SCALE ADAPTATION IN CONFORMAL PREDICTION: EFFICIENCY AND COVERAGE ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides prediction intervals with finite-sample marginal coverage under exchangeability. However, valid intervals can still be unnecessarily wide when prediction errors are asymmetric or prediction difficulty varies across inputs. We study two related procedures to improve interval efficiency. Efficient Conformal Prediction with Flexible Quantile Allocation (ECP-FQ) chooses how to divide the error budget between the lower and upper residual tails. This adjusts interval placement, but gives every input the same width. ECP-FQ+ extends this approach by using estimated prediction-error magnitude to adjust interval centers and widths across inputs. Our main procedure keeps the center at the point prediction and selects one of five exponents to control how strongly width responds to this estimated error magnitude. Both methods select their interval shape on held-out data and use a separate calibration sample to preserve marginal coverage. Under an oracle location–scale model with varying scale and suitable regularity and integrability conditions, we show that slightly weakening proportional width adjustment reduces expected length even after finite-sample calibration. We also analyze imperfect scale estimates and characterize how scale adaptation affects interval length and conditional coverage. Empirically, compared with choosing between constant-width intervals and widths proportional to estimated prediction-error magnitude, allowing intermediate adjustments yields clear width reductions in 18 of 24 simulation settings with learned predictors. At 90% nominal coverage, the five-exponent selector reduces width by 1.8–12.5% in six of ten real dataset–learner settings, with pointwise paired 95% confidence intervals excluding zero; the remaining four settings show small, statistically inconclusive differences.

1 INTRODUCTION

Regression models are widely used to support decisions, but a point prediction alone does not reveal how uncertain the outcome is. Prediction intervals provide a range of plausible outcomes around that prediction. To be useful, these intervals should contain the outcome with the intended frequency while remaining narrow enough to inform a decision. Conformal prediction addresses the coverage requirement: under exchangeability, it constructs intervals with finite-sample marginal coverage (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018). This guarantee holds on average across future inputs and outcomes, but does not ensure the same coverage at every input.

However, coverage alone does not determine how informative an interval is. There are several methods can attain the same marginal coverage and produce very different average lengths (Gneiting & Katzfuss, 2014; Yang & Kuchibhotla, 2024). Efficiency can be improved by learning from related source data when target samples are limited (Zhang et al., 2026), but it also depends on how prediction errors are converted into interval endpoints. Split conformal prediction (SCP) uses absolute prediction errors to place the same radius on either side of a point prediction. This construction is simple, but it does not use the direction of the errors. When residuals are asymmetric, shifting the interval can retain coverage while reducing its length. Figure 1 shows how adjusting interval placement can produce a shorter interval with the same coverage.

We begin with Efficient Conformal Prediction with Flexible Quantile Allocation (ECP-FQ). It uses signed residuals to choose how much of the error budget lies below and above the interval. A held-

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.

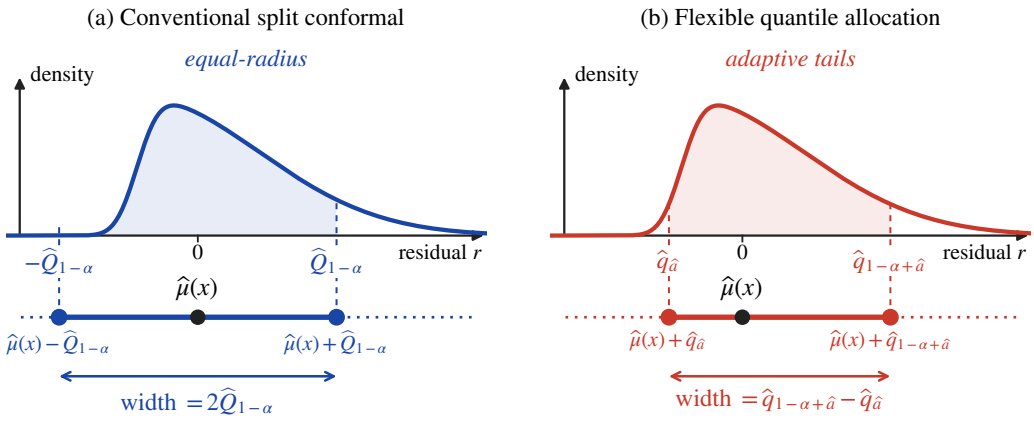

Figure 1: **Interval placement under asymmetric residuals.** Ordinary SCP uses equal geometric radii around a point prediction; flexible signed endpoints can place the same target probability in a shorter interval. Equal radii do not imply equal tail-error probabilities. Both procedures compute their final endpoints using calibration data; the drawing illustrates the population geometry.

out allocation sample selects the tail budgets to obtain a short interval, and a separate calibration sample determines the final endpoints. Thus, interval placement is learned independently of final calibration. Nevertheless, ECP-FQ gives every input the same width. ECP-FQ+ uses fitted scales to adjust centers through a shift parameter and widths through a scale exponent. Both select interval shape before independent calibration, but ECP-FQ+ cannot reproduce every ECP-FQ interval when scales vary across inputs.

Our main question is how strongly interval width should respond to estimated prediction difficulty. When the center remains at the point prediction, $\gamma = 0$ gives ordinary SCP with constant width, while $\gamma = 1$ gives fully normalized SCP with width proportional to the estimated error scale. Intermediate exponents allow widths to vary across inputs without following the scale estimate as strongly. Although full normalization is natural when seeking similar coverage across inputs, it need not minimize average length at a fixed marginal coverage level. For example, in the two-group Gaussian model in Appendix A.11, reducing the exponent can shorten intervals on average while decreasing coverage in the noisier group and increasing it in the less noisy group.

Previous work recognizes the tradeoff between average length and coverage at individual inputs (Kiyani et al., 2024; Xie et al., 2024). We explain how partial normalization can reduce average length and how estimation, selection, and finite calibration affect these gains. Our experiments compare five exponents against a choice between ordinary and fully normalized intervals, using the same fitted predictors and data split. Our contributions are:

- **Method.** ECP-FQ allocates the error budget between residual tails to adjust interval placement. ECP-FQ+ uses estimated error scales to adjust centers and widths across inputs. Both select interval parameters before independent calibration. Our main procedure keeps the center at the point prediction and selects among five scale exponents.
- **Theory.** For estimated scales, we derive a criterion based on densities at the interval endpoints that identifies when a small change in the exponent reduces expected length. When the true location and scale functions are known, we prove that slightly reducing the exponent below one strictly decreases expected length. This result holds with a finite calibration sample, provided that scale varies across inputs and the stated assumptions hold. We also explain how coverage changes across inputs and bound the effects of predictor estimation, the candidate grid, and parameter selection.
- **Empirical evidence.** We assess intermediate exponents and center adjustments, compare with boosted conformal prediction, and examine interval width and coverage when calibration is performed separately within groups defined by estimated prediction difficulty.

2 RELATED WORK

Learning interval endpoints. Conformalized quantile regression (CQR) calibrates estimated conditional quantiles (Romano et al., 2019), with scaled variants compared by Sesia & Candès (2020). Neural networks, collaborating networks, relaxed quantile regression, and percentile methods offer alternative endpoints (Kivaranovic et al., 2020; Zhou et al., 2021; Pouplin et al., 2024; Zou et al., 2026). Distributional, histogram, and interquantile methods capture more of the conditional response distribution (Chernozhukov et al., 2021; Sesia & Romano, 2021; Luo & Zhou, 2025; Guo et al., 2026), while density-based regions may be disconnected (Izbicki et al., 2022). We study contiguous intervals with point and scale predictors fixed before shape selection and calibration.

Selecting conformal scores. Independent score selection followed by calibration preserves validity under exchangeability (Yang & Kuchibhotla, 2024). However, tuning on the final calibration data can introduce coverage bias (Zeng et al., 2025). Sequential Predictive Conformal Inference (SPCI) estimates residual quantiles for time series (Xu & Xie, 2023), while Wang (2026) studies independent tail selection and finite-sample calibration ranks. We follow this separation of selection and calibration, focusing on how interval width should respond to fitted scales.

Optimizing marginal length. Conformal Prediction with Length-Optimization (CPL) constructs short prediction sets under coverage constraints (Kiyani et al., 2024). EffOrt learns interval centers, while its adaptive extension Ad-EffOrt also adjusts widths (Le Bars & Humbert, 2025). Boosted conformal prediction (BoostedCP) optimizes scores and discusses undercoverage in regions with greater noise (Xie et al., 2024). Further work considers structured sets, local asymmetry, and joint center and radius adjustment (Gao et al., 2025; Marques F. & Graziadei, 2026; Su et al., 2026). Within a simple scale family, we link length reduction under imperfect scales to densities at interval endpoints and derive an exact expected-length derivative accounting for finite calibration. Appendix Q provides further comparisons.

3 PROBLEM SETTING AND INTERVAL CONSTRUCTION

3.1 PREDICTION TARGET AND DATA ROLES

For observations $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, with input X_i in the input space $\mathcal{X}$ and real response Y_i, we seek a contiguous interval with short expected length and marginal coverage at least $p = 1-\alpha$, where $\alpha \in (0,1)$ is the error budget. Split the data independently of their values into fitting, allocation, and final-calibration samples D_F, D_A, D_C, of sizes n_F, n_A, m; the reserve budget is $N = n_A + m$. Fit preprocessing, a point predictor $\widehat{\mu}$, and a positive scale proxy $\widehat{\sigma}$ on D_F. Candidate grids and tie rules may use D_F but are fixed before observing D_A, D_C.

Ordinary split conformal prediction (SCP) calibrates absolute prediction errors and adds the same radius to both sides of $\widehat{\mu}(x)$. This symmetry need not assign equal probabilities to the two tails. We instead select interval shape on D_A, then independently calibrate it on D_C. Coverage requires exchangeability of D_C and the test observation conditional on D_F, D_A. Efficiency results additionally assume conditional i.i.d. sampling and stated regularity conditions. The target is marginal coverage, rather than coverage at each input or within each group.

3.2 ECP-FQ: CHOOSING THE TWO RESIDUAL TAILS

Signed residuals retain the direction of prediction error. Let $R = Y - \widehat{\mu}(X)$, with conditional distribution function $F_R(r) = \mathbb{P}(R \le r \mid D_F)$ and quantile function at probability level u, $Q_R(u) = \inf\{r : F_R(r) \ge u\}$. For $0 < a < \alpha$, allocating error budgets a below and $\alpha - a$ above the interval gives population width $W(a)$. ECP-FQ selects this allocation by its empirical width:

$$W(a) = Q_R(p+a) - Q_R(a), \qquad \widehat{a} \in \operatorname*{arg\,min}_{a \in \mathcal{G}} \{\widehat{Q}_A(p+a) - \widehat{Q}_A(a)\}. \tag{1}$$

Here $\widehat{Q}_A(u)$ is the $\lceil n_A u \rceil$th ordered residual on D_A, and $\mathcal{G}$ contains 25 equally spaced points in $[0.01\alpha, 0.99\alpha] \cap [1/(m+1), \alpha - 1/(m+1)]$. An empty range returns $\mathbb{R}$. Ties favor the allocation nearest $\alpha/2$, then the smaller value.

To see why unequal tails can help, suppose F_R has positive continuous density $f_R = F_R'$ near both endpoints. At an interior optimum, $W'(a) = 1/f_R\{Q_R(p+a)\} - 1/f_R\{Q_R(a)\} = 0$: the endpoint densities balance. This density assumption explains efficiency; it is unnecessary for coverage.

However, selection and calibration serve different roles: D_A chooses the tail budget, while D_C supplies the final endpoints. Freeze $\widehat{a}$ and order the final-calibration residuals $R_i = Y_i - \widehat{\mu}(X_i)$ as $R_{(1)} \le \cdots \le R_{(m)}$. We set $R_{(0)} = -\infty$ and $R_{(m+1)} = +\infty$ and define the lower and upper ranks with a finite-sample correction as follows:

$$j_L = \lfloor (m+1)\widehat{a} \rfloor, \quad j_U = \lceil (m+1)(p+\widehat{a}) \rceil, \qquad \widehat{C}_{\mathrm{FQ}}(x) = [\widehat{\mu}(x) + R_{(j_L)}, \widehat{\mu}(x) + R_{(j_U)}]. \tag{2}$$

Fixing $a = \alpha/2$ gives the signed equal-tail control, which differs from absolute-residual SCP.

3.3 ECP-FQ+: CHOOSING HOW INTERVALS VARY WITH THE INPUT

Although FQ adjusts placement, its width is constant across inputs. ECP-FQ+ extends the shape-selection idea to input-dependent centers and widths. For a center shift $b \in \mathbb{R}$ and scale exponent $\gamma \in [0, 1]$, define the center c_b and nonnegative score $S_{b,\gamma}$ by

$$c_b(x) = \widehat{\mu}(x) + b\widehat{\sigma}(x), \qquad S_{b,\gamma}(x, y) = \frac{|y - c_b(x)|}{\widehat{\sigma}(x)^\gamma}. \tag{3}$$

Thresholding this score gives radius proportional to $\widehat{\sigma}(x)^\gamma$; smaller exponents reduce relative width variation. The name denotes a conceptual extension: FQ and FQ+ are not nested when scale varies. Neither represents arbitrary conditional interval shapes.

Let Λ be a finite grid of K candidate pairs $\lambda = (b, \gamma)$. For n scores, $Q^+_{n,\lambda}(p)$ denotes their $\lceil (n+1)p \rceil$th order statistic, appending $+\infty$ at rank $n+1$. On D_A, compute a pilot threshold $\widetilde{q}_\lambda$ and average width multiplier $\widetilde{v}_\lambda$, then select their product:

$$\widetilde{q}_\lambda = Q^+_{n_A,\lambda}(p), \qquad \widetilde{v}_\lambda = \frac{2}{n_A} \sum_{i \in D_A} \widehat{\sigma}(X_i)^\gamma, \qquad \widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \widetilde{J}_\lambda, \quad \widetilde{J}_\lambda = \widetilde{q}_\lambda \widetilde{v}_\lambda. \tag{4}$$

The objective $\widetilde{J}_\lambda$ is the average interval width on D_A computed using the pilot threshold $\widetilde{q}_\lambda$. It estimates population average length, but is not an exact expression for expected length after calibration on m observations. When candidates tie, we choose the smallest γ, followed by the smallest b. This rule also applies when all objective values are infinite.

After selecting $\widehat{\lambda} = (\widehat{b}, \widehat{\gamma})$, we keep these parameters fixed and replace the pilot threshold with a new threshold $\widehat{q} = Q^+_{m,\widehat{\lambda}}(p)$ computed only from the independent calibration sample D_C. The resulting prediction interval is

$$\widehat{C}_+(x) = \left[c_{\widehat{b}}(x) - \widehat{q}\,\widehat{\sigma}(x)^{\widehat{\gamma}},\ c_{\widehat{b}}(x) + \widehat{q}\,\widehat{\sigma}(x)^{\widehat{\gamma}} \right]. \tag{5}$$

Calibration on an independent sample preserves marginal coverage after interval-shape selection, but does not guarantee a reduction in interval length.

3.4 PRACTICAL CHOICES AND COMPUTATION

Our primary *scale-only* rule fixes $b = 0$ and searches $\Gamma_5 = \{0, 0.25, 0.5, 0.75, 1\}$. Its endpoints recover ordinary and normalized SCP. The matched *two-exponent selector* searches $\Gamma_2 = \{0, 1\}$ with the same fitted functions, allocation data and criterion, and final-calibration sample. Thus their comparison isolates intermediate exponents.

To study placement, we also search 25 shifts in $[-1.5, 1.5]$, crossed with Γ_5. A second, residual-based grid rescales those shifts using out-of-fold fitting errors so that center resolution reflects prediction accuracy. Both joint grids have 125 candidates; fixing $\gamma = 1$ isolates center selection. Appendix K defines the residual grid in equation 45.

The scale proxy comes from a log-error regressor trained on three-fold out-of-fold errors in D_F. Its predictions are exponentiated, median-normalized, and clipped to $[.1, 10]$; it need not estimate conditional standard deviation. Intervals are reported in original response units. With cached residuals and scales, selection costs $O(Kn_A)$ time and $O(n_A)$ memory, followed by $O(m)$ calibration. Appendix B gives preprocessing, fitting, and baseline details.

4 COVERAGE AND EFFICIENCY ANALYSIS

In this section, we establish coverage after interval-shape selection, identify when a smaller scale exponent reduces expected interval length, and quantify the effects of fitting and selection errors. We distinguish three settings: intervals based on population quantiles, fixed interval parameters calibrated on a finite sample, and the full procedure with fitted predictors and interval parameters selected from data.

4.1 INDEPENDENT CALIBRATION

Once fitting and selection are complete, the score is fixed. Independent final calibration then supplies the coverage guarantee.

Theorem 1 (Independent final calibration)**.** *Conditional on D_F, D_A, suppose D_C and the test observation are exchangeable. Both equation 2 and equation 5 satisfy*

$$\mathbb{P}\{Y_{\text{new}} \in \widehat{C}(X_{\text{new}}) \mid D_F, D_A\} \geq p. \tag{6}$$

Writing $k = \lceil (m+1)p \rceil$, the coverage of ECP-FQ+ equals $k/(m+1)$ if its selected scores are almost surely distinct. For equation 2, distinct residuals give coverage $(j_U - j_L)/(m+1)$.

Why independence matters. Without ties, the future score has a uniform rank among $m+1$ scores; ties make coverage conservative. This argument permits imperfect predictors because D_C stays outside selection (Figure 5). The probability averages over calibration and the future observation. It does not promise coverage conditional on the realized D_C or at every input x.

4.2 SCALE RESPONSE WITH FITTED PREDICTORS

Full normalization makes interval width proportional to the fitted scale. However, an imperfect scale may warrant a weaker response. Fix a fitted center $c(x)$ and positive scale proxy $t(x)$. Conditional on D_F, define the absolute-residual distribution $H_x(r) = \mathbb{P}\{|Y - c(x)| \leq r \mid X = x, D_F\}$; its derivative h_x, when it exists, is the corresponding density. Choose the population threshold q_γ so that $\mathbb{E}H_X(q_\gamma t(X)^\gamma) = p$, and let $W(\gamma) = 2q_\gamma \mathbb{E}t(X)^\gamma$ be expected width. Write $\pi_\gamma(x) = H_x(q_\gamma t(x)^\gamma)$ for coverage at input x. Expectations here are over a fresh input. Taking $c = \widehat{\mu} + b\widehat{\sigma}$ and $t = \widehat{\sigma}$ recovers Section 3.

Theorem 2 (Sensitivity for a fixed fitted scale)**.** *Suppose $0 < t_- \leq t(X) \leq t_+ < \infty$ and $q = q_1 > 0$. Assume H_x has a continuous local derivative h_x, with an integrable envelope as specified in Appendix A.1, and $\mathbb{E}w(X) > 0$, where $w(x) = t(x)h_x(qt(x))$. Put $\ell_w = \mathbb{E}\{w(X)\log t(X)\}/\mathbb{E}w(X)$. Then q_γ is differentiable near one, and*

$$W'(1) = 2q\big[\mathbb{E}\{t(X)\log t(X)\} - \mathbb{E}t(X)\ell_w\big], \tag{7}$$

$$\partial_\gamma \pi_\gamma(x)|_{\gamma=1} = q\, w(x)\{\log t(x) - \ell_w\}, \tag{8}$$

Here w measures how coverage responds to the threshold, and ℓ_w is its weighted mean of log scale. Thus width depends on residual density at the endpoints, beyond the proxy's ordering of difficult inputs. Under the input law reweighted by $t(X)/\mathbb{E}t(X)$, $W'(1) > 0$ exactly when $h_X(qt(X))$ and $\log t(X)$ have negative covariance. A sufficiently small decrease in γ then shortens intervals, lowers coverage above $\exp(\ell_w)$ where $w > 0$, and raises it below that threshold where $w > 0$. No correct location–scale model is needed. However, a correctly ordered proxy can have the opposite width derivative, and a favorable derivative need not justify a coarse grid step (Appendix A.4).

4.3 THE LOCAL GAIN SURVIVES FINITE CALIBRATION

Does this opportunity remain when calibration estimates the threshold? For a fixed exponent under an exact scale model, it does, without a density lower bound or large-sample approximation.

Theorem 3 (Finite-calibration length and coverage sensitivity)**.** *Suppose $Y = \mu(X) + \sigma(X)\varepsilon$, with $\varepsilon \perp X$ and nonconstant $0 < s_- \leq \sigma(X) \leq s_+ < \infty$. Fix $b \in \mathbb{R}$, let $(X_i, \varepsilon_i)_{i=1}^m$ be i.i.d. calibration pairs, and suppose $T_i = |\varepsilon_i - b|$ has an atomless distribution. For $k = \lceil (m+1)p \rceil \leq m$, let $\widehat{q}_\gamma$*

be the kth order statistic of $T_i\sigma(X_i)^{1-\gamma}$. Let $T_{(k)}$ be the kth smallest T_i. If $\mathbb{E}T_{(k)} < \infty$, define $\overline{W}_m(\gamma) = 2\mathbb{E}\widehat{q}_\gamma\,\mathbb{E}\sigma(X)^\gamma$, the expected length averaged over calibration and a fresh input. Then

$$\overline{W}'_m(1) = 2\mathbb{E}T_{(k)}\,\mathrm{Cov}\{\sigma(X), \log\sigma(X)\} > 0. \tag{9}$$

Consequently, every sufficiently small fixed $h > 0$ gives $\overline{W}_m(1-h) < \overline{W}_m(1)$, while each fixed exponent has exact marginal coverage $k/(m+1)$. If additionally $H_b(t) = \mathbb{P}(T_i \le t)$ is continuously differentiable on $(0,\infty)$ and satisfies the integrable envelope in Appendix A.3, then for almost every x, the calibration-averaged coverage $\overline{\pi}_m(x;\gamma) = \mathbb{E}_{D_C} H_b\{\widehat{q}_\gamma\sigma(x)^{\gamma-1}\}$ satisfies

$$\overline{\pi}'_m(x;1) = C_m\{\log\sigma(x) - \mathbb{E}\log\sigma(X)\}, \qquad C_m = \mathbb{E}[T_{(k)}H'_b(T_{(k)})] > 0, \tag{10}$$

A sufficient envelope condition is $\sup_{t>0} tH'_b(t) < \infty$.

Why the gain survives calibration. At full normalization, the scores depend only on standardized errors. Let I_k index the observation with the kth smallest score. Its scale is independent of $T_{(k)}$ and retains the original scale distribution. Distinct scores keep their order under a sufficiently small exponent change within each sample, giving threshold derivative $-T_{(k)}\log\sigma(X_{I_k})$. Averaging separates $T_{(k)}$ from log scale, while differentiating the future width multiplier gives $\mathbb{E}\{\sigma(X)\log\sigma(X)\}$. Combining these terms yields the covariance in equation 9. Calibration changes the local gain's magnitude through $\mathbb{E}T_{(k)}$, while scale variation determines its sign. Appendix A.2 justifies differentiation under the expectation.

Reducing the exponent slightly lowers calibration-averaged coverage above the geometric-mean scale $\exp\{\mathbb{E}\log\sigma(X)\}$ and raises it below. At $\gamma = 1$, this coverage equals $k/(m+1)$ at every input; the derivative in equation 10 averages to zero. These are averages for fixed candidates using the true predictors, rather than guarantees for each calibration sample or a selected fitted exponent.

4.4 Selection, fitting, and grid error

For the selected rule, condition on D_F and define $J_\lambda = 2q_\lambda\mathbb{E}\widehat{\sigma}(X)^\gamma$, where q_λ is the candidate score's population p-quantile. Let $\widehat{J}$ be expected future width conditional on all observed data and $\delta \in (0,1)$ the error probability. Under all hypotheses of Theorem 7, including conditional i.i.d. sampling, bounded multipliers, and its density and rank neighborhood conditions, with probability at least $1-\delta$,

$$\widehat{J} \le \min_{\lambda\in\Lambda} J_\lambda + \mathcal{R}_\delta, \qquad \mathcal{R}_\delta = O\left(\sqrt{\frac{\log(8K/\delta)}{n_A}} + \sqrt{\frac{\log(4/\delta)}{m}} + \frac{1}{n_A} + \frac{1}{m}\right), \tag{11}$$

with explicit constants in Appendix A.5. The remainder $\mathcal{R}_\delta$ measures statistical error. Searching K candidates costs a $\log K$ term in selection; independent final calibration needs none. For global FQ, one residual empirical CDF controls the entire allocation grid (Proposition 11).

Theorem 8 compares with the unrestricted population optimum W^*_{opt} by adding family approximation A_Λ and fitting error E_F:

$$\widehat{J} \le W^*_{\mathrm{opt}} + A_\Lambda + E_F + \mathcal{R}_\delta. \tag{12}$$

For a grid step h below normalization, Corollary 9 compares the gain $Dh - Lh^2/2$ with $2E_F + \mathcal{R}_\delta$; D is the oracle width derivative and L bounds curvature. A coarse grid or noisy selection can therefore erase the local opportunity. Even the population power family need not attain the Gaussian optimum (Appendix A.11). These efficiency bounds require stronger assumptions than coverage; some small-sample experiments do not satisfy them. Appendix A.18 gives computable binomial alternatives, which can still be inconclusive. Neither bound guarantees sample-by-sample dominance. Appendix A provides the full assumptions, constants, and proofs.

5 Empirical analysis

In this section, we evaluate whether the proposed adjustments yield shorter intervals when predictors and interval parameters are learned from data. We first examine the benefits of intermediate scale exponents and center adjustments, then compare performance with BoostedCP. Finally, we assess how requiring coverage within predefined groups affects interval width.

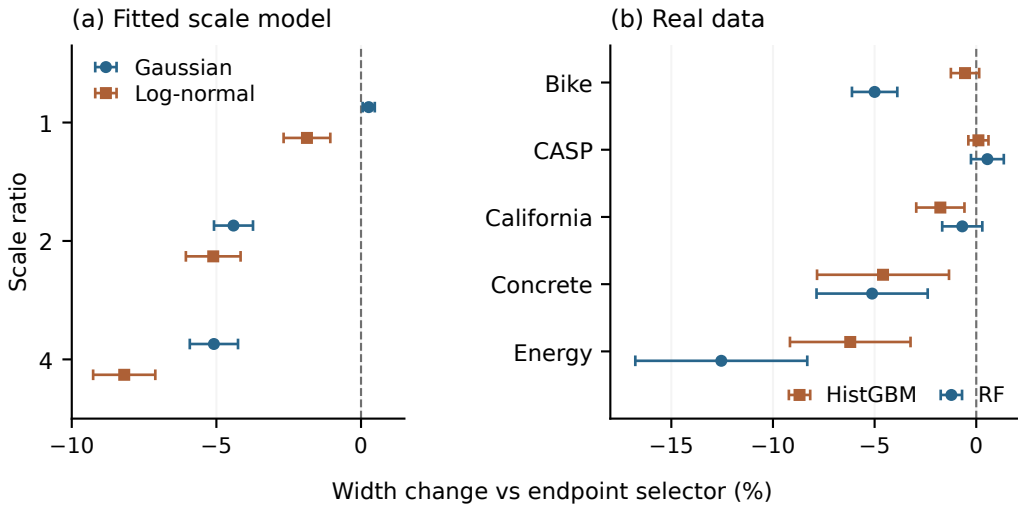

Figure 2: **Value of intermediate scale exponents.** Paired percentage width change of the five-exponent rule relative to the matched $\{0, 1\}$ selector; negative values favor five exponents. Left: all six Gaussian/log-normal and scale-ratio cells with $n_F = n_A = m = 1000$, over 50 independent repetitions. Right: all ten real dataset–learner settings at 90%, over 20 splits. Bars are pointwise 95% paired t intervals. Appendix O includes the remaining 18 fitted settings and the other coverage levels.

5.1 Protocol and matched controls

We use CASP, Concrete, Bike Sharing, California Housing, and Energy Efficiency (Rana, 2013; Yeh, 1998; Fanaee-T, 2013; Pace & Barry, 1997; Tsanas & Xifara, 2012), with HistGBM and random-forest predictors, 20 random splits, and nominal levels 80%, 90%, and 95%. Splits assign 50%/15%/15%/20% to fitting/allocation/calibration/testing. Large datasets use a fixed 6,000-row subsample; Concrete and Energy retain 1,030 and 768 rows. Bike random splits assess randomized sampling, not forecasting. Appendix B gives exact counts and implementation details.

Our primary scale-only ECP-FQ+ rule fixes $b = 0$ and searches five exponents. Its $\{0, 1\}$ control shares predictors, D_A, D_C, and test observations, isolating the value of intermediate exponents. Fixed pooled rules calibrate on $D_A \cup D_C$, accounting for the observations selected rules spend on allocation. We average within-repetition percentage width changes and construct paired 95% t intervals. Simulation repetitions are independent; real-data intervals describe split variation conditional on the fixed dataset. All intervals are pointwise, without multiplicity adjustment.

5.2 Intermediate exponents after fitting

We first test whether intermediate exponents help when both location and scale must be learned. Each simulation refits location and out-of-fold log-error scale under

$$Y = 2\sin(\pi X_1) + X_2^2 + 0.5X_3 + 0.5G + \{1 + (r-1)G\}\varepsilon. \tag{13}$$

Here $G \sim \text{Bernoulli}(0.5)$ and $X_j \sim \text{Uniform}[-1, 1]$, $j = 1, \ldots, 5$, are mutually independent; ε is independent Gaussian or centered, unit-variance log-normal noise. We cross scale ratios $r \in \{1, 2, 4\}$, $n_F \in \{300, 1000\}$, and $n_A = m \in \{300, 1000\}$, with 50 repetitions and 5,000 test observations per cell. Varying both r and n_F helps distinguish the effects of true scale variation from those of estimation error.

The five-exponent rule improves over the two-candidate control in 18 of 24 cells, has two small deteriorations, and is inconclusive in four by the paired intervals. Intermediate exponents are selected in 989 of 1,200 runs. At $n_F = n_A = m = 1000$, the Gaussian and log-normal ratio-four reductions are 5.1% and 8.2%; however, the homoskedastic Gaussian case becomes .27% wider. Intermediate

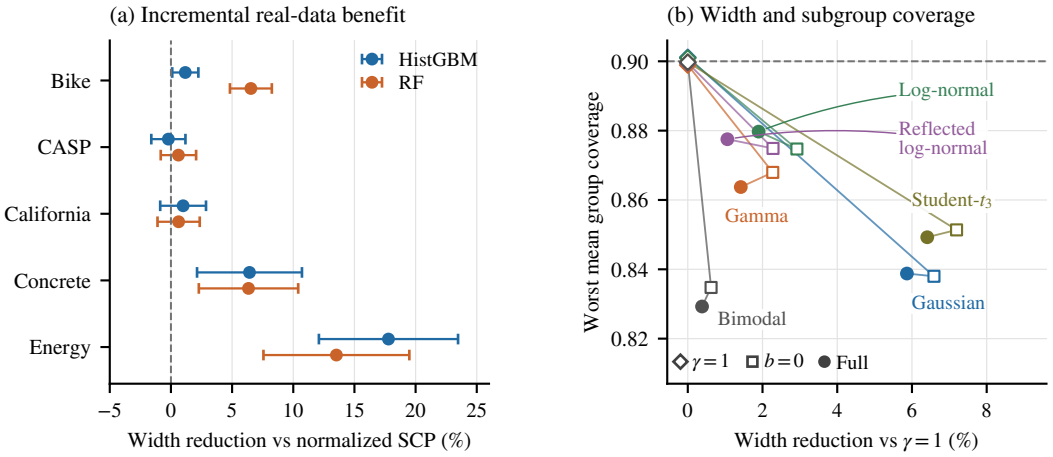

Figure 3: **Width reduction and coverage allocation at 90% nominal coverage.** Left: original-grid joint FQ+ versus pooled normalized SCP on all ten dataset–learner settings; points show mean paired width reductions and bars show pointwise 95% t intervals over 20 splits. Right: six known-predictor error families with $n_A = m = 500$ and 300 repetitions; width reduction uses the ratio of mean widths, and coverage is the smaller of the two mean group coverages. Diamonds, squares, and circles denote the $\gamma = 1$ restriction, scale-only, and the original joint family. Positive reductions indicate shorter intervals; marginal validity does not imply 90% coverage in each group.

exponents thus add value beyond accepting or discarding normalization, although extra selection can be unnecessary without scale variation. Appendix O reports every cell and exact-CDF coverage.

Figure 2 shows six real-data reductions with intervals below zero at 90%, from 1.8% on California/HistGBM to 12.5% on Energy/RF. The other four are inconclusive; none clearly favors two candidates. Intermediate exponents occur in 157 of 200 fits, and mean coverage ranges from .892 to .912. There are six clear reductions at 80% and two at 95%, so the benefit depends on the coverage level and available calibration information.

5.3 The incremental role of placement

In this subsection, we evaluate whether learning interval centers can add learning scale response. Figure 10 in Appendix L compares joint ECP-FQ+ with pooled normalized SCP and with scale-only selection. The residual-grid joint rule reduces width by 3.2–16.5% against normalized SCP across the 24 fitted settings. However, it clearly beats scale-only in just one cell and is clearly worse in six. None of the ten 90% real-data settings shows a clear additional gain; Concrete/RF becomes 2.7% wider, with interval $[.4, 5.0]\%$. Finer shifts are selected frequently, so coarse center resolution alone does not explain the limited benefit.

Global ECP-FQ instead learns how to divide error between signed residual tails. Comparing it with fixed signed equal tails isolates that choice. It has no clear advantage over pooled signed equal tails in the ten real settings. In controlled-proxy studies with $n_A = m = 1000$, reductions against calibration-matched signed tails are 7.9–11.7% for Gamma and 24.1–28.4% for log-normal errors, while Gaussian intervals become 0.8–1.1% wider. Placement therefore helps mainly with asymmetric errors in these controls; scale-only remains the supported default for the studied fitted problems (Appendices G and K). The controlled-group results in Figure 3 show that shorter intervals can leave one group less well covered, motivating the group-calibration comparison below.

5.4 Comparison with BoostedCP and Computational Cost

To test competitiveness with a more flexible score, we compare with BoostedCP's local-score optimizer (Xie et al., 2024), using the same HistGBM initialization. It selects a rate from $\{0.1, 20\}$ and up to 500 update rounds by five-fold CV within D_F, then calibrates on unused $D_A \cup D_C$. Thus it

Table 1: **Comparison with training-sample BoostedCP-local.** Twenty HistGBM splits per dataset at 90% nominal coverage. BoostedCP selects rate and round by five-fold D_F CV and calibrates on unused $D_A \cup D_C$; scale-only selects on D_A and calibrates on D_C. Entries are mean width (coverage). Changes are paired percentages for scale-only relative to BoostedCP, with pointwise 95% split intervals.

Dataset	Scale-only	BoostedCP, pooled	Width change (%)
Bike	126.467 (0.898)	127.854 (0.900)	-1.08 [-2.16, -0.01]
CASP	13.578 (0.893)	13.680 (0.897)	-0.77 [-1.99, +0.46]
California	1.428 (0.900)	1.437 (0.901)	-0.64 [-2.23, +0.96]
Concrete	16.952 (0.912)	18.419 (0.910)	-7.54 [-11.99, -3.10]
Energy	2.023 (0.899)	1.928 (0.895)	+5.59 [+0.33, +10.84]

can learn from the larger fitting sample under the same total label budget. Appendix N documents the implementation and a calibration-matched variant.

Table 1 shows a 7.5% scale-only reduction on Concrete and a small 1.1% reduction on Bike, whose interval barely excludes zero; CASP and California are inconclusive. On Energy, however, scale-only is 5.6% wider, with coverage .899 versus .895. This reverses the allocation-trained result and shows sensitivity to the optimization and calibration protocol. The optimizer's average within-split minimum tercile coverage ranges from .833 to .882, versus .838 to .869 for scale-only; neither guarantees group coverage.

Median post-fitting time is 107.85 s for the two-rate search versus .145 ms for scale-only; shared base fitting takes .339 s. These measured times include CV and candidate refits but exclude shared prediction, cache access, and one-time JIT startup. They depend on hardware and implementation. Appendix N gives update diagnostics; Appendix M retains the allocation-trained comparison. Other baselines also reveal limits: shortest-interval CQR is narrower than the full family on CASP, and scaled CQR is shorter in fitted designs C–E (Appendix I).

5.5 Marginal and group calibration

As Figure 3 illustrates, average coverage can hide missed observations within difficulty groups. To measure the width cost of protecting such groups, we divide fitted scales into terciles using D_F alone. Normalized Mondrian calibration gives each tercile its own corrected rank from D_C. Conditional on the fitted partition, exchangeability guarantees coverage at least p in every positive-probability group, averaging over calibration and a future observation in that group. This protects groups, not individual inputs.

On Bike/RF, high-tercile coverage rises from .817 for scale-only to .902 under matched Mondrian calibration, with a 9.3% width increase $[7.5, 11.2]\%$. On Energy/RF it rises from .837 to .902, at a 13.7% increase $[6.5, 21.0]\%$. Across all ten settings, mean paired width increases range from .6% to 23.8%. Empirical group minima remain noisy. These width costs are specific to normalized Mondrian calibration and its smaller groupwise samples; Appendix P provides further controls.

6 Discussion

This paper shows that slightly weakening full scale normalization can reduce expected interval length under an oracle location–scale model, even with finite calibration. We also explain how this adjustment changes conditional coverage and how estimation and parameter selection can limit its benefits. Our experiments show that including intermediate exponents often improves on choosing only between constant-width and fully normalized intervals. However, some settings show small losses, adjusting interval centers offers limited benefits, and comparisons with a more flexible optimizer are mixed. These results support selecting the scale exponent as a simple, inexpensive starting point for the studied predictors. When coverage is required within predefined groups, group calibration may produce wider intervals. Our findings explain the benefits and limits of scale adaptation within the interval family and tabular settings studied here.

AI USE STATEMENT

We used generative AI tools to assist with literature search and comparison, manuscript drafting and revision, the formulation and checking of theoretical claims, and experimental design and implementation. The authors take responsibility for the final manuscript, including its theoretical claims, empirical results, and accompanying code.

ETHICS STATEMENT

Our methods aim to reduce average interval length while preserving marginal coverage. However, shorter average intervals can be accompanied by lower coverage in some groups, particularly those with greater noise. We therefore report subgroup coverage diagnostics and distinguish average efficiency from coverage within groups. Applications requiring coverage guarantees for specific groups should use calibration procedures designed for that purpose, since marginal coverage alone does not ensure the target coverage in every group.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/~boyd/cvxbook/.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://link.springer.com/article/10.1023/A:1010933404324.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023. doi: 10.1093/jrsssb/qkac004. URL https://academic.oup.com/jrsssb/article/85/1/24/7008653.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction, 2021. URL https://arxiv.org/abs/1909.07889.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. URL https://jmlr.org/papers/v20/18-616.html.

Lu Deng and Xinzhu Liu. Features of extreme PM2.5 pollution and its influencing factors: Evidence from china. *Environmental Monitoring and Assessment*, 196(10):892, 2024. doi: 10.1007/s10661-024-12990-8. URL https://doi.org/10.1007/s10661-024-12990-8.

Dagmara Dudek and Anna Kuczmaszewska. Some practical and theoretical issues related to the quantile estimators. *Statistical Papers*, 65(6):3917–3933, 2024. doi: 10.1007/s00362-024-01543-3. URL https://link.springer.com/article/10.1007/s00362-024-01543-3.

Hadi Fanaee-T. Bike sharing. UCI Machine Learning Repository, 2013. URL https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18495–18527, 2025. URL https://proceedings.mlr.press/v267/gao25k.html.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Backward conformal prediction, 2025. URL https://arxiv.org/abs/2505.13732.

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

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008. URL https://doi.org/10.1093/jrsssb/qkaf008.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-062713-085831.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023. doi: 10.1093/biomet/asac040. URL https://doi.org/10.1093/biomet/asac040.

Naixin Guo, Rui Luo, and Zhixin Zhou. Fast conformal prediction using conditional interquantile intervals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21468–21476, 2026. doi: 10.1609/aaai.v40i26.39294. URL https://ojs.aaai.org/index.php/AAAI/article/view/39294.

Neha Gupta and Sagar Rohidas Chavan. A comprehensive decision support system for the characterization of probability distribution tails for daily precipitation. *Journal of Hydrology*, 626: 130282, 2023. doi: 10.1016/j.jhydrol.2023.130282. URL https://doi.org/10.1016/j.jhydrol.2023.130282.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025. doi: 10.1093/jrsssb/qkae103. URL https://doi.org/10.1093/jrsssb/qkae103.

Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022. URL https://www.jmlr.org/papers/v23/20-797.html.

Danijel Kivaranovic, Kory D. Johnson, and Hannes Leeb. Adaptive, distribution-free prediction intervals for deep networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4346–4356, 2020. URL https://proceedings.mlr.press/v108/kivaranovic20a.html.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction, 2024. URL https://arxiv.org/abs/2406.18814.

Abhishek Kumar, Harikrishna Narasimhan, and Andrew Cotter. Implicit rate-constrained optimization of non-decomposable objectives. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5861–5871. PMLR, 2021. URL https://proceedings.mlr.press/v139/kumar21b.html.

Antonije Lazic, J Mick Tilford, Victor P Davis, and Clare C Brown. Association of copayments with healthcare utilization and expenditures among medicaid enrollees with a substance use disorder. *Journal of Substance Use and Addiction Treatment*, 161:209314, 2024. doi: 10.1016/j.josat.2024.209314. URL https://doi.org/10.1016/j.josat.2024.209314.

Batiste Le Bars and Pierre Humbert. On volume minimization in conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3030–3053. PMLR, 2025. URL https://proceedings.mlr.press/v267/bars25a.html.

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. URL https://arxiv.org/abs/1604.04173.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *Artificial Intelligence Applications and Innovations*, volume 437 of *IFIP Advances in Information and Communication Technology*, pp. 261–270. Springer, 2014. doi: 10.1007/978-3-662-44722-2_28. URL https://link.springer.com/chapter/10.1007/978-3-662-44722-2_28.

Rui Luo and Zhixin Zhou. Conformal thresholded intervals for efficient regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19216–19223, 2025. doi: 10.1609/aaai.v39i18.34115. URL https://ojs.aaai.org/index.php/AAAI/article/view/34115.

Paulo C. Marques F. and Helton Graziadei. Skew-adaptive conformal prediction, 2026. URL https://arxiv.org/abs/2605.16145.

P. Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 18(3), 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006. URL https://jmlr.org/papers/v7/meinshausen06a.html.

Maria Gabriela Meirelles and Helena Cristina Vasconcelos. Physical–statistical characterization of PM10 and PM2.5 concentrations and atmospheric transport events in the azores during 2024. *Earth*, 6(2):54, 2025. doi: 10.3390/earth6020054. URL https://doi.org/10.3390/earth6020054.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition, 2006. doi: 10.1007/978-0-387-40065-5. URL https://link.springer.com/book/10.1007/978-0-387-40065-5.

Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024. URL https://jmlr.org/papers/v25/23-1553.html.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33:291–297, 1997. URL https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12 (85):2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Thomas Pouplin, Alan Jeffares, Nabeel Seedat, and Mihaela Van Der Schaar. Relaxed quantile regression: Prediction intervals for asymmetric noise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40951–40981. PMLR, 2024. URL https://proceedings.mlr.press/v235/pouplin24a.html.

Prashant Rana. Physicochemical properties of protein tertiary structure. UCI Machine Learning Repository, 2013. URL https://archive.ics.uci.edu/dataset/265/physicochemical+properties+of+protein+tertiary+structure.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Claudia See, Scott Grubman, Nisarg Shah, Jiun-Ruey Hu, Michael Nanna, James V Freeman, and Karthik Murugiah. Healthcare expenditure on atrial fibrillation in the united states. *JACC: Advances*, 4(5):101716, 2025. doi: 10.1016/j.jacadv.2025.101716. URL https://doi.org/10.1016/j.jacadv.2025.101716.

Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9:e261, 2020. doi: 10.1002/sta4.261. URL https://arxiv.org/abs/1909.05433.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6304–6315, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/31b3b31a1c2f8a370206f111127c0dbd-Abstract.html.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL https://jmlr.org/papers/v9/shafer08a.html.

Xiaoyi Su, Zhixin Zhou, and Rui Luo. Co-optimization for adaptive conformal prediction, 2026. URL https://arxiv.org/abs/2603.01719.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Athanasios Tsanas and Angeliki Xifara. Energy efficiency. UCI Machine Learning Repository, 2012. URL https://archive.ics.uci.edu/dataset/242/energy+efficiency.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256. URL https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.

Jasper Velthoen, Clément Dombry, Juan-Juan Cai, and Sebastian Engelke. Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667, 2023. doi: 10.1007/s10687-023-00473-x. URL https://link.springer.com/article/10.1007/s10687-023-00473-x.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715. URL https://link.springer.com/book/10.1007/b106715.

Tianying Wang. Geometry of tail allocation in conformal prediction intervals, 2026. URL https://arxiv.org/abs/2604.25202v2. Version 2, 15 July 2026.

Alexander Weyant, Alexander Gershunov, Anna K Panorska, Tomasz J Kozubowski, and Julie Kalansky. A holistic stochastic model for precipitation events. *Scientific Reports*, 15(1): 4595, 2025. doi: 10.1038/s41598-024-77031-3. URL https://doi.org/10.1038/s41598-024-77031-3.

Ran Xie, Rina Foygel Barber, and Emmanuel J. Candès. Boosted conformal prediction intervals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 71868–71899. Curran Associates, Inc., 2024. doi: 10.52202/079017-2296. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/842714f78c95096e20ac7d2591c5a24b-Paper-Conference.pdf.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569, 2021. URL https://proceedings.mlr.press/v139/xu21h.html.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38707–38727, 2023. URL https://proceedings.mlr.press/v202/xu23r.html.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets, 2024. URL https://arxiv.org/abs/2104.13871v3. Version 3; first circulated in 2021 as Finite-sample Efficient Conformal Prediction.

I-Cheng Yeh. Concrete compressive strength. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength.

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. URL https://proceedings.mlr.press/v162/zaffran22a.html.

Hao Zeng, Kangdao Liu, Bingyi Jing, and Hongxin Wei. Parametric scaling law of tuning bias in conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74133–74156, 2025. URL https://proceedings.mlr.press/v267/zeng25e.html.

Ce Zhang, Ting Li, Jinhan Xie, Linglong Kong, and Bei Jiang. Transfer conformal predictive inference for regression. *Journal of Machine Learning Research*, 27(90):1–68, 2026.

Junwen Zhou, Claire Williams, Mi Jun Keng, Runguo Wu, and Borislava Mihaylova. Estimating costs associated with disease model states using generalized linear models: a tutorial. *Pharmacoeconomics*, 42(3):261–273, 2024. doi: 10.1007/s40273-023-01319-x. URL https://doi.org/10.1007/s40273-023-01319-x.

Tianhui Zhou, Yitong Li, Yuan Wu, and David Carlson. Estimating uncertainty intervals from collaborating networks. *Journal of Machine Learning Research*, 22(257):1–47, 2021. URL https://jmlr.org/papers/v22/20-1100.html.

Ran Zou, Wanrong Zhu, and Bin Nan. Conformalized percentile interval: Finite sample validity and improved conditional performance, 2026. URL https://arxiv.org/abs/2605.03233.

A PROOFS AND ADDITIONAL THEORY

The proofs follow the questions in Section 4. We first hold the fitted predictors fixed and differentiate the population calibration equation. We then replace its threshold by the calibration sample's kth smallest score, showing why the local gain survives finite calibration and how coverage changes across inputs. The remaining results account for selection, fitting, and the candidate grid, and compare the power family with an unrestricted Gaussian benchmark. Finally, observable binomial bands provide comparisons that do not need a density lower bound. Each comparison specifies whether it concerns a population threshold, an average over calibration samples, or the final rule conditional on all observed data.

A.1 SENSITIVITY WITH FITTED PREDICTORS HELD FIXED

The first question is whether reducing the influence of a fitted scale can shorten intervals while preserving population coverage. Condition throughout on D_F, so the center c and positive scale proxy t are fixed measurable functions. All expectations and conditional distributions in this subsection use the resulting law of a new observation. Write $H_x(r) = \mathbb{P}\{|Y - c(x)| \le r \mid X = x, D_F\}$ for the conditional absolute-residual CDF and h_x for its density wherever the derivative exists. The population threshold q_γ satisfies $\mathbb{E}H_X\{q_\gamma t(X)^\gamma\} = p$, and $q = q_1$ denotes its value at full normalization. Neither predictor must be correct, and the conditional response distributions need not share a location–scale form.

Local regularity. To differentiate average coverage, we need to bound its change near the interval endpoints. The precise condition for Theorem 2 is as follows. There are an open interval J containing $q > 0$ and a number $\eta > 0$ such that, for almost every x, H_x is continuously differentiable on an open interval containing

$$\{zt(x)^\gamma : z \in J,\ |\gamma - 1| < \eta\}.$$

A single integrable function bounds the corresponding densities throughout this neighborhood:

$$\sup_{z\in J,\,|\gamma-1|<\eta} h_x\{zt(x)^\gamma\} \le M(x), \qquad \mathbb{E}M(X) < \infty.$$

We may shrink J and η so their closures are compact and $J \subset (0, \infty)$. These conditions concern only neighborhoods of the relevant endpoints; they impose no positive density lower bound at every input. To connect h_x to the response distribution, if $Y \mid X = x$ has a continuous density at both endpoints, then

$$h_x(r) = f_{Y|X=x}\{c(x) + r\} + f_{Y|X=x}\{c(x) - r\},$$

Thus the absolute-residual density adds the response density at the two endpoints: these are the places where a small expansion adds coverage.

Proof of Theorem 2. We first find how the threshold must change to keep average coverage fixed, then substitute that change into width and inputwise coverage. Let $G(z, \gamma) = \mathbb{E}H_X\{zt(X)^\gamma\} - p$, the coverage error at threshold z. The positive lower and finite upper bounds on t bound t^γ and $\log t$ uniformly nearby. The integrable envelope therefore bounds both partial derivatives of the integrand, so dominated convergence gives continuous partial derivatives

$$G_z(z, \gamma) = \mathbb{E}[t^\gamma h_X(zt^\gamma)], \qquad G_\gamma(z, \gamma) = z\mathbb{E}[t^\gamma h_X(zt^\gamma)\log t].$$

Set $w(x) = t(x)h_x\{qt(x)\}$ and $\ell_w = \mathbb{E}(w \log t)/\mathbb{E}w$. Here w measures the local response of coverage to the threshold, and ℓ_w is the corresponding weighted mean of log scale. At $(q, 1)$, $G_z = \mathbb{E}w > 0$ and $G_\gamma = q\mathbb{E}(w \log t)$. Because coverage increases with the threshold here, the implicit function theorem gives a continuously differentiable solution q_γ near one with

$$q_1' = -q\frac{\mathbb{E}(w \log t)}{\mathbb{E}w} = -q\ell_w. \tag{14}$$

The positivity of G_z persists in a sufficiently small neighborhood. The solution therefore crosses coverage level p strictly and is the population p-quantile. Applying the product rule to expected width $2q_\gamma\mathbb{E}t^\gamma$ gives equation 7. For coverage at a fixed input, write $\pi_\gamma(x) = H_x\{q_\gamma t(x)^\gamma\}$. For almost every x, the chain rule gives

$$\partial_\gamma H_x(q_\gamma t(x)^\gamma)|_1 = h_x(qt(x))t(x)\{q_1' + q\log t(x)\},$$

Substituting the threshold derivative proves equation 8. Its average is zero: coverage gained at some inputs exactly offsets the first-order loss at others.

For the sign of the width change, give an input weight proportional to its proxy scale. Formally, define the probability measure P_t by $dP_t/dP_X = t/\mathbb{E}t$, where P_X is the input law, and let $\mathbb{E}_t, \mathrm{Cov}_t$ denote its expectation and covariance. Then

$$\ell_w = \frac{\mathbb{E}_t[h_X(qt)\log t]}{\mathbb{E}_t h_X(qt)}, \qquad \mathbb{E}_t h_X(qt) = \frac{\mathbb{E}w}{\mathbb{E}t} > 0.$$

Consequently,

$$W'(1) = -\frac{2q(\mathbb{E}t)^2}{\mathbb{E}w}\,\mathrm{Cov}_t\{h_X(qt), \log t\}. \tag{15}$$

The strictly negative prefactor shows that a negative covariance gives $W'(1) > 0$ and hence a local width reduction when the exponent decreases. However, a zero covariance leaves the direction of any improvement undecided at first order. □

Recovery of the oracle identity. Suppose now that the predictors are the true functions in $Y = \mu(X) + \sigma(X)\varepsilon$, with $\varepsilon \perp X$. For a fixed shift b, define $H_b(t) = \mathbb{P}(|\varepsilon - b| \le t)$ and $q_b = H_b^{-1}(p)$. Under the regularity conditions in Corollary 6, taking $c = \mu + b\sigma$ and $t = \sigma$ gives $H_x(r) = H_b(r/\sigma(x))$ and $h_x(q_b\sigma(x)) = H_b'(q_b)/\sigma(x)$. Hence $w(x) = H_b'(q_b)$ is constant and $\ell_w = \mathbb{E}\log\sigma$. The general identities equation 7–equation 8 reduce to that corollary's oracle formulas. With an imperfect proxy, however, even the correct ordering of true scales need not make the width derivative positive.

A.2 Finite-calibration sensitivity with a fixed fitted scale

The next question is whether estimating the threshold from finitely many observations changes the local width effect. We first retain an arbitrary fixed fitted scale, allowing its normalized residual to depend on the input. In the proposition below, $A_{(k)}$ denotes the kth smallest normalized residual, I_k identifies the observation that supplies it, and the two expectations in expected width average over calibration and a fresh input, respectively. Independence under the oracle model will then simplify the formula.

Proposition 4 (General finite-calibration derivative)**.** *Condition on the fitting data and fix c, t, with $0 < t_- \le t(X) \le t_+ < \infty$. Let $(X_i, Y_i)_{i=1}^m$ be i.i.d. calibration observations and let a future input be independent of them. Put $A_i = |Y_i - c(X_i)|/t(X_i)$, assume its distribution is atomless, and take $k = \lceil (m+1)p \rceil \le m$ with $\mathbb{E}A_{(k)} < \infty$. Write I_k for the index of the kth smallest A_i, and let $\widehat{q}_\gamma$ be the kth order statistic of $A_i t(X_i)^{1-\gamma}$. The expected calibrated width $\overline{W}_m(\gamma) = 2\mathbb{E}\widehat{q}_\gamma\,\mathbb{E}t(X)^\gamma$ is differentiable at one, with*

$$\overline{W}'_m(1) = 2\left[\mathbb{E}A_{(k)}\,\mathbb{E}\{t(X)\log t(X)\} - \mathbb{E}\{A_{(k)}\log t(X_{I_k})\}\,\mathbb{E}t(X)\right]. \tag{16}$$

If, for a fixed exponent γ, the score $A_i t(X_i)^{1-\gamma}$ is also atomless, its calibrated interval has exact marginal coverage $k/(m+1)$; without this additional atomlessness, coverage is at least $k/(m+1)$.

Proof. The proof follows the observation that supplies the calibration threshold. We differentiate its score while the sample ordering stays fixed, then justify averaging this derivative. Write $t_i = t(X_i)$, $L = \max\{|\log t_-|, |\log t_+|\}$, and $Q(\gamma) = (A_i t_i^{1-\gamma})_{(k)}$, where the subscript denotes the kth smallest value. At $\gamma = 1$, $Q(1) = A_{(k)}$. Atomlessness means that the A_i are distinct with probability one. Since there are only finitely many continuous score curves, their ordering stays unchanged in some neighborhood of one for each such sample. Therefore

$$Q'(1) = -A_{(k)}\log t_{I_k}.$$

Although this neighborhood depends on the sample, a bound that holds for every sample allows us to take expectations. For any real h,

$$e^{-L|h|}A_i \le A_i t_i^{-h} \le e^{L|h|}A_i.$$

Increasing every score cannot decrease its kth smallest value. Applying this monotonicity to the bounds above gives

$$e^{-L|h|}A_{(k)} \le Q(1+h) \le e^{L|h|}A_{(k)}. \tag{17}$$

For $0 < |h| \le 1$, these inequalities give

$$\left|\frac{Q(1+h) - Q(1)}{h}\right| \le \frac{e^{L|h|} - 1}{|h|} A_{(k)} \le L e^{L} A_{(k)}.$$

The assumption $\mathbb{E}A_{(k)} < \infty$ makes this bound integrable. In particular, it is enough to control the expected threshold; the expected largest score need not be finite. Dominated convergence now gives

$$\frac{d}{d\gamma}\mathbb{E}Q(\gamma)\bigg|_1 = -\mathbb{E}[A_{(k)} \log t_{I_k}].$$

The future input is independent of calibration, so expected width factors as $\overline{W}_m(\gamma) = 2\mathbb{E}Q(\gamma)\mathbb{E}t^{\gamma}$. The scale bounds also allow differentiation of the second factor. The product rule proves equation 16. Separately, the exchangeable-rank argument proves coverage: it is exact for atomless scores and conservative when scores tie. □

Equation equation 16 is exact for finite calibration samples. Its second term uses the input X_{I_k} whose score sets the threshold. If normalized scores still depend on the proxy, that input need not behave like a fresh draw, so the term need not reduce to a covariance under the original input law. A separate subtlety concerns ties: an atomless score at exponent one can have atoms at another exponent. This is why exact coverage for another exponent needs its own atomlessness condition.

Proof of Theorem 3. For the oracle specialization, independence removes the dependence just described. Take $c = \mu + b\sigma$ and $t = \sigma$, so $A_i = T_i = |\varepsilon_i - b|$, and write $\sigma_i = \sigma(X_i)$. The residual vector $(T_1, \dots, T_m)$ is independent of the scale vector $(\sigma_1, \dots, \sigma_m)$ and determines I_k. Conditioning on the residual vector therefore gives

$$\mathbb{E}[\log \sigma_{I_k} \mid T_1, \dots, T_m] = \mathbb{E}\log\sigma.$$

Thus $\mathbb{E}[T_{(k)} \log \sigma_{I_k}] = \mathbb{E}T_{(k)}\, \mathbb{E}\log\sigma$: the log scale attached to the threshold has the ordinary mean after conditioning on the residuals. Substitution in Proposition 4 gives

$$\overline{W}'_m(1) = 2\mathbb{E}T_{(k)}\{\mathbb{E}(\sigma\log\sigma) - \mathbb{E}\sigma\, \mathbb{E}\log\sigma\}.$$

Both factors are strictly positive. For an independent copy σ', the covariance equals $\frac{1}{2}\mathbb{E}[(\sigma - \sigma')(\log\sigma - \log\sigma')]$; the increasing logarithm makes this positive when σ is nonconstant. Also, an atomless nonnegative T_i is strictly positive almost surely, so $0 < \mathbb{E}T_{(k)} < \infty$. A positive width derivative gives the strict decrease for a sufficiently small reduction in the exponent. For coverage, let $T = |\varepsilon - b|$ denote a fresh absolute error. For each fixed γ, independence of T and σ makes $T\sigma^{1-\gamma}$ a mixture of scaled atomless distributions, hence atomless. The rank argument then gives coverage exactly $k/(m+1)$. □

The gain concerns expected width across calibration samples, so some realized samples can give a different width ordering. The exponent must be fixed independently of calibration. Although the identity is exact at finite m, it is local in the exponent: the allocation-sample search and a specified finite grid step still need separate analysis.

A.3 Conditional coverage after finite calibration

Having found the width change, we now ask which inputs gain or lose coverage. Smoothness is needed for this coverage derivative; the width identity above still requires only atomless errors and a finite expectation of the kth smallest error. Let H_b be the CDF of $|\varepsilon - b|$, let $h_b = H_b'$ be its density, and put $D = \log(s_+/s_-)$. The sufficient envelope condition is that, for some $\eta > 0$,

$$\mathbb{E}\left[T_{(k)} \sup_{u \in [e^{-D\eta}T_{(k)},\, e^{D\eta}T_{(k)}]} h_b(u)\right] < \infty. \tag{18}$$

One easy way to meet this condition is $M = \sup_{u>0} u h_b(u) < \infty$. It holds, for example, for absolute Gaussian and Student error magnitudes around any fixed center b. Under this simpler condition, the coverage derivative itself needs no finite-moment assumption on $T_{(k)}$, although expected width still does.

Proof of the additional identity. We first differentiate coverage with the calibration sample held fixed, then average over that sample. Fix an input x satisfying the scale bounds. Write $\sigma_i = \sigma(X_i)$, abbreviate the kth smallest calibration error by $T = T_{(k)}$, and let I_k identify the observation supplying it. Conditional on this input and the realized calibration sample, coverage is

$$\pi_{m,D_C}(x;\gamma) = H_b\{\widehat{q}_\gamma \sigma(x)^{\gamma-1}\}.$$

As in the width proof, atomlessness makes the calibration errors distinct almost surely and keeps their ordering fixed in a sample-dependent neighborhood of one. Thus $\widehat{q}_1 = T$ and $\widehat{q}_1' = -T\log\sigma_{I_k}$. The argument of H_b is the radius measured in true error units. Calling this argument $R_h = \widehat{q}_{1+h}\sigma(x)^h$ gives

$$R_0' = T\{\log\sigma(x) - \log\sigma_{I_k}\}.$$

To average this derivative, we need a bound valid for every sample. Observe that R_h is the kth smallest value among $T_i\{\sigma(x)/\sigma_i\}^h$. Monotonicity of that value gives bounds depending only on scale ratios:

$$e^{-D|h|}T \le R_h \le e^{D|h|}T.$$

For $0 < |h| \le \eta$, the mean-value theorem converts the radius bounds into the following bound on the absolute difference quotient of coverage $H_b(R_h)$:

$$De^{D\eta}T \sup_{u\in[e^{-D\eta}T,\,e^{D\eta}T]} h_b(u),$$

Condition equation 18 makes this quantity integrable. Dominated convergence therefore allows differentiation of $\overline{\pi}_m(x;\gamma) = \mathbb{E}_{D_C}\pi_{m,D_C}(x;\gamma)$ and gives

$$\begin{aligned}\overline{\pi}_m'(x;1) &= \mathbb{E}\left[Th_b(T)\{\log\sigma(x) - \log\sigma_{I_k}\}\right] \\ &= \mathbb{E}[Th_b(T)]\{\log\sigma(x) - \mathbb{E}\log\sigma(X)\}.\end{aligned}$$

For the last equality, condition on $(T_1,\dots,T_m)$. This fixes T and I_k while leaving the independent calibration scales unchanged in law. The envelope makes the coefficient finite. To see that it is positive, a proper continuously differentiable distribution on $(0,\infty)$ has positive density on some open interval with $0 < H_b < 1$. The density of its kth smallest sample value is positive there too, so $\mathbb{E}[Th_b(T)] > 0$.

Under the simpler bound $M < \infty$, $z \mapsto H_b(e^z)$ is M-Lipschitz, so the difference quotient is directly bounded by MD without a moment assumption. At full normalization, $\overline{\pi}_m(x;1) = \mathbb{E}H_b(T_{(k)}) = k/(m+1)$ for every such input. Averaging the derivative in equation 10 over X gives zero, consistent with exact marginal coverage for every fixed exponent. □

Deterministic numerical verification. The script `code/verify_finite_calibration_extension.py` checks these identities numerically by integrating over the Beta distribution of the transformed calibration threshold. For Gaussian errors and an asymmetric Gamma construction with a nonzero center shift, it compares the formulas with central finite differences of integrated width and group coverage at $m = 19, 115, 500$. The largest discrepancies are 1.90×10^{-8} for width and 1.67×10^{-8} for coverage; integrated rank coverage differs by at most 3.24×10^{-13}. These checks support the specified calculations, while the proof supplies the result for all admissible distributions. The saved coefficient and rank-band tables record integration tolerances, parameters, and package versions.

Conditional coverage with an imperfect fitted scale. The same proof also shows why an imperfect scale can change the coverage pattern. Use Proposition 4's notation and write $\pi_{\gamma,D_C}(x) = H_x\{\widehat{q}_\gamma t(x)^\gamma\}$ for coverage conditional on $X = x$ and the realized calibration sample. Put $D_t = \log(t_+/t_-)$, and assume for almost every x that H_x is continuously differentiable on $(0,\infty)$ and that, for some $\eta > 0$,

$$\mathbb{E}\left[A_{(k)} \sup_{u\in[e^{-D_t\eta}t(x)A_{(k)},\,e^{D_t\eta}t(x)A_{(k)}]} h_x(u)\right] < \infty.$$

Differentiating the sample threshold and applying this envelope gives, for almost every such x,

$$\left.\frac{d}{d\gamma}\mathbb{E}_{D_C}\pi_{\gamma,D_C}(x)\right|_1 = t(x)\mathbb{E}\left[A_{(k)}h_x\{t(x)A_{(k)}\}\{\log t(x) - \log t(X_{I_k})\}\right]. \tag{19}$$

The residual setting the threshold can depend on its attached proxy. Consequently, the inputs that gain or lose coverage need not be separated by the oracle geometric-mean scale. As in the population formula of Theorem 2, the density of absolute residuals at the interval radius helps determine the change.

A.4 A MISSPECIFIED PROXY WITH AN EXPLICIT SIGN CRITERION

A two-group example separates correct ordering from a useful scale response. Let the groups be equally likely, with $Y \mid G = g \sim N(0, \sigma_g^2)$, $\sigma_1 = 1$, and $\sigma_2 = 3$. Fix $c = 0$ and take $t_g = \sigma_g^\beta$. Every $\beta > 0$ identifies the harder group correctly, although only $\beta = 1$ reproduces the true scale ratio. With Φ denoting the standard normal CDF, the population threshold is the unique positive solution of

$$\frac{1}{2}\sum_{g=1}^{2}\left[2\Phi\{q_\gamma \sigma_g^{\beta\gamma-1}\} - 1\right] = p. \tag{20}$$

At $\gamma = 1$, put $q = q_1$ and let ϕ denote the standard normal density. The absolute-residual densities at the interval radii are

$$h_g(qt_g) = \frac{2}{\sigma_g}\phi\{q\sigma_g^{\beta-1}\}.$$

For two groups, the covariance has the sign of the product of the differences between their two values. Since $t_2 > t_1$ when $\beta > 0$, criterion equation 15 becomes

$$\begin{aligned} W'(1) > 0 \quad &\iff \quad h_2(qt_2) < h_1(qt_1) \\ &\iff \quad \log 3 + \frac{q^2}{2}\left(3^{2(\beta-1)} - 1\right) > 0. \end{aligned} \tag{21}$$

Taking the logarithm of the Gaussian density ratio gives the last equivalence. The inequality holds for every $p \in (0, 1)$ when $\beta = 1$. However, it can fail when the proxy compresses the difference between group scales, despite preserving their ordering.

Table 2 shows the consequences at $p = .9$. For $\beta = .5$, the derivative is negative: the proxy already gives the higher-scale group too little relative width, so weakening its influence increases average width. For $\beta = .75$, the derivative is positive, although the finite step from $\gamma = 1$ to .75 slightly increases width. A local derivative therefore cannot replace evaluating the available grid step. For the exact scale, $\beta = 1$, that same step reduces population width from 6.57941 to 6.29133.

Table 2: Population scale sensitivity in the two-group Gaussian example, with $t = \sigma^\beta$ and $p = .9$. Negative percentage change favors $\gamma = .75$. Quantiles are obtained by deterministic root solving. A positive derivative ensures only a sufficiently small local improvement; the $\beta = .75$ row demonstrates that the default-sized step need not improve width.

Proxy power β	$W'(1)$	$W(1)$	$W(.75)$	Change (%)
0.50	-0.49953	6.34700	6.52584	+2.818
0.75	0.37271	6.29133	6.29741	+0.097
1.00	1.80706	6.57941	6.29133	-4.379

Exact finite-calibration expectations. To replace these population thresholds by finite calibration, we integrate over the distribution of the sample threshold. After transforming continuous scores by their CDF, their kth smallest value has a Beta distribution. This standard identity gives exact expectations without simulation or a uniform concentration bound.

Proposition 5 (Order-statistic integration)**.** *For fixed c, t, γ, assume that the calibration pairs are i.i.d. and the future pair is an independent observation from the same law. Let F_γ be the continuous CDF of $|Y - c(X)|/t(X)^\gamma$, and write $Q_\gamma(u) = F_\gamma^{-1}(u)$. For $k = \lceil (m+1)p \rceil \le m$, put $U \sim \mathrm{Beta}(k, m+1-k)$. Then the calibrated expected width is*

$$\overline{W}_m(\gamma) = 2\mathbb{E}t(X)^\gamma \int_0^1 Q_\gamma(u) f_{\mathrm{Beta}(k,m+1-k)}(u)\, du, \tag{22}$$

and expected conditional coverage, averaged over calibration samples, is

$$\mathbb{E}_{D_C}\pi_{\gamma,D_C}(x) = \int_0^1 H_x\{Q_\gamma(u)t(x)^\gamma\} f_{\mathrm{Beta}(k,m+1-k)}(u)\,du. \tag{23}$$

The width identity holds as an extended nonnegative expectation if the integral is infinite.

Proof. Apply F_γ to each calibration score. Continuity makes the transformed scores i.i.d. uniform variables, whose kth smallest value is $\mathrm{Beta}(k, m+1-k)$. Applying the quantile function back shows that the calibrated threshold has the distribution of $Q_\gamma(U)$. Independence from a fresh input then gives the width integral equation 22; holding that input fixed gives the coverage integral equation 23. Averaging the latter over X returns $\mathbb{E}U = k/(m+1)$, the exact marginal rank coverage. □

For the Gaussian mixture, replacing p by u in equation 20 gives $Q_\gamma(u)$. Deterministic quadrature of equation 22–equation 23 produces Table 3. The step to .75 reduces expected width at $m = 115, 154, 500$, even where the uniform finite-dictionary bound gives no useful comparison. The lower high-scale-group coverage also persists after finite calibration. These numbers evaluate fixed candidates under the exact model; they leave the cost of choosing γ on D_A to the selection analysis.

Table 3: Finite-calibration expectations in the same example with the exact proxy $t = \sigma$. Both fixed-exponent intervals have marginal coverage $k/(m+1)$; the final column is the high-scale-group coverage for $\gamma = .75$, averaged over calibration samples. Widths and group coverages numerically integrate the exact Beta order-statistic formula. These are fixed-candidate calculations, with no parameter selection or Monte Carlo approximation.

m	k	Marginal coverage	$\overline{W}_m(1)$	$\overline{W}_m(.75)$	High-scale coverage
115	105	0.90517	6.74579	6.46252	0.86241
154	140	0.90323	6.68955	6.40482	0.85975
500	451	0.90020	6.59724	6.31002	0.85551

Numerical verification and scope. The script `code/verify_proxy_theory.py` reproduces both tables and records derivatives, finite-calibration expectations, and numerical checks for ten proxy powers between -1 and 2. Against centered finite differences, the largest discrepancy is 1.1×10^{-9} for analytic width derivatives and 5.3×10^{-11} for conditional-coverage derivatives. Nine finite-calibration derivative checks agree with equation 16, including its oracle specialization equation 9, to within 6.3×10^{-9}. Across eighteen fixed-candidate calibration calculations, integrated marginal coverage differs from the exact rank ratio by at most 3.3×10^{-13}. These deterministic checks test the formulas in specified models; they add neither fitted-data benchmarks nor a uniform efficiency guarantee.

A.5 Population geometry and selection bounds

A.6 Scale sensitivity

We now collect the population and sampling comparisons used in the main text. Start with true predictors and a population threshold: can even exact normalization be improved? The next corollary specializes the fitted-proxy calculation. It shows both the local reduction in average width and the coverage transferred between inputs.

Corollary 6 (Scale sensitivity at normalization)**.** *Suppose $Y = \mu(X) + \sigma(X)\varepsilon$, with $\varepsilon \perp X$, $0 < s_- \le \sigma(X) \le s_+ < \infty$, and nonconstant $\sigma(X)$. Fix $b \in \mathbb{R}$ and put $H_b(t) = \mathbb{P}(|\varepsilon - b| \le t)$. Assume $q_b = H_b^{-1}(p) > 0$, H_b is continuously differentiable near q_b, and $H_b'(q_b) > 0$. Define $q_{b,\gamma}$ near $\gamma = 1$ through*

$$\mathbb{E}H_b\{q_{b,\gamma}\sigma(X)^{\gamma-1}\} = p.$$

The interval $\mu(x) + b\sigma(x) \pm q_{b,\gamma}\sigma(x)^\gamma$ has marginal coverage p. Its expected width $W_b(\gamma) = 2q_{b,\gamma}\mathbb{E}\sigma(X)^\gamma$ satisfies

$$W_b'(1) = 2q_b\,\mathrm{Cov}\{\sigma(X), \log\sigma(X)\} > 0. \tag{24}$$

Hence $W_b(1-h) < W_b(1)$ for all sufficiently small $h > 0$. For conditional coverage $\pi_{b,\gamma}(x)$,

$$\partial_\gamma \pi_{b,\gamma}(x)|_{\gamma=1} = q_b H_b'(q_b)\{\log \sigma(x) - \mathbb{E}\log \sigma(X)\}. \tag{25}$$

Thus exact scale normalization need not minimize expected width under a marginal coverage constraint. Decreasing γ slightly raises coverage below the geometric-mean scale $\exp\{\mathbb{E}\log\sigma(X)\}$ and lowers it above that scale. The average gain comes with a redistribution of miscoverage. Although the result is strict for nonconstant scale, extending it to fitted predictors or a larger exponent step requires the additional comparisons below.

How far could a richer radius rule improve? In the Gaussian benchmark of Appendix A.11, the unrestricted optimal centered radius is $s\sqrt{2\log(t^*/s)}$ for $s < t^*$ and zero otherwise, where s is the true scale and t^* is chosen to reach marginal coverage p. Its proportional response to scale varies with s, so a fixed power s^γ need not attain it. This established marginal-optimization principle (Kiyani et al., 2024) supplies a target against which to assess the fitted family's restrictions and estimation error.

A.7 Selection and calibration error

A population opportunity can be lost when finite data choose and calibrate the rule. Following work on short conformal sets (Yang & Kuchibhotla, 2024; Kiyani et al., 2024; Le Bars & Humbert, 2025), we separate these two sources of error by comparing with the best fitted candidate. Condition on D_F. For $\lambda = (b, \gamma)$, let F_λ be its score CDF, $q_\lambda = F_\lambda^{-1}(p)$ its population threshold, $v_\lambda(x) = 2\widehat{\sigma}(x)^\gamma$ its width per unit threshold, and $V_\lambda = \mathbb{E}v_\lambda(X)$ its mean over a fresh input. Then $J_\lambda = V_\lambda q_\lambda$ is that candidate's expected width using a population threshold. The theorem compares this target with the calibrated selected rule.

Theorem 7 (Finite-dictionary width comparison)**.** *Suppose allocation, calibration, and test observations are i.i.d., conditional on D_F, and $|\Lambda| = K$. Assume uniformly over Λ that $0 < v_\lambda(x) \le B$ and $q_\lambda \le Q$. For $\delta \in (0,1)$ and some $\eta > 0$, let $[p-\eta, p+\eta] \subset (0,1)$ and suppose F_λ has continuous density at least $c > 0$ between its quantiles at these two levels. Set*

$$e_A = \sqrt{\frac{\log(8K/\delta)}{2n_A}}, \quad e_C = \sqrt{\frac{\log(4/\delta)}{2m}}, \quad d_n = \frac{\lceil (n+1)p \rceil}{n} - p.$$

Assume $\lceil (n+1)p \rceil \le n$ for $n = n_A, m$, $e_A + d_{n_A} < \eta$, and $e_C + d_m < \eta$. With probability at least $1-\delta$ over D_A, D_C, the conditional expected future width $\widehat{J} = \mathbb{E}[\mathrm{len}\{\widehat{C}_+(X_{\mathrm{new}})\} \mid D_F, D_A, D_C]$ satisfies

$$\widehat{J} \le \min_{\lambda\in\Lambda} J_\lambda + 2B\left\{\frac{e_A + d_{n_A}}{c} + Q e_A\right\} + \frac{B}{c}(e_C + d_m). \tag{26}$$

The fitted predictors may be misspecified. Searching K candidates appears in the selection error through $\log K$; by the time independent final calibration begins, its score is fixed and needs no such term. The two corrections $d_n < 2/n$ account for rounding the conformal ranks.

Scope of the bound. Call the sampling remainder in equation 26 $\mathcal{R}_\delta$. Including $(0,0)$ and $(0,1)$ in the candidate set allows comparisons with the population-threshold widths of ordinary and normalized SCP. It does not establish dominance over their realized calibrated widths. The density lower bound turns a probability error into a quantile error; atoms can violate it and heavy tails can make the resulting width bound uninformative. These assumptions are therefore stronger than those for coverage. At $p = .9$, $K = 125$, and $\delta = .05$, $e_A + d_{n_A}$ is approximately .221, .188, and .102 for $n_A = 115, 154, 500$, above the required $\eta < .1$. The bound consequently gives no 95% efficiency conclusion at those sizes, although Theorem 1 still guarantees coverage.

Global tail allocation admits a simpler uniform comparison: one residual empirical CDF controls every allocation, so the bound needs no factor for grid size. Appendix A.14 develops that width bound and allocation consistency. At a fixed total budget, however, allocating more observations to selection leaves fewer for calibration. Appendix A.15 describes this tradeoff; the unknown error constants prevent a universal numerical split recommendation.

A.8 From oracle geometry to fitted intervals

The best fitted candidate is still a restricted target. To connect it to the local population gain, we compare fitted and true predictors and then account for the family and grid. Under Corollary 6's location–scale model, define the oracle score $S^0_\lambda = \sigma^{1-\gamma}|\varepsilon - b|$, its p-quantile q^0_λ, and its expected width $W^0_\lambda = 2q^0_\lambda \mathbb{E}\sigma^\gamma$. Let W^*_{opt} be the infimum expected length over contiguous intervals having marginal coverage at least p. The difference $A_\Lambda = \min_{\lambda\in\Lambda} W^0_\lambda - W^*_{\mathrm{opt}} \geq 0$ measures what the power family and finite grid cannot represent, even with true predictors and population thresholds.

Theorem 8 (Oracle-to-fitted width comparison)**.** *Let $\Lambda \subset [-b_{\max}, b_{\max}] \times [0,1]$ be finite and $0 < s_- \leq \sigma(X) \leq s_+ < \infty$. Condition on a fitting sample satisfying, almost surely in X,*

$$\frac{|\widehat{\mu}(X) - \mu(X)|}{\sigma(X)} \leq \varepsilon_\mu, \qquad \left|\log \frac{\widehat{\sigma}(X)}{\sigma(X)}\right| \leq \eta_s. \tag{27}$$

Write $a = e^{\eta_s}$, $T = \max(1, s_+)$, $W^0_{\max} = \max_{\lambda\in\Lambda} W^0_\lambda$, and

$$E_F = (a^2-1)W^0_{\max} + 2a^2T^2\{\varepsilon_\mu + b_{\max}(a-1)\}. \tag{28}$$

Then $\max_\lambda |J_\lambda - W^0_\lambda| \leq E_F$. Under the additional hypotheses of Theorem 7, with conditional probability at least $1-\delta$ over D_A, D_C,

$$\widehat{J} \leq W^*_{\mathrm{opt}} + A_\Lambda + E_F + \mathcal{R}_\delta, \tag{29}$$

where $\mathcal{R}_\delta$ is the remainder in equation 26.

The next corollary asks whether a step actually present in the grid is small enough for the local gain to overcome these errors. A curvature bound controls the gap between the derivative's prediction and the finite-step width change.

Corollary 9 (A represented grid step)**.** *Under Theorem 8, suppose $0 < h \leq 1$ and $(b_0, 1), (b_0, 1-h) \in \Lambda$, Corollary 6 applies at b_0, and W_{b_0} is twice continuously differentiable with $|W''_{b_0}| \leq L$ on $[1-h, 1]$. Set $D = 2q_{b_0}\,\mathrm{Cov}(\sigma, \log\sigma)$ and $\Delta_h = Dh - Lh^2/2$. On the joint concentration event used in Theorem 7,*

$$\widehat{J} \leq J_{(b_0,1)} - \Delta_h + 2E_F + \mathcal{R}_\delta. \tag{30}$$

Thus a certificate requires the guaranteed population gain Δ_h to exceed the fitting, selection, and calibration costs. The comparison target $J_{(b_0,1)}$ uses a population threshold. A baseline calibrated on the pooled reserve has its own threshold error, accounted for in Appendix A.16. Neither the default step $h = .25$ nor an arbitrary fitted scale is guaranteed to meet the criterion. The relative predictor-error bounds are sufficient efficiency conditions, not coverage requirements or assumptions verified for every dataset. Appendix A.16 supplies the proofs, explicit curvature conditions, and pooled-baseline comparison.

A.9 Proof of Theorem 1

The key step is to condition on the data used to fit the predictors and select the interval shape. Given D_F, D_A, the selected score is fixed, so its m calibration scores and test score are exchangeable. The rank argument therefore gives $\mathbb{P}(S_{\mathrm{new}} \leq S_{(k)}) \geq k/(m+1) \geq p$. In the absence of ties, the first probability equals $k/(m+1)$. The sentinel $S_{(m+1)} = +\infty$ handles the case $k = m+1$. Since $\widehat{\sigma}(x) > 0$, solving $S_{\widehat{b},\widehat{\gamma}}(x,y) \leq S_{(k)}$ for y gives equation 5.

For corrected ECP-FQ, apply the same exchangeability argument separately to the lower and upper tails of the signed residuals. It gives

$$\mathbb{P}(R_{\mathrm{new}} < R_{(j_L)}) \leq \frac{j_L}{m+1} \leq \widehat{a}, \qquad \mathbb{P}(R_{\mathrm{new}} > R_{(j_U)}) \leq \frac{m+1-j_U}{m+1} \leq \alpha - \widehat{a}.$$

These are the two ways the interval can miss the response. They are disjoint, and their probabilities sum to at most α. Without ties, the rank probabilities are exact, giving coverage $(j_U - j_L)/(m+1)$. If the implementation's finite-allocation range is empty, the rule instead returns $\mathbb{R}$, which has coverage one. The nonempty-range formula does not apply to that fallback. □

Rounding and finite endpoints. The rank convention determines both whether the endpoints are finite and how much coverage rounding adds. For $M = m+1$, the signed ranks give finite endpoints exactly when $1/M \le a \le \alpha - 1/M$. Clipping a rank inward to force a finite endpoint would lose the preceding coverage argument. To compare the rounding, write $z = Ma$ and $k = \lceil Mp \rceil$. Then

$$j_U - j_L = \lceil Mp + \{z\} \rceil \in \{k, k+1\},$$

where $\{z\} = z - \lfloor z \rfloor$. When a belongs to the rank-compatible grid $\{j/M\}$, this difference is k. Single-score calibration can therefore avoid an extra rounding rank. However, an appropriately designed two-tail allocation grid achieves the same rounding. This rank comparison alone says nothing about interval widths.

A.10 Proof of Corollary 6

Changing the exponent changes the score distribution, so its population threshold must move to keep marginal coverage fixed. We first differentiate that threshold, then substitute its derivative into the width and conditional coverage formulas. Suppress X in $\sigma(X)$ and define $G(q, \gamma) = \mathbb{E}H_b(q\sigma^{\gamma-1}) - p$; the equation $G = 0$ expresses marginal coverage p. The positive lower and finite upper bounds on σ allow differentiation under the expectation near $(q_b, 1)$. At that point,

$$G_q = H_b'(q_b) > 0, \qquad G_\gamma = q_b H_b'(q_b)\mathbb{E}\log\sigma.$$

The implicit function theorem gives a differentiable local quantile curve, with $q_{b,1} = q_b$ and $q_{b,1}' = -q_b\mathbb{E}\log\sigma$. Differentiating $W_b(\gamma) = 2q_{b,\gamma}\mathbb{E}\sigma^\gamma$ gives

$$W_b'(1) = 2q_b\{\mathbb{E}(\sigma\log\sigma) - \mathbb{E}\sigma\,\mathbb{E}\log\sigma\}.$$

For an independent copy σ',

$$\mathrm{Cov}(\sigma, \log\sigma) = \tfrac{1}{2}\mathbb{E}[(\sigma - \sigma')(\log\sigma - \log\sigma')] > 0.$$

The strict inequality follows because logarithm is strictly increasing and σ is nondegenerate. Thus the width derivative at one is positive, and differentiability gives a strict local reduction when the exponent moves below one. To see where coverage moves, write the conditional coverage as $\pi_{b,\gamma}(x) = H_b\{q_{b,\gamma}\sigma(x)^{\gamma-1}\}$. Differentiating this expression and substituting $q_{b,1}'$ gives equation 25. □

Shift and shortest residual intervals. The shift b controls a separate part of the interval geometry. At $\gamma = 1$, minimizing expected width over b is equivalent to minimizing q_b, because the scale multiplier is then fixed. Let $F_\varepsilon, f_\varepsilon$ denote the CDF and density of ε. If f_ε is continuous near the endpoints, implicit differentiation of $F_\varepsilon(b + q_b) - F_\varepsilon(b - q_b) = p$ gives

$$\frac{dq_b}{db} = \frac{f_\varepsilon(b - q_b) - f_\varepsilon(b + q_b)}{f_\varepsilon(b - q_b) + f_\varepsilon(b + q_b)}.$$

An interior stationary point therefore has equal endpoint densities, with lower-tail allocation $F_\varepsilon(b - q_b)$. This classical shortest-interval characterization explains how a shift corresponds to a tail allocation. It is not a new optimality principle. The validity theorem does not require unimodality; even when unimodality holds, it does not imply that the chosen interval is nested inside a symmetric one.

A.11 Gaussian radius-allocation benchmark

The local derivative shows why reducing scale adjustment can shorten an interval. It does not identify the best radius at each input. The Gaussian benchmark below solves that broader population problem and shows which features a power radius can approximate.

Oracle setting. Consider $Y = \mu(X) + \sigma(X)\varepsilon$, with $\varepsilon \sim N(0,1)$ independent of X, known μ, σ, and $0 < s_- \le \sigma(X) \le s_+ < \infty$. For a measurable radius $r(x) \ge 0$, let $C_r(x) = [\mu(x) - r(x), \mu(x) + r(x)]$. We minimize $2\mathbb{E}r(X)$ subject to $\mathbb{P}\{Y \in C_r(X)\} \ge p$. This population benchmark specializes marginal length optimization (Kiyani et al., 2024); it isolates the radius geometry without estimation or calibration error. Let Φ denote the standard normal CDF.

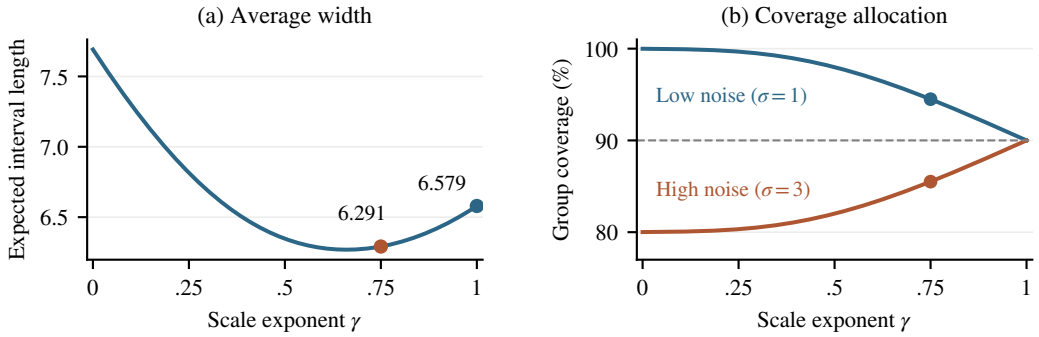

Figure 4: **Scale response changes width and coverage allocation.** Two equally likely Gaussian groups have noise scales one and three. For each fixed exponent, the population threshold is recalibrated to exactly 90% marginal coverage. Reducing γ from one to .75 shortens average width by 4.38% and changes group coverages to 94.5% and 85.5%. These are oracle population curves; the finite-calibration theory and fitted experiments address the subsequent statistical costs.

Theorem 10 (Optimal Gaussian radius allocation)**.** *In the oracle setting, the unique optimal radius, up to null sets under the law of X, is $r^*(x) = r_{t^*}\{\sigma(x)\}$, where*

$$r_t(s) = \begin{cases} s\sqrt{2\log(t/s)}, & 0 < s < t, \\ 0, & s \geq t, \end{cases} \qquad \mathbb{E}\left[2\Phi\left(\frac{r_{t^*}\{\sigma(X)\}}{\sigma(X)}\right) - 1\right] = p. \tag{31}$$

The last equation has a unique solution $t^* > 0$. On $0 < s < t^*$, the log-scale elasticity is

$$\frac{d\log r_{t^*}(s)}{d\log s} = 1 - \frac{1}{2\log(t^*/s)} < 1. \tag{32}$$

If $\sigma(X)$ is nonconstant, this optimum is strictly shorter in expectation than the normalized oracle radius $\Phi^{-1}\{(1+p)/2\}\sigma(X)$.

The optimal allocation equalizes conditional density at its nonzero endpoints. Its elasticity measures the proportional change in radius for a proportional change in scale. A power radius $q\sigma^\gamma$ has constant elasticity γ, so learning $\gamma < 1$ can approximate the optimal response locally. However, the optimal elasticity varies with scale and can become negative; sufficiently large scales receive zero radius. The $\gamma \in [0, 1]$ family therefore need not attain this benchmark. Moreover, the benchmark constrains marginal coverage and does not ensure subgroup validity.

A.12 Proof of Theorem 10

The proof first finds the threshold that meets the coverage constraint. It then uses a pointwise length–coverage tradeoff to prove global optimality, and finally compares the resulting radius with full normalization. Write $\sigma_X = \sigma(X)$ and let φ be the standard normal density. For $s > 0$, define the conditional coverage of a centered interval of radius r by $u_s(r) = 2\Phi(r/s) - 1$, $r \geq 0$. For each fixed s, this function is continuous and strictly increasing, with $u_s(0) = 0$ and $u_s(r) \to 1$ as $r \to \infty$.

Existence and uniqueness of the threshold. Let $G(t) = \mathbb{E}u_{\sigma_X}\{r_t(\sigma_X)\}$ for $t > 0$. For each fixed s, $u_s\{r_t(s)\}$ is continuous and nondecreasing in t, is zero for $t \leq s$, and is strictly increasing for $t > s$. Bounded convergence makes G continuous. The lower bound on σ_X gives $G(t) = 0$ for $t \leq s_-$, whereas $G(t) \to 1$ as $t \to \infty$. Thus a finite threshold $t^* > 0$ with $G(t^*) = p$ exists for $p \in (0, 1)$. If $G(t_1) > 0$ and $t_2 > t_1$, then $\mathbb{P}(\sigma_X < t_1) > 0$. The integrand increases strictly on this event and never decreases elsewhere, so $G(t_2) > G(t_1)$. Hence the threshold solving $G(t) = p$ is unique.

Pointwise optimization and global optimality. Use the threshold just constructed to set $\nu = t^*/\varphi(0) > 0$. At a fixed scale s, the function $\ell_s(r) = 2r - \nu u_s(r)$ trades interval length against

conditional coverage. Its derivative for $r \geq 0$ is

$$\ell'_s(r) = 2 - \frac{2\nu}{s}\varphi(r/s).$$

This derivative is strictly increasing in $r \geq 0$, approaches 2, and equals $2(1 - t^*/s)$ at zero. If $s \geq t^*$, the unique minimizer is zero, including the case $s = t^*$. If $s < t^*$, the derivative changes sign exactly once; its zero satisfies

$$\exp\{-r^2/(2s^2)\} = s/t^*,$$

giving $r = r_{t^*}(s)$. The optimum is finite and integrable; for example, $0 \leq r_{t^*}(s) \leq t^*/\sqrt{e}$ for every $s > 0$.

We can now integrate the pointwise comparison. For any other measurable, feasible radius $r(X)$ of finite expected length, pointwise optimality gives

$$2\mathbb{E}r(X) - \nu\mathbb{E}u_{\sigma_X}\{r(X)\} \geq 2\mathbb{E}r_{t^*}(\sigma_X) - \nu p.$$

Since feasibility means $\mathbb{E}u_{\sigma_X}\{r(X)\} \geq p$, it follows that $2\mathbb{E}r(X) \geq 2\mathbb{E}r_{t^*}(\sigma_X)$. Candidates with infinite expected length cannot improve the finite optimum. If equality in expected length holds, the nonnegative pointwise Lagrangian gap must have expectation zero and the coverage constraint must hold with equality. The unique pointwise minimizers imply $r(X) = r_{t^*}(\sigma_X)$ almost surely. This proves global optimality and uniqueness without assuming that a competing radius depends on X only through σ_X.

Elasticity and strict comparison. For $0 < s < t^*$, $\log r_{t^*}(s) = \log s + \frac{1}{2}\log\{2\log(t^*/s)\}$. Differentiating with respect to $\log s$ yields equation 32. Put $z_p = \Phi^{-1}\{(1+p)/2\} > 0$. The normalized oracle radius $z_p\sigma_X$ has conditional coverage p at every X, and hence is feasible. If it coincided almost surely with the optimum, then $\sigma_X < t^*$ almost surely and $z_p = \sqrt{2\log(t^*/\sigma_X)}$ almost surely, forcing σ_X to be constant. For nonconstant σ_X, uniqueness therefore yields a strict expected-length improvement. □

Interpretation and scope. At an active endpoint, the conditional response density is $\varphi\{r_{t^*}(s)/s\}/s = \varphi(0)/t^*$, independent of s. This is the common-density condition from marginal length optimization (Kiyani et al., 2024), here evaluated explicitly for Gaussian errors. The elasticity is nonnegative only for $s \leq t^*/\sqrt{e}$ and becomes negative for $t^*/\sqrt{e} < s < t^*$. The optimum can consequently decrease to zero for high scales; at a zero-radius singleton, conditional Gaussian coverage is zero. The marginal constraint permits this loss of coverage at some inputs in exchange for shorter average length. This feature is part of the unrestricted optimum and need not be attainable within the restricted power family.

A.13 Proof of Theorem 7

There are two sampling errors to control: allocation selects a candidate using estimated widths, and final calibration estimates the selected candidate's threshold. Condition on D_F, so the entire fitted dictionary is fixed. In this proof, $J_\lambda = V_\lambda q_\lambda$ is the expected width obtained with a population threshold, while $\widehat{J}$ uses the observed final threshold and conditions on all observed data.

For allocation, let $\widetilde{F}_\lambda$ be the empirical score CDF on D_A, and let $\widetilde{v}_\lambda$ be the empirical mean of $v_\lambda(X) = 2\widehat{\sigma}(X)^\gamma$ on that sample. The Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) gives

$$\mathbb{P}\left(\sup_t |\widetilde{F}_\lambda(t) - F_\lambda(t)| > e\right) \leq 2\exp(-2n_A e^2).$$

Since $v_\lambda(X) \in [0, B]$, Hoeffding's inequality gives

$$\mathbb{P}(|\widetilde{v}_\lambda - V_\lambda| > Be) \leq 2\exp(-2n_A e^2).$$

A union bound across K candidates and these two events, with $e = e_A$, shows that with probability at least $1 - \delta/2$, both inequalities hold simultaneously for every candidate.

The pilot conformal order statistic $\widetilde{q}_\lambda$ is the empirical inverse-CDF quantile at $p + d_{n_A}$, where d_{n_A} records the conformal rank adjustment. To convert CDF accuracy into threshold accuracy, use the

following quantile comparison: if an empirical CDF is within e of a continuous population CDF, its quantile at u is bracketed by population quantiles at $u-e$ and $u+e$. The density lower bound limits the distance between those quantiles. Integrating it over the relevant interval gives, on the preceding event,

$$|\widetilde{q}_\lambda - q_\lambda| \le (e_A + d_{n_A})/c.$$

This argument is uniform over the dictionary, so it applies to a subsequently selected candidate. It follows that

$$\begin{aligned}|\widetilde{q}_\lambda \widetilde{v}_\lambda - J_\lambda| &\le \widetilde{v}_\lambda |\widetilde{q}_\lambda - q_\lambda| + q_\lambda |\widetilde{v}_\lambda - V_\lambda| \\ &\le B\{(e_A + d_{n_A})/c + Qe_A\} =: r_A.\end{aligned}$$

Although the empirical mean and empirical quantile use the same D_A, the bound holds because both concentration events hold together; their independence is unnecessary. Comparing the empirical minimizer with a population minimizer λ^* of J_λ now gives

$$J_{\widehat{\lambda}} \le \widetilde{q}_{\widehat{\lambda}} \widetilde{v}_{\widehat{\lambda}} + r_A \le \widetilde{q}_{\lambda^*} \widetilde{v}_{\lambda^*} + r_A \le J_{\lambda^*} + 2r_A.$$

For final calibration, condition also on D_A. Only the selected score now needs to be controlled, so a single additional DKW bound on D_C gives a CDF error at most e_C with probability at least $1-\delta/2$. The final threshold is its empirical quantile at $p + d_m$, and the same density argument yields

$$|\widehat{q} - q_{\widehat{\lambda}}| \le (e_C + d_m)/c.$$

Given all observed data, the threshold and selected candidate are fixed; only the future input is averaged in $\widehat{J} = V_{\widehat{\lambda}} \widehat{q}$. Since $V_{\widehat{\lambda}} \le B$, the final threshold error gives the final width error. Combining the allocation and calibration events proves equation 26 with probability at least $1-\delta$. □

A.14 Uniform theory for corrected global allocation

For global tail allocation, all candidates use quantiles of the same signed residual distribution. A single empirical CDF can therefore control the full allocation grid.

Condition on the fitted regression rule. Let F_R and Q_R be its signed residual CDF and inverse. Restrict the lower-tail allocation to $\mathcal{A} = [\tau, \alpha - \tau]$, with $0 < \tau < \alpha/2$, so that both tail levels stay away from the endpoints. Define the population width by $W(a) = Q_R(p+a) - Q_R(a)$ and its minimum on this domain by $W^* = \min_{a \in \mathcal{A}} W(a)$. Let $\mathcal{G}_h \subset \mathcal{A}$ contain a point within h of every element of $\mathcal{A}$. The allocation rule minimizes empirical width on $\mathcal{G}_h$, as in equation 1. Write $\mathcal{A}^* = \arg\min_{a \in \mathcal{A}} W(a)$ for the set of population minimizers. For any $a^* \in \mathcal{A}^*$, the quantity

$$\omega_*(h) = \sup_{a \in \mathcal{A}:\, |a - a^*| \le h} |W(a) - W^*|.$$

measures the largest width increase within distance h of that minimizer. It will bound the cost of replacing the minimizer by a nearby grid point.

Proposition 11 (Global width regret and calibration error)**.** *Suppose allocation and calibration observations are i.i.d. conditional on D_F, and the residual density is continuous and at least $c_R > 0$ between its quantiles at $\tau - \eta_R$ and $1 - \tau + \eta_R$, where $0 < \eta_R < \tau$. Fix $\delta \in (0,1)$ and let*

$$u_A = \sqrt{\frac{\log(4/\delta)}{2n_A}}, \qquad u_C = \sqrt{\frac{\log(4/\delta)}{2m}}.$$

Assume $u_A < \eta_R$, $u_C + 2/m < \eta_R$, and $1/(m+1) \le \tau$. With probability at least $1-\delta$,

$$0 \le W(\widehat{a}) - W^* \le 4u_A/c_R + \omega_*(h), \tag{33}$$

$$\widehat{W}_C(\widehat{a}) \le W^* + 4u_A/c_R + \omega_*(h) + (2u_C + 3/m)/c_R. \tag{34}$$

Here $\widehat{W}_C$ is the width after the corrected final ranks. Moreover $\omega_*(h) \le 2h/c_R$.

Proof. We first bound allocation error, then add the error from final calibration. Two DKW bounds give uniform empirical CDF errors at most u_A and u_C with joint probability at least $1-\delta$. Let

$\widetilde{W}(a)$ be the allocation-sample width. On the first event, applying the density lower bound to each endpoint gives

$$\sup_{a\in\mathcal{A}} |\widetilde{W}(a) - W(a)| \le 2u_A/c_R.$$

Select a grid point a_h with $|a_h - a^*| \le h$. Empirical minimization and two applications of the displayed bound give

$$W(\widehat{a}) \le \widetilde{W}(\widehat{a}) + 2u_A/c_R \le \widetilde{W}(a_h) + 2u_A/c_R \le W^* + \omega_*(h) + 4u_A/c_R.$$

On the second event, uniformly over $\mathcal{A}$,

$$\left|\frac{j_L(a)}{m} - a\right| \le \frac{1}{m}, \qquad 0 \le \frac{j_U(a)}{m} - (p+a) < \frac{2}{m}.$$

These rank discrepancies are the rounding contribution to calibration error. The assumptions keep both ranks finite and all quantile levels within the region where the density lower bound holds. Adding CDF error to rank error bounds the two endpoint errors by $(u_C + 1/m)/c_R$ and $(u_C + 2/m)/c_R$. Consequently, $|\widehat{W}_C(a) - W(a)| \le (2u_C + 3/m)/c_R$ uniformly over a. This uniformity allows us to substitute the random $\widehat{a}$ and proves equation 34. Finally, each population quantile is $1/c_R$-Lipschitz in its level, so W is $2/c_R$-Lipschitz and $\omega_*(h) \le 2h/c_R$. □

Because one residual empirical CDF controls every allocation, the bound needs no $\log|\mathcal{G}_h|$ factor. The implementation uses trimming that depends on sample size. The finite-sample proposition still applies whenever its assumptions hold at the sample size in question. However, if the trimming level tends to zero, additional tail regularity is needed to keep the constants uniform.

Corollary 12 (Allocation consistency)**.** *Under Proposition 11 with a fixed trimmed domain, suppose the minimizer set $\mathcal{A}^*$ is well separated and $h \to 0$ as $n_A \to \infty$. Then* $\operatorname{dist}(\widehat{a}, \mathcal{A}^*) \to 0$ *in probability. If, for constants $c_0 > 0$ and $\kappa > 0$, near this set*

$$W(a) - W^* \ge c_0 \operatorname{dist}(a, \mathcal{A}^*)^\kappa,$$

then

$$\operatorname{dist}(\widehat{a}, \mathcal{A}^*) = O_P\left[\{n_A^{-1/2} + \omega_*(h)\}^{1/\kappa}\right]. \tag{35}$$

For a unique minimizer this is convergence to that allocation.

Proof. The width-regret bound first places the selected allocation near the minimizer set; the local growth condition then converts width error into allocation error. Specifically, uniform empirical width convergence and grid approximation give $W(\widehat{a}) - W^* = o_P(1)$ by equation 33. Separation means that, for every $\epsilon > 0$, the infimum of $W(a) - W^*$ over $\operatorname{dist}(a, \mathcal{A}^*) \ge \epsilon$ is positive. The selected allocation therefore lies in this set with probability tending to zero. Inside the growth neighborhood, taking the κth root of the width-regret bound gives equation 35; the probability of falling outside that neighborhood tends to zero. □

The grid contribution depends on how fast width rises near the optimum. With Lipschitz regularity alone, its contribution in equation 35 is $h^{1/\kappa}$. To obtain the stronger $O_P(n_A^{-1/(2\kappa)} + h)$ rate, one also needs the local upper growth bound $\omega_*(h) = O(h^\kappa)$; lower growth alone does not supply it. An interior twice continuously differentiable minimizer with positive curvature satisfies both growth bounds with $\kappa = 2$, giving the coarse objective-based rate $O_P(n_A^{-1/4} + h)$. This argument asserts no root-n_A allocation rate.

A valid population comparison with SCP. A population comparison requires a strict width gap. Let q_{abs} be the population p quantile of $|R|$, so the symmetric population interval has width $2q_{\mathrm{abs}}$. If $W^* < 2q_{\mathrm{abs}}$ and the corrected calibrated widths consistently estimate these respective targets, then $\mathbb{P}\{\operatorname{len}(\widehat{C}_{\mathrm{FQ}}) < \operatorname{len}(\widehat{C}_{\mathrm{SCP}})\} \to 1$: the random width difference converges in probability to the strictly negative gap.

Asymmetry alone does not imply this gap. One sufficient condition is that the symmetric interval has exact probability p, its lower-tail allocation $a_s = F_R(-q_{\mathrm{abs}})$ lies in the interior of $\mathcal{A}$, and $f_R(q_{\mathrm{abs}}) \neq f_R(-q_{\mathrm{abs}})$ with positive continuous endpoint densities. Then $W(a_s) = 2q_{\mathrm{abs}}$ and

$$W'(a_s) = \frac{1}{f_R(q_{\mathrm{abs}})} - \frac{1}{f_R(-q_{\mathrm{abs}})} \neq 0,$$

so moving the allocation a short distance in the direction of decreasing width gives a strict improvement. However, trimming can exclude this allocation, and the argument gives no universal finite-sample ordering.

A.15 Allocation and calibration at a fixed budget

Using more observations for allocation improves shape selection but leaves fewer observations for final calibration. For fixed confidence and dictionary size, the leading remainder in equation 26 makes this tradeoff explicit: it has the form $C_A n_A^{-1/2} + C_C m^{-1/2}$, where C_A and C_C are the allocation and calibration constants, and C_A contains a $\sqrt{\log K}$ factor. Holding the reserve budget $N = n_A + m$ fixed and differentiating this expression in $n_A \in (0, N)$ gives its minimum at

$$\frac{n_A}{m} = \left(\frac{C_A}{C_C}\right)^{2/3}, \qquad \frac{n_A}{N} = \frac{C_A^{2/3}}{C_A^{2/3} + C_C^{2/3}}.$$

The constants depend on quantile density and scale bounds and are generally unknown. The calculation therefore motivates examining split proportions, but supplies no universally optimal numerical split. The exact bound also includes $O(1/n_A + 1/m)$ rounding terms.

A fixed score can use all N reserve observations for calibration, so a comparison with that baseline must account for its larger calibration sample. Although selecting a shape can create a strict population width gain, that gain may be too small to outweigh the additional estimation error at the available budget.

A.16 Proof of Theorem 8

We first compare each fitted score with its oracle counterpart, then transfer the comparison to quantiles and expected widths. The finite-dictionary bound will add selection and calibration error at the last step.

Throughout this proof, the fitting sample is fixed. Write $r(X) = \widehat{\sigma}(X)/\sigma(X)$ for the scale ratio and $u(X) = \{\widehat{\mu}(X) - \mu(X)\}/\sigma(X)$ for the center error measured in true scale units. Recall that $a = e^{\eta_s}$ and $T = \max(1, s_+)$. Here T is a deterministic scale bound, distinct from the residual order statistic denoted by T in the finite-calibration argument. By assumption, $a^{-1} \leq r \leq a$ and $|u| \leq \varepsilon_\mu$. For every candidate $\lambda = (b, \gamma)$, the fitted and oracle scores are

$$S_\lambda = \sigma^{1-\gamma} r^{-\gamma} |\varepsilon - b - u - b(r-1)|,$$
$$S_\lambda^0 = \sigma^{1-\gamma} |\varepsilon - b|.$$

The fitted score has a multiplicative scale error and an additive center error. The reverse triangle inequality and $\gamma \in [0, 1]$ bound the additive part as follows:

$$|S_\lambda - r^{-\gamma} S_\lambda^0| \leq \sigma^{1-\gamma} r^{-\gamma} \{|u| + |b||r-1|\} \leq aT\{\varepsilon_\mu + b_{\max}(a-1)\} =: d.$$

The last step uses $\sigma^{1-\gamma} \leq T$, $r^{-\gamma} \in [a^{-1}, a]$, and $|r-1| \leq a-1$; these bounds hold simultaneously over the dictionary. The constant d therefore bounds the additive score error for every candidate. Since scores are nonnegative,

$$a^{-1} S_\lambda^0 - d \leq S_\lambda \leq a S_\lambda^0 + d. \tag{36}$$

If random variables satisfy $U \leq V$ almost surely, their generalized inverse-CDF quantiles satisfy $Q_U(p) \leq Q_V(p)$. Moreover, for constants $c > 0$ and $d' \in \mathbb{R}$, $Q_{cU+d'}(p) = cQ_U(p) + d'$. Applying these facts to equation 36 gives

$$a^{-1} q_\lambda^0 - d \leq q_\lambda \leq a q_\lambda^0 + d.$$

Quantile monotonicity requires no density or continuity assumption. To compare widths, define the oracle width multiplier $V_\lambda^0 = 2\mathbb{E}\sigma(X)^\gamma$. Its fitted counterpart is $V_\lambda = 2\mathbb{E}\widehat{\sigma}(X)^\gamma$. Because $r^\gamma \in [a^{-1}, a]$,

$$a^{-1}V_\lambda^0 \le V_\lambda \le aV_\lambda^0, \qquad V_\lambda^0 \le 2T.$$

For the upper bound, multiplication of nonnegative quantities yields

$$J_\lambda = V_\lambda q_\lambda \le a^2 W_\lambda^0 + aV_\lambda^0 d.$$

For the lower bound, separating the nonnegative quantile term avoids multiplying a possibly negative lower quantile bound by a lower bound on the width factor:

$$J_\lambda \ge V_\lambda q_\lambda^0/a - V_\lambda d \ge a^{-2}W_\lambda^0 - aV_\lambda^0 d.$$

It follows, using $1 - a^{-2} \le a^2 - 1$, that

$$|J_\lambda - W_\lambda^0| \le (a^2-1)W_\lambda^0 + aV_\lambda^0 d \le (a^2-1)W_{\max}^0 + 2a^2T^2\{\varepsilon_\mu + b_{\max}(a-1)\} = E_F.$$

This proves the uniform perturbation bound for population-threshold expected widths. The constants in it are finite: for example, if $q_{|\varepsilon|}(p)$ is the generalized p-quantile of $|\varepsilon|$, then

$$q_\lambda^0 \le T\{q_{|\varepsilon|}(p) + b_{\max}\}, \qquad W_{\max}^0 \le 2T^2\{q_{|\varepsilon|}(p) + b_{\max}\} < \infty.$$

It remains to add sampling error and identify the oracle comparison target. By Theorem 7, on an event with conditional probability at least $1 - \delta$,

$$\widehat{J} \le \min_{\lambda\in\Lambda} J_\lambda + \mathcal{R}_\delta \le \min_{\lambda\in\Lambda} W_\lambda^0 + E_F + \mathcal{R}_\delta.$$

Every oracle interval $\mu(x) + b\sigma(x) \pm q_\lambda^0\sigma(x)^\gamma$ is contiguous and has marginal coverage at least p, by the definition of its score quantile. Its expected width is W_λ^0. Thus the infimum over all feasible contiguous intervals satisfies $W_{\text{opt}}^* \le \min_\lambda W_\lambda^0$, and their difference is the approximation error A_Λ. Substituting its definition proves equation 29. □

Probability over fitting. The result conditions on a fit satisfying equation 27 and the hypotheses of Theorem 7. These are assumptions on the quality of the fit, not conclusions about a particular learner. If they hold on an event in D_F of probability at least $1 - \delta_F$, using deterministic bounds on that event gives an unconditional success probability at least $(1 - \delta_F)(1 - \delta) \ge 1 - \delta_F - \delta$. The fitting constants are generally unknown, and no convergence rate for a particular fitted learner is asserted.

Relation to the Gaussian benchmark. In the Gaussian setting of Theorem 10, allowing arbitrary interval centers cannot improve its optimum. To see this, fix x, an interval half-length $r \ge 0$, and a standardized center offset t. Its conditional probability is $\Phi(t + r/\sigma(x)) - \Phi(t - r/\sigma(x))$. For $r > 0$ the derivative in t is positive for $t < 0$ and negative for $t > 0$, because the standard normal density is symmetric and strictly decreases in absolute argument. Hence centering the interval at $\mu(x)$ maximizes conditional probability for its length. Replacing every interval by its centered version therefore preserves expected length and cannot reduce marginal coverage. The centered optimum in Theorem 10 consequently equals the unrestricted contiguous-interval infimum W_{opt}^* used in equation 29. The approximation term A_Λ can thus be evaluated against that benchmark in Gaussian diagnostics, but is generally unknown in fitted real-data problems.

A.17 Proof of Corollary 9

The corollary asks whether a grid step captures enough of the local population gain to survive fitting and sampling error. Taylor's theorem with integral remainder gives

$$W_{b_0}(1-h) = W_{b_0}(1) - hW_{b_0}'(1) + \int_{1-h}^{1}\{t - (1-h)\}W_{b_0}''(t)\,dt \le W_{b_0}(1) - Dh + Lh^2/2.$$

Since $(b_0, 1-h) \in \Lambda$, the best oracle dictionary width is no larger than this value. On the event from Theorem 7,

$$\begin{aligned}\widehat{J} &\le \min_{\lambda\in\Lambda} J_\lambda + \mathcal{R}_\delta \\ &\le W_{b_0}(1-h) + E_F + \mathcal{R}_\delta \\ &\le W_{b_0}(1) - \Delta_h + E_F + \mathcal{R}_\delta \\ &\le J_{(b_0,1)} - \Delta_h + 2E_F + \mathcal{R}_\delta.\end{aligned}$$

The last step uses the two-sided perturbation bound for $(b_0, 1)$. The fitting error appears twice because both the selected rule's oracle comparison and the fitted baseline must be related to their population counterparts. The claimed sufficient condition follows. □

A fixed-score baseline using the entire reserve. The preceding comparison uses the fitted baseline's population threshold. For a comparison with an observed calibrated baseline, let $N = n_A + m$ and calibrate the fixed candidate $(b_0, 1)$ using $D_A \cup D_C$. Write $\widehat{J}_{\mathrm{pool}}$ for its expected future width conditional on all observed data. Both $\widehat{J}$ and $\widehat{J}_{\mathrm{pool}}$ then average over a future input while keeping their observed thresholds fixed. In addition to the conditions already imposed, assume $\lceil (N+1)p \rceil \le N$ and $e_0 + d_N < \eta$, where

$$e_0 = \sqrt{\frac{\log(4/\delta)}{2N}}, \qquad r_0 = \frac{B}{c}(e_0 + d_N).$$

Here B, c, η, d_N are the constants from Theorem 7, and the theorem's allocation and calibration conditions are evaluated at failure probability $\delta/2$. A DKW bound for the fixed baseline score yields $|\widehat{J}_{\mathrm{pool}} - J_{(b_0,1)}| \le r_0$ with probability at least $1 - \delta/2$. Applying Corollary 9 at level $\delta/2$ and taking a union bound gives

$$\widehat{J} \le \widehat{J}_{\mathrm{pool}} - \Delta_h + 2E_F + \mathcal{R}_{\delta/2} + r_0 \tag{37}$$

with probability at least $1 - \delta$. Although the rules share the reserve sample, the union bound remains valid because it requires no independence between the two events. The result gives a sufficient condition for comparing conditional expected widths. It does not imply pointwise interval containment or domination on every sample. The density and fitting constants are generally unknown, and conservative values can make the comparison uninformative at the experimental sample sizes.

Sufficient regularity for a finite curvature bound. A finite grid step needs a bound on how much the width derivative can change over that step. The following sufficient conditions yield such a curvature bound from distributional quantities. For a fixed b_0, suppose the calibration equation has a unique solution $0 < q_\gamma \le Q_0$ for every $\gamma \in [1 - h_0, 1]$, where $0 < h_0 \le 1$. Suppose H_{b_0} is twice continuously differentiable on an open neighborhood of all arguments $q_\gamma s^{\gamma - 1}$, $s \in [s_-, s_+]$, and on that neighborhood

$$0 < c_H \le H'_{b_0} \le M_1, \qquad |H''_{b_0}| \le M_2.$$

Set $\ell = \max(|\log s_-|, |\log s_+|)$, $U = e^{h_0 \ell}$, and

$$\begin{aligned} A &= c_H / U, & C_1 &= M_1 Q_0 U \ell / A, \\ B_{qq} &= M_2 U^2, & B_{q\gamma} &= (M_2 Q_0 U + M_1) U \ell, \\ B_{\gamma\gamma} &= (M_2 Q_0^2 U^2 + M_1 Q_0 U) \ell^2, & C_2 &= (B_{qq} C_1^2 + 2 B_{q\gamma} C_1 + B_{\gamma\gamma})/A. \end{aligned}$$

For $G(q, \gamma) = \mathbb{E} H_{b_0}(q \sigma^{\gamma - 1}) - p$, $G_q \ge A > 0$. Bounded scale and the displayed derivative bounds justify differentiation under the expectation. The implicit function theorem gives a twice continuously differentiable quantile curve, and

$$q'_\gamma = -G_\gamma / G_q, \qquad q''_\gamma = -\{G_{qq}(q'_\gamma)^2 + 2 G_{q\gamma} q'_\gamma + G_{\gamma\gamma}\}/G_q.$$

The constants above bound the first and second quantile derivatives: $|q'_\gamma| \le C_1$ and $|q''_\gamma| \le C_2$. Recall that $T = \max(1, s_+)$ bounds σ^γ over the exponent range. Differentiating $W_{b_0}(\gamma) = 2 q_\gamma \mathbb{E} \sigma^\gamma$ twice then shows

$$|W''_{b_0}(\gamma)| \le 2T\{C_2 + 2\ell C_1 + Q_0 \ell^2\} =: L.$$

These distributional conditions establish a finite L. However, the bound is deliberately conservative and its inputs need not be known. It gives no distribution-free guarantee that a particular coarse grid step improves width.

A.18 Observable width bounds without a density lower bound

The preceding oracle inequality turns CDF error into width error through a density lower bound. We can avoid that lower bound by using observed order statistics to bracket the population score quantile directly. Multiplying those brackets by bounds on the width multiplier gives a computable

diagnostic for the fitted candidate family. The resulting bounds may still be too wide to establish an improvement.

We first bound expected widths at population thresholds. We then show how the same multiplier bounds can audit the observed final thresholds; that second calculation concerns expected widths conditional on all data.

Condition on D_F, assume allocation observations and the future observation are i.i.d., and let $n = n_A$. Suppose each of the K candidate score distributions F_λ is continuous. Write $q_\lambda = F_\lambda^{-1}(p)$, $v_\lambda = 2\mathbb{E}\widehat{\sigma}(X)^\gamma$, and $J_\lambda = q_\lambda v_\lambda$. Here the lowercase v_λ denotes the expected width multiplier, called V_λ above; it is a scalar, not a function of the future input. Thus J_λ is the population-threshold expected width for the fitted candidate.

Let $S_{\lambda,(j)}$ denote that candidate's allocation-sample order statistics, with $S_{\lambda,(0)} = 0$ and $S_{\lambda,(n+1)} = +\infty$. These sentinel values allow an unbounded confidence interval when the sample does not support finite endpoints. For a quantile failure budget $\delta_q \in (0,1)$ and a binomial random variable $B \sim \mathrm{Bin}(n,p)$, choose

$$\ell = \max\{j \in \{0,\dots,n\} : \mathbb{P}(B < j) \le \delta_q/(2K)\}, \tag{38}$$

$$u = \min\{j \in \{1,\dots,n+1\} : \mathbb{P}(B \ge j) \le \delta_q/(2K)\}. \tag{39}$$

Continuity makes the number of allocation scores below q_λ binomial. The ranks above bound its two tails; a union bound across candidates then gives, simultaneously over all candidates, with probability at least $1 - \delta_q$,

$$S_{\lambda,(\ell)} \le q_\lambda \le S_{\lambda,(u)}. \tag{40}$$

The candidates may use the same observations; no independence across them is needed. The upper endpoint is finite whenever $p^n \le \delta_q/(2K)$. For example, at $p = .9$, $K = 5$, and $\delta_q = .05$, the respective rank pairs are $(95, 112)$, $(128, 148)$, and $(432, 467)$ for $n = 115, 154, 500$. Although these endpoints are finite, the resulting width bounds may still overlap too much to establish an improvement.

If the expected multiplier v_λ is known, multiplying equation 40 by it gives a simultaneous width interval $[L_\lambda, U_\lambda]$. If it is unknown, use a second simultaneous confidence event. Define its allocation-sample estimate by $\widetilde{v}_\lambda = n^{-1}\sum_{i\in D_A} 2\widehat{\sigma}(X_i)^\gamma$. For example, the implemented bounds $\widehat{\sigma} \in [t_-, t_+]$ and $\gamma \in [0,1]$ place each summand between a_λ and d_λ. For a multiplier failure budget $\delta_v \in (0,1)$, set

$$a_\lambda = 2t_-^\gamma, \qquad d_\lambda = 2t_+^\gamma, \qquad r_\lambda = (d_\lambda - a_\lambda)\sqrt{\frac{\log(2K/\delta_v)}{2n}}.$$

Here r_λ is the sampling-error allowance for the empirical multiplier. Hoeffding's inequality and a union bound give $v_\lambda \in [v_\lambda^-, v_\lambda^+]$ for every candidate with probability at least $1 - \delta_v$, where clipping to the known range gives

$$v_\lambda^- = \max(a_\lambda, \widetilde{v}_\lambda - r_\lambda), \qquad v_\lambda^+ = \min(d_\lambda, \widetilde{v}_\lambda + r_\lambda).$$

In particular, $\gamma = 0$ has the exact multiplier two. Since score quantiles and multipliers are nonnegative, combining their two confidence events gives

$$L_\lambda = S_{\lambda,(\ell)} v_\lambda^- \le J_\lambda \le S_{\lambda,(u)} v_\lambda^+ = U_\lambda \quad \text{for every } \lambda \tag{41}$$

with probability at least $1 - \delta_q - \delta_v$. The union bound does not require the quantile and multiplier events to be independent, so they may use the same allocation observations. Because the result holds for all candidates at once, it also applies to any allocation-selected $\widehat{\lambda}$, giving the observable comparison

$$J_{\widehat{\lambda}} - \min_\lambda J_\lambda \le U_{\widehat{\lambda}} - \min_\lambda L_\lambda.$$

If $U_{\widehat{\lambda}} < L_{(0,1)}$, the selected candidate has a smaller *population-threshold* expected width than full normalization on this event. This compares the two fitted candidates using their population thresholds. Overlapping bounds leave the comparison inconclusive. Computing the certificate requires neither choosing a new exponent nor changing final calibration.

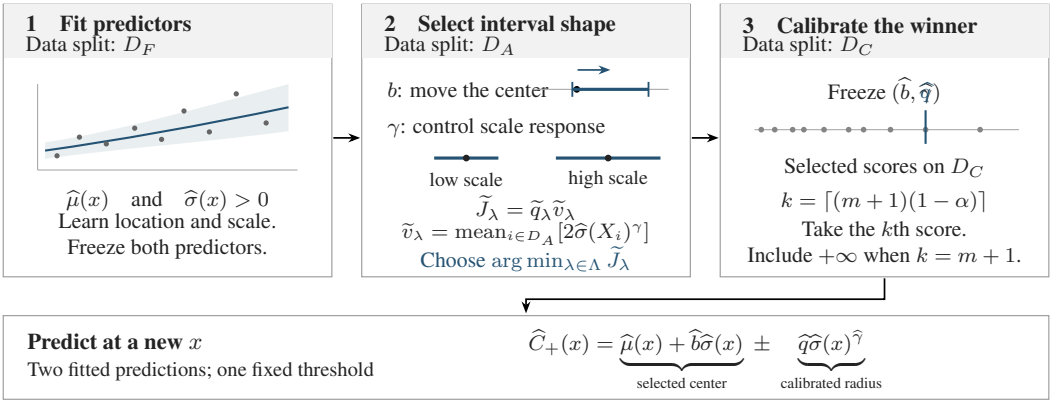

Figure 5: Fit, select, and independently calibrate. The pilot threshold helps select the shape and is discarded before final calibration.

Auditing the final calibrated intervals. After selection and calibration, the threshold $\widehat{q}$ is observed and can be treated as fixed when averaging over a future input. The resulting expected future width, conditional on all observed data, is $\widehat{J} = \widehat{q}v_{\widehat{\lambda}}$. There is no population quantile to estimate in this calculation: the only unknown factor is the expected multiplier. The multiplier event alone therefore gives

$$\widehat{q}v_{\widehat{\lambda}}^- \le \widehat{J} \le \widehat{q}v_{\widehat{\lambda}}^+. \tag{42}$$

For a fixed fitted baseline $\lambda_0 \in \Lambda$ with observed calibrated threshold $\widehat{q}_0$, the inequality $\widehat{q}v_{\widehat{\lambda}}^+ < \widehat{q}_0 v_{\lambda_0}^-$ is thus a sufficient certificate of a smaller conditional expected width, with failure probability at most δ_v. This probability controls the multiplier confidence event; the widths being compared condition on all observed data.

The baseline may use the pooled reserve. On the simultaneous multiplier event, the bounds hold for every observed threshold value, so the two thresholds need not be independent. These certificates audit intervals that have already been selected and calibrated. Switching between those intervals on the basis of the diagnostic would introduce a new selection step, which is not covered by Theorem 1.

B Implementation and Experimental Protocol

The experiments follow three data roles: the fitting sample learns the predictors, the allocation sample selects the interval shape, and the final calibration sample sets its coverage threshold. These samples are denoted by D_F, D_A, and D_C, respectively. Figure 5 summarizes this sequence, and Algorithm 1 gives the steps for ECP-FQ+.

The implementation is organized into `code/ecpfq_core.py`, the fitted-data experiment scripts, the oracle diagnostic script, and the distributional/interquantile sidecars. Every reported table is generated from saved repetition-level metrics. The supplied scripts, seeds, data provenance, and raw metrics specify the numerical experiments. The opening figure illustrates the procedure conceptually; the documented implementations and saved results provide the performance comparisons.

Algorithm 1 (ECP-FQ+: select shape, then calibrate)**. Input:** disjoint D_F, D_A, D_C, miscoverage α, ordered dictionary Λ.

1. Fit and freeze $\widehat{\mu}$ and positive $\widehat{\sigma}$ on D_F.
2. Evaluate each pilot objective $\widetilde{q}_\lambda \widetilde{v}_\lambda$ on D_A.
3. Freeze the first minimizing $(\widehat{b}, \widehat{\gamma})$ and discard its pilot threshold.
4. On D_C, compute the selected scores and take rank $\lceil (m+1)(1-\alpha) \rceil$, retaining the $+\infty$ sentinel when necessary.

Output: the interval in equation 5.

B.1 Fitting, grids, and numerical conventions

Preprocessing uses D_F alone so that later samples retain their separate selection and calibration roles. We standardize the response by subtracting its fitting-sample mean and dividing by its fitting-sample standard deviation, bounded below by 10^{-8}. Missing feature values are imputed by their D_F medians, and the tree models need no feature scaling. Before measuring width, we transform intervals back to the original response units. This transformation preserves the shift/scale family, with the corresponding common rescaling of its calibrated radius.

We use two point learners. The histogram gradient-boosting learner uses 100 iterations, learning rate .08, at most 15 leaves, L_2 regularization 1, and no early stopping. The random forest uses 80 trees, minimum leaf size 5, all features eligible at a split, and one worker.

To learn a scale proxy, we first predict each fitting observation with a model that did not train on that observation. For each $i \in D_F$, let $\mathrm{fold}(i)$ be its assigned fold in a three-fold partition, and let $\widehat{\mu}^{\mathrm{std}}_{-\mathrm{fold}(i)}$ be the point predictor fitted on the other two folds. Write $Y_i^{\mathrm{std}} = (Y_i - \overline{Y}_F)/s_F$, where $\overline{Y}_F$ and s_F are the fitting-sample response mean and clipped standard deviation. The scale regressor learns the log absolute residual, with target $\log\{\max(|Y_i^{\mathrm{std}} - \widehat{\mu}^{\mathrm{std}}_{-\mathrm{fold}(i)}(X_i)|, .01)\}$. It uses 70 iterations, learning rate .06, at most 10 leaves, and L_2 regularization 3. We exponentiate its predictions, divide them by their median on D_F, and clip them to $[.1, 10]$. The final point predictor is then refitted on all of D_F. Thus both predictors are fixed before any D_A, D_C, or test responses are used. These choices use scikit-learn (Pedregosa et al., 2011).

The signed-tail allocation grid is given in Section 3. Its selection criterion uses inverse empirical CDF quantiles. Width ties choose the allocation closest to $\alpha/2$, then the smaller value. For the shift/exponent search, we evaluate γ in ascending order and, within each exponent, b in ascending order; we retain the first minimizing pair. The two restricted variants fix either $b = 0$ to study scale selection alone or $\gamma = 1$ to study center selection at full normalization. These restrictions help distinguish the two adjustments. The FQ and FQ+ grids nevertheless describe different interval families: a global offset from FQ need not equal $b\widehat{\sigma}(x)$ when scale varies with x, so neither realized finite-sample family is assumed to contain the other.

The rank calculations implement the stated floor/ceiling conventions exactly, allowing a 10^{-12} tolerance only for binary floating-point error at integer-valued products. Endpoints at ranks 0 or $m+1$ remain infinite, and the code checks the total rank gap for every two-tail interval. Likewise, for a sample of size n, a one-sided score quantile at rank $n+1$ is $+\infty$. These cases remain part of the results: infinite widths stay infinite in averages. Any diagnostic restricted to finite intervals is labeled and reported separately from the probability and width summaries.

Calibration resolution. Small samples limit which corrected ranks are available. For a sample of size n, the corrected quantile of finite scores is finite exactly when $n \geq \lceil 1/\alpha - 1 \rceil$. At 90% target coverage, ECP-FQ+ therefore needs at least nine allocation observations for finite pilot objectives and nine final-calibration observations for a finite threshold. If every pilot objective is infinite, the first-candidate rule still selects a shape; if the final threshold is infinite, the method returns $\mathbb{R}$. FQ needs more calibration observations for two finite endpoints: $(m+1)\alpha \geq 2$, or $m \geq 19$ at 90% coverage. These conditions make finite intervals possible. Reliable shape selection also requires controlling the statistical error studied in Section 4.

B.2 Equivariance to response units

A change of response units should transform the interval in the same way as the response. The calibration construction has this property when the fitted functions transform correspondingly. For $Y' = uY + d$, $u > 0$, use fitted functions $\widehat{\mu}' = u\widehat{\mu} + d$, $\widehat{\sigma}' = u\widehat{\sigma}$. Then

$$S'_{b,\gamma} = u^{1-\gamma} S_{b,\gamma}, \qquad \widetilde{q}'_{b,\gamma}\widetilde{v}'_{b,\gamma} = u\widetilde{q}_{b,\gamma}\widetilde{v}_{b,\gamma}.$$

Every pilot width is therefore multiplied by the same positive factor, so the selected pair is unchanged. The final interval transforms as $\widehat{C}'_+(x) = u\widehat{C}_+(x) + d$. This statement requires the fitted models and the scale-clipping thresholds to transform consistently with the response units.

B.3 BASELINE DEFINITIONS

The baselines make the cost of selection visible through their calibration budgets. Pooled SCP and normalized SCP use all $n_A + m$ reserve observations. Matched SCP and the fixed signed-tail allocations $\alpha/2$ and $\alpha/4$ use only D_C, matching the final-calibration sample of the adaptive methods. Corrected FQ and all FQ+ variants use D_A for selection and D_C for calibration.

The sample-reuse diagnostic intentionally combines those last two roles. In fitted-data experiments, FQ selects and calibrates on the pooled reserve; in oracle diagnostics, it reuses D_C alone. Both lack the independent-selection rank argument. Their empirical coverage helps assess the effect of sample reuse and is not used to justify a width gain for a valid method.

All quantile-based comparators use a common histogram gradient-boosting learner fitted on D_F at levels

$$.025, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, .975.$$

Each quantile fit uses 100 iterations, learning rate .08, at most 15 leaves, and L_2 regularization 1. We share these fits and predictions across point-learner labels and coverage levels. Consequently, the quantile comparisons evaluate the same histogram-boosting benchmark under both labels; the forest label does not indicate a forest-specific CQR fit.

Let $\widehat{l}, \widehat{u}$ be the fitted quantiles at levels $\alpha/2, 1-\alpha/2$. We order the two predictions to remove crossings before calibration. Standard CQR then calibrates $\max\{\widehat{l}(X) - Y, Y - \widehat{u}(X)\}$ on the pooled reserve and returns $[\widehat{l}(x) - q, \widehat{u}(x) + q]$ (Romano et al., 2019). CQR-r divides this score by $d(x) = \max\{\widehat{u}(x) - \widehat{l}(x), 10^{-6}\}$ and returns $[\widehat{l}(x) - qd(x), \widehat{u}(x) + qd(x)]$ (Sesia & Candès, 2020). The correction q can be negative, in which case calibration shrinks the interval. If the corrected endpoints cross, the output is an empty set with width zero. To examine a fixed unequal tail allocation, one-sided CQR calibrates $\widehat{l}(X) - Y$ at lower error budget $\alpha/4$ and $Y - \widehat{u}(X)$ at upper budget $3\alpha/4$, using their respective corrected ranks. It keeps the same fitted endpoint quantiles, so this comparison changes the calibration budgets without optimizing their allocation.

Distributional conformal prediction. For distributional conformal prediction (DCP), we first sort the fitted quantiles and join them to form a monotone piecewise-linear quantile function on $[.025, .975]$. To extend this function beyond the fitted levels, the primary baseline attaches exponential tails. Let q_L, q_{L2} be the fitted .025 and .05 quantiles, and q_{U2}, q_U the .95 and .975 quantiles after strict monotone adjustment. Define

$$a_L = (q_{L2} - q_L)/\log 2, \qquad a_U = (q_U - q_{U2})/\log 2.$$

For $u < .025$ use $\widehat{Q}(u \mid x) = q_L + a_L \log(u/.025)$; for $u > .975$ use $q_U - a_U \log\{(1-u)/.025\}$. A negligible monotone spacing ensures positive slopes. Inverting this quantile function gives the fitted CDF $\widehat{F}(\cdot \mid x)$, which has full support. We calibrate

$$S_{\mathrm{DCP}}(x, y) = -\log\big[2\min\{\widehat{F}(y \mid x), 1 - \widehat{F}(y \mid x)\}\big]$$

on the pooled reserve, then invert the score at its calibrated threshold to obtain the interval. Extreme-tail scores and inverses are evaluated directly in log probability to avoid numerical underflow. The score is a monotone transformation of the standard distributional conformity ordering (Chernozhukov et al., 2021); Appendix I also evaluates the shortest-interval adjustment in optimized DCP. To examine sensitivity to the fitted tails, we include a version with finite linear extrapolation to levels zero and one. Its occasional infinite intervals arise from that fitted-support convention.

Nested greedy interquantile comparison. The CIR-style comparator builds a nested sequence from the common fitted quantile function. It first interpolates that function onto an equally spaced probability grid, starts with the shortest adjacent cell, and then adds the shorter available neighboring cell at each step. The score records the first step that contains the response, and final calibration uses the corrected score rank. We use the same sequence to invert the calibrated score. A response outside all finite cells has infinite score, and an infinite threshold returns $\mathbb{R}$.

We examine 20 and 100 cells to assess the effect of finer resolution. This greedy implementation follows the idea in Guo et al. (2026), while the full published CIR/CIR+ pipeline and its conditional optimality are outside this comparison. Every finite output is connected, and the discrete scores can make coverage conservative.

B.4 DATASETS AND SPLITTING

Table 4: Real-data analysis samples and response units. The cap is selected once without using outcomes for method tuning.

Dataset	Original n	Used n	Features	Response
CASP	45,730	6,000	9	RMSD
Concrete	1,030	1,030	8	Strength (MPa)
Bike Sharing	17,379	6,000	12	Hourly count
California Housing	20,640	6,000	8	Value (10^5 USD)
Energy Efficiency	768	768	8	Heating load

For the capped datasets in Table 4, we select a uniform sample of row indices once, using seed 20260909. We then use split seeds 31000–31019 to assign $\lfloor .5n \rfloor$ observations to fitting and $\lfloor .15n \rfloor$ each to allocation and calibration, with the remaining rows assigned to testing. In that order, Concrete uses 515/154/154/207 observations, Energy uses 384/115/115/154, and each capped dataset uses 3000/900/900/1200. We save the original row indices and every split index so that all methods can be compared on the same observations.

CASP, Concrete, Bike Sharing, and Energy come from the UCI repository. For Energy, heating load is the response and cooling load is excluded from the predictors. For Bike, we use the hourly data and remove `casual` and `registered` because their sum is the target. We also remove the raw date and row identifier, while retaining calendar and weather predictors. The random splits assess sampling from the observed data; they do not evaluate prediction of a future time series.

California Housing comes from the OpenML 537 mirror of the original StatLib data. Although a historical description mentions a logarithmic regression, the raw response values are dollar amounts. We divide them by 10^5 and reconstruct the standard eight predictors: median income, median age, rooms per household, bedrooms per household, population, occupants per household, latitude, and longitude. The supplied provenance file records dataset URLs, checksums, variable names, and the exact transformation.

B.5 INFERENCE, TIMING, AND REPRODUCIBILITY CHECKS

The uncertainty summaries reflect how each experiment is repeated. Simulations use independent repetitions. Real-data experiments repeatedly split a fixed selected dataset, so their summaries describe split variation rather than uncertainty from collecting a new population sample. We use Student-t intervals for repetition-level means and form each paired width difference or ratio within a repetition before averaging. These descriptive intervals are not simultaneous confidence bands. The supplementary coverage-check file also reports exploratory one-sided tests against nominal coverage, with Holm adjustments within experiment suites. Those tests assess the recorded experiments; the finite-sample guarantee comes from the proof.

Selection/calibration runtimes measure the work after nuisance fitting and shared prediction. In the oracle diagnostics, full and restricted variants share the full grid computation, so their timing fields cannot measure any speed advantage from restricting the search. Fitted-data logs separately record nuisance fitting and quantile fitting times. Shared fits and duplicated quantile-comparator rows count as the same computation, rather than as independent timing experiments.

The numerical verification script checks that the implementation handles the cases used in the analysis: normal-distribution rank coverage, ties, infinite endpoints, score inversion, sign reflection, selection independence from calibration/test labels, and known-CDF evaluations. The DCP sidecar additionally checks extreme-tail score inversion. These checks connect the code to the mathematical construction. Reproduction commands and environment versions are provided in the package README.

C ADDITIONAL DIAGNOSTIC RESULTS

Population illustration of scale adaptation. Figure 6 makes the tradeoff in Corollary 6 concrete. Consider two equally frequent Gaussian groups with standard deviations 1 and 4. Full normalization gives 90% coverage in each group and average length 8.22. Partial normalization lowers average length to approximately 7.66 at the same marginal coverage, but group coverages become approximately 97% and 83%. Thus the width reduction comes with a change in where errors occur. This example uses known distributions and population thresholds. The experiments below add the uncertainty from selecting and calibrating intervals with finite samples.

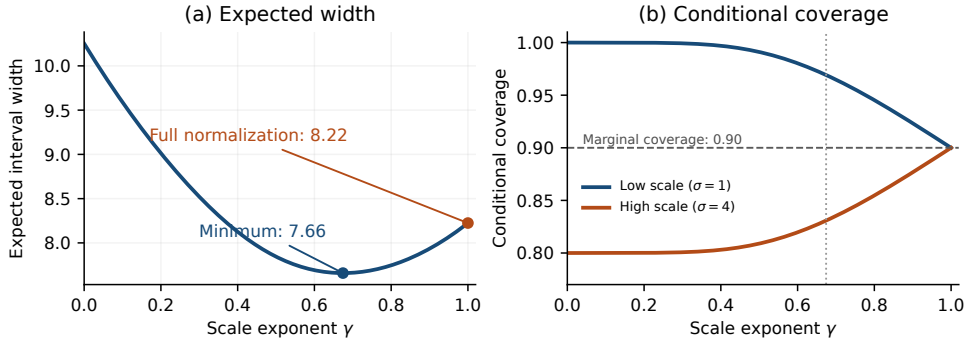

Figure 6: Scale response and coverage allocation in a population example. Each exponent is calibrated to 90% marginal coverage for two equally frequent Gaussian groups with scales 1 and 4. Partial normalization shortens average length while lowering coverage in the high-scale group.

Finite-sample diagnostic settings. We use six error distributions to examine different reasons why interval shape may matter. In the first five cases, standardized errors are independent of the scale, which is .5 or 2 with equal probability. Gaussian errors are standard normal. Gamma errors have shape 2, and log-normal errors have unit log standard deviation; both are centered and divided by their population standard deviation. The bimodal error is an equal mixture of $N(-1.5, .4^2)$ and $N(1.5, .4^2)$ divided by $\sqrt{2.41}$. Student errors have three degrees of freedom and are divided by $\sqrt{3}$. The opposing-skew case instead reverses the log-normal orientation between scale groups. Its standardized errors therefore depend on scale, making it a stress test outside the assumptions of Corollary 6.

All coverage-level, sample-size, and allocation-fraction sweeps use Gamma errors. Their common central setting is $n_A = m = 500$, $\alpha = .1$. The size sweep increases allocation and calibration samples together. In contrast, the fraction sweep holds the total reserve size $n_A + m = 1000$ fixed and changes how it is divided. This separates the value of more observations from the choice of how many to spend on selection.

Table 5: central normal. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.901 [0.899, 0.902]	5.146 [5.121, 5.170]	1.000	0.801
SCP matched	0.901 [0.899, 0.902]	5.158 [5.122, 5.193]	1.000	0.801
Equal tail	0.900 [0.899, 0.902]	5.170 [5.134, 5.206]	1.000	0.801
ECP-FQ	0.901 [0.899, 0.903]	5.241 [5.204, 5.278]	1.000	0.802
Reused FQ (diagnostic)	0.894 [0.893, 0.896]	5.071 [5.035, 5.106]	1.000	0.789
ECP-FQ+ $\gamma = 1$	0.901 [0.900, 0.903]	4.168 [4.147, 4.188]	0.901	0.901
ECP-FQ+ $b = 0$	0.901 [0.900, 0.903]	3.893 [3.872, 3.914]	0.964	0.838
ECP-FQ+	0.901 [0.900, 0.903]	3.923 [3.903, 3.944]	0.964	0.839

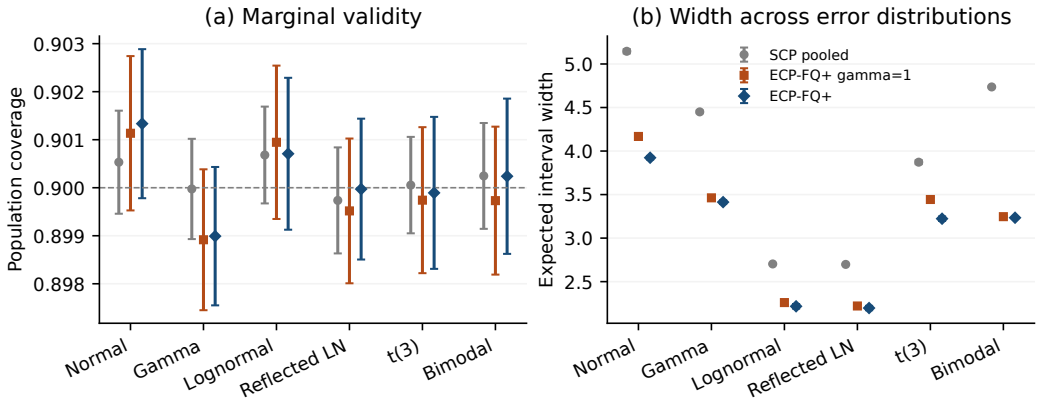

Figure 7: Independent oracle-nuisance diagnostics across error families. Intervals reflect variation over independent allocation/calibration repetitions. Coverage and width must be read together; the sample-reusing diagnostic has no exact rank guarantee.

Table 6: central gamma. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	4.450 [4.435, 4.465]	0.998	0.802
SCP matched	0.900 [0.899, 0.902]	4.460 [4.438, 4.482]	0.998	0.803
Equal tail	0.901 [0.899, 0.902]	4.799 [4.761, 4.838]	0.999	0.803
ECP-FQ	0.900 [0.899, 0.902]	4.287 [4.260, 4.314]	0.991	0.809
Reused FQ (diagnostic)	0.896 [0.894, 0.897]	4.206 [4.181, 4.231]	0.990	0.802
ECP-FQ+ $\gamma = 1$	0.899 [0.897, 0.900]	3.463 [3.442, 3.484]	0.899	0.899
ECP-FQ+ $b = 0$	0.899 [0.898, 0.901]	3.384 [3.366, 3.402]	0.931	0.868
ECP-FQ+	0.899 [0.898, 0.900]	3.413 [3.392, 3.435]	0.934	0.864

Table 7: central lognormal. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.901 [0.900, 0.902]	2.703 [2.697, 2.709]	0.978	0.823
SCP matched	0.900 [0.898, 0.901]	2.702 [2.692, 2.711]	0.978	0.822
Equal tail	0.900 [0.899, 0.902]	3.314 [3.273, 3.354]	0.990	0.811
ECP-FQ	0.901 [0.899, 0.902]	2.495 [2.475, 2.514]	0.964	0.838
Reused FQ (diagnostic)	0.899 [0.897, 0.900]	2.478 [2.459, 2.497]	0.964	0.834
ECP-FQ+ $\gamma = 1$	0.901 [0.899, 0.903]	2.260 [2.232, 2.287]	0.901	0.901
ECP-FQ+ $b = 0$	0.901 [0.899, 0.902]	2.194 [2.167, 2.220]	0.927	0.875
ECP-FQ+	0.901 [0.899, 0.902]	2.217 [2.190, 2.243]	0.922	0.880

Table 8: central reflected lognormal. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	2.698 [2.691, 2.705]	0.978	0.822
SCP matched	0.900 [0.899, 0.902]	2.702 [2.693, 2.711]	0.978	0.822
Equal tail	0.900 [0.899, 0.902]	3.321 [3.282, 3.359]	0.990	0.811
ECP-FQ	0.899 [0.898, 0.901]	2.478 [2.459, 2.498]	0.963	0.836
Reused FQ (diagnostic)	0.898 [0.897, 0.900]	2.465 [2.446, 2.485]	0.963	0.833
ECP-FQ+ $\gamma = 1$	0.900 [0.898, 0.901]	2.220 [2.196, 2.245]	0.900	0.900
ECP-FQ+ $b = 0$	0.900 [0.899, 0.901]	2.170 [2.145, 2.194]	0.925	0.875
ECP-FQ+	0.900 [0.898, 0.901]	2.197 [2.172, 2.222]	0.922	0.878

Table 9: central student t3. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	3.872 [3.850, 3.895]	0.993	0.807
SCP matched	0.900 [0.899, 0.902]	3.892 [3.856, 3.928]	0.993	0.807
Equal tail	0.900 [0.899, 0.902]	3.910 [3.875, 3.945]	0.993	0.807
ECP-FQ	0.900 [0.898, 0.902]	3.977 [3.936, 4.018]	0.993	0.807
Reused FQ (diagnostic)	0.895 [0.893, 0.897]	3.825 [3.791, 3.858]	0.992	0.798
ECP-FQ+ $\gamma = 1$	0.900 [0.898, 0.901]	3.444 [3.419, 3.469]	0.900	0.900
ECP-FQ+ $b = 0$	0.900 [0.898, 0.902]	3.196 [3.171, 3.221]	0.949	0.851
ECP-FQ+	0.900 [0.898, 0.901]	3.223 [3.197, 3.249]	0.951	0.849

Table 10: central bimodal. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	4.736 [4.728, 4.744]	1.000	0.800
SCP matched	0.900 [0.899, 0.902]	4.738 [4.727, 4.750]	1.000	0.800
Equal tail	0.900 [0.898, 0.902]	4.741 [4.730, 4.753]	1.000	0.800
ECP-FQ	0.901 [0.899, 0.902]	4.766 [4.754, 4.779]	1.000	0.802
Reused FQ (diagnostic)	0.894 [0.892, 0.895]	4.709 [4.697, 4.721]	1.000	0.787
ECP-FQ+ $\gamma = 1$	0.900 [0.898, 0.901]	3.246 [3.240, 3.252]	0.900	0.900
ECP-FQ+ $b = 0$	0.901 [0.899, 0.902]	3.226 [3.219, 3.233]	0.967	0.835
ECP-FQ+	0.900 [0.899, 0.902]	3.234 [3.226, 3.242]	0.971	0.829

Table 11: alpha 0.05. $n_A = 500, m = 500, \alpha = 0.05$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.950 [0.949, 0.951]	5.545 [5.507, 5.583]	0.999	0.901
SCP matched	0.950 [0.949, 0.951]	5.617 [5.557, 5.677]	0.999	0.901
Equal tail	0.952 [0.951, 0.953]	6.372 [6.313, 6.431]	1.000	0.905
ECP-FQ	0.952 [0.950, 0.953]	5.586 [5.542, 5.630]	0.999	0.904
Reused FQ (diagnostic)	0.951 [0.950, 0.952]	5.561 [5.519, 5.603]	0.999	0.903
ECP-FQ+ $\gamma = 1$	0.950 [0.949, 0.951]	4.334 [4.299, 4.369]	0.950	0.950
ECP-FQ+ $b = 0$	0.950 [0.949, 0.951]	4.549 [4.497, 4.601]	0.977	0.923
ECP-FQ+	0.950 [0.949, 0.952]	4.330 [4.291, 4.369]	0.965	0.935

Table 12: alpha 0.2. $n_A = 500, m = 500, \alpha = 0.2$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.800 [0.799, 0.801]	3.241 [3.226, 3.256]	0.989	0.611
SCP matched	0.800 [0.798, 0.802]	3.245 [3.223, 3.266]	0.989	0.611
Equal tail	0.800 [0.797, 0.802]	3.221 [3.194, 3.248]	0.987	0.613
ECP-FQ	0.802 [0.800, 0.804]	3.120 [3.098, 3.142]	0.957	0.647
Reused FQ (diagnostic)	0.795 [0.793, 0.797]	3.040 [3.021, 3.059]	0.954	0.636
ECP-FQ+ $\gamma = 1$	0.801 [0.799, 0.804]	2.625 [2.609, 2.641]	0.801	0.801
ECP-FQ+ $b = 0$	0.800 [0.798, 0.802]	2.584 [2.571, 2.597]	0.898	0.702
ECP-FQ+	0.801 [0.799, 0.803]	2.510 [2.494, 2.526]	0.878	0.724

Table 13: ncal 200. $n_A = 200, m = 200, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.902]	4.461 [4.437, 4.486]	0.998	0.803
SCP matched	0.901 [0.899, 0.903]	4.484 [4.448, 4.520]	0.998	0.804
Equal tail	0.901 [0.898, 0.903]	4.880 [4.816, 4.945]	0.999	0.803
ECP-FQ	0.904 [0.901, 0.906]	4.431 [4.379, 4.482]	0.991	0.817
Reused FQ (diagnostic)	0.901 [0.899, 0.904]	4.363 [4.317, 4.409]	0.991	0.811
ECP-FQ+ $\gamma = 1$	0.901 [0.898, 0.903]	3.593 [3.552, 3.634]	0.901	0.901
ECP-FQ+ $b = 0$	0.901 [0.899, 0.904]	3.473 [3.436, 3.510]	0.931	0.872
ECP-FQ+	0.901 [0.898, 0.903]	3.539 [3.497, 3.580]	0.937	0.865

Table 14: ncal 1000. $n_A = 1000, m = 1000, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.900, 0.901]	4.452 [4.442, 4.463]	0.998	0.803
SCP matched	0.901 [0.900, 0.902]	4.468 [4.453, 4.484]	0.998	0.805
Equal tail	0.901 [0.900, 0.902]	4.812 [4.785, 4.839]	0.999	0.802
ECP-FQ	0.902 [0.901, 0.903]	4.295 [4.275, 4.315]	0.992	0.813
Reused FQ (diagnostic)	0.899 [0.898, 0.900]	4.238 [4.220, 4.257]	0.991	0.807
ECP-FQ+ $\gamma = 1$	0.901 [0.900, 0.902]	3.450 [3.435, 3.465]	0.901	0.901
ECP-FQ+ $b = 0$	0.901 [0.900, 0.902]	3.383 [3.371, 3.396]	0.931	0.871
ECP-FQ+	0.901 [0.900, 0.902]	3.406 [3.391, 3.422]	0.934	0.868

Table 15: ncal 3000. $n_A = 3000, m = 3000, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.900]	4.445 [4.438, 4.451]	0.998	0.802
SCP matched	0.900 [0.899, 0.901]	4.448 [4.439, 4.457]	0.998	0.802
Equal tail	0.900 [0.899, 0.900]	4.766 [4.750, 4.782]	0.999	0.800
ECP-FQ	0.900 [0.900, 0.901]	4.243 [4.232, 4.254]	0.991	0.810
Reused FQ (diagnostic)	0.899 [0.898, 0.900]	4.220 [4.209, 4.230]	0.991	0.807
ECP-FQ+ $\gamma = 1$	0.900 [0.900, 0.901]	3.420 [3.411, 3.428]	0.900	0.900
ECP-FQ+ $b = 0$	0.900 [0.899, 0.900]	3.357 [3.350, 3.364]	0.935	0.865
ECP-FQ+	0.900 [0.899, 0.900]	3.364 [3.356, 3.372]	0.934	0.866

Table 16: alloc fraction 0.25. $n_A = 250, m = 750, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	4.450 [4.435, 4.465]	0.998	0.802
SCP matched	0.900 [0.899, 0.901]	4.452 [4.435, 4.469]	0.998	0.803
Equal tail	0.902 [0.901, 0.903]	4.807 [4.776, 4.837]	0.999	0.804
ECP-FQ	0.901 [0.900, 0.902]	4.320 [4.295, 4.344]	0.992	0.810
Reused FQ (diagnostic)	0.897 [0.896, 0.899]	4.213 [4.193, 4.232]	0.990	0.804
ECP-FQ+ $\gamma = 1$	0.900 [0.899, 0.901]	3.506 [3.486, 3.526]	0.900	0.900
ECP-FQ+ $b = 0$	0.901 [0.899, 0.902]	3.408 [3.392, 3.424]	0.932	0.869
ECP-FQ+	0.900 [0.899, 0.901]	3.460 [3.439, 3.482]	0.935	0.865

Table 17: alloc fraction 0.75. $n_A = 750, m = 250, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	4.446 [4.431, 4.461]	0.998	0.802
SCP matched	0.901 [0.899, 0.903]	4.477 [4.446, 4.509]	0.998	0.804
Equal tail	0.904 [0.902, 0.906]	4.962 [4.906, 5.018]	0.999	0.809
ECP-FQ	0.905 [0.903, 0.907]	4.383 [4.344, 4.422]	0.991	0.819
Reused FQ (diagnostic)	0.902 [0.900, 0.905]	4.345 [4.307, 4.383]	0.992	0.813
ECP-FQ+ $\gamma = 1$	0.901 [0.899, 0.903]	3.515 [3.483, 3.547]	0.901	0.901
ECP-FQ+ $b = 0$	0.901 [0.899, 0.903]	3.421 [3.392, 3.449]	0.932	0.870
ECP-FQ+	0.901 [0.899, 0.903]	3.452 [3.420, 3.485]	0.935	0.867

Table 18: stress opposing skew. $n_A = 500, m = 500, \alpha = 0.1$. Each bracket is a 95% interval for the repetition mean. Group coverage is conditional on the low or high scale.

Method	Coverage [CI]	Width [CI]	Low	High
SCP pooled	0.900 [0.899, 0.901]	2.698 [2.691, 2.704]	0.978	0.822
SCP matched	0.900 [0.899, 0.902]	2.706 [2.696, 2.716]	0.978	0.823
Equal tail	0.900 [0.898, 0.901]	3.151 [3.107, 3.195]	0.976	0.823
ECP-FQ	0.900 [0.899, 0.902]	2.450 [2.409, 2.492]	0.983	0.818
Reused FQ (diagnostic)	0.899 [0.898, 0.901]	2.314 [2.285, 2.342]	0.981	0.818
ECP-FQ+ $\gamma = 1$	0.901 [0.899, 0.903]	2.365 [2.330, 2.400]	0.901	0.901
ECP-FQ+ $b = 0$	0.900 [0.899, 0.902]	2.201 [2.172, 2.230]	0.926	0.875
ECP-FQ+	0.901 [0.899, 0.902]	2.207 [2.175, 2.239]	0.937	0.864

D ORIGINAL REGRESSION DESIGNS AND RERUNS

The five original regression designs allow us to compare interval construction after fitting a point predictor. We preserve their data-generating distributions, centering each Gaussian mixture error by subtracting its mixture expectation. Designs A–E have 10, 12, 20, 18, and 15 covariates, respectively; design C is therefore not a high-dimensional setting. Each experiment uses $n_F = 900$, $n_A = m = 1500$, and 1000 independent test observations. Pooled baselines use all 3000 reserve observations for calibration. Results average 30 independent repetitions for each learner and design. The location function is denoted by μ below, and SD denotes standard deviation.

For A, the covariates are independent standard Gaussians and $\mu = 2.6X_1 - 1.9X_2 + 1.5X_3 - X_4$. The mixture has weights $(.56, .24, .20)$, means $(0, 2.6, 6.4)$, and common SD .09. For B, covariates are independent uniform on $[-2, 2]$, $\mu = 2.2X_1 - 1.7X_2 + 1.3X_3 - .9X_4 + .6X_5$, and the mixture has weights $(.56, .24, .20)$, means $(0, 2.5, 6.2)$, and SD .08. For C, Gaussian covariates have covariance $.85^{|j-k|}$, $\mu = 2.4X_1 - 1.8X_2 + 1.4X_3 - X_4 + .7X_5$, and the mixture has weights $(.56, .24, .20)$, means $(0, 2.5, 6.1)$, and SD .08.

For D, covariates are independent standard Gaussians, $\mu = 1.8X_1 - 1.2X_2 + .7X_3 + .8\mathbf{1}\{X_1 > 0\} - .5\mathbf{1}\{X_4 > 0\}$, and the mixture has weights $(.62, .23, .15)$, means $(0, 2.8, 7.8)$, and SD .07. For E, covariates are independent uniform on $[-2, 2]$, $\mu = .9\sin(\pi X_1) + .25X_2^2 + .15X_3$, and the mixture has weights $(.58, .23, .19)$, means $(0, 2.3, 5.6)$, and within-component SD $.08(1 + .10|X_1|)$. The last modification produces only mild within-component heteroscedasticity; it does not vary mixture probabilities or modes with covariates.

Table 19: Original regression distributions: HistGBM. Entries are coverage/mean width. Quantile comparators share fits across point-learner labels. Full repetition metrics, confidence intervals and paired differences are supplied as CSV files.

Method	A	B	C	D	E
SCP pooled	0.900/9.249	0.897/8.827	0.899/8.654	0.898/11.218	0.898/8.008
SCP matched	0.901/9.267	0.899/8.855	0.900/8.670	0.899/11.244	0.898/7.993
Normalized SCP	0.898/9.613	0.899/9.197	0.900/9.058	0.897/10.800	0.899/8.194
ECP-FQ	0.902/8.586	0.895/8.254	0.899/8.150	0.901/9.777	0.898/7.273
ECP-FQ+	0.900/8.688	0.898/8.337	0.897/8.212	0.897/9.695	0.898/7.312
ECP-FQ+ ($b = 0$)	0.900/9.255	0.899/8.861	0.901/8.694	0.896/10.797	0.900/7.988
ECP-FQ+ ($\gamma = 1$)	0.898/9.193	0.898/8.822	0.901/8.774	0.897/9.982	0.900/7.709
Fixed equal tail	0.901/8.578	0.896/8.256	0.899/8.153	0.900/9.791	0.897/7.247
Fixed quarter tail	0.901/8.670	0.896/8.303	0.901/8.192	0.902/9.789	0.900/7.309
Reused FQ (diagnostic)	0.899/8.548	0.894/8.220	0.897/8.096	0.899/9.737	0.897/7.228
CQR	0.898/9.369	0.897/9.455	0.899/7.804	0.898/9.216	0.901/6.268
CQR-r	0.896/9.489	0.898/9.547	0.899/7.813	0.898/9.170	0.901/6.259
Two-tail CQR	0.898/9.353	0.897/9.440	0.899/7.804	0.898/9.228	0.898/6.296
DCP finite support	0.914/∞	0.945/∞	0.922/∞	0.902/∞	0.900/6.917
DCP exponential tails	0.897/10.997	0.900/11.523	0.899/8.311	0.899/10.228	0.900/6.917
CIR greedy (100 cells)	0.917/∞	0.947/∞	0.927/∞	0.909/∞	0.907/7.012

Table 20: Original regression distributions: RF. Entries are coverage/mean width. Quantile comparators share fits across point-learner labels. Full repetition metrics, confidence intervals and paired differences are supplied as CSV files.

Method	A	B	C	D	E
SCP pooled	0.896/9.264	0.895/8.849	0.901/8.543	0.897/11.239	0.900/7.856
SCP matched	0.897/9.283	0.894/8.840	0.902/8.554	0.898/11.253	0.900/7.846
Normalized SCP	0.900/9.722	0.897/9.269	0.902/8.942	0.896/10.782	0.903/7.996
ECP-FQ	0.898/8.675	0.898/8.454	0.903/7.938	0.897/9.394	0.899/6.793
ECP-FQ+	0.899/8.827	0.897/8.485	0.902/7.997	0.896/9.379	0.900/6.857
ECP-FQ+ ($b = 0$)	0.898/9.311	0.897/8.851	0.902/8.584	0.897/10.781	0.903/7.845
ECP-FQ+ ($\gamma = 1$)	0.899/9.302	0.897/8.931	0.901/8.570	0.896/9.735	0.902/7.352
Fixed equal tail	0.897/8.642	0.898/8.445	0.902/7.922	0.896/9.390	0.897/6.785
Fixed quarter tail	0.898/8.799	0.896/8.475	0.904/7.977	0.895/9.399	0.900/6.817
Reused FQ (diagnostic)	0.895/8.611	0.896/8.400	0.901/7.895	0.895/9.351	0.894/6.758
CQR	0.898/9.369	0.897/9.455	0.899/7.804	0.898/9.216	0.901/6.268
CQR-r	0.896/9.489	0.898/9.547	0.899/7.813	0.898/9.170	0.901/6.259
Two-tail CQR	0.898/9.353	0.897/9.440	0.899/7.804	0.898/9.228	0.898/6.296
DCP finite support	0.914/∞	0.945/∞	0.922/∞	0.902/∞	0.900/6.917
DCP exponential tails	0.897/10.997	0.900/11.523	0.899/8.311	0.899/10.228	0.900/6.917
CIR greedy (100 cells)	0.917/∞	0.947/∞	0.927/∞	0.909/∞	0.907/7.012

Table 21: Paired percentage width changes relative to pooled SCP in the original designs. Brackets are 95% intervals over independent repetitions. Negative changes indicate shorter intervals.

Design	Learner	ECP-FQ change [CI]	ECP-FQ+ change [CI]
A	HistGBM	-7.15 [-7.80, -6.50]	-6.06 [-6.74, -5.37]
B	HistGBM	-6.46 [-7.14, -5.79]	-5.53 [-6.24, -4.82]
C	HistGBM	-5.80 [-6.46, -5.14]	-5.10 [-5.82, -4.38]
D	HistGBM	-12.79 [-13.74, -11.85]	-13.54 [-14.60, -12.48]
E	HistGBM	-9.12 [-10.19, -8.05]	-8.65 [-9.62, -7.69]
A	RF	-6.35 [-7.17, -5.52]	-4.71 [-5.41, -4.01]
B	RF	-4.43 [-5.32, -3.53]	-4.08 [-5.07, -3.09]
C	RF	-7.05 [-7.81, -6.29]	-6.36 [-7.12, -5.59]
D	RF	-16.35 [-17.38, -15.32]	-16.48 [-17.61, -15.36]
E	RF	-13.48 [-14.42, -12.54]	-12.66 [-13.69, -11.63]

E FULL REAL-DATA RESULTS

The following tables report both point learners at all three nominal coverage levels. Each entry averages 20 randomized splits, and interval widths use the original response units. Repeated rows for the common quantile learner make comparisons easier; they represent the same fitted model. The raw and summary CSV files also retain tail errors, confidence intervals, selected parameters, finite-width frequencies, and timings. We include finite-support DCP and 20-cell CIR to examine sensitivity to tail support and grid resolution. Their infinite widths or conservative coverage describe those particular implementations, rather than every member of the corresponding method families.

These quantile models use fixed hyperparameters. Although they provide useful comparisons, they do not exhaust the possible tuning of published methods. Pooled SCP controls and the shift/exponent restrictions give the most direct tests of our own design choices.

Table 22: HistGBM, nominal coverage 95%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.949/18.72	0.953/22.29	0.951/211.69	0.951/2.15	0.941/3.24
SCP matched	0.946/18.57	0.955/22.44	0.952/212.07	0.951/2.16	0.951/3.61
Normalized SCP	0.949/17.53	0.958/25.52	0.951/160.38	0.950/1.88	0.948/3.27
ECP-FQ	0.948/18.57	0.958/25.01	0.951/215.01	0.953/2.20	0.952/3.90
ECP-FQ+	0.946/16.93	0.965/24.48	0.949/158.66	0.952/1.90	0.958/2.77
ECP-FQ+ ($b = 0$)	0.946/17.18	0.964/24.05	0.949/158.66	0.951/1.89	0.958/2.77
ECP-FQ+ ($\gamma = 1$)	0.946/17.19	0.961/28.33	0.950/159.15	0.950/1.89	0.953/3.44
Fixed equal tail	0.949/18.59	0.966/25.55	0.952/213.48	0.953/2.21	0.959/3.93
Fixed quarter tail	0.948/18.57	0.962/26.56	0.951/231.15	0.953/2.21	0.952/4.07
Reused FQ (diagnostic)	0.948/18.40	0.952/22.74	0.952/210.74	0.950/2.12	0.943/3.32
CQR	0.951/16.57	0.948/35.08	0.948/282.91	0.948/2.36	0.952/12.50
CQR-r	0.951/16.56	0.950/35.66	0.947/283.77	0.948/2.37	0.953/12.51
Two-tail CQR	0.952/16.58	0.955/36.45	0.948/282.41	0.951/2.34	0.961/13.11
DCP finite support	0.972/∞	0.997/∞	0.955/∞	0.960/∞	0.955/∞
DCP exponential tails	0.952/17.73	0.949/65.24	0.948/311.16	0.949/2.68	0.950/11.52
CIR greedy (100 cells)	0.974/∞	0.997/∞	0.960/∞	0.961/∞	0.964/∞
CIR greedy (20 cells)	0.975/∞	0.997/∞	0.965/∞	0.964/∞	0.969/∞

F COVERAGE CHECKS AND REPRODUCIBILITY ARTIFACTS

The file `results/coverage_checks.csv` gives an exploratory check for average coverage below the nominal target. It applies a one-sided Student-t calculation to repetition-level coverage and reports both unadjusted and within-suite Holm-adjusted values for the three experiment suites. The uncertainty comes from repetitions or random splits, rather than treating all test predictions as independent Bernoulli observations. In particular, real-data checks condition on the fixed observed dataset. These approximate checks help identify implementation or selection problems; the coverage guarantee comes from the conformal rank argument. The reused-sample FQ diagnostic illustrates why using the same observations for selection and calibration can fail.

The package contains the source scripts, numerical results, dataset provenance, and split indices needed to trace these comparisons. Numerical checks examine distributional rank identities and change labels that the procedure should not use. This tests data separation as well as interval arithmetic. Nevertheless, software checks cannot establish the independent-sampling and regularity assumptions used in the efficiency proofs.

Table 23: HistGBM, nominal coverage 90%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.899/15.28	0.907/16.98	0.902/158.47	0.900/1.57	0.901/2.52
SCP matched	0.896/15.18	0.907/17.30	0.900/156.99	0.900/1.57	0.909/2.72
Normalized SCP	0.898/13.70	0.906/18.40	0.899/127.97	0.902/1.44	0.890/2.50
ECP-FQ	0.896/15.11	0.904/18.63	0.901/161.49	0.903/1.61	0.898/2.81
ECP-FQ+	0.894/13.73	0.913/17.18	0.898/126.47	0.900/1.43	0.899/2.02
ECP-FQ+ ($b = 0$)	0.893/13.58	0.912/16.95	0.898/126.47	0.900/1.43	0.899/2.02
ECP-FQ+ ($\gamma = 1$)	0.894/13.70	0.915/19.39	0.898/127.17	0.901/1.45	0.900/2.58
Fixed equal tail	0.896/15.16	0.907/17.82	0.900/157.05	0.901/1.62	0.910/2.76
Fixed quarter tail	0.897/15.31	0.921/19.63	0.903/172.16	0.902/1.61	0.913/3.15
Reused FQ (diagnostic)	0.895/15.05	0.901/16.94	0.900/157.96	0.898/1.57	0.896/2.48
CQR	0.899/14.61	0.907/24.04	0.897/207.33	0.900/1.78	0.890/6.40
CQR-r	0.899/14.61	0.904/24.74	0.899/211.78	0.899/1.80	0.890/6.44
Two-tail CQR	0.899/14.55	0.900/24.44	0.899/206.33	0.908/1.77	0.899/6.59
DCP finite support	0.901/15.53	0.906/∞	0.898/218.84	0.900/1.96	0.898/6.47
DCP exponential tails	0.901/15.53	0.896/32.74	0.898/218.84	0.900/1.96	0.898/6.47
CIR greedy (100 cells)	0.907/14.66	0.918/∞	0.903/179.75	0.903/1.69	0.904/3.38
CIR greedy (20 cells)	0.923/15.82	0.928/∞	0.911/195.30	0.918/1.90	0.917/4.25

Table 24: HistGBM, nominal coverage 80%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.795/11.48	0.799/11.51	0.802/106.81	0.800/1.06	0.799/1.52
SCP matched	0.796/11.52	0.806/11.74	0.799/105.64	0.800/1.07	0.796/1.52
Normalized SCP	0.793/10.30	0.807/12.73	0.799/96.80	0.801/1.06	0.790/1.89
ECP-FQ	0.798/11.59	0.816/12.52	0.798/106.02	0.803/1.07	0.817/1.80
ECP-FQ+	0.796/10.38	0.799/11.52	0.797/93.60	0.802/1.04	0.794/1.45
ECP-FQ+ ($b = 0$)	0.795/10.33	0.798/11.45	0.797/93.60	0.803/1.03	0.794/1.45
ECP-FQ+ ($\gamma = 1$)	0.794/10.33	0.809/13.06	0.795/96.00	0.804/1.07	0.793/1.91
Fixed equal tail	0.798/11.69	0.821/12.27	0.800/106.34	0.802/1.08	0.810/1.75
Fixed quarter tail	0.798/11.63	0.811/13.18	0.798/114.59	0.802/1.12	0.808/2.11
Reused FQ (diagnostic)	0.794/11.45	0.796/11.61	0.797/106.08	0.796/1.06	0.802/1.53
CQR	0.796/11.73	0.808/15.73	0.800/150.67	0.794/1.22	0.798/3.73
CQR-r	0.795/11.76	0.807/16.84	0.798/155.33	0.794/1.25	0.802/3.80
Two-tail CQR	0.797/11.48	0.804/16.18	0.801/152.12	0.798/1.23	0.804/3.87
DCP finite support	0.796/12.82	0.805/19.43	0.799/154.16	0.794/1.34	0.798/3.68
DCP exponential tails	0.796/12.82	0.805/19.43	0.799/154.16	0.794/1.34	0.798/3.68
CIR greedy (100 cells)	0.802/10.34	0.819/15.85	0.805/121.50	0.803/1.12	0.803/1.96
CIR greedy (20 cells)	0.816/10.86	0.838/17.92	0.823/130.21	0.823/1.21	0.824/2.23

Table 25: RF, nominal coverage 95%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.949/18.66	0.947/25.89	0.952/261.28	0.951/2.41	0.949/4.31
SCP matched	0.948/18.56	0.948/26.68	0.951/258.31	0.950/2.41	0.960/4.90
Normalized SCP	0.948/17.41	0.954/28.66	0.951/186.90	0.952/2.08	0.944/3.73
ECP-FQ	0.949/18.61	0.953/28.88	0.953/264.91	0.953/2.45	0.952/5.39
ECP-FQ+	0.944/16.77	0.950/26.10	0.951/182.19	0.952/2.07	0.952/3.54
ECP-FQ+ ($b = 0$)	0.946/17.02	0.954/26.15	0.950/180.91	0.952/2.08	0.951/3.49
ECP-FQ+ ($\gamma = 1$)	0.945/17.08	0.958/30.05	0.948/184.49	0.952/2.09	0.946/3.91
Fixed equal tail	0.949/18.54	0.961/29.65	0.952/261.12	0.952/2.45	0.966/6.09
Fixed quarter tail	0.950/18.68	0.959/30.93	0.951/284.31	0.952/2.45	0.955/6.92
Reused FQ (diagnostic)	0.950/18.39	0.945/25.93	0.951/260.29	0.949/2.37	0.947/4.32
CQR	0.951/16.57	0.948/35.08	0.948/282.91	0.948/2.36	0.952/12.50
CQR-r	0.951/16.56	0.950/35.66	0.947/283.77	0.948/2.37	0.953/12.51
Two-tail CQR	0.952/16.58	0.955/36.45	0.948/282.41	0.951/2.34	0.961/13.11
DCP finite support	0.972/∞	0.997/∞	0.955/∞	0.960/∞	0.955/∞
DCP exponential tails	0.952/17.73	0.949/65.24	0.948/311.16	0.949/2.68	0.950/11.52
CIR greedy (100 cells)	0.974/∞	0.997/∞	0.960/∞	0.961/∞	0.964/∞
CIR greedy (20 cells)	0.975/∞	0.997/∞	0.965/∞	0.964/∞	0.969/∞

Table 26: RF, nominal coverage 90%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.896/15.10	0.903/20.53	0.902/193.07	0.903/1.78	0.894/2.95
SCP matched	0.893/15.00	0.905/20.66	0.899/190.14	0.903/1.79	0.899/3.04
Normalized SCP	0.894/13.44	0.904/21.74	0.901/150.05	0.904/1.60	0.893/2.90
ECP-FQ	0.896/15.10	0.901/21.10	0.900/192.93	0.904/1.82	0.905/3.49
ECP-FQ+	0.893/13.35	0.898/20.31	0.895/140.22	0.903/1.59	0.893/2.47
ECP-FQ+ ($b = 0$)	0.892/13.36	0.900/19.85	0.895/140.22	0.903/1.59	0.893/2.47
ECP-FQ+ ($\gamma = 1$)	0.894/13.40	0.906/22.00	0.897/147.57	0.902/1.60	0.896/2.95
Fixed equal tail	0.894/15.00	0.905/21.49	0.899/189.80	0.902/1.82	0.906/3.51
Fixed quarter tail	0.894/15.22	0.916/23.31	0.901/204.92	0.902/1.82	0.907/5.06
Reused FQ (diagnostic)	0.892/14.90	0.896/20.39	0.900/191.58	0.900/1.78	0.899/3.01
CQR	0.899/14.61	0.907/24.04	0.897/207.33	0.900/1.78	0.890/6.40
CQR-r	0.899/14.61	0.904/24.74	0.899/211.78	0.899/1.80	0.890/6.44
Two-tail CQR	0.899/14.55	0.900/24.44	0.899/206.33	0.908/1.77	0.899/6.59
DCP finite support	0.901/15.53	0.906/∞	0.898/218.84	0.900/1.96	0.898/6.47
DCP exponential tails	0.901/15.53	0.896/32.74	0.898/218.84	0.900/1.96	0.898/6.47
CIR greedy (100 cells)	0.907/14.66	0.918/∞	0.903/179.75	0.903/1.69	0.904/3.38
CIR greedy (20 cells)	0.923/15.82	0.928/∞	0.911/195.30	0.918/1.90	0.917/4.25

Table 27: RF, nominal coverage 80%. Entries are coverage/mean width over 20 splits.

Method	CASP	Concrete	Bike	California	Energy
SCP pooled	0.798/11.23	0.799/15.03	0.804/126.90	0.801/1.21	0.796/1.92
SCP matched	0.799/11.29	0.796/14.98	0.801/125.63	0.804/1.22	0.798/1.95
Normalized SCP	0.793/10.00	0.800/15.75	0.801/111.99	0.801/1.16	0.792/2.05
ECP-FQ	0.799/11.36	0.804/15.51	0.801/127.20	0.805/1.22	0.795/2.06
ECP-FQ+	0.792/10.02	0.800/15.14	0.800/103.80	0.806/1.16	0.796/1.79
ECP-FQ+ ($b = 0$)	0.791/9.97	0.799/14.84	0.800/103.80	0.804/1.15	0.796/1.79
ECP-FQ+ ($\gamma = 1$)	0.793/10.03	0.799/15.99	0.797/110.86	0.806/1.18	0.793/2.06
Fixed equal tail	0.797/11.34	0.803/15.44	0.800/125.59	0.804/1.24	0.812/2.08
Fixed quarter tail	0.795/11.45	0.811/15.99	0.800/137.24	0.806/1.25	0.807/2.63
Reused FQ (diagnostic)	0.794/11.18	0.793/14.98	0.800/126.01	0.800/1.20	0.794/1.90
CQR	0.796/11.73	0.808/15.73	0.800/150.67	0.794/1.22	0.798/3.73
CQR-r	0.795/11.76	0.807/16.84	0.798/155.33	0.794/1.25	0.802/3.80
Two-tail CQR	0.797/11.48	0.804/16.18	0.801/152.12	0.798/1.23	0.804/3.87
DCP finite support	0.796/12.82	0.805/19.43	0.799/154.16	0.794/1.34	0.798/3.68
DCP exponential tails	0.796/12.82	0.805/19.43	0.799/154.16	0.794/1.34	0.798/3.68
CIR greedy (100 cells)	0.802/10.34	0.819/15.85	0.805/121.50	0.803/1.12	0.803/1.96
CIR greedy (20 cells)	0.816/10.86	0.838/17.92	0.823/130.21	0.823/1.21	0.824/2.23

Table 28: ECP-FQ+ at 90% nominal coverage: coverage and paired percentage width change from pooled SCP, with 95% random-split intervals.

Dataset	Learner	Coverage [CI]	Width change [CI]
CASP	HistGBM	0.894 [0.888, 0.901]	-10.14 [-11.49, -8.80]
CASP	RF	0.893 [0.886, 0.899]	-11.57 [-13.13, -10.00]
Concrete	HistGBM	0.913 [0.902, 0.924]	1.19 [-3.15, 5.54]
Concrete	RF	0.898 [0.883, 0.912]	-1.12 [-4.29, 2.05]
Bike	HistGBM	0.898 [0.890, 0.905]	-20.09 [-21.42, -18.77]
Bike	RF	0.895 [0.888, 0.902]	-27.31 [-29.07, -25.56]
California	HistGBM	0.900 [0.892, 0.908]	-8.88 [-10.89, -6.88]
California	RF	0.903 [0.897, 0.909]	-10.83 [-12.54, -9.12]
Energy	HistGBM	0.899 [0.886, 0.912]	-9.26 [-19.49, 0.97]
Energy	RF	0.893 [0.873, 0.912]	-15.06 [-21.23, -8.90]

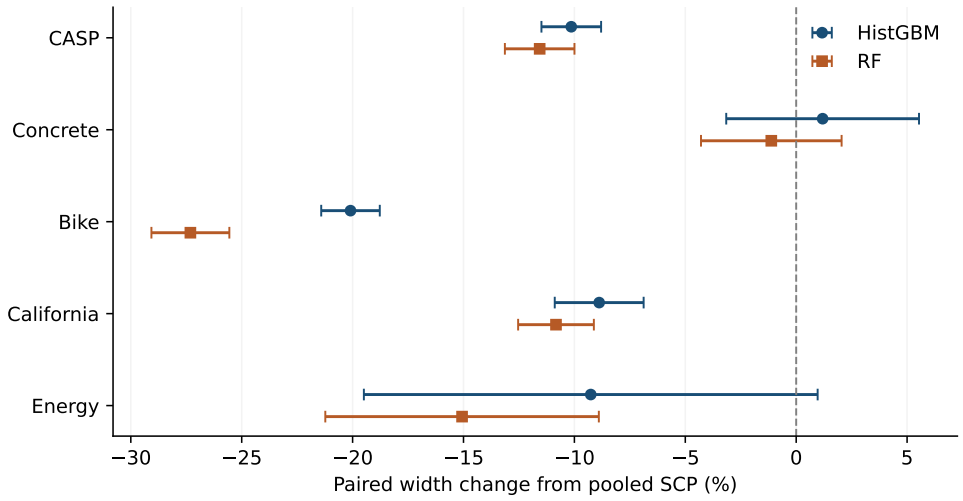

Figure 8: Paired percentage width change of ECP-FQ+ from pooled SCP at 90% nominal coverage. Error bars are 95% random-split intervals. Concrete and Energy illustrate why the uncertainty in a gain must be reported.

G CONTROLLED SCALE-PROXY AND SAMPLE-SIZE STUDY

Fitted-data results combine two effects: the quality of the predictors and the choice of interval shape. To separate them, this experiment supplies the location and scale proxy directly. We can then vary scale misspecification without also changing how a learner is trained. The study explains how each interval parameter responds to a prescribed proxy; it does not measure how accurately a scale model can be learned.

Design and data roles. Let $G \sim \text{Bernoulli}(1/2)$, $\mu(G) = 0$, and $Y = \sigma_G \varepsilon$, with $(\sigma_0, \sigma_1) = (1, r)$ and $r \in \{1, 2, 4\}$. The independent error ε is either standard Gaussian, $(V - 2)/\sqrt{2}$ for $V \sim \text{Gamma}(2, 1)$, or $(\exp Z - e^{1/2})/\sqrt{e(e-1)}$ for $Z \sim N(0, 1)$. All three errors have mean zero and variance one. The supplied scale proxy is $(1, r)$ (exact), $(1, 1)$ (flat), or $(r, 1)$ (reversed). We use $\alpha = .1$ and independent allocation and calibration samples with $n_A = m \in \{200, 1000\}$, giving 54 design cells and 100 repetitions per cell. Common random numbers are used across the heterogeneity and proxy conditions; repetitions are independent. At $r = 1$, all three proxies coincide, so these cells are identical controls rather than three independent bodies of evidence.

Procedures and evaluation. The primary FQ+ family uses the original 25-point b grid on $[-1.5, 1.5]$ and $\gamma \in \{0, .25, .5, .75, 1\}$, including the $b = 0$ and $\gamma = 1$ restrictions. We also evaluate ordinary SCP, normalized SCP, global FQ, and fixed signed equal-tail intervals. Fixed SCP, normalized SCP, and the pooled equal-tail baseline use all $n_A + m$ reserve observations; a second equal-tail baseline uses the same m calibration observations as FQ to isolate the benefit of allocation selection. Adaptive methods use allocation labels only for parameter selection and independent calibration labels for their final thresholds. All procedures call the released core functions, including their original rank and tie conventions. A sensitivity variant augments the exponent grid with $.85, .90, .95$ and otherwise uses the same procedure and observations. Let F_ε be the known distribution function of the standardized error. For each selected and calibrated interval with endpoints L_g, U_g in group g, we evaluate $\Pr\{L_g \le Y \le U_g \mid G = g\} = F_\varepsilon(U_g/\sigma_g) - F_\varepsilon(L_g/\sigma_g)$. Because the error distribution is known, coverage and average width can be evaluated exactly after the two reserve samples are observed. No finite test sample is needed for this evaluation. Uncertainty intervals quantify variation across the 100 independent repetitions. Percentage differences are averaged within-repetition ratios, $100(W_{\text{method}}/W_{\text{reference}} - 1)$, with pointwise 95% t intervals; they are not multiplicity-adjusted.

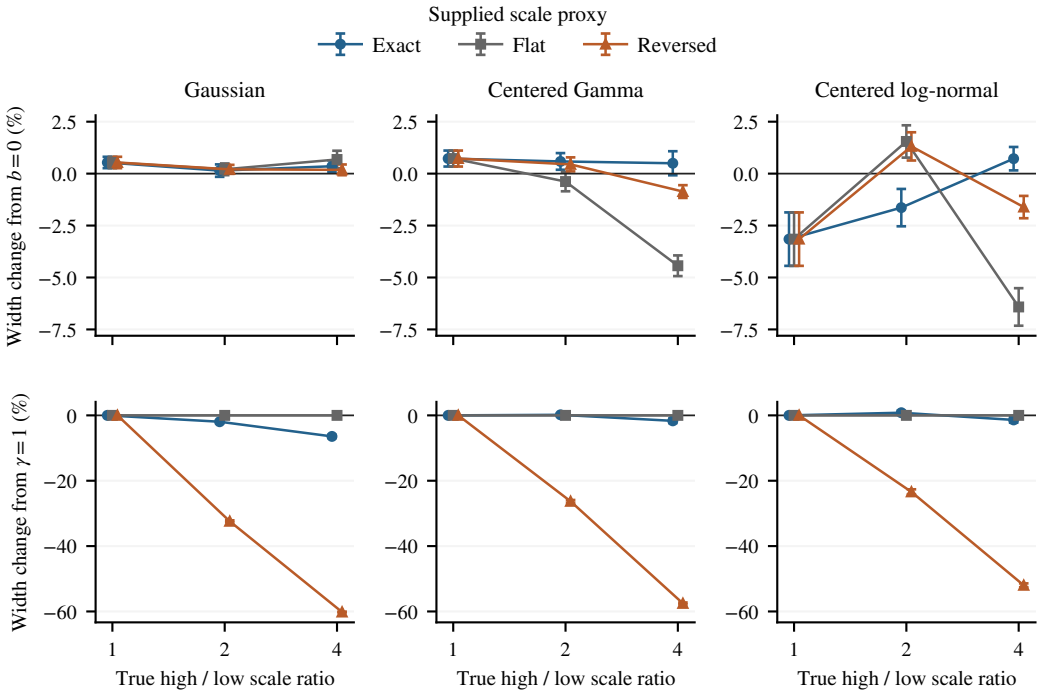

Figure 9: Incremental effects of shift and exponent selection in the controlled study, $n_A = m = 1000$. Each point is the mean paired percentage width change of full FQ+ from the indicated restriction, with pointwise 95% intervals over 100 repetitions. The top row isolates the added shift parameter; the bottom row isolates exponent selection. Negative values favor the full family. All error laws and proxy conditions are shown, including regimes where added flexibility increases width. The large reductions for reversed proxies compare with a deliberately misspecified fully normalized rule.

Allocation and shift address different sources of asymmetry. At $n_A = m = 1000$, global FQ shortens intervals relative to matched fixed signed tails by 7.9–11.7% for Gamma errors and 24.1–28.4% for log-normal errors across the three heterogeneity levels. For Gaussian errors it instead increases width by 0.8–1.1%. Therefore allocation selection has value when the residual distribution calls for unequal tail probabilities, but can incur an estimation cost when equal tails are appropriate. These comparisons complement the fitted examples, where fixed signed endpoints capture nearly all of the observed gain; they do not establish a symmetry property for those fitted residuals. For the exact scale proxy, learning b improves on the $b = 0$ restriction for log-normal errors at $r = 1$ by 3.15% (95% interval 1.87–4.44% reduction) and at $r = 2$ by 1.64% (0.74–2.54% reduction). At $r = 4$, however, the full procedure is 0.72% wider (0.16–1.28%). At $n_A = m = 200$, the mean incremental width change from learning b is positive in every exact-proxy cell. The shift parameter is consequently useful in some regimes, and the simpler restriction can be more efficient when selection noise dominates.

Scale information and the coverage tradeoff. For Gaussian errors and $r = 4$, full FQ+ with the exact proxy is 6.44% shorter than its $\gamma = 1$ restriction (95% interval 5.95–6.92% reduction). Its high-scale coverage is .837, compared with .899 for the restriction, while marginal coverage is .898 versus .899. This is the coverage redistribution predicted by the population analysis. A reversed proxy causes full FQ+ to select $\gamma = 0$ in all repetitions of the $r = 4$, $n_A = m = 1000$ cells, avoiding the severe width inflation caused by imposing $\gamma = 1$. It does not recover the performance obtained with a correct proxy: for example, Gaussian mean width is 10.30 with the reversed proxy and 7.72 with the exact proxy. Thus the large reductions under reversal show that selection can avoid harmful normalization. They do not mean that it recovers the true scale.

Table 29: Controlled scale-proxy study at scale ratio four and $n_A = m = 1000$ (100 independent repetitions). Width entries are paired percentage changes of FQ+ relative to each restriction, with pointwise 95% t intervals. The last column is exact high-scale coverage averaged across fitted intervals. Nuisance functions are supplied, not learned. Negative width changes favor FQ+.

Error law	Scale proxy	vs. $b = 0$ (%)	vs. $\gamma = 1$ (%)	High-scale cov.
Gaussian	Exact	+0.4 [+0.1, +0.6]	-6.4 [-6.9, -6.0]	0.837
Gaussian	Reversed	+0.2 [-0.1, +0.4]	-60.2 [-60.3, -60.1]	0.800
Gamma	Exact	+0.5 [-0.1, +1.1]	-1.7 [-2.3, -1.0]	0.868
Gamma	Reversed	-0.9 [-1.2, -0.6]	-57.5 [-57.8, -57.3]	0.811
Log-normal	Exact	+0.7 [+0.2, +1.3]	-1.4 [-2.2, -0.5]	0.881
Log-normal	Reversed	-1.6 [-2.1, -1.1]	-52.0 [-52.6, -51.4]	0.841

Table 30: Incremental real-data comparisons at $1-\alpha = .9$. Entries are mean paired percentage width changes (negative is shorter), with 95% Student-t intervals across 20 random splits. These quantify split variability conditional on each dataset, not independent new-dataset uncertainty; intervals are unadjusted. Normalized SCP uses $D_A \cup D_C$; the $b = 0$ restriction uses the same selection/calibration split as the full method.

Dataset	Learner	Coverage	vs norm. SCP	vs $b = 0$	$b = 0$ freq.
Bike	HistGBM	0.898	$-1.2\,[-2.2, -0.1]$	$+0.0\,[+0.0, +0.0]$	100%
Bike	RF	0.895	$-6.5\,[-8.2, -4.8]$	$+0.0\,[+0.0, +0.0]$	100%
CASP	HistGBM	0.894	$+0.2\,[-1.2, +1.6]$	$+1.2\,[+0.1, +2.2]$	40%
CASP	RF	0.893	$-0.6\,[-2.1, +0.8]$	$-0.0\,[-1.1, +1.0]$	45%
California	HistGBM	0.900	$-1.0\,[-2.9, +0.9]$	$+0.1\,[-0.1, +0.4]$	95%
California	RF	0.903	$-0.6\,[-2.4, +1.1]$	$+0.2\,[-0.3, +0.8]$	95%
Concrete	HistGBM	0.913	$-6.4\,[-10.7, -2.1]$	$+1.3\,[-0.6, +3.2]$	90%
Concrete	RF	0.898	$-6.3\,[-10.4, -2.3]$	$+2.3\,[-0.5, +5.0]$	70%
Energy	HistGBM	0.899	$-17.8\,[-23.5, -12.1]$	$+0.0\,[+0.0, +0.0]$	100%
Energy	RF	0.893	$-13.5\,[-19.5, -7.6]$	$+0.0\,[+0.0, +0.0]$	100%

Grid sensitivity. Adding three candidates near one does not uniformly improve finite-sample performance. With the exact proxy and $n_A = m = 1000$, mean changes relative to the original grid range from a 0.53% reduction to a 0.16% increase. The clearest reductions occur at $r = 2$ for Gamma (0.53%, interval 0.19–0.86%) and log-normal errors (0.42%, interval 0.13–0.71%). Results for flat and reversed proxies are identical for the two grids. A finer grid therefore serves two competing roles: it offers more useful exponents, but also adds choices that must be distinguished from finite data.

H PAIRED COMPONENT COMPARISONS AND COMPUTATION

The overall comparison does not by itself show which adjustment helps. We therefore compare methods within the same dataset, learner, nominal level, and repetition before averaging their percentage width changes. This separates gains from tail allocation, center shift, and scale response. The script `code/analyze_revision_evidence.py` uses the supplied run-level results and reproduces their archived summary means exactly, without refitting the benchmarks. Uncertainty is measured across repetitions. For real data, the pointwise intervals describe random-split variation on the fixed observed dataset and are not adjusted for multiple comparisons.

Selection frequencies. Table 31 reports how often the full search chooses zero shift, how often it chooses an intermediate exponent, and the mean selected exponent. All 90% Bike and Energy runs choose zero shift. Their improvement over normalized SCP therefore comes from changing scale response. However, this pattern depends on the nominal level: Bike/RF chooses a nonzero shift in one of twenty 95% runs. The supplied CSV files retain the full parameter distributions at every level.

Table 31: Selected full-family parameters at 90% nominal coverage, 20 random splits.

Dataset	Learner	$b = 0$ fraction	$0 < \gamma < 1$ fraction	Mean γ
Bike	HistGBM	1.00	0.65	0.825
Bike	RF	1.00	1.00	0.738
CASP	HistGBM	0.40	0.55	0.863
CASP	RF	0.45	0.55	0.838
California	HistGBM	0.95	0.80	0.750
California	RF	0.95	0.80	0.738
Concrete	HistGBM	0.90	0.95	0.338
Concrete	RF	0.70	1.00	0.375
Energy	HistGBM	1.00	0.95	0.475
Energy	RF	1.00	1.00	0.562

Table 32: Archived component timing medians. Fitting is in seconds; selection plus calibration is in milliseconds. These are not full end-to-end runtimes.

Dataset	Learner	Point/scale fit (s)	Full search/cal. (ms)	$b = 0$ search/cal. (ms)
Bike	HistGBM	0.460	2.491	0.119
Bike	RF	1.768	2.541	0.115
CASP	HistGBM	0.465	2.678	0.151
CASP	RF	3.229	2.668	0.155
California	HistGBM	0.467	2.079	0.122
California	RF	2.832	2.142	0.125
Concrete	HistGBM	0.306	1.599	0.089
Concrete	RF	0.491	1.348	0.073
Energy	HistGBM	0.182	0.947	0.055
Energy	RF	0.309	0.944	0.053

Fitted residuals and fixed signed tails. In the original mixture simulations, three mean paired changes favor FQ over matched equal signed tails, but every reduction is below .15% and every interval includes zero. The other seven changes are positive. Appendix G shows controlled settings where tail allocation produces substantial gains. Together, these comparisons show that an opportunity created by asymmetry does not always survive fitting and selection. Small differences among candidate allocations and estimation error can limit the gain even when fitted residuals are asymmetric.

Subgroup interpretation. The right panel of Figure 3 first averages coverage within each of the two groups and then takes the smaller average. This differs from averaging the minimum within each repetition. The panel includes all six central distributions and both restrictions. Per-group coverage, tail errors, and paired widths remain available in `results/revision_diagnostic_group_tradeoffs.csv` and the original diagnostic results. These comparisons show how a marginal objective affects groups when no subgroup constraint is imposed.

Computation. Table 32 separates the recorded components of computation at 90% coverage. Point-and-scale fitting includes three out-of-fold point fits, a final point fit, and the log-error scale learner. The shared prediction timer also includes quantile-model predictions, even for methods that do not use them. It therefore cannot measure standalone FQ+ prediction time. Likewise, shared quantile fitting is not part of fitting FQ+. The logs support the low cost of the selection step, but do not identify complete end-to-end latency for each method. The diagnostic restrictions share the full search and have no separate timing measurement.

Reproduction check. The additional script `code/verify_reproduction.py` reloads the original Concrete data, reconstructs split zero using seed 31000, refits the shared quantile learners and both point/scale learners, and evaluates all fourteen original methods at 90% nominal coverage.

All 280 metric and parameter comparisons agree with the archived values at absolute and relative tolerances 10^{-9}; the maximum absolute difference is 3.55×10^{-15}. This check verifies that the recorded execution of one split can be reproduced. It does not independently reimplement the methods or rerun every benchmark. Hashes, environment information, and individual comparisons are stored in `results/revision_reproduction_check.json`.

I Optimized distributional and endpoint comparisons

The original quantile baselines do not explicitly choose the shortest fitted interval. We therefore add two comparisons that do. Both reuse all 100 supplied real-data quantile caches: five datasets, twenty splits each, and $\alpha \in \{.05, .1, .2\}$. Fitted quantiles, response standardization, and imputation use only D_F. The same quantile predictions are compared with both point learners, so they are not separate quantile fits. Both added methods can use the entire reserve $D_A \cup D_C$ for calibration.

Fitted shortest-interval allocation. The thirteen fitted quantiles from Appendix B are rearranged, linearly interpolated, and supplied with the same full-support exponential tails as the original DCP comparison. Write $\widehat{Q}(u \mid x)$ for the resulting fitted quantile at probability u. For each input x, choose a lower-tail probability $b_\alpha(x)$ and its probability-space midpoint $c_\alpha(x)$ by computing

$$b_\alpha(x) \in \operatorname*{arg\,min}_{0 \le u \le \alpha} \{\widehat{Q}(p + u \mid x) - \widehat{Q}(u \mid x)\}, \qquad c_\alpha(x) = b_\alpha(x) + p/2. \tag{43}$$

The fitted quantile function allows a direct global search. The code checks its piecewise-linear knots and every analytic stationary point in the exponential-tail segments. Infinite endpoint quantiles put the minimizer in the interior. Ties within relative tolerance 10^{-12} favor the allocation nearest $\alpha/2$. The implementation supports $0 < \alpha \le .5$, including every level used here.

Optimal split DCP. Following Algorithm S1 and equations (14), (16) of Chernozhukov et al. (2021), define

$$S_{\mathrm{ODCP}}(x, y) = |\widehat{F}(y \mid x) - c_\alpha(x)|. \tag{44}$$

Here $\widehat{F}(y \mid x)$ is the fitted conditional distribution function. Holding it and the nominal level fixed, increasing the score threshold produces nested prediction sets. Let q be the corrected pooled calibration quantile. Inversion gives

$$[\widehat{Q}\{c_\alpha(x) - q \mid x\}, \widehat{Q}\{c_\alpha(x) + q \mid x\}],$$

with $-\infty$ or $+\infty$ when a probability argument reaches the corresponding boundary. These endpoints are retained, including in average widths. The procedure implements the published score using the supplied histogram-boosting quantile model; it does not reproduce the original authors' entire empirical pipeline. However, intervals optimized separately at different nominal levels need not be nested.

Shortest-interval CQR. The second comparator sets $\widehat{L}(x) = \widehat{Q}\{b_\alpha(x) \mid x\}$ and $\widehat{U}(x) = \widehat{Q}\{p + b_\alpha(x) \mid x\}$, then applies ordinary response-space CQR calibration:

$$S_{\mathrm{SI-CQR}}(x, y) = \max\{\widehat{L}(x) - y, y - \widehat{U}(x)\}, \qquad C(x) = [\widehat{L}(x) - q, \widehat{U}(x) + q].$$

Here q is the corrected pooled calibration quantile of $S_{\mathrm{SI-CQR}}$. This comparator combines the shortest endpoints under the common fitted learner with the standard CQR correction (Romano et al., 2019). It is distinct from optimal DCP, CPL, and Ad-EffOrt. Negative correction can produce an empty interval, assigned width zero. Ten of 237,660 evaluated test intervals are empty and none is infinite. For empty intervals the two endpoint-error indicators can overlap, so coverage and empty-set frequency are the authoritative set metrics.

Validity and verification. Both methods preserve marginal coverage for the same reason as SCP. Conditioning on D_F fixes their endpoints, CDFs, and allocation functions. The reserve and future test scores are then exchangeable, so the corrected calibration rank gives the guarantee even if the fitted distribution is imperfect. Numerical checks include 64,000 score/inversion membership comparisons, CDF/quantile round trips, comparison of the global optimizer with a

Table 33: Additional common-quantile-model baselines, twenty splits per dataset. Widths use original response units. The finite fraction is the fraction of test predictions with finite width, averaged over splits; one infinite width makes the all-split mean infinite.

Nominal	Dataset	DCP cov.	DCP finite	DCP width	SI-CQR cov.	SI-CQR width
95%	CASP	0.950	0.489	∞	0.951	16.01
95%	Concrete	0.953	0.905	∞	0.949	27.62
95%	Bike	0.948	0.997	∞	0.948	243.39
95%	California	0.951	0.920	∞	0.950	2.09
95%	Energy	0.940	1.000	6.06	0.948	8.41
90%	CASP	0.894	0.586	∞	0.894	13.27
90%	Concrete	0.911	0.807	∞	0.906	19.99
90%	Bike	0.899	0.999	∞	0.897	184.14
90%	California	0.897	0.922	∞	0.897	1.54
90%	Energy	0.900	1.000	3.86	0.892	4.71
80%	CASP	0.798	0.802	∞	0.798	9.86
80%	Concrete	0.810	0.848	∞	0.808	13.29
80%	Bike	0.797	1.000	123.71	0.798	127.52
80%	California	0.797	0.913	∞	0.796	1.07
80%	Energy	0.792	1.000	1.99	0.794	2.22

Table 34: FQ+ versus shortest-interval CQR at 90% nominal coverage. Changes are means of within-split percentage width changes, with unadjusted 95% t intervals over twenty paired splits. Negative values favor FQ+.

Dataset	Point learner	FQ+ coverage	SI-CQR coverage	Width change (%)
Bike	HistGBM	0.898	0.897	-31.30 [-32.70, -29.89]
Bike	RF	0.895	0.897	-23.83 [-25.67, -22.00]
CASP	HistGBM	0.894	0.894	+3.43 [+1.83, +5.03]
CASP	RF	0.893	0.894	+0.64 [-1.46, +2.73]
California	HistGBM	0.900	0.897	-7.38 [-9.68, -5.09]
California	RF	0.903	0.897	+2.93 [+1.50, +4.37]
Concrete	HistGBM	0.913	0.906	-13.98 [-18.87, -9.09]
Concrete	RF	0.898	0.906	+1.65 [-2.81, +6.10]
Energy	HistGBM	0.899	0.892	-56.28 [-60.35, -52.22]
Energy	RF	0.893	0.892	-46.83 [-51.26, -42.41]

2,001-point grid on 160 synthetic fitted-quantile rows, and the infinite-rank sentinel. An independent simulation of 4,000 misspecified-CDF calibration/test samples gives coverage .898 at nominal .9. Every real-data prediction is additionally checked against its score sublevel set. The files `code/run_optimized_dcp.py`, `code/run_shortest_interval_cqr.py`, and `code/OPTIMIZED_BASELINES.md` provide complete formulas and commands.

Results and interpretation. Learning shorter fitted endpoints improves this baseline: at 90% nominal coverage, shortest-interval CQR is narrower than the original CQR comparator on all five datasets, with mean coverage between .892 and .906. Table 34 compares FQ+ directly with this stronger endpoint baseline. FQ+ is wider on CASP/HistGBM and California/RF, and the Concrete/RF difference is uncertain. Its advantages on other datasets concern this fixed quantile learner, not every possible implementation of quantile regression.

The optimal-DCP comparison also reveals sensitivity to the fitted tails. Its interval has an infinite endpoint when the calibrated probability range reaches zero or one. They can occur even though $b_\alpha(x)$ is interior: at 90% coverage, the mean across CASP splits of the minimum test allocation is about 10^{-5}, and the mean score-quantile correction above $p/2$ is .0087. Very small fitted outer quantile gaps can yield narrow exponential-tail scales. Thus this result diagnoses sensitivity of the evaluated distribution model and probability-space correction; it is not a universal failure of optimal

DCP. Per-split files retain allocation quantiles, correction sizes, endpoint infinity frequencies, and finite-width fractions. In particular, Bike has a finite fraction of .9987 but one of twenty split means is infinite at 90% coverage, so its overall width is correctly reported as infinite.

These additions test shortest fitted intervals and the published optimized-DCP score under a common learner. However, they do not exhaust tuning choices or length-optimization methods; CPL and AdEffOrt were not evaluated.

J REAL-DATA COVERAGE ACROSS FITTED-SCALE GROUPS

The theory describes true conditional scale, but a fitted scale can be inaccurate. To see how this distinction affects practice, we audit coverage across fitted-scale groups on the original five datasets, twenty random splits, both point learners, and 90% nominal coverage. For all 200 dataset/split/learner combinations, we refit the point and scale models from the original data, saved split indices, and seeds. We do not refit conditional quantile models because the audited procedures do not use them.

Groups and data separation. We define difficulty groups before evaluating coverage. Let t_1, t_2 be the one-third and two-thirds empirical quantiles of $\widehat{\sigma}(X)$ on D_F, using linear interpolation. The low, middle, and high groups are $\widehat{\sigma}(x) \le t_1$, $t_1 < \widehat{\sigma}(x) \le t_2$, and $\widehat{\sigma}(x) > t_2$. Ties stay together. The fitting sample alone determines both the predictor and these cutoffs; allocation, calibration, and test responses do not enter the grouping. The groups serve only to evaluate the intervals and do not change selection or calibration. Their test sizes can differ. Empty groups would retain zero counts and undefined metrics, with the number of contributing splits disclosed; none occurred here. These groups describe a fitted proxy, rather than known noise levels or demographic categories.

Metrics and uncertainty. For full FQ+, its two restrictions, pooled normalized SCP, and pooled ordinary SCP, we save each group's size, coverage, and mean width in every split. This produces 3,000 subgroup records. Summaries give equal weight to each split and use unadjusted 95% Student-t intervals across its twenty values. Coverage differences and percentage width changes are paired within splits. The CSV files retain all three groups, overall and groupwise widths, and sample counts. Because test observations can recur across splits, we do not pool them as independent Bernoulli trials. Table 35 reports overall width change beside high-group coverage, keeping the two targets distinct.

Table 35: Real-data coverage in the highest fitted-scale group at $1 - \alpha = .9$. Tercile cutoffs are computed from fitted-scale predictions on D_F only. Coverage entries are means across 20 random splits; brackets give unadjusted 95% Student-t intervals. ΔW is the mean paired percentage change in *overall* width of full ECP-FQ+ relative to pooled normalized SCP (negative is shorter). These groups describe the fitted proxy, not known true noise levels.

Dataset	Learner	Norm. SCP cov.	Full cov.	$b = 0$ cov.	ΔW (%)
Bike	HistGBM	0.871 [0.862,0.880]	0.841 [0.827,0.855]	0.841	−1.2
Bike	RF	0.881 [0.868,0.894]	0.817 [0.804,0.831]	0.817	−6.5
CASP	HistGBM	0.932 [0.922,0.942]	0.918 [0.907,0.929]	0.914	+0.2
CASP	RF	0.929 [0.920,0.938]	0.913 [0.900,0.926]	0.914	−0.6
California	HistGBM	0.890 [0.882,0.898]	0.859 [0.846,0.872]	0.861	−1.0
California	RF	0.885 [0.873,0.897]	0.859 [0.845,0.873]	0.857	−0.6
Concrete	HistGBM	0.959 [0.943,0.974]	0.891 [0.866,0.915]	0.887	−6.4
Concrete	RF	0.944 [0.926,0.961]	0.860 [0.831,0.888]	0.867	−6.3
Energy	HistGBM	0.942 [0.920,0.964]	0.876 [0.850,0.903]	0.876	−17.8
Energy	RF	0.938 [0.919,0.957]	0.837 [0.802,0.872]	0.837	−13.5

Results and limits. The shorter marginal intervals come with lower coverage in the high fitted-scale group: full FQ+ has lower mean coverage there than pooled normalized SCP in every dataset/learner comparison. The mean paired decrease ranges from approximately 1.4 to 10.1 percentage points. Bike/RF gives .817 versus .881, Concrete/HistGBM gives .891 versus .959, and

Energy/RF gives .837 versus .938. In CASP the high-group coverage remains above .9, while lower fitted-scale groups have lower coverage; a fitted scale ordering need not coincide with the ordering of calibration difficulty. Moreover, pooled normalized SCP itself does not ensure .9 coverage in every group. Thus marginal calibration does not ensure equal reliability across these fitted groups. Neither procedure has a subgroup guarantee in this audit.

Reproduction and implementation. The script `code/run_real_subgroup_audit.py` loads the supplied raw data, checks their hashes and subset indices, and reconstructs all five audited methods through the original core implementation. All 5,000 comparisons of full-sample coverage, width, and selected parameters against archived results pass at absolute and relative tolerances 10^{-9}; the maximum absolute difference is 5.68×10^{-14}. This reproduces all point/scale fits required for the audit. It does not refit the quantile comparators or independently reimplement the methods. Two workers with one numerical thread per fit took approximately 90 seconds in the recorded environment. Configuration, software versions, code hashes, per-split records, paired differences, and verification output are included in `results/real_subgroup_*`.

K Additional real-data controls

K.1 Center resolution and the cost of learning extra parameters

The original grid often selects no center shift, $b = 0$. However, this can have two explanations: changing the center may offer little benefit, or the grid may be too coarse to detect a useful small change. The fitted scale has median approximately one on D_F, while residual size depends on prediction accuracy. Consequently, the original spacing 0.125 can be large relative to the errors on an accurately predicted task. We need a finer, error-based comparison to distinguish these explanations.

We address this issue with one additional grid defined entirely from the fitting sample. Let Y_i^{std} denote the response standardized using D_F, and let $\widehat{\mu}_{-k(i)}(X_i)$ be the existing three-fold out-of-fold point prediction within D_F. Define

$$r_F = \frac{\mathrm{median}_{i \in D_F} |Y_i^{\mathrm{std}} - \widehat{\mu}_{-k(i)}(X_i)|}{\mathrm{median}_{i \in D_F} \widehat{\sigma}(X_i)}, \qquad \mathcal{B}_F = r_F\{-1.5, -1.375, \ldots, 1.5\}. \tag{45}$$

Here r_F compares a typical out-of-fold prediction error with a typical scale prediction. The implementation floors it at 10^{-8} for stability. Multiplying the original shift coordinates by this ratio retains 25 center candidates and five exponents, but adjusts the resolution to the fitting errors. Only D_F determines this grid, so the independent-calibration guarantee still applies. The two grids need not be nested, and finer resolution need not improve finite-sample performance.

We evaluate both grids on all 200 original dataset–learner splits at each of the three coverage levels, refitting the original point and scale learners. All 21,600 checks of the historical methods agree with their recorded values, with maximum absolute difference 8.53×10^{-14}. The original core code and archived results remain unchanged. The added grids were specified before their test outcomes were computed.

At 90% nominal coverage, the residual-unit grid selects a nonzero center on 11–19 of the 20 splits in every dataset–learner setting. In particular, the counts are 17 and 18 on Bike and 14 and 19 on Energy for HistGBM and RF, respectively, although the original grid selected zero on every Bike and Energy split. Thus the original zero selections cannot by themselves establish that smaller center adjustments are unnecessary. Nevertheless, better center resolution does not establish an efficiency advantage over scale-only selection: none of the ten 90% paired intervals shows a clear width reduction relative to $b = 0$, and the refined full method is wider on Concrete/RF by 2.7% (interval $[0.4, 5.0]\%$). Energy/RF has the largest favorable mean change, -1.5%, but its interval $[-4.4, 1.3]\%$ includes zero. These results support using the five-candidate scale-only method as a parsimonious default for the studied fitted predictors, with joint center selection retained as an optional extension. They do not imply that center adjustment is unhelpful for every residual distribution.

The narrower marginal intervals continue to redistribute errors across fitted scale groups. For example, the refined method's upper-tercile coverage is 0.821 on Bike/RF and 0.836 on Energy/RF, compared with 0.881 and 0.938 for pooled normalized SCP. The overall coverage means are 0.894

Table 36: Fitting-only residual-unit center-grid audit at 90% marginal coverage. ΔW gives the mean paired percentage width change of residual-grid FQ+ relative to each indicated reference (negative is shorter), with unadjusted 95% Student-t intervals over 20 splits. $\widehat{C}$ is overall coverage. Both full searches use 25 center candidates and five exponents; the residual grid scales centers using D_F only.

Dataset	Learner	$\widehat{C}$	ΔW: old grid (%)	ΔW: $b = 0$ (%)
Bike	HistGBM	0.900	+0.5 [-0.3,+1.3]	+0.5 [-0.3,+1.3]
Bike	RF	0.894	+0.3 [-0.4,+1.0]	+0.3 [-0.4,+1.0]
CASP	HistGBM	0.896	-0.5 [-1.2,+0.2]	+0.6 [-0.3,+1.6]
CASP	RF	0.894	+0.8 [+0.2,+1.4]	+0.8 [-0.2,+1.8]
California	HistGBM	0.901	+0.3 [-0.6,+1.2]	+0.4 [-0.3,+1.2]
California	RF	0.902	+0.2 [-0.8,+1.2]	+0.4 [-0.7,+1.5]
Concrete	HistGBM	0.912	+0.0 [-2.7,+2.7]	+1.3 [-1.8,+4.4]
Concrete	RF	0.904	+0.6 [-1.9,+3.1]	+2.7 [+0.4,+5.0]
Energy	HistGBM	0.899	+1.8 [-1.0,+4.7]	+1.8 [-1.0,+4.7]
Energy	RF	0.888	-1.5 [-4.4,+1.3]	-1.5 [-4.4,+1.3]

Table 37: Pooled signed-tail control and the coverage cost of center-grid refinement at 90% nominal coverage. Width change compares honest global FQ against fixed equal signed tails calibrated on $D_A \cup D_C$. High-group coverage uses the upper fitted-scale tercile with cutoffs fixed on D_F; these are proxy-defined groups, not true noise strata.

Dataset	Learner	FQ ΔW vs pooled tails (%)	Norm. SCP high cov.	Refined FQ+ high cov.
Bike	HistGBM	+1.9 [+0.3,+3.6]	0.871	0.843
Bike	RF	+0.1 [-1.8,+2.0]	0.881	0.821
CASP	HistGBM	-0.7 [-2.1,+0.7]	0.932	0.920
CASP	RF	+0.4 [-1.3,+2.0]	0.929	0.913
California	HistGBM	+0.3 [-2.3,+2.8]	0.890	0.862
California	RF	+0.1 [-1.7,+2.0]	0.885	0.857
Concrete	HistGBM	+10.0 [-2.6,+22.5]	0.959	0.891
Concrete	RF	+2.8 [-0.5,+6.2]	0.944	0.872
Energy	HistGBM	+13.6 [+0.4,+26.7]	0.942	0.871
Energy	RF	+12.6 [-5.3,+30.4]	0.938	0.836

and 0.888, respectively; finite-split empirical coverage is not the same quantity as the exchangeability guarantee averaged over calibration and a new observation. The group cutoffs are fixed from D_F and describe the fitted scale proxy, not the unknown true conditional noise. Center refinement therefore does not remove the coverage-allocation tradeoff explained by the scale-sensitivity analysis.

A pooled signed-tail control. The original fixed-tail comparison uses the same final-calibration sample as global FQ and isolates the effect of tail allocation. A fixed-tail procedure can also use the complete reserve $D_A \cup D_C$, since it performs no selection. We therefore add this pooled control while keeping the total labeled-data budget fixed. At 90% coverage, none of the ten paired comparisons shows a clear width advantage for global FQ over pooled equal signed tails. The mean width increases are 1.9% on Bike/HistGBM and 13.6% on Energy/HistGBM, with intervals $[0.3, 3.6]\%$ and $[0.4, 26.7]\%$; the other eight intervals include zero. This reinforces the distinction between an oracle opportunity to reallocate asymmetric tails and a finite-data benefit after paying for allocation selection. All intervals here are unadjusted repeated-split summaries conditional on the observed dataset.

L A MECHANISM STUDY WITH FITTED LOCATION AND SCALE

The controlled-proxy experiment explains how interval shape responds to a supplied scale. In practice, however, that scale must also be learned. This study connects the two settings by fitting both

location and scale in every repetition. We vary the error distribution, true scale ratio, fitting-sample size, and reserve-sample size separately. The protocol was fixed before computing its results, and the released files retain every cell.

Let $G = X_0 \sim \text{Bernoulli}(1/2)$ and let $X_1, \ldots, X_5$ be independent $\text{Uniform}[-1, 1]$ variables, independent of G. We generate

$$Y = 2\sin(\pi X_1) + X_2^2 + 0.5X_3 + 0.5G + \{1 + (r-1)G\}\varepsilon, \qquad r \in \{1, 2, 4\}.$$

The errors are either standard Gaussian or $\{\exp(Z) - \exp(1/2)\}/\sqrt{e(e-1)}$, with $Z \sim N(0,1)$; both have mean zero and variance one and are independent of the covariates. We use $n_F \in \{300, 1000\}$ and $n_A = m \in \{300, 1000\}$, giving 24 cells. Each cell has 50 independent repetitions and 5000 independent test observations per repetition. Common random numbers couple the scale-ratio conditions within each otherwise identical repetition. They do not couple different repetitions.

Both predictors use the benchmark fitting procedure. A HistGBM regressor estimates location. Its three-fold out-of-fold absolute residuals on D_F train the log-error regressor, whose exponentiated predictions are divided by their fitting-sample median and clipped to $[0.1, 10]$. This produces a learned scale proxy. The final location model then uses all of D_F. Neither fit uses allocation, calibration, or test labels, and this study fits no quantile models.

We compare pooled ordinary SCP, pooled normalized SCP, pooled and calibration-matched fixed signed tails, global FQ, scale-only FQ+, the original FQ+ grid, the fitting-residual grid, and its $\gamma = 1$ restriction. Fixed pooled methods use all $n_A + m$ reserve labels for calibration. Adaptive methods select on D_A and calibrate only on D_C. The original shift grid is $\{-1.5, -1.375, \ldots, 1.5\}$ in standardized response units. The additional grid multiplies these coordinates by

$$r_F = \frac{\text{median}_{i \in D_F} |\widetilde{Y}_i - \widehat{\mu}^{(-i)}(X_i)|}{\text{median}_{i \in D_F} \widehat{\sigma}(X_i)},$$

where $\widetilde{Y}$ uses the fitting-sample response standardization and $\widehat{\mu}^{(-i)}$ denotes the corresponding out-of-fold prediction. Consequently, the scale of the shift grid is fixed using D_F alone. All three exponent searches use $\{0, 0.25, 0.5, 0.75, 1\}$.

We compare methods by their percentage width change within each repetition, then average these paired changes and form pointwise 95% Student-t intervals over 50 repetitions. There is no multiplicity adjustment. Coverage is evaluated overall and in the known group $G = 1$; at $r = 1$, both groups have the same noise variance. The saved files also evaluate the known error CDF at each test input, reducing variation from sampling new responses. These CDF calculations are used only for evaluation.

To understand the fitted proxy, we report its correlation with true log-scale, its root-mean-square log-scale shape error after removing a common multiplier, and its high/low scale ratio. These diagnostics describe estimation quality; they neither tune the method nor establish the uniform predictor-error conditions assumed by the theory.

The fitting-residual version reduces mean paired width from pooled normalized SCP by 3.2–16.5% across the 24 cells, with every pointwise interval below zero; mean marginal coverage ranges from .896 to .903. These reductions have two interpretations. Under constant true scale, they primarily remove unnecessary variation in the learned proxy: for Gaussian errors and $n_A = m = 1000$, the mean selected exponent is .03 with $n_F = 300$ and .10 with $n_F = 1000$. Under genuine heterogeneity, the exponent also controls the allocation of coverage across the two groups. At ratio four, the Gaussian log-scale correlation increases from .771 to .916 as n_F increases, while the mean selected exponent increases from .485 to .685. Thus a better fitted scale permits a stronger scale response in this experiment. These fitted-data patterns complement the oracle derivative calculation; that calculation alone does not explain improvements under constant true scale.

The additional center search remains secondary. Relative to scale-only FQ+, its pointwise interval excludes zero in favor of the full method in one cell and in favor of the restriction in six cells. The favorable cell has log-normal errors, ratio four, and $n_F = n_A = m = 1000$, with a width change of -0.79% and interval $[-1.46, -0.11]\%$. These are unadjusted comparisons, so the isolated favorable result should not be read as general evidence for center learning. The new grid also has

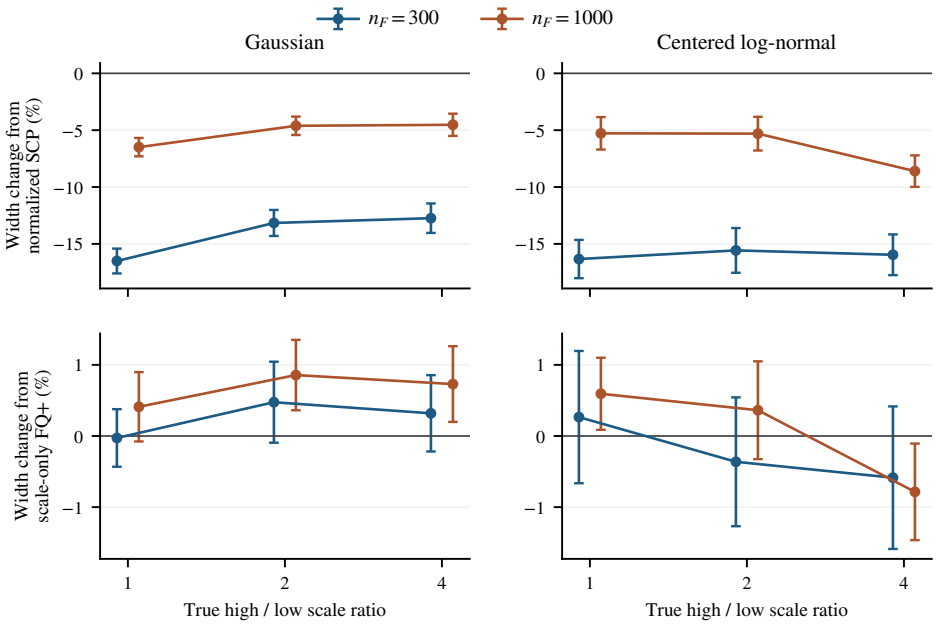

Figure 10: **Efficiency after fitting location and scale.** All 12 settings with $n_A = m = 1000$, two error laws, three scale ratios, and $n_F = 300$ or 1000. Paired width changes compare the residual-grid joint family with pooled normalized SCP (top) and scale-only (bottom). Bars are pointwise 95% t intervals over 50 repetitions. Negative values favor the joint family.

Table 38: Fitted-scale mechanism study at $n_F = n_A = m = 1000$ (50 independent repetitions). FQ+ uses the fitting-residual grid. Width changes are paired percentages with pointwise 95% t intervals. Coverage is measured on 5000 independent test observations per repetition; the final column concerns group $G = 1$. All 24 cells are supplied in the CSV results.

Error	Ratio	vs. normalized SCP	vs. $b = 0$	Cov.	High cov.
Gaussian	1	-6.5 [-7.3, -5.7]	+0.4 [-0.1, +0.9]	0.898	0.898
Gaussian	2	-4.6 [-5.4, -3.8]	+0.9 [+0.4, +1.4]	0.900	0.841
Gaussian	4	-4.5 [-5.5, -3.5]	+0.7 [+0.2, +1.3]	0.899	0.829
Log-normal	1	-5.3 [-6.7, -3.8]	+0.6 [+0.1, +1.1]	0.902	0.899
Log-normal	2	-5.3 [-6.8, -3.8]	+0.4 [-0.3, +1.0]	0.902	0.870
Log-normal	4	-8.6 [-10.0, -7.2]	-0.8 [-1.5, -0.1]	0.902	0.856

mixed, generally small changes from the old grid. Global FQ has a separate role: at $n_A = m = 1000$, it reduces width from pooled fixed signed tails by 7.2–8.2% for log-normal errors, whereas its Gaussian changes are positive. This distinguishes tail allocation from scale selection when both nuisance functions must be learned.

The group results show the practical cost of the shorter intervals. At $n_F = n_A = m = 1000$, Gaussian group-$G = 1$ coverage is .841 at ratio two and .829 at ratio four, although marginal coverage remains near .90. The corresponding log-normal values are .870 and .856. Thus a successful reduction in average width can still leave the higher-noise group less well covered.

Figure 10 displays the larger-reserve slice; the tables below retain detailed coverage and all fitted settings.

Table 39: All fitted-scale cells with Gaussian errors. FQ+ uses the fitting-residual grid; entries are paired percentage width changes with pointwise 95% t intervals across 50 independent repetitions, followed by mean test coverage. The two reserve samples have equal size. Negative width changes favor FQ+.

n_F	$n_A = m$	r	vs. norm. SCP (%)	vs. $b = 0$ (%)	Cov.	High cov.
300	300	1	-15.6 [-17.3, -14.0]	+0.7 [-0.2, +1.6]	0.901	0.901
300	300	2	-11.9 [-13.5, -10.2]	+1.0 [+0.1, +1.8]	0.899	0.834
300	300	4	-11.1 [-13.2, -9.0]	-0.2 [-1.4, +0.9]	0.896	0.813
300	1000	1	-16.5 [-17.6, -15.4]	-0.0 [-0.4, +0.4]	0.901	0.902
300	1000	2	-13.2 [-14.3, -12.0]	+0.5 [-0.1, +1.0]	0.901	0.835
300	1000	4	-12.7 [-14.0, -11.4]	+0.3 [-0.2, +0.9]	0.899	0.819
1000	300	1	-5.6 [-6.9, -4.4]	+0.6 [-0.2, +1.4]	0.903	0.904
1000	300	2	-3.2 [-4.8, -1.7]	+1.8 [+0.5, +3.2]	0.902	0.843
1000	300	4	-4.6 [-6.1, -3.1]	+1.4 [+0.1, +2.7]	0.900	0.830
1000	1000	1	-6.5 [-7.3, -5.7]	+0.4 [-0.1, +0.9]	0.898	0.898
1000	1000	2	-4.6 [-5.4, -3.8]	+0.9 [+0.4, +1.4]	0.900	0.841
1000	1000	4	-4.5 [-5.5, -3.5]	+0.7 [+0.2, +1.3]	0.899	0.829

Table 40: All fitted-scale cells with centered log-normal errors. FQ+ uses the fitting-residual grid; entries are paired percentage width changes with pointwise 95% t intervals across 50 independent repetitions, followed by mean test coverage. The two reserve samples have equal size. Negative width changes favor FQ+.

n_F	$n_A = m$	r	vs. norm. SCP (%)	vs. $b = 0$ (%)	Cov.	High cov.
300	300	1	-15.1 [-17.4, -12.7]	+0.5 [-0.8, +1.7]	0.899	0.897
300	300	2	-14.9 [-17.1, -12.7]	+0.2 [-1.6, +2.0]	0.899	0.860
300	300	4	-15.3 [-17.2, -13.5]	+0.7 [-0.8, +2.2]	0.902	0.845
300	1000	1	-16.3 [-18.0, -14.6]	+0.3 [-0.7, +1.2]	0.900	0.897
300	1000	2	-15.6 [-17.5, -13.6]	-0.4 [-1.3, +0.5]	0.899	0.858
300	1000	4	-16.0 [-17.7, -14.2]	-0.6 [-1.6, +0.4]	0.899	0.837
1000	300	1	-4.2 [-6.2, -2.3]	-0.4 [-1.5, +0.8]	0.897	0.895
1000	300	2	-3.6 [-6.1, -1.2]	+0.8 [-0.6, +2.1]	0.897	0.862
1000	300	4	-5.2 [-7.2, -3.2]	+0.7 [-0.7, +2.2]	0.902	0.857
1000	1000	1	-5.3 [-6.7, -3.8]	+0.6 [+0.1, +1.1]	0.902	0.899
1000	1000	2	-5.3 [-6.8, -3.8]	+0.4 [-0.3, +1.0]	0.902	0.870
1000	1000	4	-8.6 [-10.0, -7.2]	-0.8 [-1.5, -0.1]	0.902	0.856

M Comparison with a published score optimizer

Implementation and matched data roles. A small parameter grid is inexpensive, but a more flexible score optimizer may find better intervals. We therefore compare with the local-score length optimizer in Boosted Conformal Prediction Intervals (BoostedCP) (Xie et al., 2024). This comparison adapts the released `gradient_boost_len` function at commit `41aba95e2672c80836ee148bf4ae488b2b8d75e1`. Source files, hashes, licenses, and the complete protocol are included. We retain alternating scale and center updates, depth-one XGBoost trees, exact tree construction, the all-ones surrogate Hessian, Harrell–Davis weights, soft-sort regularization strength one, and the original gradient smoothing constant. Over four update rounds, an independent comparison with the extracted original function gives identical center and scale predictions, with maximum absolute difference zero.

This first comparison gives both methods the same allocation budget. They start from predictors fitted on D_F. BoostedCP then learns score updates and chooses its stopping round using D_A, before calibrating the fixed score on D_C. Imputation, response standardization, and scale normalization

use D_F alone. Multiplying the shared clipped relative-scale proxy by its fitting-sample normalizer restores residual units in the standardized response without changing the initial normalized conformal interval. However, the published BoostedCP benchmark learns updates on the fitting sample and uses its own base estimators. The next appendix evaluates those training-sample data roles with shared initialization.

Learning-rate and stopping selection. The primary comparison comprises all five datasets, twenty saved splits, and the HistGBM point predictor. For every split, five-fold cross-validation entirely within D_A selects the learning rate from $\{0.01, 0.1, 1, 20\}$ and the number of boosting rounds from $\{0, \ldots, 500\}$. Every rate receives the full 500-round search. We minimize the mean held-out conformal width across folds, with deterministic ties resolved by smaller rate and then earlier round. The rate-20 setting, 500-round limit, and five-fold default match the authors' example configuration; the additional rates address instability observed in allocation-fold losses under the smaller shared allocation budget. The rate grid and balanced comparison scope were fixed before the full rate benchmark. Test performance does not enter this choice.

We additionally retain the rate-20 configuration for all five datasets, both point learners, and twenty splits. A zero-round choice returns the initial normalized score calibrated on D_C alone. It therefore differs from the pooled normalized-SCP baseline, which calibrates on $D_A \cup D_C$. On Concrete and Energy, the internal validation folds contain approximately 23–31 observations, so stopping and rate selection can vary substantially across splits.

Calibration and code adaptations. We include round zero explicitly and index positive rounds consistently, correcting an off-by-one discrepancy in the released selection wrapper. Internal validation and final calibration use the order statistic at rank $\lceil (n+1)(1-\alpha) \rceil$, including $+\infty$ when appropriate, instead of a linearly interpolated sample quantile. Final score scales are $\max\{|\widehat{\sigma}(x)|, 10^{-12}\}$; the training gradients retain the authors' original smoothing constant. These choices preserve the finite-rank convention used throughout this paper. Unrelated Torch, R, and plotting imports are omitted from the length-only adapter, and every XGBoost fit uses one CPU thread.

We save the final endpoints of every fixed rate candidate so that the comparison can be reconstructed. The rate selector reads only allocation-CV curves, so computing the unselected candidates' calibration and test metrics cannot influence the selected interval. Because the rate choice does not use those outcomes, this gives the same selected interval as choosing the rate first and calibrating afterward. Numerical diagnostics record nonfinite losses, parameter selections, boundary selections, and endpoint finiteness rather than dropping unfavorable candidates.

Timing and scope. Post-fitting time includes every internal CV path and the actual candidate refits and calibrations performed by the implementation. The tuned search time sums all four rates; shared point/scale fitting is added once when reporting total elapsed time. Data loading, cache access, and one-time JIT compilation are excluded. Timings are measured with concurrent single-thread workers and depend on implementation and hardware. An implementation that defers refitting unselected rates could reduce the measured tuned-search cost.

The comparison concerns one published local-score optimizer under the stated data allocation and finite search budget. It does not establish dominance over BoostedCP's CQR or conditional-coverage variants, other hyperparameters, or other training allocations. Paired intervals describe split variation on fixed datasets and are not multiplicity-adjusted independent-dataset confidence statements. Raw per-split outcomes, all internal loss curves, fitted-score predictions, and exact split indices are supplied for reproduction.

Results and diagnostics. Table 41 reports the primary tuned comparison. Scale-only selection is shorter on Concrete, California, and Energy under the paired split analysis, while CASP and Bike do not establish a difference. On Energy the tuned optimizer has higher realized marginal coverage (.913 versus .899); the reported width difference should be read together with this discrepancy. The comparison uses a common nominal target and finite calibration ranks, rather than matching empirical coverage after seeing test outcomes. Table 42 retains the original example-rate configuration, whose allocation CV chooses zero updates throughout. This behavior concerns that rate under the present small allocation folds and is not evidence that score optimization is intrinsically ineffective.

Table 41: Published-optimizer comparison at 90% nominal coverage: twenty matched splits per dataset, HistGBM point predictor. Entries are mean width (coverage). BoostedCP uses the published local-score length updates with rate and round selected solely by five-fold allocation-sample cross-validation. Relative changes compare scale-only FQ+ with tuned BoostedCP; brackets give unadjusted paired 95% split intervals.

Dataset	Scale only	Residual grid	BoostedCP	Width change (%)
Concrete	16.952 (0.912)	17.165 (0.912)	19.063 (0.913)	-10.5 [-14.3, -6.8]
CASP	13.578 (0.893)	13.658 (0.896)	13.607 (0.895)	-0.2 [-0.7, 0.2]
Bike	126.467 (0.898)	127.109 (0.900)	127.434 (0.896)	-0.8 [-1.5, 0.0]
California	1.428 (0.900)	1.434 (0.901)	1.453 (0.903)	-1.7 [-2.8, -0.6]
Energy	2.023 (0.899)	2.059 (0.899)	2.523 (0.913)	-15.4 [-23.4, -7.3]

Table 42: Secondary BoostedCP local-score configuration using the authors' example learning rate 20, 500-round cap and five allocation-CV folds. All ten dataset/learner settings have twenty splits. A zero-round selection returns the initial score calibrated on D_C; it is not the pooled normalized-SCP baseline.

Dataset	Learner	Width	Coverage	Nonzero rounds	Stage time (s)
Bike	HistGBM	127.172	0.898	0/20	22.31
Bike	RF	147.570	0.897	0/20	22.35
CASP	HistGBM	13.566	0.894	0/20	18.96
CASP	RF	13.290	0.892	0/20	18.70
California	HistGBM	1.454	0.903	0/20	22.88
California	RF	1.598	0.903	0/20	24.14
Concrete	HistGBM	18.977	0.911	0/20	8.26
Concrete	RF	21.730	0.905	0/20	8.27
Energy	HistGBM	2.580	0.900	0/20	10.52
Energy	RF	2.949	0.896	0/20	9.58

Across the 100 matched HistGBM fits, median post-fitting elapsed time was 0.117 ms for scale-only FQ+ and 1.806 ms for the residual-grid family, compared with 90.78 s for the complete four-rate BoostedCP search. The median shared point/scale fitting time was 0.339 s. The tuned optimizer selected zero rounds on 6/100 fits and the 500-round boundary on 1/100 fits. The author-example learning rate selected zero rounds on 200/200 fits. All 400 rate paths in the primary comparison had finite endpoints; the total number of nonfinite allocation-CV losses was 0. These timings include the actual search and candidate-refit work described above and depend on hardware and implementation; the comparison does not claim a hardware-independent speed ratio.

Table 43: Overall and high-fitted-scale coverage in the primary published-optimizer comparison. Each entry reports overall coverage / high-group coverage, averaged over twenty matched HistGBM splits at nominal 90% coverage. The high group lies above the common upper fitted-scale tercile cutoff learned on D_F; it is fixed across methods within a split. These subgroup values are diagnostics, not conditional coverage guarantees.

Dataset	Scale-only FQ+	Residual-grid FQ+	Tuned BoostedCP
Concrete	0.912 / 0.887	0.912 / 0.891	0.913 / 0.956
CASP	0.893 / 0.914	0.896 / 0.920	0.895 / 0.931
Bike	0.898 / 0.841	0.900 / 0.843	0.896 / 0.867
California	0.900 / 0.861	0.901 / 0.862	0.903 / 0.888
Energy	0.899 / 0.876	0.899 / 0.871	0.913 / 0.940

The tuned optimizer has higher mean high-tercile coverage in all five settings: for example, .956 versus .887 for scale-only selection on Concrete and .940 versus .876 on Energy. Neither method guarantees coverage in these fitted groups. These differences show why the marginal-width comparison should not be interpreted as improvement in every reliability criterion.

N SCORE OPTIMIZATION ON THE FITTING SAMPLE

The allocation-trained comparison in Appendix M asks what can be gained from a common post-fitting selection sample. The following experiment addresses a different question: whether the scale-only rule remains competitive when the published optimizer can learn updates and select its stopping rule on the larger fitting sample. It uses all five datasets, twenty saved splits, HistGBM initialization, and 90% nominal coverage.

Fidelity and initialization. The implementation uses the same audited BoostedCP-local update adapter, author commit, depth-one exact trees, Harrell–Davis weights, soft-sort regularization, gradient smoothing, and finite-rank conventions described above. The authors' Algorithm 1 (Xie et al., 2024) places boosting and cross-validation on the training sample. We follow those data roles while supplying the common HistGBM point predictor and clipped residual-scale proxy. The authors' local-score example instead initializes with a 1,000-tree random forest with leaf size 40 and a random-forest deviation model. This is therefore a comparison using the published training-sample optimization protocol with shared initialization; it is not a reproduction of the authors' complete benchmark or every BoostedCP variant.

To match the published training procedure, we fit the baseline predictors once on the full training sample and then subset their fixed predictions inside boosting cross-validation. We follow that detail: the initial point predictions on D_F are in-sample predictions, while the scale model is trained on out-of-fold point residuals. Each CV validation fold is held out of update fitting but not out of the initial base-model fitting. Thus the folds separate update fitting from update validation, but not base-model fitting from validation. This can affect stopping selection. Nevertheless, final conformal coverage follows from the separate calibration sample. CV validation folds contain approximately 77–600 observations, compared with 23–180 in the allocation-trained protocol.

Learning-rate selection without test selection. Our experiment with the authors' default rate evaluates 500 rounds at rate 20 with five-fold CV entirely inside D_F, including round zero. Early training-CV diagnostics selected zero rounds in 28 of the first 30 completed cases. We therefore specified an additional two-rate sensitivity protocol, with the rationale recorded before executing the added rate and the execution timing documented in the protocol. Every one of the 100 dataset–split cases receives the full 500-round path at each rate in $\{0.1, 20\}$. The selected rate and round minimize the finite mean held-out conformal width over the five training folds, breaking exact ties by smaller rate and then earlier round. The selected update count is refitted on all of D_F.

The rate selector reads only the two training-CV curves before reading the chosen candidate's outcome records. Neither D_A, D_C, nor test labels enter boosting or rate selection. Candidate endpoints and test metrics are saved for auditability, not used to choose the reported rate. Both rate candidates and the original fixed-rate results remain available, including unfavorable and zero-round cases. The added rate is a disclosed response to optimization diagnostics; the study is not presented as prospectively preregistered.

Calibration and equal data budgets. The training-sample optimizer leaves both reserve samples unused, so we evaluate two calibration choices for its selected score. The matched version uses only D_C, the same calibration size as scale-only selection. The pooled version uses $D_A \cup D_C$ and is the primary comparison under the same total label budget. Both choices were specified in the protocol; test results do not choose between them. Scale-only selection still spends D_A on its exponent and therefore calibrates only on D_C. All methods share the fitted predictors, split identities, fitting-only preprocessing, and nominal level.

Why is pooling valid here? For a fixed boosted score s, the calibration threshold is its order statistic at $k = \lceil (n+1)p \rceil$, with an infinite threshold when $k > n$. After conditioning on fitting and optimizer selection, the unused calibration observations and future observation remain exchangeable. This argument works for the pooled version because D_A has not helped choose the boosted score. In contrast, the allocation-trained optimizer and selected scale exponent already use D_A for selection, so the same pooling argument does not apply to them.

Outcomes and uncertainty. We report overall coverage, mean length, and coverage in the three common fitted-scale terciles. To summarize the least-covered group, $C_{\min}$ takes the minimum of

the three empirical coverages within each split and then averages those minima. It differs from taking the minimum after averaging each group's coverage. This is a noisy diagnostic, and neither optimizer guarantees group coverage. Width comparisons use within-split percentage changes. Their pointwise 95% Student-t intervals describe variation across splits of the fixed datasets, without multiplicity adjustment.

Post-fitting runtime counts every CV path at both rates, both candidate refits and calibrations, and selection. Total runtime adds shared base-model fitting once. Data loading, cached base prediction, and one-time import/JIT startup are excluded. Both calibration variants reuse the same boosted score; each method's reported time counts its fitting work once. These are measured concurrent wall times for the implemented search, not hardware-independent complexity comparisons.

Results and optimization diagnostics. Table 1 reports the primary pooled-reserve comparison. Table 44 separates calibration-matched and pooled width effects and reports the group-minimum diagnostic; Table 45 retains every method's mean width, marginal and tercile coverage, and runtime. The tuned optimizer uses nonzero updates in 93/100 cases, with median selected round 208 and four selections at the 500-round cap. Rate 0.1 is selected in 97 cases and rate 20 in three. Seven cases select round zero; their smaller-rate tie choice is not evidence that an update helped.

The primary comparison has mixed outcomes. Scale-only is shorter on Concrete by 7.54% with interval $[-11.99, -3.10]\%$. Its Bike change is -1.08% with interval $[-2.16, -0.01]\%$; the unrounded upper limit is $-.007\%$, so this is a small and statistically fragile difference. CASP and California are inconclusive. On Energy, scale-only is 5.59% wider with interval $[.33, 10.84]\%$. Optimization on D_F therefore reverses the earlier allocation-trained result on that dataset. Realized coverage is .899 for scale-only and .895 for the pooled optimizer on Energy, and .912 versus .910 on Concrete. The comparison targets the same nominal coverage rather than equalizing empirical coverage using test labels.

The group diagnostics also resist a single ordering. The pooled optimizer has higher average within-split minimum tercile coverage on Bike and California, while scale-only has higher values on Concrete and Energy. The optimizer's high-tercile coverage is higher in all five settings, but that tercile is not always the least covered. These are different reliability summaries, and neither is guaranteed by marginal calibration. The matched-D_C variant separates calibration size from the optimized score. The fixed-rate outputs are also retained in the supplement; rate 20 selects zero updates in 89/100 cases, illustrating the sensitivity to learning-rate choice under this initialization.

Saved evidence and checks. All 200 rate candidates are complete. The two fixed-rate reconstruction audits pass 24,054 and 23,562 checks, respectively. The final two-rate audit passes 58,868 checks over 100 selected cases and 1,600 overall/group records, reconstructing rate choices solely from D_F CV, calibration ranks and endpoints from saved arrays, and summaries and paired effects from complete raw records. Checksums accompany the raw, selection, summary, and paired files. The verification scripts, cached candidate arrays, full training-CV curves, source provenance, and commands are supplied in `TRAINING_OPTIMIZER.md`. These computational checks establish consistency of the stored implementation and evidence; they do not establish statistical dominance or reproduce every author benchmark.

O MATCHED TWO-EXPONENT AND FIVE-EXPONENT SELECTION

The simplest way to avoid a harmful scale predictor is to choose between ordinary SCP and fully normalized SCP. We ask whether intermediate exponents provide an additional benefit. Both selectors fix $b = 0$, use the same fitted location and scale, minimize the same estimated length on D_A, and calibrate on untouched D_C. Only the candidate set changes: $\Gamma_2 = \{0, 1\}$ versus $\Gamma_5 = \{0, .25, .5, .75, 1\}$. Both resolve ties by ascending exponent. Calibration and test outcomes do not select the grid, predictor, or reported operating point.

O.1 FITTED SIMULATIONS

The response model is equation 13; Appendix L gives its complete data-generating process. The 24 settings cross two error laws, three scale ratios, two fitting sizes, and two reserve sizes. The compar-

Table 44: Training-sample BoostedCP-local with the common HistGBM initialization at 90% nominal coverage. The optimizer uses D_F for five-fold stopping selection and update fitting (500 rounds, rates $\{0.1, 20\}$ selected by the same D_F CV). Scale-only selects on D_A. ΔW is the mean paired percentage change in scale-only width relative to the stated optimizer, with unadjusted 95% intervals over 20 splits. The pooled optimizer calibrates on unused $D_A \cup D_C$. $C_{\min}$ averages each split's minimum coverage across the three common fitted-scale terciles.

	ΔW (%)		$C_{\min}$	
Dataset	vs. D_C	vs. pooled	Boosted, pooled	Scale-only
Bike	-0.53 [-1.11,+0.06]	-1.08 [-2.16,-0.01]	0.867	0.840
CASP	+0.30 [-0.25,+0.85]	-0.77 [-1.99,+0.46]	0.869	0.869
California	-1.32 [-2.34,-0.30]	-0.64 [-2.23,+0.96]	0.882	0.860
Concrete	-9.87 [-14.26,-5.48]	-7.54 [-11.99,-3.10]	0.850	0.869
Energy	+2.99 [-0.97,+6.96]	+5.59 [+0.33,+10.84]	0.833	0.838

Table 45: Detailed training-sample optimizer comparison with rates 0.1 and 20 selected by fitting-sample CV. Widths are in original response units. Low/mid/high groups are common fitted-scale terciles with cutoffs fixed on D_F. t is median optimization and calibration time in seconds; the shared fitted-model cost is excluded here and recorded separately in the supplied CSV. S denotes scale-only and E the two-endpoint selector. B and BP denote training-sample BoostedCP with D_C and pooled calibration, respectively.

Dataset	Method	Width	Coverage	Low	Mid	High	t (s)
Bike	S	126.467	0.898	0.957	0.897	0.841	< 0.01
Bike	E	127.172	0.898	0.936	0.889	0.869	< 0.01
Bike	B	127.136	0.898	0.940	0.888	0.867	111.56
Bike	BP	127.854	0.900	0.941	0.890	0.869	111.56
CASP	S	13.578	0.893	0.884	0.880	0.914	< 0.01
CASP	E	13.566	0.894	0.873	0.882	0.929	< 0.01
CASP	B	13.542	0.894	0.873	0.882	0.928	120.34
CASP	BP	13.680	0.897	0.875	0.885	0.932	120.34
California	S	1.428	0.900	0.931	0.909	0.861	< 0.01
California	E	1.454	0.902	0.913	0.905	0.890	< 0.01
California	B	1.447	0.903	0.918	0.904	0.887	113.64
California	BP	1.437	0.901	0.916	0.902	0.885	113.64
Concrete	S	16.952	0.912	0.938	0.916	0.887	< 0.01
Concrete	E	17.904	0.908	0.926	0.922	0.879	< 0.01
Concrete	B	19.021	0.911	0.854	0.909	0.960	33.27
Concrete	BP	18.419	0.910	0.856	0.909	0.955	33.27
Energy	S	2.023	0.899	0.943	0.875	0.876	< 0.01
Energy	E	2.163	0.899	0.954	0.907	0.838	< 0.01
Energy	B	1.973	0.897	0.925	0.860	0.903	27.05
Energy	BP	1.928	0.895	0.921	0.855	0.906	27.05

ison uses the same 50 independent repetitions per setting and 5,000 independent test observations in each repetition. Location and scale are refitted in every run. The companion raw file contains six methods, including matched and pooled fixed-exponent controls, giving $24 \times 50 \times 6 = 7{,}200$ records. The associated original nine-method study contains $10{,}800$ records.

Tables 46 and 47 report all 24 comparisons. The width statistic is the mean of the within-repetition percentage change, rather than a ratio of mean widths. High-group coverage evaluates the known conditional response CDF at each predicted interval and averages over independent test inputs with $G = 1$. This removes response-sampling variation while retaining variation from fitting, selection, calibration, and test covariates. It is distinct from a guarantee conditional on the realized calibration sample.

Eighteen pointwise paired intervals lie below zero, two lie above zero, and four cross zero. The two increases occur for Gaussian errors with constant true scale and matched fitting/reserve sizes 300 or 1,000. The mean increases are .40% and .27%, respectively. Intermediate exponents are chosen in 989 of the 1,200 fitted runs. Intermediate exponents therefore help in many of these settings, although the extra selection can cost more than it gains when true scale is constant.

Table 46: Gaussian errors: all twelve fitted settings. Negative paired percentage width changes favor five exponents. Interior frequency and known-CDF high-group coverage use the same 50 repetitions.

r	n_F	$n_A = m$	Width change (%) [95% CI]	Interior (%)	High: two	High: five
1	300	300	+0.40 [+0.06, +0.74]	28	0.901	0.900
1	300	1000	+0.31 [-0.08, +0.70]	18	0.902	0.902
1	1000	300	+0.28 [-0.57, +1.12]	58	0.902	0.903
1	1000	1000	+0.27 [+0.06, +0.48]	34	0.899	0.899
2	300	300	-0.77 [-1.97, +0.44]	72	0.817	0.833
2	300	1000	-1.34 [-1.92, -0.76]	88	0.820	0.836
2	1000	300	-4.05 [-5.22, -2.88]	92	0.843	0.842
2	1000	1000	-4.41 [-5.08, -3.73]	98	0.845	0.839
4	300	300	-9.58 [-11.16, -7.99]	98	0.834	0.822
4	300	1000	-11.32 [-12.35, -10.28]	100	0.835	0.822
4	1000	300	-5.97 [-7.37, -4.57]	94	0.883	0.833
4	1000	1000	-5.09 [-5.92, -4.25]	100	0.875	0.828

Table 47: Log-normal errors: all twelve fitted settings. Negative paired percentage width changes favor five exponents. Interior frequency and known-CDF high-group coverage use the same 50 repetitions.

r	n_F	$n_A = m$	Width change (%) [95% CI]	Interior (%)	High: two	High: five
1	300	300	+0.19 [-1.07, +1.45]	64	0.897	0.897
1	300	1000	-1.05 [-1.77, -0.34]	82	0.900	0.897
1	1000	300	-1.43 [-2.52, -0.34]	82	0.897	0.897
1	1000	1000	-1.87 [-2.68, -1.06]	96	0.900	0.898
2	300	300	-2.82 [-4.42, -1.22]	94	0.854	0.863
2	300	1000	-3.34 [-4.26, -2.42]	96	0.851	0.860
2	1000	300	-3.49 [-4.58, -2.39]	92	0.866	0.866
2	1000	1000	-5.11 [-6.06, -4.16]	100	0.868	0.871
4	300	300	-9.17 [-10.84, -7.50]	96	0.833	0.845
4	300	1000	-7.96 [-8.99, -6.93]	100	0.824	0.839
4	1000	300	-6.54 [-8.07, -5.01]	96	0.872	0.859
4	1000	1000	-8.19 [-9.26, -7.11]	100	0.880	0.858

O.2 Real-data comparison across coverage levels

The real-data audit reuses all 200 frozen fitting/allocation/calibration/test caches. The fitting-defined residual grids and scale-tercile boundaries are unchanged. For each of the three nominal coverage levels, the two selectors are evaluated on identical data. Tables 48–50 report every dataset–learner comparison. Cov. two and Cov. five are mean empirical test coverages, in percent. Six intervals favor five exponents at 80%, six at 90%, and two at 95%. No interval clearly favors the two-exponent selector. These intervals describe the repeated-split experiment conditional on each fixed dataset and are not simultaneous confidence bands or evidence from ten independent datasets.

Computation and data verification. The companion scripts are `code/run_scale_selection_audit.py`, `code/run_fitted_mechanism.py`, and `code/analyze_fitted_scale_selection.py`. Each preserves the split or repetition identifiers, selected parameters, and calibration thresholds. The real audit contains 6,600 overall records and 19,800 group records across eleven methods. Reopened-file checks verify the complete

Table 48: Real-data endpoint-versus-five comparison at 80% nominal coverage, with twenty matched splits per setting.

Data	Learner	Width change (%) [95% CI]	Interior (%)	Cov. two	Cov. five
Bike	HistGBM	-2.66 [-3.58, -1.74]	100	79.5	79.7
Bike	RF	-6.34 [-7.62, -5.05]	100	79.7	80.0
CASP	HistGBM	+0.48 [-0.20, +1.17]	60	79.2	79.5
CASP	RF	-0.07 [-0.84, +0.71]	55	79.1	79.1
California	HistGBM	-3.51 [-4.60, -2.42]	100	80.4	80.3
California	RF	-2.10 [-2.90, -1.29]	100	80.4	80.4
Concrete	HistGBM	-1.45 [-2.80, -0.10]	75	80.1	79.8
Concrete	RF	-1.33 [-2.84, +0.18]	85	79.3	79.9
Energy	HistGBM	-1.57 [-4.18, +1.05]	80	79.4	79.4
Energy	RF	-7.44 [-11.23, -3.65]	80	79.0	79.6

Table 49: Real-data endpoint-versus-five comparison at 90% nominal coverage, with twenty matched splits per setting.

Data	Learner	Width change (%) [95% CI]	Interior (%)	Cov. two	Cov. five
Bike	HistGBM	-0.55 [-1.25, +0.14]	65	89.8	89.8
Bike	RF	-5.00 [-6.11, -3.88]	100	89.7	89.5
CASP	HistGBM	+0.11 [-0.39, +0.60]	50	89.4	89.3
CASP	RF	+0.55 [-0.26, +1.36]	50	89.2	89.2
California	HistGBM	-1.77 [-2.95, -0.58]	75	90.2	90.0
California	RF	-0.69 [-1.67, +0.30]	80	90.3	90.3
Concrete	HistGBM	-4.59 [-7.83, -1.34]	85	90.8	91.2
Concrete	RF	-5.12 [-7.86, -2.39]	85	90.7	90.0
Energy	HistGBM	-6.20 [-9.16, -3.23]	95	89.9	89.9
Energy	RF	-12.54 [-16.77, -8.31]	100	89.2	89.3

Table 50: Real-data endpoint-versus-five comparison at 95% nominal coverage, with twenty matched splits per setting.

Data	Learner	Width change (%) [95% CI]	Interior (%)	Cov. two	Cov. five
Bike	HistGBM	-0.28 [-1.35, +0.80]	55	95.0	94.9
Bike	RF	-1.91 [-3.37, -0.45]	70	94.8	95.0
CASP	HistGBM	-0.87 [-1.86, +0.12]	75	94.7	94.6
CASP	RF	-0.78 [-2.05, +0.49]	75	94.5	94.6
California	HistGBM	+0.69 [-0.41, +1.79]	65	95.0	95.1
California	RF	+0.16 [-0.97, +1.29]	65	95.2	95.2
Concrete	HistGBM	-1.62 [-5.82, +2.58]	75	96.1	96.4
Concrete	RF	-3.49 [-8.94, +1.97]	75	95.4	95.4
Energy	HistGBM	-3.52 [-9.06, +2.03]	95	94.4	95.8
Energy	RF	-9.71 [-16.88, -2.54]	85	94.4	95.1

protocol keys, uniqueness, finite or explicitly infinite metrics, and agreement of unchanged methods with the original records. Raw-file checksums are saved with the verification output.

P EXPLICIT GROUP CALIBRATION AND COVERAGE ALLOCATION

The preceding experiments optimize average length subject to marginal coverage. To ask what changes when each declared group needs protection, we also use groupwise calibration. Define a partition $g_F(x) \in \{1, 2, 3\}$ using the fitted-scale terciles on D_F only. Let m_j be the number of calibration observations in group j. Normalized Mondrian intervals use

$$q_j = Q^+_{m_j}(p; \{|Y_i - \widehat{\mu}(X_i)|/\widehat{\sigma}(X_i) : i \in D_C,\ g_F(X_i) = j\}), \qquad C_{\mathrm{M}}(x) = \widehat{\mu}(x) \pm q_{g_F(x)}\widehat{\sigma}(x).$$

Each group receives its own normalized-SCP threshold. If its corrected rank is unavailable, including when the calibration group is empty, the threshold is $+\infty$. The matched version uses D_C. The pooled version can use $D_A \cup D_C$ because its fitted score and partition do not require allocation tuning.

Coverage target. Conditional on D_F, assume calibration observations and the future observation are i.i.d. For every group j with positive probability, further condition on the number m_j of calibration observations in that group. The group calibration scores and a test score conditional on $g_F(X_{\mathrm{new}}) = j$ are exchangeable. The rank argument therefore gives

$$\mathbb{P}\{Y_{\mathrm{new}} \in C_{\mathrm{M}}(X_{\mathrm{new}}) \mid D_F,\ g_F(X_{\mathrm{new}}) = j,\ m_j\} \geq p.$$

Averaging over m_j preserves the bound. Ties make it conservative, and an unavailable rank gives the full real line. This standard stratified calibration argument (Vovk et al., 2005) guarantees coverage within the declared fitted groups, not at every input or conditional on the realized calibration sample. It does not apply to a group selected afterward because its test coverage is smallest.

Observed tradeoffs. Table 51 contrasts the five-exponent rule with matched normalized Mondrian calibration at 90%. High denotes the highest fitted-scale tercile. The latter is not necessarily the group with the lowest coverage. For example, CASP already has high-tercile coverage above its marginal coverage under the five-exponent rule, so group calibration can lower high-tercile coverage toward the nominal level. Mean paired width increases range from .6% to 23.8%. Thus these observations quantify a change of objective rather than establish one universally preferable procedure.

Table 51: Matched group-calibration comparison at 90%. All coverages are percentages. Width change is Mondrian relative to the five-exponent selector. The last column is the minimum calibration group count across all twenty splits in that setting.

Data	Learner	Five cov.	Five high	Mond. cov.	Mond. high	Δ width (%)	Min. m_g
Bike	HistGBM	89.8	84.1	89.8	90.5	+5.9	273
Bike	RF	89.5	81.7	89.9	90.2	+9.3	267
CASP	HistGBM	89.3	91.4	89.7	89.9	+1.3	255
CASP	RF	89.2	91.4	89.1	89.1	+0.6	260
California	HistGBM	90.0	86.1	90.4	90.5	+4.0	267
California	RF	90.3	85.7	90.6	90.0	+3.0	260
Concrete	HistGBM	91.2	88.7	91.6	91.8	+8.5	36
Concrete	RF	90.0	86.7	91.3	91.0	+6.8	32
Energy	HistGBM	89.9	87.6	90.7	90.1	+23.8	21
Energy	RF	89.3	83.7	90.0	90.2	+13.7	28

Both normalized Mondrian versions and every fixed exponent are retained in the raw and summary files. In the reported data, all thresholds are finite; the smallest matched calibration group has 21 observations. Small test groups, particularly in Concrete and Energy, make empirical group minima variable even for a group-valid procedure. Coverage within a group is therefore reported together with its definition, calibration sample size, and the distinction between a finite-sample guarantee and a particular empirical realization.

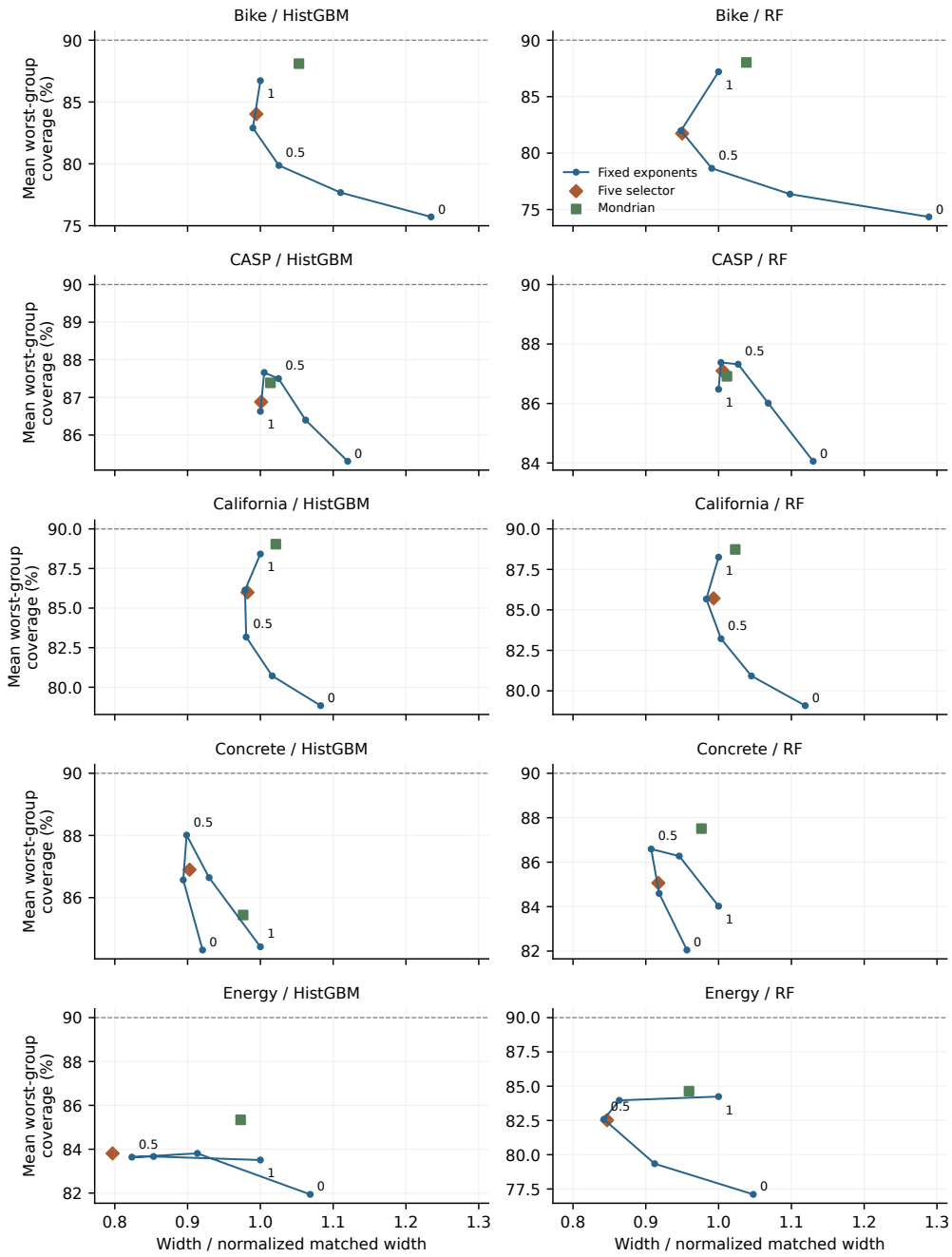

Figure 11: Fixed-exponent width and empirical group coverage at 90%. Connected points follow $\gamma = 0, .25, .5, .75, 1$; labels identify $0, .5, 1$. Width is divided by matched normalized-SCP width within each split before averaging. The vertical coordinate is the mean across splits of the minimum empirical coverage among the three fitted groups; this minimum is downward noisy and is not a direct estimate of the minimum population group coverage. Diamonds show five-exponent selection and squares matched normalized Mondrian calibration. Every method is fixed or selected without test outcomes; these curves are diagnostics, not test-selected operating points.

Q BROADER CONTEXT

Section 2 reviews the closest interval methods. Broader work distinguishes improvements to interval shape from methods designed for stronger coverage targets or different sampling assumptions.

Coverage and informativeness. Conformal prediction provides a framework for calibrating predictive uncertainty under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023; Lei et al., 2018). A coverage guarantee is only one part of an informative prediction: the resulting uncertainty should also be sufficiently concentrated to support decisions (Gneiting & Katzfuss, 2014). Earlier comparisons of conformal classifiers likewise examine efficiency as well as validity (Linusson et al., 2014). In regression, interval length provides a direct measure of this concentration. For example, jackknife+ reuses data through resampling (Barber et al., 2021). Our explicit fitting, allocation, and calibration split makes the cost of learning interval shape visible.

Asymmetric and extreme outcomes. Asymmetric outcomes motivate deliberate interval placement. Studies of healthcare expenditure (Zhou et al., 2024; Lazic et al., 2024; See et al., 2025), pollutant concentrations (Deng & Liu, 2024; Meirelles & Vasconcelos, 2025), and precipitation (Gupta & Chavan, 2023; Weyant et al., 2025) illustrate why asymmetric or extreme outcomes deserve attention in predictive analysis. In such problems, an interval centered at a point prediction can allocate its length poorly even when its marginal coverage is adequate. These examples motivate the question; our empirical claims concern the simulations and benchmarks in Section 5.

Prediction learners and quantile estimation. Interval geometry depends on what the fitted predictors have already captured. Our experiments use histogram gradient boosting and random forests as point learners, drawing on gradient boosting and forest methodology (Friedman, 2001; Breiman, 2001) through scikit-learn (Pedregosa et al., 2011). Quantile regression forests (Meinshausen, 2006) and boosting for extreme quantiles (Velthoen et al., 2023) provide further approaches to learning conditional endpoints. The quantile comparators evaluated here share histogram-boosting fits, so their results should be interpreted for that common learner. This distinguishes adjustment of existing predictors from learning a richer conditional distribution.

Discrete selection and optimization. After fitting, our allocation objective is a function of empirical quantiles. Quantile conventions affect the resulting finite ranks (Dudek & Kuczmaszewska, 2024), which is why the implementation specifies inverse-CDF quantiles and conformal rank corrections explicitly. General constrained optimization (Boyd & Vandenberghe, 2004; Nocedal & Wright, 2006) and methods for non-differentiable or implicit rate constraints (Cotter et al., 2019; Kumar et al., 2021) offer broader tools for optimizing predictive objectives. Our finite-sample efficiency analysis uses quantile regularity and empirical-CDF concentration (van der Vaart, 1998; Massart, 1990) to connect population width comparisons with selection from finite samples. Backward conformal prediction considers a complementary formulation in which a size rule is fixed and coverage is adapted (Gauthier et al., 2025); here we fix the marginal coverage target and compare interval lengths.

Finer coverage targets. Average length does not describe how errors are distributed across inputs. Localized conformal prediction (Guan, 2023), randomized local weighting (Hore & Barber, 2025), and methods targeting conditional moments (Gibbs et al., 2025) pursue more detailed coverage properties. Our marginal-length objective can favor shorter intervals in high-noise regions relative to full normalization, and the scale-bin analyses report the resulting coverage tradeoff.

Changing distributions and structured observations. Other extensions address the sampling assumptions themselves. Covariate-shift weighting (Tibshirani et al., 2019) and conformal calibration beyond exchangeability (Barber et al., 2023; Oliveira et al., 2024) study changing distributions. Adaptive conformal inference (Gibbs & Candès, 2021; Zaffran et al., 2022) and sequential residual methods (Xu & Xie, 2021; 2023) address temporal structure, while conformalized survival analysis treats censoring (Candès et al., 2023). Our results concern exchangeable regression. Although placement and scale selection can improve interval efficiency, they do not by themselves correct distribution shift, temporal dependence, or censoring.