell_i(q) < \ell_i(q')$, a point with i-th coordinate strictly between them and other coordinates inside $B(q)$ would lie in $B(q) \setminus B(q')$. Hence $\{q : \ell_i(q) \le y_i\}$ and $\{q : y_i \le u_i(q)\}$ are up-sets, with infima $\tau_{i,-}(y_i)$ and $\tau_{i,+}(y_i)$ as in (4). Now

$$y \in B(q) \iff \forall i:\ \ell_i(q) \le y_i \le u_i(q) \iff \forall i:\ \tau_{i,-}(y_i) \le q \text{ and } \tau_{i,+}(y_i) \le q,$$

using attainment coordinatewise. A real number dominates a finite family iff it dominates its maximum, so the set of admissible q is the up-set generated by $\max_i\{\tau_{i,-}(y_i), \tau_{i,+}(y_i)\}$, whose infimum is that maximum. This is (5); the maximum is attained because the family is finite.

(ii) Completeness of the invariant. Formula (5) depends on the native path only through (ℓ, u). Let $R^{(1)}, R^{(2)}$ be two native paths whose induced envelopes coincide, $\ell^{(1)} \equiv \ell^{(2)}$ and $u^{(1)} \equiv u^{(2)}$. Then $s_{B^{(1)}} \equiv s_{B^{(2)}}$ as functions on $\mathbb{R}^N$, so for any outcome sequence the two score sequences are equal; any calibration procedure that is a function of the observed scores alone therefore returns the same threshold, and the emitted boxes $B^{(1)}(\widehat{q})$ and $B^{(2)}(\widehat{q})$ coincide as sets. This covers split conformal, rolling and weighted quantiles, full conformal on the box score, and the online update of (10), since each depends on the data only through $\{s_B(Y_j)\}$.

(iii) Score domination and sharpness. Suppose $R(q) \subseteq B(q)$ for all q and let y have $s_R(y) = q^\star < \infty$. By attainment $y \in R(q^\star) \subseteq B(q^\star)$, so $s_B(y) \le q^\star = s_R(y)$; if $s_R(y) = \infty$ the inequality is trivial. Let $\Gamma : \mathbb{R}^n \to \mathbb{R}$ be coordinatewise nondecreasing—empirical quantiles, order statistics, weighted quantiles and maxima all are. Then $\widehat{q}_B = \Gamma(s_B(Y_1), \dots, s_B(Y_n)) \le \Gamma(s_R(Y_1), \dots, s_R(Y_n)) = \widehat{q}_R$, and $B(\widehat{q}_B) \subseteq B(\widehat{q}_R)$ by nesting. □

Remark 6 (What (iii) does and does not say). It says the exactly recalibrated box is a subset of the box obtained by calibrating natively and projecting, *deterministically, on every calibration sample*. It does not say the exactly recalibrated box is smaller than the native region, which is false in general, nor that its coverage is higher. Stage C is not uniformly preferable to stage A; it is preferable to stage B, which is the comparison a box interface actually forces.

A.4 Homothetic, asymmetric and non-homothetic families

Proposition 7 (Support-function form)**.** *Let $R(q) = \mu + qK$ with $K \subseteq \mathbb{R}^N$ bounded and star-shaped about 0, and let $h_K(w) = \sup_{z \in K} \langle w, z \rangle$ be its support function. The coordinate envelope of $R(q)$ is $[\mu_i - q\, h_K(-e_i),\ \mu_i + q\, h_K(e_i)]$, so with $b_i^+ = h_K(e_i)$ and $b_i^- = h_K(-e_i)$, both assumed positive, Theorem 2 gives* (6)*. If K is centrally symmetric then $b^+ = b^- =: b$ and $s_B(y) = \max_i |y_i - \mu_i|/b_i$.*

Only the axial values $h_K(\pm e_i)$ enter. Two consequences are used in the main text. For a *Minkowski sum* $K = K_1 + K_2$—the natural geometry of a factor-plus-remainder construction—support functions add, so $b_i = b_i^{(1)} + b_i^{(2)}$: a structured region built by combining a low-rank part and a coordinatewise part is, at the box interface, exactly the coordinatewise sum of their axial supports, and the factor loadings enter only through $\|U_i\|$-type quantities. For a *linear image* $K = DB_p$ the axial support is the $\ell_{p'}$ norm of the i-th row of D; in particular DB_2 gives $\|D_{i\cdot}\|_2$, which for $D = \Sigma^{1/2}$ is $\sqrt{\Sigma_{ii}}$, and DB_1 gives $\max_j |D_{ij}|$. This last pair is the ellipsoid-versus-cross-polytope construction of Figure 1(a) and of the T2 check in §F.

For *non-homothetic* paths nothing above is needed: Theorem 2 already covers arbitrary nested (ℓ, u), including paths whose shape changes with q and paths that are not derived from any region at all. What homothety buys is that the envelope is described by $2N$ numbers instead of $2N$ functions.

Corollary 8 (Fitted covariance collapse)**.** *For $K_\Sigma = \{z : z^\top \Sigma^{-1} z \le 1\}$ with $\Sigma \succ 0$, $h_{K_\Sigma}(e_i) = \|\Sigma^{1/2} e_i\|_2 = \sqrt{\Sigma_{ii}}$. Hence any two covariance estimates with the same diagonal—sample, Ledoit–Wolf shrunk, POET-factor, graphical-lasso—induce the identical exact box predictor, for every outcome sequence and every score-driven calibration rule.*

A.5 Rank deficiency

Proof of Corollary 3. Let $h \in \ker(A)$. Since $A \succeq 0$, $\ker(A) = \ker(A^+)$ and $\operatorname{ran}(A^+) = \operatorname{ran}(A) \perp \ker(A)$, so $A^+ h = 0$ and $z^\top A^+ h = 0$ for every z. Therefore $(z+th)^\top A^+ (z+th) = z^\top A^+ z$ for all $t \in \mathbb{R}$, and $R(q)$ is invariant under translation along $\ker(A)$. If $r < N$ then $\ker(A) \neq \{0\}$, so $R(q)$ is unbounded whenever it is nonempty; some coordinate i has $h_i \neq 0$, and along that direction the i-th coordinate is unbounded above and below. No finite box contains an unbounded set, so any finite compilation E satisfies $R(q) \not\subseteq E$, and Lemma 1 applies. □

The practical form of this is worth separating from the mathematics. A practitioner who fits a rank-r covariance and reads off coordinate scales gets the box with half-widths $q\sqrt{(\Sigma_r)_{ii}}$, which is finite and looks unremarkable. It is not the bounding box of $\{z : z^\top \Sigma_r^+ z \le q^2\}$, which is infinite in every dropped direction. Both objects fail, and Lemma 1 separates the two failures: the first is finite and non-containing, hence unguaranteed; the second is containing and vacuous. §F measures the first directly.

A.6 The Gaussian outer box

Proof of Theorem 4. Coverage. $Y_N \notin E_N$ requires $|Y_{N,i} - \mu_{N,i}| > q_{N,\alpha}\sigma_{N,i}$ for some i. Each marginal is exactly $\mathcal{N}(\mu_{N,i}, \sigma_{N,i}^2)$ regardless of the correlations, so a union bound gives $\mathbb{P}(Y_N \notin E_N) \le 2N\overline{\Phi}(q_{N,\alpha})$, which is (7). Since $q_{N,\alpha}^2 = N + \sqrt{2N} z_{1-\alpha} + O(1)$, we have $q_{N,\alpha} \sim \sqrt{N}$, and $2N\overline{\Phi}(\sqrt{N}) \le 2Ne^{-N/2} \to 0$; the bound is uniform over correlation structures because only marginals were used.

Width. The Bonferroni box $\prod_i [\mu_i \pm c_{N,\alpha}^{\mathrm{Bon}} \sigma_i]$ is valid by the same union bound, so the narrowest valid centered box satisfies $W_N^\star \le c_{N,\alpha}^{\mathrm{Bon}} \bar{\sigma}_N$ with $\bar{\sigma}_N = N^{-1} \sum_i \sigma_{N,i}$, while $W(E_N) = q_{N,\alpha} \bar{\sigma}_N$ exactly; dividing gives the first inequality of (8). The standard tail expansion $\Phi^{-1}(1-u) = \sqrt{2\log(1/u)}(1 + o(1))$ as $u \downarrow 0$ with $u = \alpha/2N$ gives $c_{N,\alpha}^{\mathrm{Bon}} = (1+o(1))\sqrt{2\log(2N/\alpha)}$, and combining with $q_{N,\alpha} = (1+o(1))\sqrt{N}$ yields the stated rate.

The independent optimum. Under $\Sigma_N = I_N$ a centered box $\prod_i [-a_i, a_i]$ has coverage $\prod_i (2\Phi(a_i) - 1)$. The map $a \mapsto \log(2\Phi(a) - 1)$ is concave on $a > 0$, since its second derivative is $-[2a\phi(a)(2\Phi(a) - 1) + 4\phi(a)^2]/(2\Phi(a) - 1)^2 < 0$, so Jensen gives $N^{-1} \sum_i \log(2\Phi(a_i) - 1) \le \log(2\Phi(\bar{a}) - 1)$ with $\bar{a} = N^{-1} \sum_i a_i$: replacing any box by the cube of its own mean half-width weakly increases coverage. The narrowest valid box is therefore the cube of radius $c_{N,\alpha}^{\mathrm{iid}}$ with $2\Phi(c) - 1 = (1-\alpha)^{1/N}$, i.e. $c_{N,\alpha}^{\mathrm{iid}} = \Phi^{-1}((1 + (1-\alpha)^{1/N})/2)$. Since $1 - (1-\alpha)^{1/N} = \alpha/N + O(N^{-2})$, the same tail expansion gives $c_{N,\alpha}^{\mathrm{iid}} = (1+o(1))\sqrt{2\log N}$ and hence the ratio $\sqrt{N/(2\log N)}$. □

Both statements in (8) are needed for different reasons. The Bonferroni comparison is valid for arbitrary Σ_N and is what makes the blow-up a statement about correlation-free geometry; the independent constant is exact and is what Figure 2(b) is checked against. Substituting the asymptotic rate for the exact ratio would be a material error at these dimensions: at $N = 2{,}048$ the exact ratio is 11.4 while $\sqrt{N/2 \log N}$ is visibly different, which is why the two curves are drawn separately.

A.7 Exact split validity of the delivered box

Proposition 9. *Let $(X_j, Y_j)_{j=1}^{n+1}$ be exchangeable conditionally on the design block, let $v_t, \widehat{\mu}_t$ be functions of the design block and X_t only, and let S_j be as in* (9)*. With $k = \lceil (n+1)(1-\alpha) \rceil$ and $\widehat{q} = S_{(k)}$,*

$$1 - \alpha \;\le\; \mathbb{P}\big(Y_{n+1} \in C_{n+1}(\widehat{q})\big) \;\le\; 1 - \alpha + \tfrac{1}{n+1}$$

with the upper bound under almost-sure distinctness of the scores.

Proof. Predictability makes $S_1, \dots, S_{n+1}$ exchangeable. By (9), $Y_{n+1} \in C_{n+1}(\widehat{q}) \iff S_{n+1} \le \widehat{q}$, which is the standard split-conformal rank statement; the exact-membership equivalence is what lets it be applied to the delivered box rather than to a region containing it. □

A.8 The online identity, with projection

Proposition 10. *For $\lambda_{t+1} = \Pi_{[\lambda_-, \lambda_+]}[\lambda_t + \eta(\alpha - M_t)]$ and r_t as in* (10)*, identity* (11) *holds pathwise for every sequence $(M_t) \in \{0,1\}^T$, with no assumption on how M_t is generated. If the projection is inactive then $r_t \equiv 0$ and $|M_T - \alpha| \le D/(\eta T)$ with $D = \lambda_+ - \lambda_-$.*

Proof. Summing $\lambda_{t+1} - \lambda_t = \eta(\alpha - M_t) + r_t$ over $t \le T$ telescopes to $\lambda_{T+1} - \lambda_1 = \eta(T\alpha - \sum_t M_t) + \sum_t r_t$; divide by ηT and rearrange. If no projection occurs then $r_t = 0$ and $|\lambda_1 - \lambda_{T+1}| \le D$. □

In our implementation $[\lambda_-, \lambda_+] = [-\frac{1}{2}, \frac{3}{2}]$, wider than $[0, 1]$, so the level may leave the unit interval without being clipped. At $\lambda_t \le 0$ the threshold is infinite and the emitted box is the whole space, which covers and raises the level again; because one step lowers λ_t by at most $\eta < \frac{1}{2}$, the lower bound can never bind. At $\lambda_t \ge 1$ the threshold is zero and the emitted box is the envelope at $q = 0$—for Online ExactBox the point forecast, which misses and lowers the level. The upper bound binds only when the envelope at $q = 0$ already covers more often than the target, as for the sampled hull of PCP in the single exception of Appendix D; then $r_t < 0$, and (11) places the realized miss rate below α, in the safe direction. The cost of the wide interval is paid in width rather than hidden in r_t: an unbounded step has infinite width. We therefore report the unbounded-output fraction alongside every width, and it is the reason width is summarized by a median over all steps rather than a mean over bounded ones (§2).

A.9 Scale invariance and support stability

The score (9) satisfies $S_{cv} = S_v / c$ for $c > 0$, so the emitted box is unchanged when v is rescaled: only the projective class of v matters. Let $d_{\mathrm{proj}}(v, w) = \frac{1}{2}\big(\max_i \log(v_i/w_i) - \min_i \log(v_i/w_i)\big)$ be the induced Hilbert projective metric. Then for the p-th quantile width functional W_p,

$$e^{-2d_{\mathrm{proj}}(v,w)} \;\le\; \frac{W_p(v)}{W_p(w)} \;\le\; e^{2d_{\mathrm{proj}}(v,w)}. \tag{13}$$

The proof is that $v_i/w_i \in [m, M]$ implies $S_w\, m \le S_v \cdot (\text{coordinate ratios}) \le S_w\, M$ pointwise, so quantiles and hence widths move by at most $M/m = e^{2d_{\mathrm{proj}}}$. The practical reading is that a support model needs to get the *relative* coordinate profile right to within a bounded factor, and gets global scale for free from the calibrator—which is precisely why a pooled model that badly misses a single coordinate is so costly (§5.4): one coordinate suffices to make d_{proj} large.

A.10 Clipped scores and saturation

The construction that first raised the interface question clipped a remainder score at a robustness cap τ. The following says what that does at the interface, and generalizes the specific failure recorded in §G.

Proposition 11 (Saturation makes a branch vacuous)**.** *Let a region be defined by a conjunction $R(q) = \{y : g_1(y) \le \rho_1(q)\} \cap \{y : g_2(y) \le \rho_2(q)\}$ where g_2 is clipped at τ, i.e. $g_2 = \min(\tilde{g}_2, \tau)$. If $\rho_2(q) \ge \tau$ then the second constraint holds for every y and $R(q) = \{y : g_1(y) \le \rho_1(q)\}$. The emitted box then has the coordinate supports of the first branch alone, while the reported region indicator—which is what an ACI update sees—remains that of the conjunction. Coverage measured on the indicator therefore exceeds coverage of the emitted box whenever the second branch would have bound.*

This is the mechanism behind the coverage discrepancy of §G: a capped remainder branch that has gone vacuous still contributes an indicator of 1, so the online calibrator is driven by an event that is not the event the user experiences. §G reports the measured size of the gap and the repaired, uncapped construction.

B Interface Algebra and Provenance by Method

For each family we record the native set, the native calibrated score, the coordinate-envelope formula, whether that envelope is *exact* (equal to the true coordinate extrema rather than an approximation), the behavior under rank deficiency, and the cost of computing the envelope. Throughout, e_t is the residual, $\widehat{\mu}_t$ the backbone forecast, and N the dimension. The interface label used in Table 1 follows one rule fixed in advance: *native box* if the published method emits coordinate intervals, *documented projection* if the source supplies its own coordinate summary, and *derived box* if we compute the envelope ourselves from the published region. A method is eligible for stages B–D only if its exact envelope is available in closed form or by a certified computation.

Mahalanobis and Ledoit–Wolf. Native set $\{z : z^\top \widehat{\Sigma}^{-1} z \le q^2\}$ with $\widehat{\Sigma}$ the ridge-stabilized sample covariance or the Ledoit–Wolf shrinkage estimator (Ledoit & Wolf, 2004); native score $\sqrt{e^\top \widehat{\Sigma}^{-1} e}$. By Corollary 8 the envelope is $\pm q\sqrt{\widehat{\Sigma}_{ii}}$, exact, at cost $O(N)$ once the diagonal is known. The two differ only in the estimator, hence only in the diagonal, which is why their stage-D rows in Table 1 are close and their supports frequently match exactly.

POET-factor. Native set as above with $\widehat{\Sigma} = BB^\top + \Psi$ for an r-factor loading matrix B and diagonal Ψ (Fan et al., 2013). The envelope is $\pm q\sqrt{\|B_{i\cdot}\|_2^2 + \Psi_{ii}}$, exact. This is the sharpest statement of the collapse for a factor model: the entire factor structure enters the delivered box through one scalar per coordinate.

L2-ball. Native set $\{z : \|z\|_2 \le q\}$, an isotropic template whose coordinate support is 1 in every direction; the envelope is the cube $[-q, q]^N$, exact. Its stage-B coverage of 1.000 in Table 1 is the purest instance of Theorem 4: the box of a Euclidean ball calibrated to 90% in high dimension covers essentially everything.

PCP. Native set a union of balls around M conditional samples, $\bigcup_m B(\widehat{y}^{(m)}, q)$ (Wang et al., 2023); native score the distance to the nearest sample. The exact envelope is $[\min_m \widehat{y}_i^{(m)} - q,\ \max_m \widehat{y}_i^{(m)} + q]$, exact and $O(MN)$. Because the sample spread is present at $q = 0$, this envelope has a strictly positive *coverage floor*: no downward calibration can shrink it below the sample hull. That floor causes the one overcovering method–cell pair discussed in §D, and it is a property of the envelope, not of the calibrator.

MultiDimSPCI. Native set $\{z : z^\top \widehat{\Sigma}_r^{+} z \le q^2\}$ with a rank-r truncation (Xu et al., 2024). Corollary 3 applies: the native region is an unbounded cylinder and has no finite containing box. We report the source's own documented coordinate summary, $\pm q\sqrt{(\widehat{\Sigma}_r)_{ii}}$, and label it a documented projection

precisely because it is finite and therefore necessarily non-containing. Stage A is unaffected and on target.

Bonferroni and MaxNorm. Native boxes. Bonferroni calibrates each coordinate at level α/N; MaxNorm calibrates one scalar on $\max_i |e_i|/w_i$ for a predictable scale w. The envelope is the set itself, so $B \equiv A$ and stages A and B coincide identically—this is the control row that shows the $B - A$ column measures compilation and nothing else.

CopulaCPTS. Native box obtained by coupling per-coordinate intervals through an empirical copula fitted on a held-out block (Sun & Yu, 2024). The envelope is the set itself. Its low stage-A coverage in Table 1 reflects the difficulty of estimating an N-dimensional copula at these calibration sizes, and is recovered fully once exact box membership is recalibrated at stages C and D.

DCP-MT. Native box from an optimized error budget (Schlembach et al., 2025): sort each coordinate's calibration scores and choose order statistics minimizing $\sum_i \log \alpha_i^{(n_i)}$ subject to $\sum_i \epsilon_i \leq \alpha$. We solve the resulting multiple-choice knapsack *exactly* by dynamic programming over the integer budget rather than greedily; greedy is not merely approximate here but wrong in direction, since moving further into the tail increases the marginal value of the next budget unit, so returns are increasing and the exchange argument fails. When $\lfloor \alpha(n_{\mathrm{cal}} + 1) \rfloor < N$ no allocation gives every coordinate even its widest finite interval; we then emit the sample maximum and flag the cell. This union-bound wall binds in 116 of 306 records and is the method's own behavior at these calibration sizes, not a defect of our implementation.

SPACE. Native set a time-local ellipsoid from a sample cloud (Li et al., 2026), $\Sigma_t = DC_tD$ with C_t the empirical covariance of wrapper-standardized samples. Its coordinate support is $\sqrt{(\Sigma_t)_{ii}}$, and the standardization cancels exactly: $(\Sigma_t)_{ii} = D_i^2 \mathrm{Var}(r_i) = \mathrm{Var}(y_i)$, so the emitted box depends on the cloud only through its coordinatewise standard deviation. Everything that makes the region time-local and non-spherical—the correlations, the eigenvectors, the stabilized pseudo-inverse—has no effect on the box whatsoever. We therefore treat SPACE as a sample-cloud *support path* and report it as an auxiliary comparison, because our clouds come from a pre-specified causal residual bootstrap rather than the deep probabilistic forecasters of the original paper; its own window-selection rule is replaced by the common online update, so what is tested here is its time-local shape, not its calibration. With M samples the cloud covariance has rank at most $M - 1$, so for $N > M$ the native region is unbounded off the retained subspace and this box does not contain it—the same structure as rank-truncated MultiDimSPCI, measured rather than patched.

FCP. Native set a flow pushforward $f_t(B(0, q))$ of a Euclidean ball (Lee et al., 2026). Its exact coordinate extrema require globally optimizing $\pm e_i^\top f_t(z)$ over the ball, a non-convex neural-ODE problem with no certificate; the reference implementation supplies no envelope routine. Under the eligibility rule fixed in advance, FCP is therefore reported at the native stage only. We record explicitly why the obvious workaround is refused: a hull of pushed samples is an *inner* approximation, so it would understate the true envelope, overstate stage-B coverage loss, and bias the comparison towards our own thesis. This is an ineligibility under our criterion, not a deficiency of FCP.

C Online ExactBox: Algorithm and Protocol

Algorithm 1 Online ExactBox at level α

select $v \in$ Eq. (12) minimizing validation width among candidates meeting the validation coverage threshold
$\lambda_1 \leftarrow \alpha$; window $\mathcal{W} \leftarrow \varnothing$
for $t = 1, 2, \ldots$ over the final-calibration block, then the test block **do**
    compute $\widehat{\mu}_t, v_t$ from data strictly before t; $q_t \leftarrow \widehat{Q}_{\mathcal{W}}(1 - \lambda_t)$
    emit $C_t(q_t) = \prod_i [\widehat{\mu}_{t,i} \pm q_t v_{t,i}]$
    observe Y_t; $S_t \leftarrow \max_i |Y_{t,i} - \widehat{\mu}_{t,i}|/v_{t,i}$; $M_t \leftarrow \mathbb{1}\{S_t > q_t\}$
    $\lambda_{t+1} \leftarrow \Pi_{[-1/2,\, 3/2]}[\lambda_t + \eta(\alpha - M_t)]$; append S_t to $\mathcal{W}$ (trailing W scores)
end for

Conventions. The calibrator starts afresh, with $\lambda_1 = \alpha$ and an empty window, at the start of each calibration block: the validation-calibration block when candidates are compared, and the final-calibration block before test. Scores from that block fill the window and update λ_t but are not themselves evaluated. For $0 < \lambda < 1$, $\widehat{Q}_{\mathcal{W}}(1-\lambda)$ is the empirical quantile of the window (with linear interpolation) at the finite-sample level $\min\{\lceil(|\mathcal{W}|+1)(1-\lambda)\rceil/|\mathcal{W}|, 1\}$; it is $+\infty$ when the window is empty or $\lambda \leq 0$, and 0 when $\lambda \geq 1$. An infinite threshold emits an unbounded box, counted as infinite width (§A.8). The window is $W = 500$ and the primary learning rate is $\eta = 0.05$, both fixed before the confirmation experiments; every cell also stores a run at $\eta = 0.01$ as a sensitivity check. Supports are floored at 10^{-8} to keep the score finite for dead coordinates. The calibrator never sees the score it is about to judge: at every t, q_t depends only on $S_1, \ldots, S_{t-1}$.

Support library. Static is the design-block coordinate scale, frozen. EWMA_h is a coordinatewise exponentially weighted mean of squared residuals with half-life h steps. Seasonal-EWMA_c is the same recursion applied within a time-of-day bucket, with c the per-bucket half-life in cycles. All are coordinatewise by construction, so no coordinate's support can be corrupted by another's. The pooled ablations of §5.4 are never available to the selector. FACTORVOL reduces the design-block log absolute residuals to their top r principal components and forecasts the component scores by an EWMA, with $r \in \{1, 3, 8\}$ at half-life 48 and $r = 3$ at half-life 192. SHAREDSCALE is a single ridge regression, shared by all coordinates, of the log absolute residual on its own two lags and EWMA, the cross-sectional mean at lag one, and zero or two calendar harmonics, at half-life 48.

Split geometry. Each forecast origin is cut chronologically into design, validation-calibration, validation-evaluation, final-calibration, and test blocks, disjoint and in that order. Geometry is fitted on design only; the selector reads validation-calibration and validation-evaluation only; test is scored once. The same rule picks every family's representative configuration: a candidate meets the validation coverage threshold if its validation coverage is at least $1 - \alpha - 0.01$; candidates whose validation width exceeds twenty times the median across candidates are excluded first; the narrowest candidate meeting the threshold is taken, or, if none does, the one with the highest validation coverage.

D BENCHMARK DATA AND SETUP

Table 2 lists the panels. For each, the released reader documents the source, column ordering, resampling and transformations, and checks the resulting input matrix against a recorded checksum. The solar and LOOP Seattle panels are read from the GIFT-Eval mirror (Aksu et al., 2024), and the M5 panel is 200 food items sold in California. Electricity and smart-meter loads enter as $\log(1+y)$; the Azure and Alibaba telemetry are 5-minute utilizations differenced once. `indep200` is a negative control with no shared structure: 200 mutually independent AR(1) series with coefficients drawn uniformly from $[0.3, 0.85]$ and independent, slowly varying log-volatilities.

Table 2: **Forecasting panels.** T is the length of the input series, N the dimension after the eligibility filter (a range where it varies by origin), and s the seasonal lag of the ridge-AR backbone.

Panel	Domain	T	N	s	Source
Primary benchmark					
`solar137`	energy	8,760	137	24	NREL solar (Lai et al., 2018)
`elec370`	energy	35,065	321–332	24	UCI electricity (Trindade, 2015)
`smartmeters200`	energy	20,352	200	48	London smart meters (Godahewa et al., 2021)
`azure128`	cloud	1,781	128	24	Azure VMs (Cortez et al., 2017)
`loop323`	traffic	8,760	323	24	LOOP Seattle (Cui et al., 2020)
`metrla207`	traffic	34,249	207	288	METR-LA (Li et al., 2018)
`pemsbay325`	traffic	52,093	325	288	PEMS-BAY (Li et al., 2018)
`largest2327`	traffic	8,760	2,327	24	LargeST GBA 2019 (Liu et al., 2023)
`indep200`	synthetic	12,000	200	–	Independent AR(1) control
Held-out confirmation					
`largest_gba_2020`	traffic	8,784	2,352	24	LargeST GBA 2020 (Liu et al., 2023)
`largest_sd_2019`	traffic	8,760	705–706	24	LargeST SD 2019 (Liu et al., 2023)
`alibaba314`	cloud	2,193	313–314	288	Alibaba 2018 (Guo et al., 2019)
`m5retail200`	retail	1,500	200	7	M5 (Makridakis et al., 2022)

Origins and blocks. The last 40% of each panel is divided into six consecutive evaluation windows (three for Azure); origin k is tested on the k-th window, capped at $3{,}000$ steps, and everything before that window is its history. The first half of the history is used to standardize each coordinate and to fit the backbone. Coordinates whose standard deviation there falls below 10^{-3} times the median across coordinates are removed before any method sees the panel; this is why N can be smaller than the input dimension. The backbone residuals on the second half of the history form, in order, the design (40%), validation-calibration (15%), validation-evaluation (15%) and final-calibration (30%) blocks.

Backbones. The primary backbone is a ridge autoregression fitted separately for each coordinate of the standardized panel, on the coordinate's own lags 1–3, its seasonal lag s, the cross-sectional mean at lag 1, and an intercept; it is fitted once per origin and applied one step ahead. The sensitivity backbone is persistence, $\widehat{\mu}_t = Y_{t-1}$.

The SPACE comparison. SPACE is run with the ridge-AR backbone on `solar137`, `azure128`, `metrla207` and `largest2327`, at every origin and coverage level, with sample clouds from the causal residual bootstrap of §B. Its results form the separate sample-cloud row of Table 1.

The one exception to the reliability criterion. Exactly one method–cell pair with a long test block (at least 500 steps) misses its target by more than three points: PCP on solar137 at $\alpha = 0.20$, where the stage-D coverage is 86.5% against a target of 80%. This is *over*coverage by 6.47 points, and it is structural. As noted in §B, PCP's envelope is the hull of M conditional samples inflated by q; at $q = 0$ the hull already covers 86.5% of outcomes on this panel, so no value of λ_t can bring coverage down to 80%. The calibrator drives q_t to zero and stays there, which is the correct behavior. No pair undercovers by that margin, and undercoverage is the direction that costs a user anything. We report this criterion as failed rather than redefining it, and state the estimand that is statistically meaningful: over all method–cell pairs in the 306 cells, the median absolute stage-D coverage error is 0.11 points, and $3{,}503$ of the $3{,}510$ long pairs (99.8%) are within one point.

E Held-Out Confirmation

The four confirmation panels (Table 2) are LargeST-GBA 2020 (the region of `largest2327`, but the following year, which was not used during development), LargeST-SD 2019 (a region not used during development), Alibaba cluster telemetry (non-traffic, weak cycle), and M5 retail (non-traffic, near-aperiodic). They span seasonal shares from 0.05 to 0.91. They are split and evaluated exactly as the benchmark panels (Appendix D), at six origins, both backbones and all three coverage levels.

Per-panel gains. The two non-traffic panels are where conditioning matters most, and there the choice of temporal model barely matters: on Alibaba the selector gains 84.2% over static support with ordinary and seasonal EWMA at 84.2% and 76.6%, and it takes the seasonal model in 10 of 36 cells; on M5 the corresponding figures are 97.2%, 97.4%, 97.2% and 6 of 36. Seasonality is therefore a real choice only on the two periodic panels, which is exactly where ordinary and seasonal EWMA differ.

Diagnostics. The prespecified diagnostic d_{common}—the fraction of log-absolute-residual variance explained by the top five principal components, corrected for its null expectation—separates the traffic panels (0.46, 0.43) from the non-traffic ones (0.05, 0.14) out of sample, as it was designed to. It does not, however, identify panels on which pooling helps, because there are none.

Why the pooled predictors lose. Panelwise pooled-to-selected width ratios are $1.4\times$, $3.1\times$, $14.9\times$ and $130.2\times$. The stability bound (13) explains the spread: the cost of a support model is governed by the projective distance of its coordinate profile from a good one, and a single badly under-scaled coordinate makes that distance large no matter how well the other $N - 1$ are predicted. A max-type interface has no averaging to hide behind. Both sides of each comparison meet the validation coverage threshold and both have negligible unbounded fraction, so the ratios measure width, not a coverage trade-off.

Reliability and the sampling floor. The pre-registered criterion asked for 90% of confirmation cells within one percentage point; the observed rate is 79.9% over 144 cells at a median absolute

error of 0.29 points, so the criterion fails as written. It was mis-specified, and we record why rather than replacing it after the fact. Coverage here is measured over test blocks as short as 100 steps, where a single additional miss moves coverage by a full point. A procedure calibrated *exactly* would land within one point in at most 59.2% of cells on the longest of the four panels and 31.5% on the shortest, so the criterion is unreachable in principle at these block lengths and its failure carries no information about the calibrator. The estimand that does carry information is the standardized deviation $z = (\text{cov} - \text{target})/\text{se}$: pooled over 144 cells its median is -0.01 with standard deviation 0.29 and 0 cells beyond $|z| > 2$. An unbiased, exactly calibrated procedure gives median z near zero and standard deviation near one. The observed spread is well below one because the online update pushes back after each miss, so its coverage sequence is negatively autocorrelated and varies less over a finite window than an independent binomial sequence with the same marginal rate.

F Synthetic Theory Checks

Ten dimensions $N \in \{4, 8, \ldots, 2{,}048\}$, 20,000 draws per dimension, at $\alpha = 0.10$, under independent, equicorrelated ($\rho = 0.9$), block, and low-rank-plus-diagonal covariance, plus a multivariate-t arm.

T1, exact finite-N ratios. Simulated width inflation matches $q_{N,\alpha}/c^{\text{iid}}_{N,\alpha}$ to at most 4.3×10^{-3} relative error across all dimensions.

T2, matched supports. An ellipsoid, a box and a cross-polytope are constructed with *identical* coordinate supports by three deliberately different routes—a matrix square root, a diagonal read-off, and a row maximum of the generator—and their exact box scores are compared. Agreement is limited by the conditioning of Σ, since a symmetric eigensolver has relative error $O(\kappa\varepsilon)$; the tolerance is therefore $10\kappa\varepsilon$, floored at machine precision, and the worst observed discrepancy is 7% of its own tolerance. The three templates differ in log-volume by up to 8,679 nats. Same procedure, same predictions, volumes differing by thousands of nats: this is Theorem 2 with nothing empirical in the way.

T3, one-sided bounds under arbitrary correlation. Every non-independent cell satisfies (7) and (8), confirming that the Bonferroni comparison, not the independent constant, is what holds in general.

T4, rank truncation. Truncation ranks are $1, 2, 4, \ldots, 32$, below the ambient dimension. The box built from a truncated covariance's own diagonal undercovers under all 4 structures and collapses below 50% under 4 of them, while the native region is exactly calibrated at 0.900 throughout. The deficit is created purely by the compilation step, with no change of calibration, and the recorded containment violations show directly that a finite box over an unbounded region carries no transferred guarantee. Under equicorrelation every eigenvalue after the first is the same, so a truncation at rank two or more keeps an arbitrary basis of a degenerate eigenspace, and its diagonal, hence its box coverage, depends on the eigensolver and can differ across platforms. The rank-one truncation is unique, and it alone produces both verdicts under every structure.

The multivariate-t arm. The equicorrelated structure with multivariate-t draws on five degrees of freedom, reported as an illustration of behavior under heavy tails and run separately from the main design. It is *not* used to validate the Gaussian constants of Theorem 4, which are stated for Gaussian marginals and checked only in the Gaussian cells.

G Motivating Case Study and Rejected Alternatives

The investigation began with a structured box built from a factor-plus-remainder geometry with a clipped remainder score. It is retained here as a compact case study, because the way it failed is the reason the interface question is worth asking, and because each rejected alternative rules out an obvious objection to the main text.

Cap saturation and the region-versus-box gap. In high dimension the calibrated remainder threshold exceeded the cap τ, making the remainder branch vacuous in the sense of Proposition 11.

The region indicator that drove the online update was therefore effectively factor-only, while the emitted box was the conjunction's envelope: reported coverage referred to an event the user never experienced, and the delivered box undercovered.

Repaired geometry. Removing the cap and retuning the budget split restores nominal coverage of the emitted box. Under a validation-only comparison the repaired construction does not beat a plain exact-box baseline on a predictable per-coordinate scale, and the structured geometry earns nothing at the interface—which is what Theorem 2 predicts, since after exact recalibration the two differ only through their coordinate supports.

Learned static geometry. Directly optimizing shape parameters to minimize calibrated width rediscovers coordinatewise σ rather than any richer structure, across every basis we tried, and does not meet the criterion we had set in advance for pursuing it.

Factor and shared supports, and the seasonality confound. Exploratory results initially suggested that pooled cross-series supports helped on traffic panels. Adding a coordinatewise *seasonal* control removed the effect entirely: the apparent benefit of shared structure was a calendar effect that a purely coordinatewise model captures. Because this correction was made after seeing results, the seasonal model and the learning rate were confirmed on held-out panels (§E), which is the purpose of the confirmation experiment.

The general lesson is not about this particular geometry. An apparently strong structured-box result can disappear once the operational event is measured correctly and the strongest simple baseline is tuned as carefully as the proposed method.

H Pre-Registration, Deviations, and Reproducibility

Pre-registration. Before the confirmation experiments we fixed in writing the benchmark's methods, the eligibility and interface-labeling rules, the selection rule and its degeneracy guard, the validation coverage threshold, the width metrics, the bootstrap, the learning rate and window, the confirmation panels, and a numerical criterion for each claim together with the claim it supports. The criteria carry the labels used in the released code: P1–P4 concern the reliability and width of the common online calibrator on the benchmark, C1–C3 the width of predicted supports relative to published geometries and the equivalence of matched envelopes, Q1–Q4 the conditional gain, seasonal control, pooling and reliability of the held-out confirmation, and T1–T4 the simulations of §F. This appendix summarizes the protocol and lists every departure from it.

Deviations. Each departure from the protocol is listed with its reason: the exclusion of FCP from stages B–D under the exact-envelope criterion, and the refusal to substitute a sampled hull; the residual-bootstrap sample cloud used for the SPACE comparison; the redefinition of headline width as a median over all steps with the unbounded fraction reported beside it, after we found that averaging bounded steps only would let a saturating calibrator hide unbounded outputs; the condition-aware tolerance in T2, adopted after a fixed 10^{-12} tolerance proved to be a statement about floating point rather than about the mathematics; the separation of "undercovers" from "collapses" in T4; and the mis-specified reliability criterion Q4 analyzed in §E. Two criteria were not met, and both are reported: P3, because of the PCP overcoverage of §D, and Q4.

One deviation was found by the independent reconstruction described below and changed a reported number. The rank-reversal statistic originally scored a pair by comparing two strict inequalities, which is not symmetric in the two families when their stage-D widths are equal. Such ties are not incidental here: when two coordinate envelopes coincide the emitted boxes are identical, so the widths agree bitwise—the collapse this paper reports—and 122 of the $2{,}950$ eligible pairs are exact ties, with 21 of 199 cells tied for best. The count was therefore partly a fact about the order in which families were enumerated, and eight orderings of the same results gave eight answers. A tie means the box task expressed no preference, so we now count a reversal only when both tasks order a pair strictly and oppositely, and treat the winner as the set of families attaining the best score. Reversals fall from 39.8% to 36.2%; the winner statistic is unchanged. No run was repeated, and no other number moved.

From results to paper. A result digest carries the aggregate behind every printed number, and beside it a per-cell table records, for each panel, origin, backbone, level and stage, the selected configuration and its reported metrics. The manuscript's numbers, tables and figures are generated from the digest by one script that *imports* its selection and bootstrap functions from the code that evaluates the pre-registered criteria, so a number in the paper and the same number in a criterion come from the same code path rather than from two implementations that happen to agree. The digest also records what produced it: the SHA-256 of each experiment's source, a checksum of every result file it reads, the interpreter and library versions, and a checksum of every input panel.

Sharing a code path removes disagreement but cannot detect a mistake inside the shared code, so the headline aggregates are also recomputed by a second implementation that reads the raw results and does not import the aggregation code or the script that produces the paper's numbers. It re-derives the selection rule, the criteria, the stage table, the compilation gaps, the ranking statistics and the collapse gains from the written protocol, and requires every deterministic quantity to agree to 10^{-9} relative error; the bootstrap intervals are checked for overlap under a deliberately different seed, since requiring equality there would test only that the same random draws were replayed. This is what found the ranking defect above.

Released code. The released code runs every experiment from the public data on a CPU, and its report command regenerates the digest, the per-cell table, the tables, the numbers and the figures from the new results and then runs the independent reconstruction. Its tests check, among other things, that the calibration threshold never reads the current or a future score, that fitted supports read no evaluation data, that neither the backbone fit nor the choice of coordinates reads test data, exact box membership for asymmetric envelopes, the analytic Gaussian constants, the finite-sample boundary of the split quantile, the identity behind the SPACE support, the DCP-MT allocation against exhaustive search, and that unbounded steps enter the headline width. Each input matrix is checked against the SHA-256 checksum recorded for it. A checksum is a statement about the reader's arithmetic as well as about the data—two NumPy releases can disagree in the last bit of $\exp$, which is enough to change every hash—so each panel also carries a fingerprint of its moments, and a panel that matches its fingerprint but not its hash is accepted and reported as not bitwise verified.

Five reference runs are included: a full-size cell of the synthetic control, a `solar137` cell with SPACE, a cell of the confirmation panel `m5retail200`, and the independent and equicorrelated Gaussian simulations. A comparison command diffs a recomputed run against its reference field by field, ignoring only wall-clock timings, and reports the worst relative deviation rather than only a verdict. A quick check that recomputes the synthetic cell and the independent simulation needs no downloads and takes a few minutes on a laptop CPU; on a second machine, with a different operating system and library versions, it reproduced the stored references to within 2×10^{-12} relative error. The references are regression checks, not an archive of all results; the aggregate targets reported in this paper are recorded with the code, so that a full rerun can be compared with them.