WHEN ERRORS CANCEL: RECOVERING HIDDEN ERROR STRUCTURE FOR CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Locally adaptive conformal prediction normalises the nonconformity score by a learned scale, and essentially every construction learns this scale by regressing it on the observed absolute residual. We show that this default target is corrupted when the residual is a signed aggregate of latent elementary errors that partially cancel. Detection-based counting makes this exact. Over- and under-detections subtract, so the observed count error reports what survived cancellation rather than what the model got wrong. This corrupts the regression target used to fit the scale. The error's predictable component, its latent magnitude, is masked by a survival factor that the input predicts far less well, and the intervals end up wider than they need to be. We propose cancellation-aware normalisation, which replaces the residual target by an exact factorisation into a latent error magnitude and a survival rate (the fraction of latent error mass that survives cancellation), fitted separately and multiplied. This preserves the coverage guarantee exactly, and shrinks the uncertainty intervals. Our central finding is about how this auxiliary information must enter the scale. Supplying it as an extra input feature, the standard practice in the literature, does not reproduce the gain. Fitting it as a factor of the regression target does. We verify the mechanism directly by intervening on the detector confidence threshold, which controls the survival rate, and evaluate it across three counting benchmarks at very different imaging scales.

1 INTRODUCTION

Split conformal prediction turns any point predictor into an interval predictor with a finite-sample marginal coverage guarantee (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018). The guarantee holds for any nonconformity score, so the entire design effort goes into making the resulting intervals short and input-adaptive rather than into making them valid. The dominant approach for that is normalisation. The residual, defined as the absolute error between the point prediction and the ground truth, is divided by a learned, input-dependent scale, so that the interval widens on inputs the scale judges difficult and narrows elsewhere (Papadopoulos et al., 2008; Lei et al., 2018).

Two decades of work on this construction have gone into the estimator of that scale. It has been instantiated as nearest-neighbour dispersion, as the spread of an ensemble, as a learned partition of the input space, as a localisation kernel, and as a regressor on auxiliary self-supervised features. Almost none of that work has gone into the quantity the scale is fitted to. Whatever the estimator, the scale is trained by regressing the observed absolute residual on the input. This is the natural choice under one implicit assumption, namely that the observed residual faithfully measures how badly the model made errors on that input. We identify a broad regime in which that assumption fails, called a 'signed aggregate', where the observed error is a sum of elementary errors that can share or oppose sign, so opposite-sign errors cancel before anything is observed. What the residual then measures is not the error committed but the error that survived cancellation, and the two coincide only when all elementary errors share a sign.

In detection-based counting, aggregate count metrics are understood to look good while localisation is poor, precisely because false positives and false negatives balance out (Tabernik et al., 2024). It is treated in that case as a nuisance for evaluation. Our concern is different. Cancellation is not only a mis-measurement, but a corruption of the regression target from which the uncertainty scale is learned. Inputs where the model made many errors can produce a small observed residual, so

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.

the scale is trained on a label that mixes scene difficulty with how much of the error happened to cancel, which the input predicts far less well. Because conformal validity holds whatever the scale, the damage never surfaces as a coverage failure. It surfaces as intervals that are wider than they need to be.

So, two questions follow. The first is whether the destroyed information is recoverable at all, given that the cancelled quantity is by construction absent from the residual observed at prediction time. The second, which is more general, is how such information must re-enter the construction. The usual strategy is to hand that information to the scale estimator as an additional input feature. We argue that it does not work for an identifiable class of auxiliary information, namely those that do not describe the input better but instead identify a better quantity to regress. Enriching the covariates cannot recover a target when the corruption is in the label.

We study these questions on object counting derived from detectors, used as a testbed rather than as an application domain, because there the cancellation is exact and fully observable on annotated data. Note that benchmarks exist at very different imaging scales, and nothing in what follows requires ground truth at inference step.

Our contributions are the following ones.

- **A failure mode.** To our knowledge, when the residual is a signed aggregate, cancellation corrupts the target of any locally adaptive construction. The latent error magnitude can be far more predictable from inference-time features than the observed residual that masks it. We give a cheap pre-hoc test for it (Section 3.4).
- **A remedy, and where it must act.** Replacing the target by the latent magnitude preserves the guarantee and often shrinks the intervals. Adding a fitted survival rate makes this gain safe, never meaningfully worse than the standard method. Supplying the same information as a feature instead does not reproduce the effect.

2 Related work

Normalised split conformal prediction and scale estimators. Normalised split conformal prediction divides the residual by a learned, input-dependent scale before taking the calibration quantile (Papadopoulos et al., 2002; Vovk et al., 2005; Papadopoulos et al., 2008; Lei et al., 2018), so that calibration fixes only the global level and the scale carries all the input dependence (Section 3.1). The estimators of that scale differ only in their proxy for local difficulty: nearest-neighbour dispersion of the training residuals (Papadopoulos et al., 2011), ensemble spread (Boström et al., 2017; Johansson et al., 2021), an explicit regressor of the absolute residual on the covariates (Lei et al., 2018), which we take as the canonical baseline, or that same target under a learned partition (Kiyani et al., 2024) or a localisation kernel (Guan, 2023). Romano et al. (2019) replace the scale by two fitted conditional quantiles. They note, after Lei et al. (2018), that on homoscedastic data normalisation widens intervals through the variance of estimating the scale, and they trace its weak adaptivity to a biased target, the in-sample residuals of the point predictor. Our residuals are held out, since the detector never trains on the pool, so the corruption we study is a different one; Section 5.4 tests our change of target inside their construction.

Conditional coverage. Exact conditional coverage is unattainable distribution-free (Foygel Barber et al., 2021). Mondrian or group-conditional calibration (Vovk, 2012) assigns each input to a group, or taxonomy cell in Mondrian terminology, and computes a separate quantile per cell, so coverage holds within each cell at the cost of a smaller effective calibration set per cell. Learned partitions (Kiyani et al., 2024) choose the cells rather than fixing them, and localised approaches (Guan, 2023) replace the hard partition by a smooth weighting. All act on how the quantile is calibrated, downstream of the scale. Our contribution is upstream of calibration and composes with it, and we evaluate under both regimes in Section 5.4.

Auxiliary signals for the scale. Closest to our work, Seedat et al. (2023) improve the difficulty estimator by computing a self-supervised pretext error per input and appending it to the covariates of the scale regressor, leaving the target at the observed residual. This is the natural way to inject side information and it is the design we take as our concatenation baseline. Our result is that for

cancellation this path does not work, because the auxiliary quantity here is informative about what should have been regressed rather than about which inputs are difficult.

Conformal prediction for object detection. Existing work conformalises bounding box coordinates, producing calibrated boxes around already matched objects (de Grancey et al., 2022; Andéol et al., 2023; Timans et al., 2024), so the quantity of interest is localisation conditional on a match. The aggregate count is a different quantity of interest, the one practitioners actually report, and the only one in which false positives and false negatives enter with opposite signs. To our knowledge no prior work constructs conformal intervals for detector-derived counts, or uses the decomposition of the count error into its two components as a target for the scale.

3 Problem setting and method

3.1 Notation and normalised split conformal prediction

Let $\mathcal{X}$ be the input space of images. For an input image $x \in \mathcal{X}$ a detector emits a finite set of predicted boxes, and the annotation provides a finite set of ground-truth boxes. We write $n(x) \in \mathbb{N}$ for the number of ground-truth boxes, the quantity to be predicted, and $\hat{n}(x) \in \mathbb{N}$ for the number of predicted boxes after post-processing, the point prediction. The two sets are put in correspondence by a one-to-one assignment (see Section 4.4 for details). A predicted box assigned to a ground-truth box is a true positive, an unassigned predicted box is a false positive, and an unassigned ground-truth box is a false negative, denoted $\mathrm{TP}(x)$, $\mathrm{FP}(x)$ and $\mathrm{FN}(x)$, respectively, so that $\hat{n}(x) = \mathrm{TP}(x) + \mathrm{FP}(x)$ and $n(x) = \mathrm{TP}(x) + \mathrm{FN}(x)$. The signed count error is

$$E(x) \;=\; \hat{n}(x) - n(x) \;=\; \mathrm{FP}(x) - \mathrm{FN}(x), \tag{1}$$

and $|E(x)|$, its absolute value, is the observed residual, which the standard construction uses as its regression target. We define the *latent error mass* $D(x) = \mathrm{FP}(x) + \mathrm{FN}(x)$, the total number of elementary detection errors committed on x, and the *survival rate*, denoted b for balance,

$$b(x) \;=\; |E(x)| \,/\, D(x) \;\in\; [0,1], \qquad b(x) := 1 \text{ when } D(x) = 0. \tag{2}$$

If $b(x) = 1$, all elementary errors on x shared a sign and none cancelled, whereas $b(x) = 0$ means they cancelled exactly. Both D and b require the annotation of x and are available only on folds with ground truth. Neither is needed at inference step.

To build the prediction interval, we split the data into three disjoint folds. Every regressor is fitted on the *selection* fold and then frozen, the *calibration* fold only sets the conformal quantile q, and the *test* fold only reports results. We start from a feature map $\phi : \mathcal{X} \to \mathbb{R}^p$, with p features, computable from the detector at inference step and without access to ground truth (e.g., ensemble disagreement, FPN activation statistics, confidence statistics, as reported in Section 4.4). On the selection fold, we fit a regressor called $\sigma : \mathcal{X} \to \mathbb{R}_{\geq 0}$, the *normalisation scale*, a function of $\phi(x)$ only. Fitting σ means choosing a *target*: a quantity, computed from the annotation on the selection fold, that σ is trained to reproduce from $\phi(x)$. The default choice in the literature is to regress the observed absolute residual $|E|$ on the input, that is, to fit σ so that $\sigma(x)$ approximates $|E(x)|$. Section 3.2 shows when the assumption behind this choice fails, and Section 3.3 gives the alternative target we propose.

Once σ is fit and frozen, we use the calibration fold to compute the quantile q. Let m be the number of images in the calibration fold, indexed $i \in \{1, \ldots, m\}$. For each calibration image x_i we compute a nonconformity score

$$s_i \;=\; \frac{|E(x_i)|}{\sigma(x_i) + \varepsilon}, \tag{3}$$

where $\varepsilon > 0$ is a small additive constant, the *stability floor*, that keeps the score from exploding when the fitted scale σ is close to zero. We define ε as a fixed percentile of the strictly positive values taken by σ on the selection fold, and use the tenth percentile throughout.

Actually, this choice of floor matters. Because ε is a percentile of σ's own values, replacing σ by $c\,\sigma$ for any $c > 0$ leaves the final interval unchanged. Concretely, $\varepsilon \mapsto c\,\varepsilon$, $q \mapsto q/c$, and $\sigma(x) + \varepsilon \mapsto c(\sigma(x) + \varepsilon)$, so their product is invariant. Only the relative profile of σ across inputs affects the interval width; its overall level is absorbed by the quantile. A fixed additive constant

would break this invariance. This is what makes a biased but stable estimate of σ's profile preferable to an unbiased but noisy one, a point we return to in Section 3.3. We rank the m calibration scores from smallest to largest. We denote $\alpha \in (0, 1)$ for the target miscoverage, so the target coverage is $1 - \alpha$. We set $\alpha = 0.1$ in what follows, for a target coverage of 90 percent. The conformal quantile q is the score at rank $k = \lceil (m + 1)(1 - \alpha) \rceil$ in that ordering, i.e., $q = s_{(k)}$.

At test step, for a test input x, the prediction of uncertainty intervals is

$$C_\alpha(x) \;=\; \big[\, \hat{n}(x) - q\,(\sigma(x) + \varepsilon),\;\; \hat{n}(x) + q\,(\sigma(x) + \varepsilon)\,\big], \tag{4}$$

centred on the point prediction $\hat{n}(x)$. Under exchangeability of the calibration and test folds, this interval satisfies $\mathbb{P}(n(X) \in C_\alpha(X)) \geq 1 - \alpha$ for any choice of σ (Papadopoulos et al., 2002; Lei et al., 2018), the guarantee holds regardless of what σ was trained to predict, which is what lets us change that target freely in the rest of the paper.

3.2 Error cancellation and its effect on the target

Starting from Eq. (1), we have

$$D(x) - |E(x)| \;=\; 2\min\big(\mathrm{FP}(x), \mathrm{FN}(x)\big) \;\geq\; 0, \tag{5}$$

the amount of elementary error that occurred and then disappeared from the observed residual. This quantity is zero exactly when the detector's elementary errors on x all share the same sign, in which case $D(x) = |E(x)|$. Equivalently, and this is the identity the method rests on,

$$|E(x)| \;=\; b(x) \cdot D(x) \tag{6}$$

holds exactly, by the definition of b in Eq. (2), not as an approximation or a modelling assumption. The two factors are of very different nature. The magnitude D reflects how hard the scene was for the detector and is plausibly a function of scene content, which is what a difficulty estimator is meant to capture. The rate b reflects how the detector's false positives occurred to be distributed relative to its false negatives, which we expect to be only weakly predictable from the input (Table 3). A regressor trained on $|E|$ is therefore fitting a product of a factor that plausibly depends on scene content and one that largely does not, with no mechanism separating the two. Here $\mathbb{E}[.]$ denotes the mathematical expectation. If $\mathbb{E}[b \mid x]$ is roughly constant, $\mathbb{E}[|E| \mid x] \propto \mathbb{E}[D \mid x]$: D targets the same difficulty profile as $|E|$, up to a constant the conformal quantile absorbs (Section 3.1), with far less label noise. Seedat et al. (2023) noticed more generally that residuals are hard to predict, since a predictable residual would have been predicted away by the point predictor. Cancellation gives a concrete structural reason for that difficulty in the signed-aggregate case. But unlike that argument, which offers no way to act on the difficulty it explains, our decomposition offers one, since both factors of Eq. (6) are computable wherever annotations exist.

3.3 Cancellation-aware normalisation

On the selection fold we fit two regressors of the same class with the same features ϕ. We denote $\hat{D}$ and $\hat{b}$ the estimates of D and $b = |E|/D$, respectively. Both $\hat{D}$ and $\hat{b}$ are frozen and applied in prediction mode only to the calibration and test folds. Throughout this paper, we label every construction used for comparison B (for baseline) and our constructions C, followed by a number (i.e., C1, C2, C3 for our contributions). The simplest proposal, C1, uses $\hat{D}$ alone, and the safeguarded one, C3, is their product,

$$\sigma_{\mathsf{C3}}(x) \;=\; \hat{b}\big(\phi(x)\big) \cdot \hat{D}\big(\phi(x)\big), \tag{7}$$

substituted into Eq. (4), with predictions clipped below at zero. The same calibration fold, the same rank k and the same coverage guarantee apply, because the guarantee is indifferent to how σ was obtained.

Why a biased plug-in is admissible. The product of two separately fitted regressors, $\hat{b}$ and $\hat{D}$, is a biased estimator of $\mathbb{E}[|E| \mid x]$: in general $\mathbb{E}[b \cdot D \mid x] \neq \mathbb{E}[b \mid x] \cdot \mathbb{E}[D \mid x]$, so Eq. (7) does not estimate the same quantity the standard method targets. By the scale invariance of Section 3.1, only the relative profile of σ affects the width, so a biased but lower-variance estimator of that profile can be strictly preferable to an unbiased but noisy one.

Reduction to the standard method. When cancellation is absent, so that $b(x) \to 1$ for all x, we have $D \to |E|$ from Eq. (6), the target of $\hat{b}$ becomes constant, and σ_{C3} converges to the residual-targeted scale by construction. The method is therefore a strict generalisation of the standard one and recovers it exactly where the standard one is already correct (See Section 5.3 for an empirical demonstration).

3.4 Testing whether D is more predictable than $|E|$

The quantities D, $|E|$, and every plausible feature ϕ all grow with the number of objects n. Comparing how well ϕ predicts each target without accounting for this would be misleading, since a good score could just mean that images with more objects have both a larger target and a larger ϕ, with nothing to do with difficulty. We therefore factor n out first. For each target Y (one of D, $|E|$ or b), we regress Y on n alone to obtain a residual, the part of Y that n does not already explain, then regress that residual on the features ϕ. What ϕ recovers at this second stage is, by construction, information beyond n.

We evaluate both stages with two-fold cross-fitting (Chernozhukov et al., 2018): fit both steps on one half of the data, score on the other, then swap, so no model is ever tested on data it was trained on. This yields a score $R^2_{\text{resid}}(Y) \in (-\infty, 1]$ (exact formula given in Appendix B), where 0 means the features add nothing beyond object count. We compute it for $Y = D$, $|E|$ and b (the latter on images with $D > 0$), with the same features and folds.

The diagnostic is the gap $G(D, |E|) = R^2_{\text{resid}}(D) - R^2_{\text{resid}}(|E|)$. A positive gap means the features explain something about the latent error mass that they do not explain about the observed residual — the mechanism this paper claims. We test whether $R^2_{\text{resid}}(D)$ exceeds chance against a permutation null: we permute D across images, keeping ϕ and n paired, and recompute $R^2_{\text{resid}}(D)$ from scratch on the permuted labels, using the same procedure above. Repeating this yields a null distribution of $R^2_{\text{resid}}(D)$ under no true D-ϕ relationship, against which we locate the observed value.

4 Experimental setup

4.1 Metrics

Our primary metric is **mean interval width**, the mean of $2q(\sigma(x) + \varepsilon)$ over the test fold. Because every method is calibrated to the same target coverage by construction, width is the only axis on which they can usefully differ, and a lower width at equal coverage is a strictly better predictor. We report **empirical coverage**, the fraction of test inputs with $n(x) \in C_\alpha(x)$, to certify that width comparisons are made at equal validity rather than bought by under-coverage, flagging a cell when a one-sample t-test over splits rejects equality with 0.9 at $p < 0.05$.

As in Section 3.4 we report the **cross-fit residual** R^2 of Eq. (8) (see formula in Appendix B), on a scale where zero means the features add nothing beyond object count and negative values are possible because the score is out of fold. For b the score is computed on images with $D > 0$ only, since $b := 1$ at $D = 0$ is a convention, not a measurement. Widths of two methods are always compared on the same splits, with a **paired t-test** over splits. The diagnostic of Section 3.4 has no parametric null, so its **permutation p-value** is the fraction of 200 permutation draws whose cross-fit R^2 reaches the observed one (same pool across splits, so these p-values are anti-conservative and overstate significance).

4.2 Datasets for experiments

We use three counting benchmarks that differ in imaging scale and object density, since a claim about a mechanism should not rest on one imaging regime. **LIVECell** (Edlund et al., 2021) is phase-contrast microscopy of cultured cells, the densest of the three, with on the order of three hundred annotated objects per image on average. **GWHD** (David et al., 2021) is ground-level RGB imagery of wheat heads for agricultural phenotyping, with the fewest objects per image of the three. **VisDrone2019-DET** (Zhu et al., 2022) is aerial surveillance imagery of vehicles and people, the macroscopic end, with intermediate density of objects, multi-class and heterogeneous scene composition. For LIVECell and VisDrone2019-DET the pool from which the three folds are drawn is the

Table 1: The five constructions. All share the regressor class and the features ϕ of Section 4.4, and the nonconformity score is $|E(x)|/(\sigma(x)+\varepsilon)$ throughout. Only the *target* of σ, and for C2 its input, change across rows. The ladder doubles as the ablation of Section 5.2.

		Target of σ	Input to σ	Role
	B0 marginal	—	—	Constant-width split conformal ($\sigma \equiv 1$). The floor below which adaptivity does not repay its estimation cost.
	B1 residual	$\lvert E(x)\rvert$	$\phi(x)$	Canonical locally adaptive predictor (Papadopoulos et al., 2008; Lei et al., 2018): the method we replace, and the reference for every comparison below.
	C1 magnitude	$D(x)$	$\phi(x)$	Simplest. Changes the target only, isolating it from the factorisation. Biased upward since $D \geq \lvert E\rvert$ by Eq. (5), tests whether that bias makes the target easier to predict.
ours	C2 concat.	$\lvert E(x)\rvert$	ϕ + $\hat{D}_{\text{oof}}$	Augmented. Changes the input only. The standard way of injecting auxiliary information into a conformal scale (Seedat et al., 2023), adapted to our case.
	C3 factorised	$b(x)$ and $D(x)$	$\phi(x)$	Safeguarded. Two regressors, fit independently and multiplied, $\sigma_{\text{C3}} = \hat{b} \cdot \hat{D}$ (see Eq. (7)).

official test split. For GWHD the official test split is too small to support our three-way partition, so we draw the pool from the official validation split instead, which provides enough data to allocate across selection, calibration and test while remaining reproducible from the public release. Within that pool each dataset is partitioned into 25 % selection, 25 % calibration and 50 % test, stratified on deciles of n so the three folds carry comparable count distributions. Datasets sizes are reported in Table 2.

4.3 Construction methods

Table 1 lists the two baselines (B0, B1) and three constructions (C1-C3) we compare, in the sense of target given in Section 3.1. Here C2 and C3 receive exactly the same auxiliary information. The only difference is how that information enters the scale estimator. It arrives as an input feature for C2, and as a regression target for C3. Here $\hat{D}_{\text{oof}}$ is a five-fold cross-fitted prediction of D computed within the selection fold, so no training row sees its own D.

4.4 Implementation details

We use YOLOv8 detectors across all three benchmarks, ensembled where possible and a single fine-tuned checkpoint for VisDrone2019-DET. Detections below a confidence of 0.25 are discarded; the survivors are matched to ground truth by the Greedy algorithm at an intersection-over-union threshold of 0.5, identically across all three datasets. We consider four families of inference-time features ϕ (ENS_DISAGREE, FPN, CONF, NMS_DENSITY) and fix one family per dataset in advance of all conformal experiments, choosing the one with the highest $R^2_{\text{resid}}(D)$ by the criterion of Section 3.4 (ranking holds under each method's own best family/regressor, Appendix C). The other three families were tested as well, Appendix C reports them and shows the mechanism holds there too. Each method is instantiated with two regressor classes, ridge and gradient boosting, so that no conclusion depends on one inductive bias, and every experiment is repeated over 100 random three-way splits. Full detector, matching, feature and regressor details are in Appendix A.

5 Results

5.1 Cancellation destroys error mass and predictability

Table 2 reports cancellation statistics for all three datasets. Cancellation destroys 63.4, 65.6 and 37.8 percent of the elementary error mass on GWHD, LIVECell and VisDrone. On LIVECell, about two out of every three detection errors the model commits are invisible in $|E|$.

Table 2: Cancellation statistics per dataset. % of D cancelled: the share of elementary error mass invisible in $|E|$.

Dataset	Images	$\sum D$	$\sum \lvert E \rvert$	% of D cancelled
GWHD	977	9,254	3,385	63.4%
LIVECell	1512	191,052	65,731	65.6%
VisDrone	1610	45,263	28,161	37.8%

Table 3: Predictability of D, $|E|$ and b from image features, per feature family and dataset (cross-fit protocol of Section 3.4; column definitions in Appendix B). Bold: candidate selected for Table 4.

	GWHD				LIVECell				VisDrone			
Candidate	$R^2(D)$	Gap	ρ	$R^2(b)$	$R^2(D)$	Gap	ρ	$R^2(b)$	$R^2(D)$	Gap	ρ	$R^2(b)$
CONF	$+\mathbf{0.369}^*$	$+0.328$	$+0.836$	$+0.068^*$	$+0.482^*$	$+0.470$	$+0.919$	$+0.034^*$	$+0.261^*$	$+0.002$	$+0.925$	$+0.089^*$
FPN	$+0.230^*$	$+0.291$	$+0.782$	-0.026	$+\mathbf{0.548}^*$	$+0.556$	$+0.906$	$+0.056^*$	$+\mathbf{0.376}^*$	-0.001	$+0.892$	$+0.127^*$
NMS_DENSITY	$+0.009^*$	$+0.087$	$+0.712$	-0.081	$+0.117^*$	$+0.122$	$+0.845$	$+0.025^*$	$+0.156^*$	-0.194	$+0.904$	$+0.114^*$
ENS_DISAGREE	$+0.083^*$	$+0.052$	$+0.746$	-0.061	$+0.354^*$	$+0.363$	$+0.927$	$+0.038^*$	—	—	—	—

Table 3 applies the pre-hoc test of Section 3.4 and shows that cancellation destroys predictability as well as mass. On GWHD and LIVECell the latent magnitude carries substantial residual signal while the observed residual carries essentially none, for gaps of $+0.328$ and $+0.556$ on the selected feature family. The survival rate b is significantly but only weakly predictable (R^2_{resid} of 0.08, 0.056 and 0.127), far below D. On VisDrone the gap is null, and we report it as such. Section 5.2 shows that VisDrone nonetheless benefits from the method, so the test identifies where the mechanism is strongest rather than where it is present, and should be read as indicative rather than necessary. Every feature family passes the permutation test for D on every dataset.

5.2 Main comparison and ablation

The standard adaptive method does not reliably outperform no adaptation. Table 4 reports mean interval widths at the 90 percent target across three datasets and two regressor classes. B1 improves on the constant-width B0 by 10.8 percent on GWHD and 14.5 percent on LIVECell, but on VisDrone with ridge it is 12.0 percent worse, at 89.16 against 79.62. Normalising by a scale fitted to a target that cancellation has scrambled can cost more in estimation variance than it recovers in adaptivity, the failure mode Romano et al. (2019) anticipate.

Changing the target alone recovers most of the width reduction. Both constructions that regress the latent magnitude largely repair the VisDrone regression and improve substantially elsewhere. Relative to B1, C3 reduces width by 13.3 and 8.5 percent on LIVECell and by 7.9 and 6.4 percent on VisDrone, all with $p < 10^{-7}$, and by 1.2 percent with $p = 0.02$ and 0.4 percent not significantly on GWHD, so it improves in all six cells. C1, which discards the cancellation factor entirely, improves in four and is the best of the four adaptive methods on LIVECell under both regressors and on VisDrone under ridge, at 183.16, 168.46 and 81.44. The latent magnitude alone, with no correction for sign, is therefore a strong enough target to outperform the observed residual. It is the strongest evidence that the problem of Section 5.1 is real.

The cancellation factor buys stability, not peak width. C1 looses to B1 on GWHD under both regressors, at 14.36 and 14.67 against 13.78 and 13.97, where b is weakly but significantly predictable (Table 3). C3 does not, because $\hat{b}$ corrects the scale profile wherever cancellation varies across inputs. Over all 22 ridge and gradient-boosting cells (Appendix C), C1 is wider than B1 in seven, by up to 7.4%, and C3 never by more than 0.3%. C1 is thus the better choice when cancellation is heavy and roughly uniform, C3 the safe default otherwise.

The feature path does not substitute for the target path. C2 receives exactly the auxiliary information C3 receives, cross-fitted, in the standard way. Measured as the fraction of C3's absolute width reduction that it recovers, C2 recovers between -64.6 and $+37.5$ percent on five of six cells, and on LIVECell with ridge, the cell with the largest available gain, it is markedly worse than B1 itself, at 238.03 against 213.01, while C3 reaches 184.64. The exception is VisDrone with

Table 4: Mean interval width at the 90% coverage target, for constructions B0 to C3 (Table 1), on each dataset's selected feature family (in parentheses it is the selection criterion in Section 4.4), under two regressor classes: ridge regression (Ridge) and gradient boosting (GBM). The MLP regressor is omitted here as numerically fragile (Appendix C). Lower width is better at equal coverage. Bold: lowest width per column. Last row: paired width gain of C3 over B1, with the paired t-test p-value in parentheses.

		GWHD (conf)		LIVECell (fpn)		VisDrone (fpn)	
	Method	Ridge	GBM	Ridge	GBM	Ridge	GBM
	B0	15.44*	15.44*	249.09	249.09	79.62*	79.62*
	B1	13.78†	13.97	213.01†	194.17	89.16†	80.29
ours	C1	14.36†	14.67	**183.16**†	**168.46**	**81.44**†	76.28
ours	C2	13.78†	13.96	238.03†	193.35	87.08†	80.06
ours	C3	**13.61**†	**13.92**	184.64†	177.57	82.12†	**75.17**
Gain: C3 vs. B1		+1.2% (1.6e-02)	+0.4% (n.s.)	+13.3% (1.4e-19)	+8.5% (5.2e-23)	+7.9% (2.0e-10)	+6.4% (4.2e-14)

* coverage significantly deviates from the 90% target ($p < 0.05$). Empirical coverage lies in $[0.897, 0.905]$ across every cell. † scale numerically fragile on $\geq 10\%$ of seeds.

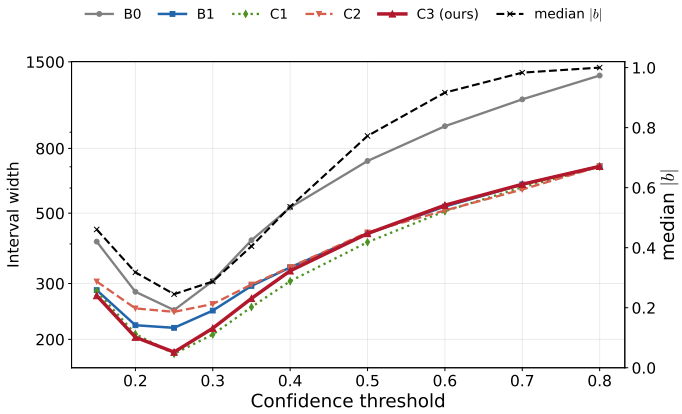

Figure 1: Mean interval width against the detector confidence threshold, on LIVECell, with the median survival rate b on the right axis. B1 to C3 differ most where b is low and coincide as $b \to 1$ (see Eq. (6)).

ridge, where C2 reaches 87.08 and recovers 29.5 percent of C3's gain. Concatenation enriches the covariates while the label remains $|E|$, and the cancellation noise lives in the label, so auxiliary information that identifies a better target cannot in general be consumed through the feature path. We regard this as the most transferable result in the paper and as what most sharply distinguishes our proposal from Seedat et al. (2023).

5.3 Mechanism: intervening on the survival rate

A comparison across three benchmarks cannot establish what drives the gain, because the benchmarks differ in every respect at once. Within a dataset the detector confidence threshold is a direct handle on b. Raising it removes false positives but turns some detections into false negatives. Lowering it does the reverse, adding false positives that coexist with existing false negatives and cancel. This is a controlled intervention on the variable our account says is causal, run inside each dataset with the detector and annotations held fixed.

Figure 1 illustrates that the advantage of the factorisation is largest at low b and shrinks monotonically as b rises, vanishing when b approaches one and the two constructions coincide by definition. The same curve appears on all three datasets, including the one where the pre-hoc test was null. This is consistent with a mechanistic account, and it delivers a practical property. In the no-cancellation regime the method does not degrade, it becomes the standard method.

5.4 ROBUSTNESS

Under Mondrian calibration (Vovk, 2012) within five quantile bins of predicted count (Table 11, Appendix E.2), all intervals widen, as expected from calibrating on fewer points per cell, and the C3 advantage survives on LIVECell, moving from $+13.3$ to $+9.1$ percent under ridge and $+8.5$ to $+7.6$ under gradient boosting, and grows on VisDrone, from $+7.9$ to $+9.0$ and $+6.4$ to $+6.8$, all with $p < 10^{-10}$. On GWHD it disappears, at -0.3 and -0.1 percent, neither significant, consistent with GWHD's marginal gain already being the smallest of the three and only marginally significant.

Under conformalised quantile regression (CQR) (Romano et al., 2019), two fitted conditional quantiles of E set both the centre and the shape of the interval. Dividing E by $\sigma + \varepsilon$ before fitting them, with σ the scale of C3 rather than B1, narrows the interval in 7 of 12 rows of Table 16 (Appendix E.7), against 9 of 12 for the symmetric construction. CQR without a scale is up to 25 percent narrower than symmetric C3 on VisDrone, while symmetric C3 outperforms all three CQR variants on LIVECell. Overall, some construction using σ_{C3}, symmetric or CQR, is narrowest in 7 of 12 rows, 7 of 8 excluding GWHD (exception: VisDrone FPN ridge, where plain CQR wins). On GWHD itself, plain CQR wins 3 of 4 rows, the last going to CQR$/\sigma_{\mathsf{B1}}$.

The stability floor ε (Section 3.1) is the one free quantity in the normalisation. Swept from the 1st to the 50th percentile with ridge and thirty splits, C3's advantage over B1 stays positive on VisDrone and LIVECell, and on GWHD it is positive up to the default tenth percentile and slightly negative beyond (Figure 3, Appendix E.3). We report the sweep rather than tune ε per dataset.

Read as coverage rather than width, with C3's α chosen per split so its mean width matches B1's width at $\alpha = 0.1$, C3 reaches 90.5 to 91.5 percent against B1's 90 percent. An image-level analysis shows that in five of six cells C3 covers a different set of images than B1 at matched width, gaining on some and losing on others rather than simply covering a superset. On the label-conditional bin of highest D every construction under-covers, B0 most, and C3 is not better than B1 in this case (Table 14): its gain is one of efficiency. Finally, since the gain ordering across datasets tracks mean latent error mass, we tested a density explanation within each dataset, where detector, resolution and annotation protocol are fixed. It holds on VisDrone (Spearman $\rho = +1.00$, exact permutation $p = 2/120 \approx 0.017$) but is absent on GWHD and LIVECell (both $+0.10$, n.s.), so object density does not generally explain when the method helps.

6 CONCLUSION

The method applies when the error is a signed aggregate of latent elementary errors (**R1**) and the latent magnitude is computable from ground truth on a fitting split, never at inference step (**R2**). A third condition, that this magnitude is the more predictable target, is indicative but not necessary; Section 3.4 makes it cheap to test, and Section 5.2 shows the method still helps on VisDrone where the test is null. Detection-derived counts satisfy R1 and R2 exactly, by Eq. (1). Other aggregate-of-predictions quantities plausibly satisfy them as well, for instance demand or load forecasting and portfolio-level profit and loss, where regional, hourly or positional over- and under-shoots offset in the reported net. Extending the construction to a non-vision signed aggregate, with the diagnostic run first, is the most natural next step.

Two decades of work on locally adaptive conformal prediction have improved the estimator of the uncertainty scale while leaving its regression target untouched. We showed that when the residual is a signed aggregate, that target is broken by cancellation, which destroys most of the elementary error mass in our benchmarks and, with it, most of the predictable signal. Regressing the latent magnitude instead recovers it, and is often the strongest scale on its own. Factorising the residual into magnitude and survival rate keeps that gain while never doing meaningfully worse than the residual, and both changes inherit the coverage guarantee.

The finding we expect to travel furthest is the negative one. Handing the same auxiliary signal to the estimator as an input feature, which is how the literature currently injects such signals, does not reproduce the effect and can make matters worse. When auxiliary information tells you what you should have been regressing, it belongs in the label, not the covariates.

REPRODUCIBILITY STATEMENT

The full pipeline that produces every table and figure, including the appendix, is submitted as supplementary code. It has two stages. Stage 1 runs the detectors, matches predictions to ground truth, and caches one row per image with the counts, features and cancellation statistics of Section 3.1. It needs a GPU and the detector checkpoints, and can be skipped by using the released per-image tables directly. Stage 2 reads only those cached tables, needs neither a GPU nor the images, and produces every reported number in well under an hour for all three datasets. An optional third stage reruns the threshold-sweep intervention of Section 5.3 by rematching the cached detections at a grid of confidence thresholds, without touching the detector or the annotations.

Sections 3 and 4 specify the estimators, the three-way stratified split, the number of splits, the conformal quantile with its finite-sample correction, and the stability floor. These constants are centralised in a single configuration module rather than duplicated across scripts. Section 3.4 specifies the cross-fitting and permutation-null protocols. The supplementary code includes invariant tests covering the exactness of the factorisation, the destroyed error mass, and the coverage guarantee, and equivalence tests confirming the results match the original implementation bit for bit. All three datasets are public (David et al., 2021; Edlund et al., 2021; Zhu et al., 2022), as is the VisDrone detector checkpoint (mshamrai, 2023). Results downstream of extraction are deterministic given the seeds. The appendix reports per-family results for every dataset, the Mondrian, equal-width and density tables summarised in Section 5.4.

ETHICS STATEMENT

This work studies uncertainty quantification for object counting. One of the datasets is drone imagery containing people and vehicles in public spaces. We use only released annotations and derive no identity information. We note that better-calibrated uncertainty for aerial person counting is dual-use, and that the intended application is to make automated counts auditable rather than to extend surveillance capability.

AI USE STATEMENT

We used generative AI tools to help phrase some passages more clearly and to assist with literature search, with all retrieved literature read by the authors before citing. We also used generative AI tools to help draft parts of the analysis code, with all such code reviewed by the authors. We used generative AI tools for research ideation and execution, specifically to test hypotheses and implement exploratory analyses during auditability audits. We did not use generative AI tools to generate the core research ideas or to design the principal experiments. All numbers reported in the paper were produced by the submitted pipeline and checked by the authors against its raw output. We take responsibility for the final content of this work.

REFERENCES

Léo Andéol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 36–55. PMLR, 2023.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, 81 (1):125–144, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto, Shahameh Shafiee, Izzat S. A. Tahir, Hisashi Tsujimoto, Shuhei

Nasuda, Bangyou Zheng, Norbert Kirchgessner, Helge Aasen, Andreas Hund, Pouria Sadhegi-Tehran, Koichi Nagasawa, Goro Ishikawa, Sébastien Dandrifosse, Alexis Carlier, Benjamin Dumont, Benoit Mercatoris, Byron Evers, Ken Kuroki, Haozhou Wang, Masanori Ishii, Minhajul A. Badhon, Curtis Pozniak, David Shaner LeBauer, Morten Lillemo, Jesse Poland, Scott Chapman, Benoit de Solan, Frédéric Baret, Ian Stavness, and Wei Guo. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, 2021: 9846158, 2021. doi: 10.34133/2021/9846158.

Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and David Vigouroux. Object detection with probabilistic guarantees: A conformal prediction approach. In *Computer Safety, Reliability, and Security. SAFECOMP 2022 Workshops*, pp. 316–329. Springer, 2022.

Christoffer Edlund, Timothy R. Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. LIVECell — a large-scale dataset for label-free live cell segmentation. *Nature Methods*, 18(9):1038–1045, 2021. doi: 10.1038/s41592-021-01249-6.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Ulf Johansson, Henrik Boström, and Tuve Löfström. Investigating normalized conformal regressors. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, 2021.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24749–24769. PMLR, 2024.

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.

mshamrai. yolov8l-visdrone. `https://huggingface.co/mshamrai/yolov8l-visdrone`, 2023.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pp. 64–69, 2008.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011. doi: 10.1613/jair.3198.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32*, 2019.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10160–10177. PMLR, 2023.

Domen Tabernik, Jon Muhovič, and Danijel Skočaj. Dense center-direction regression for object counting and localization with point supervision. *Pattern Recognition*, 153:110540, 2024. doi: 10.1016/j.patcog.2024.110540.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. In *Computer Vision – ECCV 2024*, pp. 363–398. Springer, 2024.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the 4th Asian Conference on Machine Learning*, pp. 475–490, 2012.

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

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.

Each appendix section is self-contained and follows the order of the main text. Section A gives the implementation details promised in Section 4.4, Section B the diagnostic formula of Section 3.4, and Section C the per-family results that check the choice of feature family. Section D extends the intervention of Section 5.3 to all three datasets, and Section E reports the robustness checks summarised in Section 5.4, one subsection per check, each with a short reading of its table.

A Implementation details

Detectors. We use a single detector family, YOLOv8, across all three benchmarks rather than one architecture per dataset. The claim under test concerns the error structure of detection-based counting, not the ranking of detectors, so holding the architecture fixed removes a confound while keeping the implementation tractable. For GWHD and LIVECell we train YOLOv8s and form deep ensembles by varying the training seed, with five members for GWHD and three for LIVECell. For VisDrone2019-DET we use a single model, the publicly available `mshamrai/yolov8l-visdrone` checkpoint (mshamrai, 2023), a YOLOv8l fine-tuned on VisDrone2019-DET from COCO-pretrained weights and used off the shelf. The single-model case is retained on purpose, so that the comparison includes a detector for which ensemble-based features are undefined.

Matching and counting. Detections below a confidence of 0.25 are discarded. Survivors are matched to ground-truth boxes by the Greedy algorithm, applied independently within each image at an intersection-over-union threshold of 0.5. Both settings are identical across all three benchmarks, and the counts of Section 3.1 follow from that assignment. For ensembles the per-member counts are averaged.

Features. We consider four families computable at inference time from the detector alone. ENS_DISAGREE holds two cross-member agreement statistics, the mean cross-member intersection over union and the fraction of unmatched boxes across members, following the ensemble-spread estimators of Section 2 (Boström et al., 2017; Johansson et al., 2021). It is undefined for the single-model detector, which excludes it from cross-dataset selection. FPN holds channel statistics of the feature pyramid network activations, with 195 dimensions for YOLOv8s and 222 for YOLOv8l, in the spirit of the self-supervised pretext signal of Seedat et al. (2023) discussed in Section 2. CONF holds five statistics of the confidence distribution of the retained detections, namely the mean, the standard deviation, the entropy, the fraction of detections whose confidence lies near the retention threshold, and the number of retained boxes. Low or dispersed confidence is a standard proxy for detector uncertainty, so we include it as a candidate family. NMS_DENSITY holds two statistics of the pre-suppression candidate set, the mean pairwise overlap and the number of candidates, and is intended to capture crowding, a plausible driver of detection error in dense scenes. For each dataset we fix one family in advance of all conformal experiments, choosing the one with the highest $R^2_{\mathrm{resid}}(D)$ among those that pass the permutation test and are defined on all three datasets.

Regressors. Every method is instantiated with two regressor classes so that no conclusion depends on one inductive bias, namely ridge regression with the regularisation strength chosen by cross-validation, and gradient boosting with maximum depth two, subsampling, and early stopping on an internal validation split of the selection fold. A small multilayer perceptron was also evaluated and

is unstable on the high-dimensional FPN family. Two cells diverge outright: on LIVECell it reaches 2411.81 and 2618.27 for C1 and C3, against 183.16 and 184.64 with ridge, a factor of 13.2 and 14.2. The remaining FPN cells stay between 0.98 and 2.26 times ridge, and on the low-dimensional families the ratio stays between 0.97 and 1.31. This is what one expects from selection folds of 244 to 402 images against 195 to 222 fitted input dimensions, and not a general shortfall of the class. We therefore keep it out of the main tables and report it per family in Section C, where its fragility can be read against the two regressors used in the main text.

Protocol. Features are standardised using the selection fold only. Every experiment is repeated over 100 random three-way splits and we report means over them. Every regressor is fitted once on the selection fold and then frozen, so no calibration or test information reaches any fitted object, and the cross-fitted prediction used by C2 is computed within the selection fold. The stability floor ε is fixed on the same fold and before the calibration scores exist, which is what keeps the $m + 1$ conformal scores exchangeable. We additionally record, per split, whether a fitted scale produced near-zero calibration predictions large enough to dominate the floor. This affects B1 to C3 alike and does not affect coverage; the per-cell rates are reported in Section E.1. Two markers recur on every width table from here on: $*$ flags a cell whose empirical coverage significantly deviates from the 90% target (one-sample t-test over splits, $p < 0.05$), and † flags a cell numerically fragile in the sense just defined (near-zero calibration-fold predictions dominating the normalization floor) on at least 10% of seeds. This paragraph is the one place their meaning is stated; individual table captions only use the symbols.

B DIAGNOSTIC SCORE

The score used in Section 3.4 and reported in Table 3 is

$$R^2_{\text{resid}}(Y) \;=\; 1 - \frac{\sum_j \big(r_j - g_F(\phi(x_j))\big)^2}{\sum_j \big(r_j - \bar{r}\big)^2}, \tag{8}$$

where $r_j = Y_j - g_N(n_j)$ is the part of Y that object count misses for example j, $g_F(\phi(x_j))$ is the features' prediction of that missed part, $\bar{r}$ is the mean of r, and j ranges over all examples, each scored out of fold. Both stages are fitted on one half of the data and scored on the other, then swapped, so no model is ever tested on data it was trained on. A score of zero means the features add nothing beyond object count, and negative values are possible because the score is out of fold.

Table 3 reports this score for $Y \in \{D, |E|, b\}$, same features and folds throughout. Its columns are: $R^2(\hat{D})$, the score for $Y = D$; $G = R^2_{\text{resid}}(D) - R^2_{\text{resid}}(|E|)$, the gap between the two targets; ρ, the Spearman correlation between the out-of-fold $\hat{D}$ and the true D; and $R^2(b)$, the same score for $Y = b$, computed only on images with $D > 0$ ($b := 1$ at $D = 0$ is a convention, Eq. (2), not a measurement of cancellation). An asterisk marks a candidate passing the permutation-null test of Section 3.4 on that column's target ($p < 0.05$).

C PER-FAMILY RESULTS

Section 4.4 fixes one feature family per dataset, by a criterion that never looks at an interval width. The four tables below repeat the comparison of Table 4 with the family fixed to the same one on every dataset, so the ranking of the constructions can be read off a family that was not selected. The ordering of Table 4 is not an artefact of that selection. Over the ridge and gradient boosting cells of these four tables, C1 is wider than B1 in seven and C3 never by more than 0.3 percent, which is the trade Section 5.2 describes. Each table also reports the MLP regressor alongside ridge and gradient boosting, so the numerical fragility described in Section A can be compared directly against the two regressors used in the main text. A stricter check lets each method pick its own best family and regressor, restricted to cells whose empirical coverage does not significantly deviate from the 90% target. On GWHD this selects CONF/ridge for both methods, exactly the cell already reported in Table 4 ($+1.2\%$, $p = 1.6 \times 10^{-2}$). The width-blind selection of Section 4.4 was already the best case for B1. On LIVECell the two methods' best cells differ (NMS_DENSITY/ridge for B1

Table 5: Appendix: same comparison as Table 4, feature candidate fixed to CONF on every dataset shown.

	Method	GWHD			LIVECell			VisDrone		
		Ridge	MLP	GBM	Ridge	MLP	GBM	Ridge	MLP	GBM
	B0	15.44*	15.44*	15.44*	249.09	249.09	249.09	79.62*	79.62*	79.62*
	B1	13.78†	14.14*	13.97	192.40†	188.85†	187.97	76.81†	84.06†	72.87
ours	C1	14.36†	**14.85**	14.67	173.76†	**169.51**	**165.30**	74.15†	83.59	**70.60**
ours	C2	13.78†	14.39*	13.96	193.93†	194.77†	188.15	74.69†	**82.19**	73.00
ours	C3	**13.61**†	16.35*	**13.92**	**172.53**†	187.97	170.70	**74.11**†	89.89	71.54
Gain: C3 vs. B1		+1.2% (1.6e-02)	-15.6% (1.6e-07)	+0.4% (n.s.)	+10.3% (1.1e-17)	+0.5% (n.s.)	+9.2% (2.3e-18)	+3.5% (1.3e-06)	-6.9% (8.9e-05)	+1.8% (4.3e-05)

* coverage significantly deviates from the 90% target ($p < 0.05$, one-sample t-test over 100 seeds). † scale numerically fragile on $\geq 10\%$ of seeds (near-zero calibration-fold predictions dominating the normalization floor; most common for the MLP regressor on high-dimensional candidates).

Table 6: Appendix: same comparison as Table 4, feature candidate fixed to FPN on every dataset shown. MLP is included here (unlike Table 4) for completeness; see Appendix A for its fragility at this dimensionality.

	Method	GWHD			LIVECell			VisDrone		
		Ridge	MLP	GBM	Ridge	MLP	GBM	Ridge	MLP	GBM
	B0	15.44*	15.44*	15.44*	249.09	249.09	249.09	79.62*	79.62*	79.62*
	B1	**14.69**†	18.85†	14.59	213.01†	233.87†	194.17	89.16†	107.77†	80.29
ours	C1	15.78†	**18.11**	14.95	**183.16**†	2411.81†	**168.46**	**81.44**†	**95.89**†	76.28
ours	C2	15.34†	18.37†	**14.52**	238.03†	**233.04**†	193.35	87.08†	109.60†	80.06
ours	C3	14.71†	33.31†	14.58	184.64†	2618.27†	177.57	82.12†	133.84†	**75.17**
Gain: C3 vs. B1		-0.1% (n.s.)	-76.7% (2.6e-16)	+0.1% (n.s.)	+13.3% (1.4e-19)	-1019.6% (2.7e-06)	+8.5% (5.2e-23)	+7.9% (2.0e-10)	-24.2% (6.0e-11)	+6.4% (4.2e-14)

* coverage significantly deviates from the 90% target ($p < 0.05$, one-sample t-test over 100 seeds). † scale numerically fragile on $\geq 10\%$ of seeds (near-zero calibration-fold predictions dominating the normalization floor; most common for the MLP regressor on high-dimensional candidates).

at 186.64, CONF/gbm for C3 at 170.70), a $+8.5\%$ improvement for which no paired test across families is available. On VisDrone both best cells coincide (CONF/gbm), reproducing the entry already reported for that cell ($+1.8\%$, $p = 4.3 \times 10^{-5}$).

D MECHANISM ON ALL THREE DATASETS

Section 5.3 shows the confidence-threshold intervention on LIVECell. Figure 2 repeats it on the three datasets. Nothing is re-inferred between points on the grid; the cached detections are reused and only the matching and the threshold-dependent features are recomputed, with the detector and the annotations held fixed. The reading is the same everywhere. As the threshold rises, the survival rate b approaches one, the two factors of Eq. (6) coincide, and C3 converges to B1. The gap between them is largest where b is lowest. This holds on VisDrone as well, where the pre-hoc test of Section 3.4 was null, which is why we read that test as indicative rather than necessary.

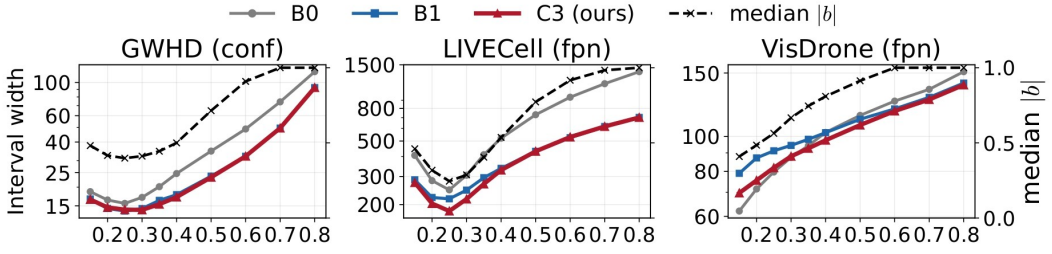

Figure 2: Mean interval width of B0, B1 and C3 against the detector confidence threshold (x-axis), on the three datasets, with the median survival rate b on the right axis.

Table 7: Appendix: same comparison as Table 4, feature candidate fixed to NMS_DENSITY on every dataset shown.

		GWHD			LIVECell			VisDrone		
	Method	Ridge	MLP	GBM	Ridge	MLP	GBM	Ridge	MLP	GBM
	B0	15.44*	15.44*	15.44*	249.09	249.09	249.09	79.62*	79.62*	79.62*
	B1	13.76*	13.71	14.10	186.64†	**181.08**	187.22	**76.26**	77.12	77.93
ours	C1	13.97*†	**13.50**	**13.69**	183.15†	232.32	**180.53**	78.00	**76.18**	**76.35**
ours	C2	**13.67***	14.06	14.10	184.98†	181.94	185.99	77.00	78.99	78.08
ours	C3	13.78*†	14.22	13.92	**178.63**†	234.36	180.95	76.38	79.22	77.07
Gain: C3 vs. B1		-0.1% (n.s.)	-3.7% (3.6e-04)	+1.3% (1.2e-02)	+4.3% (8.4e-04)	-29.4% (8.2e-11)	+3.3% (3.1e-04)	-0.1% (n.s.)	-2.7% (n.s.)	+1.1% (9.8e-03)

* coverage significantly deviates from the 90% target ($p < 0.05$, one-sample t-test over 100 seeds). † scale numerically fragile on $\geq 10\%$ of seeds (near-zero calibration-fold predictions dominating the normalization floor; most common for the MLP regressor on high-dimensional candidates).

Table 8: Appendix: same comparison as Table 4, feature candidate fixed to ENS_DISAGREE on every dataset shown.

		GWHD			LIVECell		
	Method	Ridge	MLP	GBM	Ridge	MLP	GBM
	B0	15.44*	15.44*	15.44*	249.09	249.09	249.09
	B1	15.42	15.49	15.48	221.89	237.50	235.30
ours	C1	15.40	**15.40**	15.84	**212.31**†	**218.71**	**214.95**
ours	C2	15.45	15.68	15.51	223.23	236.65	229.58
ours	C3	**15.30**	16.76*	**15.47**	217.21†	233.06	219.17
Gain: C3 vs. B1		+0.7% (4.8e-06)	-8.2% (2.0e-08)	+0.1% (n.s.)	+2.1% (8.7e-05)	+1.9% (1.9e-02)	+6.9% (3.6e-12)

* coverage significantly deviates from the 90% target ($p < 0.05$, one-sample t-test over 100 seeds). † scale numerically fragile on $\geq 10\%$ of seeds (near-zero calibration-fold predictions dominating the normalization floor; most common for the MLP regressor on high-dimensional candidates).

E ROBUSTNESS CHECKS

Each check summarised in Section 5.4 is given its own subsection, with a short reading of what the table shows and what it does not show.

E.1 COVERAGE PER CELL AND CROSS-REGRESSOR RECAP

Table 10 repeats the main comparison with the empirical coverage of every cell alongside its width, which is the check that the width comparisons of Section 5.2 are made at equal validity rather than bought by under-coverage. Every locally adaptive construction sits at the 90 percent target. The only cells that deviate are B0, on GWHD and VisDrone, at 0.905 and 0.904; B0 fits no scale at all, so this is finite-sample noise around the nominal level and is not attributable to any fitted target.

A second marker records numerical fragility, that is, a split on which the fitted scale $\sigma(x)$ was close to zero on enough of the calibration fold for the stability floor to dominate it. The pattern differs across datasets. On GWHD only C1 to C3 are occasionally affected, and only under one regressor per construction. On LIVECell and VisDrone every locally adaptive construction is affected under at least one regressor, which is consistent with the smaller selection fold relative to the higher-dimensional FPN family used there (Section 4.4). All constructions within a cell share the same percentile definition of ε, so the marker moves together across them. By the scale invariance of Section 3.1, a floor-dominated split changes the achieved width but not the coverage guarantee, so this is a note on estimation variance and not on validity. Where a table instead averages width across regressors for a fixed baseline (Table 9, below), a regressor is dropped from that cell's average when its width exceeds three times the median of the other regressors in the same cell, marked ‡ there; this is the same fragility, read off the per-regressor breakdown of Table 4.

Table 9 averages each construction over the two regressor classes of Table 4. It is descriptive only, and the per-regressor table remains the source of truth; we report it because the ordering of the constructions is easier to read from one column per dataset.

Table 9: Recap of Table 4, mean width averaged over Ridge/GBM, per baseline and dataset (descriptive; the per-regressor table is the source of truth). Bold: lowest mean width per column. ‡: a regressor was excluded from the average (Appendix E.1).

	Method	GWHD	LIVECell	VisDrone
	B0	15.44	249.09	79.62
	B1	13.88	203.59	84.72
ours	C1	14.51	**175.81**	78.86
ours	C2	13.87	215.69	83.57
ours	C3	**13.77**	181.11	**78.65**

‡ at least one regressor excluded from this cell's average (numerically fragile; see caption).

Table 10: Mean interval width and empirical coverage at the 90% target, five constructions (Table 1), two regressors, three datasets. Gain: relative width reduction of C3 vs. B1 (paired t-test p-value). Regressor-averaged view with fragility markers: next table.

		Ridge		Gradient boosting	
		w	cov.	w	cov.
GWHD (features: CONF*)*					
	B0 marginal	15.44	0.905*	15.44	0.905*
	B1 residual	13.78	0.903	13.97	0.901
ours	C1 magnitude	14.36	0.901	14.67	0.900
ours	C2 concat.	13.78	0.903	13.96	0.901
ours	C3 factorised	13.61	0.903	13.92	0.902
Gain (C3 vs. B1)		$+1.2\%$ ($p = 1.6 \times 10^{-2}$)		$+0.4\%$ ($p = 4.9 \times 10^{-1}$)	
LIVECell (features: FPN*)*					
	B0 marginal	249.09	0.902	249.09	0.902
	B1 residual	213.01	0.899	194.17	0.899
ours	C1 magnitude	183.16	0.897	168.46	0.898
ours	C2 concat.	238.03	0.901	193.35	0.898
ours	C3 factorised	184.64	0.900	177.57	0.900
Gain (C3 vs. B1)		$+13.3\%$ ($p = 1.4 \times 10^{-19}$)		$+8.5\%$ ($p = 5.2 \times 10^{-23}$)	
VisDrone (features: FPN*)*					
	B0 marginal	79.62	0.904*	79.62	0.904*
	B1 residual	89.16	0.898	80.29	0.900
ours	C1 magnitude	81.44	0.899	76.28	0.900
ours	C2 concat.	87.08	0.899	80.06	0.900
ours	C3 factorised	82.12	0.899	75.17	0.900
Gain (C3 vs. B1)		$+7.9\%$ ($p = 2.0 \times 10^{-10}$)		$+6.4\%$ ($p = 4.2 \times 10^{-14}$)	

E.2 MONDRIAN CALIBRATION

Mondrian calibration (Vovk, 2012) computes a separate conformal quantile within each of five quantile bins of the predicted count, so coverage holds within each bin at the cost of fewer calibration points per bin. It acts on how the quantile is calibrated, downstream of the scale, whereas the change of target acts upstream of it, so the two compose. Table 11 evaluates B1 and C3 under both regimes, on the same splits.

All intervals widen under Mondrian calibration, as expected from calibrating on fewer points per cell. The advantage of C3 survives on LIVECell and grows on VisDrone. On GWHD it disappears, which is consistent with GWHD's marginal gain already being the smallest of the three and only marginally significant. The change of target therefore does not depend on marginal calibration, and a practitioner already using group-conditional calibration can adopt it without giving that up.

Table 11: C3 vs. B1 under marginal and 5-bin Mondrian (Vovk, 2012) calibration on $\hat{n}$, same splits as Table 4. Gain rows: paired width reduction, C3 vs. B1, same regime.

		GWHD		LIVECell		VisDrone	
	Method	Ridge	GBM	Ridge	GBM	Ridge	GBM
	B1 marginal	13.78	13.97	213.01	194.17	89.16	80.29
	B1 Mondrian	14.20*	14.44	242.90*	231.98*	96.40	87.55*
ours	C3 marginal	13.61	13.92	184.64	177.57	82.12	75.17
ours	C3 Mondrian	14.24*	14.45*	220.81*	214.42*	87.70*	81.57*
Gain (marginal)		+1.2% (1.6e-02)	+0.4% (n.s.)	+13.3% (1.4e-19)	+8.5% (5.2e-23)	+7.9% (2.0e-10)	+6.4% (4.2e-14)
Gain (Mondrian)		-0.3% (n.s.)	-0.1% (n.s.)	+9.1% (3.7e-13)	+7.6% (1.8e-11)	+9.0% (4.0e-13)	+6.8% (1.1e-13)

* coverage significantly deviates from the 90% target ($p < 0.05$).

E.3 Stability floor sweep

The stability floor ε is the one free quantity in the normalisation, so we sweep it rather than tune it. Figure 3 reports the paired width gain of C3 over B1 at five percentiles of the positive scale values, with ridge and thirty splits. The gain stays positive across the sweep on VisDrone and LIVECell, and on GWHD it is positive through the default tenth percentile and slightly negative beyond it. The sweep uses thirty splits where Table 4 uses one hundred, which is why the value at the default percentile differs slightly from the one reported there. We read this sensitivity as a property of the normalisation protocol rather than of the mechanism, since the floor applies identically to both constructions.

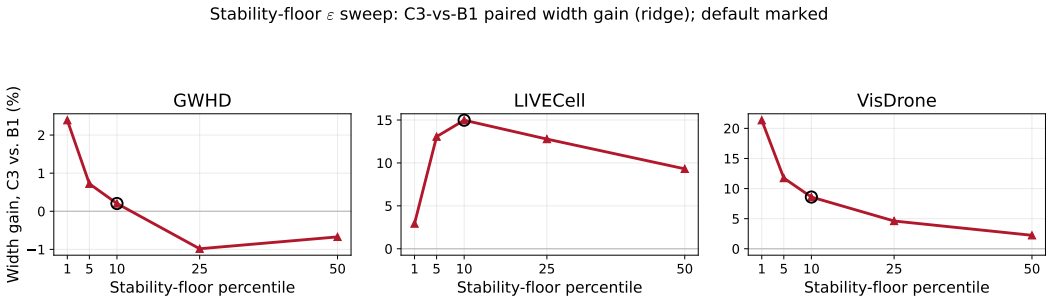

Figure 3: Paired width gain of C3 over B1 against the stability-floor percentile, on the three datasets, with ridge and thirty splits. The default percentile used everywhere else in the paper is marked.

E.4 Equal-width coverage and image-level overlap

Table 4 compares widths at a fixed coverage target. The same comparison can be read the other way round, by choosing C3's miscoverage per split so that its mean width matches B1's width at $\alpha = 0.1$, and asking what coverage it then reaches. Table 12 reports it; B1 is at 90 percent by construction at that width, and C3 sits above it in every cell. A C3 coverage meaningfully above 90 percent here means C3 converts an equal width budget into higher coverage, rather than merely being narrower at matched coverage as Table 4 already shows.

That gain is not a free addition. Table 13 splits the test images into those covered by both constructions, by one only, or by neither. In five of six cells the images covered only by B1 form a non-trivial fraction, so C3 trades which images it covers rather than covering a superset of B1's set. Figure 4 shows these categories on the one cell where C3 is close to a strict superset, projected onto the first two principal components of the test-fold features.

E.5 Coverage on the highest bin of the latent magnitude

A natural question is whether the change of target also repairs coverage on the images where cancellation was in fact severe. It cannot be answered by conditioning on D, and Table 14 is reported so that the negative result is visible rather than omitted. Conditioning on D mechanically over-selects

Table 12: Coverage C3 reaches when its α is chosen (per seed, by bisection) to match B1's mean width at the 90% target; B1 is at 90% by construction at that width (Appendix E.4). "—": regressor not evaluated for that dataset (Table 4).

Regressor	GWHD	LIVECell	VisDrone
Ridge	0.907	0.914	0.911
GBM	0.905	0.910	0.913

B1's coverage at this width is 90% by construction (not shown as a row); values above compare directly against that 90% reference.

Table 13: Overlap, at the widths of Table 12, between images B1 covers and images C3 covers, both, B1 only, C3 only, neither (mean over 100 seeds, same seeds/splits as Table 12).

Dataset	Regressor	Both	B1 only	C3 only	Neither
GWHD (conf)	Ridge	89.0%	1.3%	1.7%	8.0%
GWHD (conf)	GBM	88.1%	2.0%	2.4%	7.5%
LIVECell (fpn)	Ridge	88.9%	1.0%	2.4%	7.6%
LIVECell (fpn)	GBM	88.7%	1.1%	2.3%	7.9%
VisDrone (fpn)	Ridge	88.4%	1.4%	2.7%	7.5%
VisDrone (fpn)	GBM	88.9%	1.1%	2.4%	7.6%

large residuals, since $D \geq |E|$ always by Eq. (5), so every construction under-covers on the highest bin, B0 included. More fundamentally, no scale that is a function of $\phi(x)$ can know in advance whether cancellation will occur on a given test image, so a label-conditional bin cannot demonstrate that kind of repair whatever the outcome.

What the table does show is the ordering. All five constructions under-cover on this bin, B0 most; B1 roughly halves B0's gap, so the standard adaptive scale is not the quantity at fault here. C3 does not improve on B1 on this bin, and the last two columns give the paired difference and its p-value over the same seeds. Its gain is one of efficiency, not of conditional coverage, and we state it as such.

E.6 Density dependence

The gain ordering across datasets tracks mean latent error mass, which also tracks object density, so density is a candidate explanation for when the method helps. Across datasets the two cannot be separated, since the benchmarks differ in detector, resolution and annotation protocol at once. Table 15 therefore tests it within each dataset, where those three are fixed and only density varies, by computing the gain per quantile bin of the predicted count. It holds on VisDrone and is absent on GWHD and LIVECell, so object density does not generally explain when the method helps.

E.7 Conformalised quantile regression

Conformalised quantile regression (Romano et al., 2019) replaces the symmetric interval of Eq. (4) by two conditional quantiles of the signed error, fitted on the selection fold and conformalised on the calibration fold. It is the strongest locally adaptive construction in the related work of Section 2, and it does not fit a scale at all, so it gives the change of target a second construction to act in, built on a different principle from the one the rest of the paper uses.

Table 16 crosses the two design choices on the same splits, regressors and floor as Table 4. The first choice is the interval construction, symmetric or CQR. The second is the scale dividing the signed error before anything is fitted, none, σ_{B1} or σ_{C3}. The three CQR columns run the same code and differ only in that divisor, so the comparison that tests the claim of this paper is horizontal and inside one construction, σ_{B1} against σ_{C3} with everything else held fixed; these are the last two columns. Under the ridge rows the scale is fitted by ridge regression and the two conditional quantiles are linear with an L_1 penalty, chosen inside each split by three-fold cross-validated pinball loss on the selection fold; under the gradient boosting rows both are gradient boosting, the quantiles under the pinball loss. Widths are not comparable *across* constructions in that sense, because the shape of the interval is a separate design choice and one construction can be narrower than another for reasons

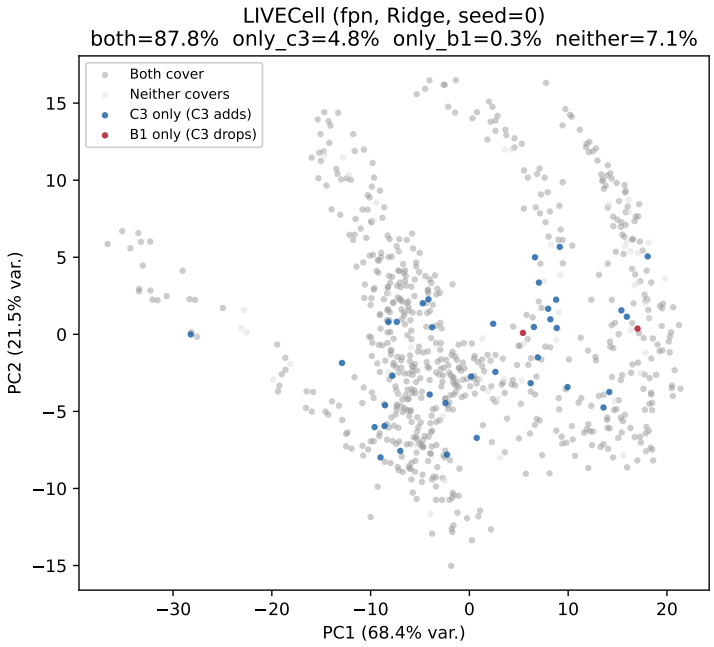

Figure 4: Overlap categories on LIVECell with ridge and a single split, projected onto two principal components of the test-fold features. Single-split percentages differ from the seed-averaged values in Table 13.

Table 14: Empirical coverage minus the 90% target (pp) on the highest quantile bin of D (Appendix E.5), mean (sd) over 20 seeds, ridge, one fixed low-dimensional feature family per dataset. Last two columns: paired C3−B1 difference and its p-value.

Dataset	B0	B1	C1	C2	C3	C3−B1	paired p
GWHD	-26.52 (7.38)	-11.03 (6.54)	-7.20 (5.09)	-9.81 (6.24)	-14.41 (7.29)	-3.37	2.6e-03
LIVECell	-34.57 (7.04)	-17.95 (3.88)	-15.81 (3.99)	-18.09 (3.85)	-18.83 (4.85)	-0.88	n.s.
VisDrone	-36.22 (6.07)	-16.47 (7.46)	-18.68 (5.50)	-20.59 (6.34)	-19.87 (5.08)	-3.41	4.4e-04

Cell format: mean (sd) over 20 seeds, ridge regressor, one fixed low-dimensional feature family per dataset (not necessarily the family selected for Table 4).

that have nothing to do with cancellation. Both feature families defined on all three datasets are reported, so that the reading does not rest on the family selected in Section 4.4. Two "target effect" columns report, for the symmetric and CQR constructions respectively, the width reduction from σ_{B1} to σ_{C3} within that construction (positive: σ_{C3} narrower), with the fraction of splits on which it is narrower in parentheses – the column-level readout of the same horizontal comparison. On each dataset's selected feature family (Section 4.4) the two symmetric columns reproduce Table 4's B1 and C3 rows exactly; the other feature family is reported here for the first time.

Two things follow. First, the change of target is not tied to the symmetric score. The same move that narrows the symmetric interval in Table 4 also narrows the conformalised quantile interval on the majority of rows, and on the rows where it does not we report the loss rather than dropping the family. Second, the construction stays free. Every fitted object is frozen on the selection fold and every variant is conformalised on the calibration fold, so all five columns carry the same marginal guarantee, and a practitioner already using CQR can change the target without giving up the interval shape they chose. This is the same composability that Section E.2 reports for group-conditional calibration; the change acts upstream of both.

Table 15: C3-vs-B1 paired width gain (%) by density bin (bin 0 = lowest $\hat{n}$; same bins as Table 11), Ridge only. Spearman ρ: bin rank vs. gain across bins.

Dataset	Bin 0	Bin 1	Bin 2	Bin 3	Bin 4	Spearman ρ
GWHD	-1.1%	+0.7%	-1.3%	+0.6%	-1.0%	+0.10
LIVECell	+1.1%	+14.3%	+12.8%	+13.5%	+3.6%	+0.10
VisDrone	-3.2%	+0.5%	+5.6%	+11.1%	+11.4%	+1.00*

Cell values: C3-vs-B1 paired width gain (%) in that density bin, mean over 100 seeds. * Spearman $p < 0.05$.

Table 16: Symmetric vs. CQR intervals, crossed with the scale dividing the signed error before fitting (none, σ_{B1}, σ_{C3}); same splits, regressors and floor as Table 4 (Appendix E.7 for what this comparison tests). Target effect: width reduction from σ_{B1} to σ_{C3} within each construction (fraction of splits narrower in parentheses). Bold: narrowest per row (deployment pick, not the claim).

		Symmetric		CQR			Target effect	
Dataset	Reg.	σ_{B1}	σ_{C3}	none	σ_{B1}	σ_{C3}	symmetric	CQR
CONF features								
GWHD	Ridge	13.78	13.61	12.49	**12.01**	12.10	+1.2% (0.61)	-0.7% (0.48)
	GBM	13.97	13.92	**11.98**	12.17	12.25	+0.4% (0.50)	-0.6% (0.44)
LIVECell	Ridge	192.40	**172.53**	215.51	218.68	238.21	+10.3% (0.89)	-8.9% (0.15)
	GBM	187.97	**170.70**	187.85	187.81	182.96	+9.2% (0.88)	+2.6% (0.69)
VisDrone	Ridge	76.81	74.11	59.35	60.14	**58.95**	+3.5% (0.73)	+2.0% (0.55)
	GBM	72.87	71.54	53.78	53.31	**52.75**	+1.8% (0.59)	+1.1% (0.66)
FPN features								
GWHD	Ridge	14.69	14.71	**13.35**	13.75	13.57	-0.1% (0.43)	+1.3% (0.48)
	GBM	14.59	14.58	**13.05**	13.34	13.41	+0.1% (0.47)	-0.5% (0.48)
LIVECell	Ridge	213.01	**184.64**	224.98	217.42	214.86	+13.3% (0.89)	+1.2% (0.56)
	GBM	194.17	**177.57**	185.85	186.41	180.72	+8.5% (0.89)	+3.1% (0.70)
VisDrone	Ridge	89.16	82.12	**62.84**	69.72	63.43	+7.9% (0.74)	+9.0% (0.85)
	GBM	80.29	75.17	58.08	57.71	**53.93**	+6.4% (0.84)	+6.6% (0.80)

Empirical coverage lies in $[0.897, 0.907]$ across every cell.

The bold entry on each row marks the narrowest of the five, which answers the deployment question and not the claim. It can be won by plain CQR, for reasons unrelated to cancellation, and we mark it as such so that the two readings are not confused.