Marginally Useful

Formalizing the Information Gap in Conformal Prediction

Peter Cotton*

June 3, 2026

Abstract

Conformal prediction gives finite-sample, distribution-free *marginal* coverage for a set. The guarantee is real, and it is often misread as evidence of forecast quality. We separate the two with one decomposition, the residual-information gap: for a fixed location predictor and a single-shape residual predictive system, the log-score regret relative to the oracle is exactly the mutual information $I(R; X)$ between the residual and the input. Conformalization re-levels coverage but cannot touch this quantity, because it is a property of the predictor's shape class and not of calibration; no recalibration that ignores X reduces it within that class. The familiar cautions about conformal prediction follow as context: marginal coverage is not conditional, validity is insensitive to sharpness, and the guarantee needs exchangeability.

1 Introduction

A practitioner maintains a probabilistic forecaster and is told, repeatedly, that conformal prediction is the principled, distribution-free way to "add uncertainty."[1] They bolt it on. The number they actually report (a proper score such as the predictive log-likelihood) does not improve. The natural diagnosis is a tuning failure. The real one is a *category error*: conformal prediction is a method for *certifying the coverage of a set*, and it is being asked to *estimate a distribution*. These are different kinds of object, graded by different instruments, and no amount of tuning turns one into the other. None of this is a criticism of the method; it is a guide to where the method belongs.

None of this impugns the mathematics. Given exchangeable data, a fitted predictor, and a nonconformity score, split conformal prediction returns a set $C_\alpha(x)$ with

$$\mathbb{P}\big(Y_{n+1} \in C_\alpha(X_{n+1})\big) \;\geq\; 1 - \alpha, \tag{1}$$

finite-sample and with no assumption on the data distribution (Vovk et al., 2005; Lei et al., 2018; Angelopoulos and Bates, 2023). The trouble is entirely in the reading of (1). The probability averages over X_{n+1} as well as Y_{n+1}: it is a statement about the *average* future input, not about

*Microprediction. peter.cotton@microprediction.com. All figures are produced by the accompanying script `figures.py` and the experiments by the benchmark scripts in the same repository; interactive versions are also available online, but the paper does not depend on them.

[1]The reading is pervasive. Applied and survey writing routinely calls the marginal-coverage guarantee "well-calibrated" or "reliable" uncertainty quantification (Bellotti and Zhao, 2025), and tutorials present conformal prediction as the way to "add uncertainty" (tds) or to "predict full probability distributions" (Manokhin). The methodological literature is more careful, separating *validity* from *efficiency*, and a critique strand makes the present point independently (Mehrtens et al., 2025; Min et al., 2026). Our quarrel is with the conflation, not with conformal prediction.

the one in front of you. It is *marginal* coverage, and as the foundational analyses warned from the start, "a good estimator must satisfy something more than marginal coverage" (Lei and Wasserman, 2014).

Contributions and structure. The one new result is in Section 5: for a single-shape residual predictive system, the log-score regret to the oracle equals the mutual information $I(R; X)$ between the residual and the input, and no recalibration that ignores X can reduce it (Proposition 3). We call this the residual-information gap.

The other sections are background and consequences. The impossibility of distribution-free conditional coverage (Section 3) is due to Lei and Wasserman (2014) and Foygel Barber et al. (2021). The orthogonality of coverage and log-score (Section 4) is folklore. The coverage–score plane (Section 6) is a way of presenting the result. Section 7 covers exchangeability and time series, Section 8 the case where coverage is the objective, and Section 9 the relation to the modern conformal literature.

2 Setup: split conformal and the marginal guarantee

Let $(X_i, Y_i)_{i=1}^{n+1}$ be exchangeable. Fix a point predictor $\hat{\mu}$ trained on a disjoint set, and a *nonconformity score* $s(x, y)$; the canonical choice in regression is the absolute residual $s(x, y) = |y - \hat{\mu}(x)|$. Compute calibration scores $s_i = s(X_i, Y_i)$ for $i = 1, \ldots, n$ and let

$$\hat{q} \;=\; s_{(k)}, \qquad k = \big\lceil (n+1)(1-\alpha) \big\rceil, \tag{2}$$

the k-th order statistic (with $\hat{q} = +\infty$ when $k > n$). The split-conformal set is

$$C_\alpha(x) = \{y :\; s(x, y) \le \hat{q}\} \;=\; [\hat{\mu}(x) - \hat{q},\; \hat{\mu}(x) + \hat{q}] \tag{3}$$

for the absolute-residual score. Exchangeability of the augmented scores makes the rank of s_{n+1} uniform, which yields (1) (Vovk et al., 2005; Lei et al., 2018). The construction is simple and the guarantee is real, and one readily verifies that empirical coverage tracks $1 - \alpha$ as the sample size, noise level, and α are varied. None of this is in dispute; the question is what it is taken to mean.

3 Three known limits of the marginal guarantee

None of the three facts below is new. We restate them, with attribution, to fix notation and to mark where the contribution of Section 5 departs from what is already established.

(i) Marginal is not conditional, and conditional is impossible for free. One would prefer *conditional* coverage, $\mathbb{P}\big(Y \in C(x) \,|\, X = x\big) \ge 1 - \alpha$ for (almost) every x. Distribution-free, this cannot be bought. Lei and Wasserman (2014) show (their Lemma 1) that any band with non-trivial finite-sample conditional validity has *infinite* expected length at almost every point of a continuous distribution. Foygel Barber et al. (2021) restate and sharpen this: a conditionally valid C_n has $\mathbb{E}\,\mathrm{leb}(C_n(x)) = \infty$ at almost all non-atomic x (their Prop. 2.2), and even the relaxed demand of coverage on every subgroup of probability $\ge \delta$ "is impossible to attain beyond the trivial solution" of inflating the marginal level to $1 - \alpha\delta$ (their Thm. 3.1). These are not loose bounds to be tightened; they are the reason the gap cannot be closed for free, and both surface as concrete finite-sample facts. The first is illustrated in Figure 1: localizing split conformal to recover per-x coverage shrinks each cell's calibration count n_b until $\lceil (n_b + 1)(1 - \alpha) \rceil > n_b$, at which point (2) gives $\hat{q} = +\infty$ and

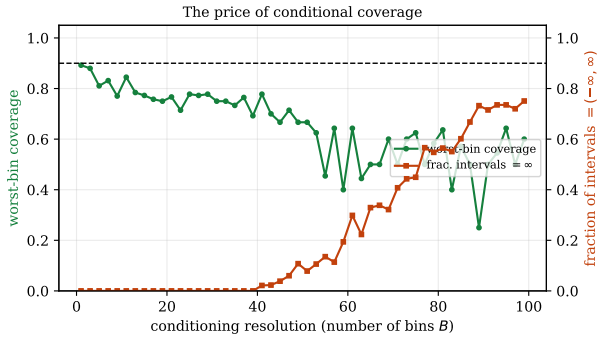

Figure 1: The price of conditional coverage (Lei and Wasserman, 2014). As the conditioning resolution B grows, per-cell calibration sets shrink and the only valid cell interval becomes $(-\infty, \infty)$; uniform (worst-bin) coverage is bought only as the fraction of infinite-length intervals rises.

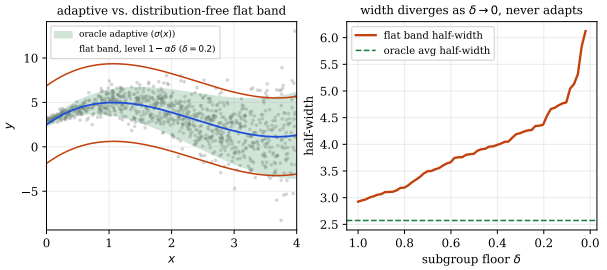

Figure 2: Subgroup coverage buys only a wider band (Foygel Barber et al., 2021). The distribution-free protection of all δ-subgroups is a flat, non-adaptive band at level $1 - \alpha\delta$ whose width diverges as $\delta \to 0$; the adaptive oracle that uses $\sigma(x)$ is far tighter but requires the distributional assumption.

the only valid cell interval is $(-\infty, \infty)$; uniform coverage is recovered only as the fraction of infinite-length intervals grows. The second is Figure 2: the distribution-free protection of all δ-subgroups is exactly the flat band at level $1 - \alpha\delta$, whose width diverges as $\delta \to 0$ while never adapting to x, at a multiplicative width cost over an assumption-driven adaptive oracle.

So the guarantee you can have distribution-free is the marginal one, and a marginally valid band "tends to overestimate the set when x is in the high density area and to underestimate for low density x" (Lei and Wasserman, 2014): it over-covers the easy inputs and under-covers the hard ones, silent by construction about the cases the model was built to get right. Figure 3 exhibits exactly this on heteroscedastic data: 90% overall, near-100% where the noise is small, well below target where it is large.

The conformal literature already separates *validity* from *efficiency*; the point here is only that the marginal-coverage certificate does not measure distributional quality.

(ii) Validity is trivially satisfiable. A predictor that returns the whole outcome space with probability $1 - \alpha$ and the empty set otherwise satisfies (1) exactly and conveys nothing. More practically, the *same* coverage attaches to an excellent $\hat{\mu}$ and to a deliberately terrible one; the latter merely yields a wider $\hat{q}$ in (2). Marginal validity is therefore not evidence of a good method; it certifies the fence, not the cattle. The next section makes this precise.

(iii) Without exchangeability, even the marginal guarantee fails. Equation (1) rests on exchangeability of the augmented sample. Under covariate shift, label shift, or temporal dependence it does not hold, and the patches that restore something do so by weakening what is guaranteed (Section 7).

4 Coverage and score are orthogonal

We first make "coverage is not a measure of forecast quality" a precise, constructive fact. Separate the two things a forecaster reports: a point predictor $\hat{\mu}(x)$ and a full predictive density $f(\cdot \mid x)$. The log-score grades f, that is, its spread and shape, its sharpness. The canonical residual score

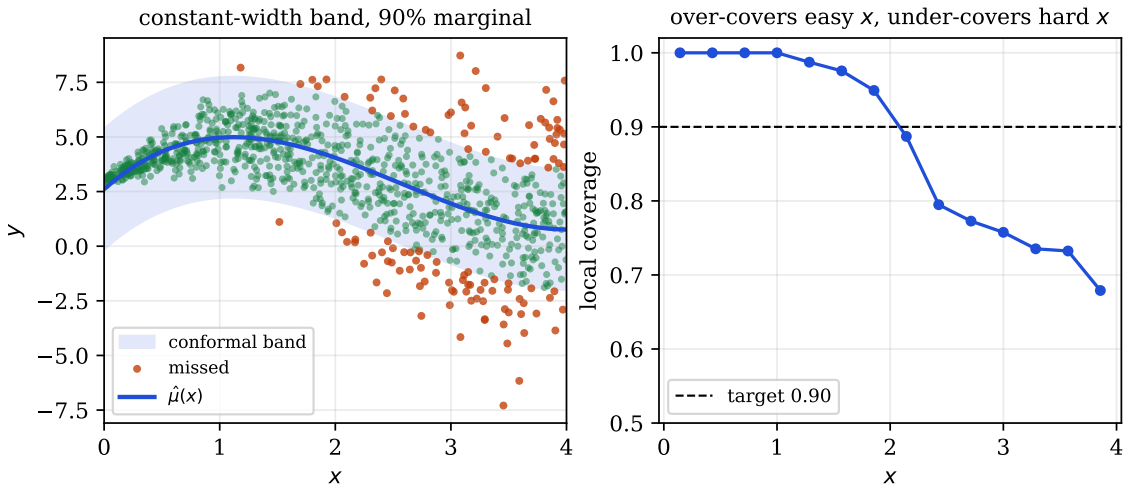

Figure 3: Marginal is not conditional. A constant-width split-conformal band (left) attains 90% *marginal* coverage on heteroscedastic data, but its *local* coverage (right) runs near 100% where the noise is small and well below target where it is large: it over-covers the easy inputs and under-covers the hard ones.

$s(x, y) = |y - \hat{\mu}(x)|$ uses only $\hat{\mu}$, the *location* of the forecast; it never sees the spread or shape of f. So the conformal set, and its coverage, are invariant to exactly the part of f that the log-score rewards. (This is specific to such scores; it does not apply to density- or quantile-based scores, which read f directly.)

Proposition 1 (Residual-score coverage does not constrain the log-score)**.** *The split-conformal set* (3) *is a measurable function of the nonconformity scores* $\{s_i\}$ *and of* $s(x, \cdot)$*; it is not a functional of the predictive density* f *unless* f *is explicitly used in the score. Consequently two forecasters with the same location* $\hat{\mu}$ *(hence, for the residual score, the same score function and calibration scores) but different predictive densities* f *have identical conformal sets, and identical marginal coverage at every level* α*, while their expected log-scores* $\mathbb{E}[\log f(Y \mid X)]$ *may differ by an arbitrary amount.*

Proof. The first claim is immediate from (2)–(3): $\hat{q}$ is an order statistic of $\{s_i\}$ and $C_\alpha(x) = \{y : s(x, y) \le \hat{q}\}$, so two forecasters agreeing on s and on $\{s_i\}$ produce the same set for every x, α, and (1) concerns only that set.

For the gap, take X degenerate (a point mass at an arbitrary x_0, i.e. a single repeated input, so there is nothing to vary across) and $Y \sim N(0, 1)$. Both forecasters share the location $\hat{\mu} \equiv 0$, so both use the absolute-residual score $s(x, y) = |y|$, which depends on $\hat{\mu}$ but not on the spread of the predictive density. They differ only in that claimed density, $f_\sigma = N(0, \sigma^2)$: for any σ the scores are $|Y_i|$, independent of σ, so the conformal set is the fixed interval $[-\hat{q}, \hat{q}]$ with $\hat{q}$ the empirical $(1 - \alpha)$-quantile of $|Y|$. Yet

$$\mathbb{E}[\log f_\sigma(Y)] \;=\; -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{\mathbb{E}[Y^2]}{2\sigma^2} \;=\; -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \;\xrightarrow[\sigma \to 0^+]{}\; -\infty,$$

and likewise $\to -\infty$ as $\sigma \to \infty$. Hence for any $M > 0$ there are σ_1, σ_2 with identical conformal sets (identical marginal coverage at every α) yet log-scores differing by more than M. □

Corollary 1. *For a residual-score conformal predictor, marginal validity certifies an order statistic of residual magnitudes. It does not identify, bound, or rank the scale, shape, or sharpness of any*

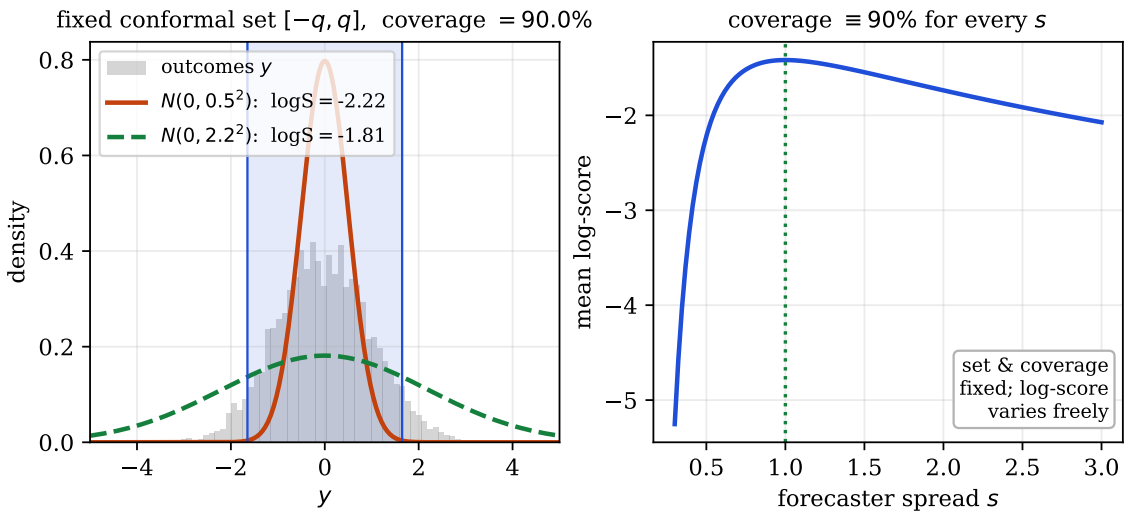

Figure 4: Coverage does not constrain the log-score (Proposition 1). With a residual score the conformal set $[-q, q]$ is fixed by the outcomes alone (left), so two predictive densities with very different log-scores share the *same* set and the same 90% coverage. Sweeping the forecaster's spread s (right) moves the log-score freely while coverage stays pinned at 90%.

predictive density that was not used to construct the score; in particular, no inference about forecast quality (as graded by a strictly proper score) can be drawn from coverage alone.

This is the formal version of a point several authors have made informally: that conformal validity "almost invites you to use garbage prediction functions" (Recht, 2024). Figure 4 shows it: the claimed density's spread slides freely while the conformal interval and its 90% coverage do not move, because the spread never entered (2).

5 The residual-information gap

Forecast quality decomposes into *calibration* and *sharpness*, and the discipline's stated goal is to maximize sharpness subject to calibration, assessed by a proper scoring rule (Gneiting et al., 2007; Gneiting and Raftery, 2007). Within the fixed-location, single-shape residual class, signed CPS is log-score optimal; the loss is the class restriction, not the conformal calibration step.

Conformal predictive systems. A conformal predictor can be made to output a distribution rather than a set, the *conformal predictive system* (CPS) of Vovk et al. (2019), built from *signed* residual scores. It is the mode in which conformal prediction estimates a full distribution, so it is the form the criticism must address. We must, however, distinguish it from the more common practice of reading the nested symmetric intervals of ordinary *absolute*-residual split conformal as if they were a distribution; the two re-level different objects, and the distinction matters for the gap below.

Proposition 2 (Re-leveling, two cases)**.** *Fix a location predictor $\hat{\mu}$ with residual $R = Y - \hat{\mu}(X)$ of marginal density g and CDF G. In the large-calibration limit (under exchangeability):*[2]

[2]The log-score comparison is for the limiting residual density, or equivalently a smoothed CPS density: a finite empirical CPS is a valid predictive distribution but not generally a Lebesgue density, so the density/log-score identity is an asymptotic (or smoothed-density) statement. The finite-sample guarantee is coverage.

(A) Signed-residual CPS. The CPS predictive distribution is $\Pi_x(y) = G\big(y - \hat{\mu}(x)\big)$ *(the calibration ECDF* $\widehat{G}_n \to G$ *by Glivenko–Cantelli), with density* $\pi_x(y) = g\big(y - \hat{\mu}(x)\big)$*: the base location forecast re-leveled by the marginal signed-residual law.*

(B) Absolute-residual intervals. Ordinary split conformal at level α *returns the symmetric interval* $\hat{\mu}(x) \pm \hat{q}_\alpha$*, with* $\hat{q}_\alpha$ *the* $(1-\alpha)$*-quantile of* $|R|$*. Read as a predictive distribution (by sweeping* α*) it is symmetric about* $\hat{\mu}(x)$ *with* $|\cdot|$ *distributed as* $|R|$*; its density is the symmetrized marginal residual law* $h_{\mathrm{sym}}(z) = \frac{1}{2}\big(g(z) + g(-z)\big)$*. This is not an additional conformal guarantee; it is the density implicitly obtained if the nested absolute-residual intervals are read as central credible intervals.*

Proof sketch. (A) The CPS at level α returns endpoints equal to the signed-residual order statistics; sweeping α traces $\hat{\mu}(x)$ plus the residual quantile function, i.e. $\widehat{G}_n^{-1}$ shifted by $\hat{\mu}(x)$, so $\Pi_x(y) = \widehat{G}_n(y - \hat{\mu}(x))$; take limits (Vovk et al., 2019). (B) The intervals depend on the data only through $|R|$ and are symmetric about $\hat{\mu}(x)$; the unique symmetric density whose absolute value has the law of $|R|$ (with signed marginal g) is $h_{\mathrm{sym}}(z) = \frac{1}{2}(g(z) + g(-z))$. □

In both cases the estimation was done by the base model and the residual sample; conformal supplies only the leveling. We now quantify the cost. Write the true conditional residual density as $r(\cdot \,|\, x)$, so the oracle density is $q^\star(y \,|\, x) = r\big(y - \hat{\mu}(x) \,|\, x\big)$, and let $\bar{r} = \mathbb{E}_X\, r(\cdot \,|\, X) = g$ be the marginal residual density.

Proposition 3 (Residual-information gap)**.** *Fix the location predictor* $\hat{\mu}$ *and write* $I(R;X) = \mathbb{E}_X \,\mathrm{KL}\big(r(\cdot \,|\, X) \,\|\, \bar{r}\big) \geq 0$*, the mutual information between the residual and the input. In the large-calibration limit of Proposition 2 (so that the signed-CPS shape is* $\bar{r}$ *and the absolute-residual shape is* h_{sym}*), and assuming the relevant densities and information quantities are finite, the expected log-score regret relative to the oracle* $q^\star$ *is*

$$\textit{(A) signed CPS:}\quad \mathbb{E}[\log q^\star(Y \,|\, X)] - \mathbb{E}[\log \pi_X(Y)] \;=\; I(R;X), \tag{4}$$

$$\textit{(B) absolute intervals:}\quad \mathbb{E}[\log q^\star(Y \,|\, X)] - \mathbb{E}[\log h_{\mathrm{sym}}(R)] \;=\; I(R;X) + \mathrm{KL}\big(\bar{r} \,\|\, h_{\mathrm{sym}}\big). \tag{5}$$

Both are ≥ 0*. The term* $I(R;X)$ *vanishes iff* $R \perp X$*; the extra term* $\mathrm{KL}(\bar{r} \,\|\, h_{\mathrm{sym}})$ *vanishes iff the marginal residual law is symmetric. Among all single-shape forecasters* $y \mapsto h(y - \hat{\mu}(x))$*,* $\mathbb{E}[\log h(R)]$ *is maximized at* $h = \bar{r}$*; the signed CPS attains this optimum. Hence no recalibration that ignores* X *can reduce* $I(R;X)$*: it can at best replace the shape by* $\bar{r}$*, removing the skewness term in* (5) *and reducing case (B) to case (A). Reducing* $I(R;X)$ *itself requires conditioning on* X*.*

Proof. Write H for the residual law with density h, and let $\mathcal{R}(h)$ be the expected log-score regret of the single-shape forecaster $y \mapsto h(y - \hat{\mu}(x))$ relative to the oracle. The oracle scores Y by $\log q^\star(Y \,|\, X) = \log r(R \,|\, X)$ and this forecaster by $\log h(R)$; the joint $P_{X,R}$ has density $p(x)\, r(\rho \,|\, x)$ and the product reference $P_X \otimes H$ has density $p(x)\, h(\rho)$, so

$$\mathcal{R}(h) = \mathbb{E}\left[\log \frac{r(R \,|\, X)}{h(R)}\right] = \mathrm{KL}\big(P_{X,R} \,\big\|\, P_X \otimes H\big) = I(R;X) + \mathrm{KL}(\bar{r} \,\|\, h), \tag{6}$$

the last step the chain rule for relative entropy along a product reference (the R-marginal of $P_{X,R}$ is $\bar{r}$, and $I(R;X) = \mathrm{KL}(P_{X,R} \,\|\, P_X P_R)$). Taking $h = \bar{r}$ kills the second term and gives (4); taking $h = h_{\mathrm{sym}}$ gives (5). The second term is ≥ 0 and is minimized at $h = \bar{r}$ (Gibbs), so no X-blind shape improves on $I(R;X)$, and $I(R;X) = \mathrm{KL}(P_{X,R} \,\|\, P_X P_R)$ is reduced only by conditioning on X. □

Remark 1 (What this says, and does not say)**.** The right-hand side of (4) is exactly the information about the residual distribution that remains in X after the location predictor is fixed. Calling this the residual-information gap makes the content of "blind to conditional residual shape" precise. Three caveats. (1) *Conformal is not dominated within its class.* The signed CPS is log-score optimal among single-shape location forecasters; the cost is the restriction, not the calibration step. The absolute-residual interval system is the one that pays the extra $\mathrm{KL}(\bar{r} \,\|\, h_{\mathrm{sym}})$, for discarding sign information; since $h_{\mathrm{sym}} \geq \frac{1}{2}\bar{r}$ this penalty is at most $\log 2$, one bit, so the structural loss is $I(R;X)$. (2) *What recalibration can and cannot do.* A recalibration that ignores X (e.g. marginal PIT recalibration) can move the shape toward $\bar{r}$ and thereby erase the skewness term, i.e. turn case (B) into case (A); it cannot touch $I(R;X)$. Reducing $I(R;X)$ requires modeling conditional residual shape: heteroscedastic models, conditional recalibration (Kuleshov et al., 2018), or conformalized quantile regression (Romano et al., 2019), which is the conformal world's way of leaving the single-shape class. (3) *The distribution-free guarantee is real.* Recalibration offers calibration only in expectation or asymptotically and carries no coverage certificate; conformal's marginal guarantee is finite-sample and assumption-light. When the certificate itself is the deliverable, that is decisive (Section 8).

Remark 2 (Four readings of the gap)**.** The identity $\mathcal{R}(\bar{r}) = I(R;X) = \mathrm{KL}(P_{X,R} \,\|\, P_X P_R)$ admits several readings, each a different angle on the same number. *(i) A false-pooling cost.* A single-shape system asserts that, after $\hat{\mu}$, one residual law fits everyone, i.e. $R \perp X$; the gap is the relative entropy from the true residual experiment to the nearest such independence model, the log-score price of that assumption. *(ii) An average log Bayes factor.* Since $I(R;X) = \mathbb{E} \log \frac{r(R \,|\, X)}{\bar{r}(R)}$, each observation contributes the log-likelihood ratio of its x-specific residual law against the pooled one, and the gap is the oracle's expected per-sample advantage. *(iii) Conditional non-uniformity of conformal ranks.* Let $U = G(R)$ be the rank of the residual under the pooled law, the PIT used by signed CPS (assuming G is continuous; use the randomized PIT otherwise). Marginally U is uniform, which is the conformal achievement; but U is an invertible transform of R, so $I(R;X) = \mathbb{E}_X \,\mathrm{KL}\big(P_{U \,|\, X} \,\|\, \mathrm{Unif}[0,1]\big)$ is the *conditional* non-uniformity of those ranks. Easy inputs concentrate U near $\frac{1}{2}$, hard inputs push it into the tails, and the single pooled shape cannot tell them apart (Figure 5). *(iv) A projection.* Equation (6) is the information projection of $P_{X,R}$ onto the single-shape product family $\{P_X \otimes H\}$. The optimizer is $H = P_R$, that is $P_X \otimes P_R$, and the projection loss is $\mathrm{KL}(P_{X,R} \,\|\, P_X P_R) = I(R;X)$. No X-blind step can cross that loss, because the missing information is dependence, not marginal shape.

Related work. The pieces this assembles are individually known. A conformal predictive system re-levels the marginal residual law (Vovk et al., 2019); the decomposition of a proper score into calibration and sharpness is standard (Gneiting et al., 2007); and an information-theoretic reading of conformal prediction has been pursued by Correia et al. (2024), though for coverage and efficiency rather than the log-score regret studied here. The core conformal literature already separates *validity* from *efficiency* and adaptivity, and a recent critique strand makes related points operationally: that valid sets can be uninformative, and that interval length can be gamed while marginal coverage is held fixed (Min et al., 2026). Our addition is a single exact identity along the score axis. For the single-shape system the log-score regret to the oracle equals the conditional residual information $I(R;X)$, with an additional symmetrisation term for absolute-residual intervals, so the quantity conformalization cannot touch has a name and a closed form. We have not seen the regret stated this way.

Figure 5 draws the gap: under heteroscedasticity the true conditional residual law $r(\cdot \,|\, x)$ varies with x, while the single-shape system uses the marginal $\bar{r}$ everywhere, and the regret is exactly the

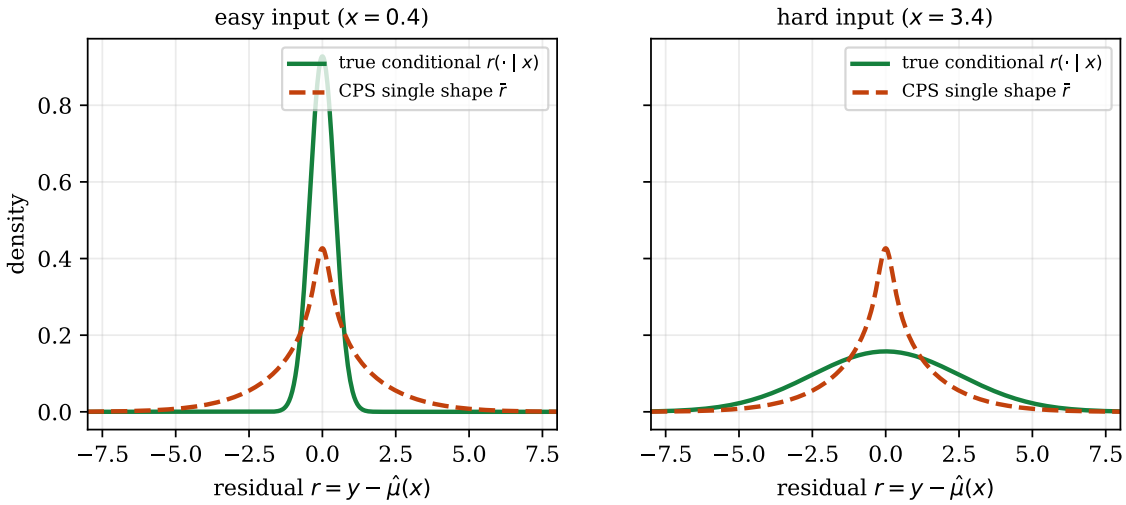

Figure 5: The residual-information gap (Propositions 2–3). A single-shape conformal predictive system uses the marginal residual law $\bar{r}$ for every input; the oracle uses the true conditional $r(\cdot \mid x)$, narrow for easy inputs (left) and wide for hard ones (right). The expected log-score regret is exactly the mutual information $I(R; X)$, which no X-ignoring recalibration can reduce.

information $I(R; X)$ that separates them. The special case $I(R; X) = 0$ ($\hat{\mu}$ correct and r a fixed Gaussian with the wrong scale) is the one where variance recalibration fit to the log-score recovers the shape and improves the score, while conformalization re-levels to exact coverage and returns a set.

6 The coverage–score plane

The preceding results suggest a single diagnostic, provided we are careful about what is being plotted. Define a *reporting system* as a set-valued report $C_\alpha(x)$ and, optionally, a density-valued report $f(\cdot \mid x)$. Its coordinates are

$$\big(C,\, S\big), \qquad C = \big|\, \mathbb{P}(Y \in C_\alpha(X)) - (1 - \alpha) \,\big|, \qquad S = \mathbb{E}[\log f(Y \mid X)], \tag{7}$$

the marginal-coverage error of the *set* and the proper-score performance of the *density*. The decomposition of forecast quality into calibration and sharpness (Gneiting et al., 2007) is, in this plane, the statement that the two coordinates are distinct, and our results read geometrically:

- Residual split conformal is a horizontal projection. It maps a report $(C_\alpha^{\text{base}}, f)$ to $(C_\alpha^{\text{conf}}, f)$: it changes the set so that $C \to 0$ while leaving the density f (and hence S) untouched (Proposition 1). It moves a point left, never up.
- Conformal predictive systems are different: they supply their own density-valued report, so they do have a vertical coordinate, but that coordinate is governed by the single-shape restriction of Proposition 3, capped at the oracle minus $I(R; X)$.
- Recalibration and better modeling move vertically. Fitting shape to a proper score raises S; conditioning on x is the only way to raise it past the ceiling set by the residual-information gap (4).

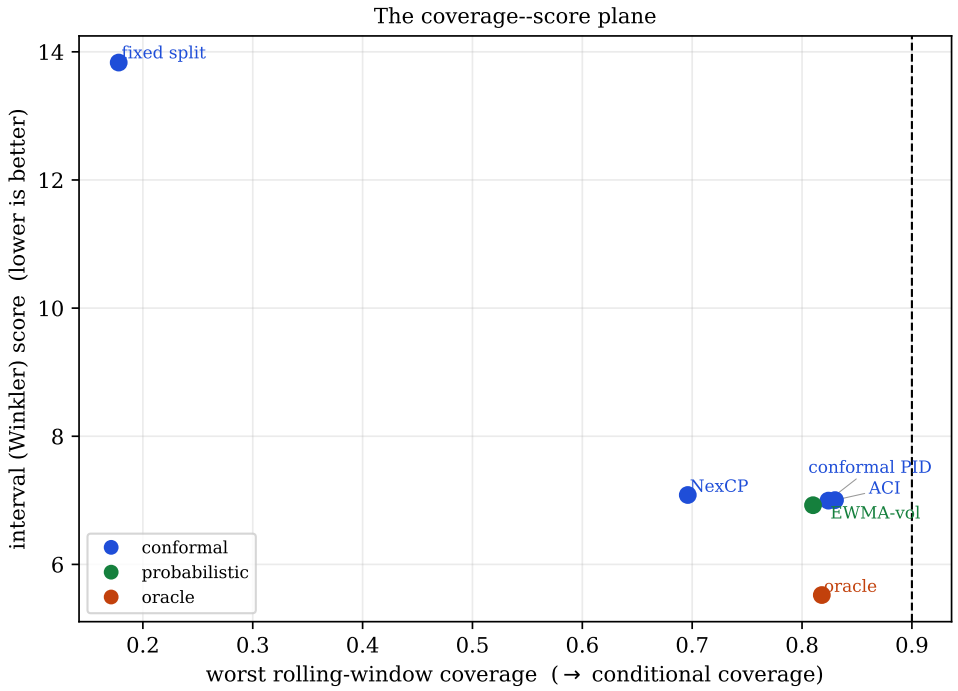

Figure 6: The coverage–efficiency plane, populated by the time-series methods of Section 7 (means over five seeds). The horizontal axis is worst rolling-window coverage (a proxy for conditional coverage); the vertical axis is the proper interval score. Coverage and forecast quality are distinct coordinates: the probabilistic volatility model and the oracle dominate on score, and no method reaches per-step conditional coverage.

A system reported by its coverage alone is reported by one coordinate of a two-coordinate quality. The plane is a presentation device (adjacent to reliability and sharpness diagrams (Gneiting et al., 2007)), but the "conformal = move left, never up" reading makes the category distinction hard to miss. Figure 6 populates the plane with the time-series methods of Section 7, using two concrete axes: worst-window coverage (a conditional-coverage proxy) and the interval (Winkler) score, which, like any proper score, is minimized at the oracle. The vertical axis is therefore oriented opposite to the schematic S above (down is better, not up), but the geometry is unchanged: conformalizing moves a point horizontally, never toward a better score.

7 Exchangeability and the time-series case

Drop exchangeability and (1) is no longer guaranteed. The literature's repairs are ingenious, and uniform in one respect: they survive by redefining what is guaranteed.

- Weighted conformal prediction (Tibshirani et al., 2019) restores marginal coverage under covariate shift when the likelihood ratio is known; with estimated weights coverage becomes approximate and the effective sample size drops.
- Adaptive Conformal Inference (Gibbs and Candès, 2021) updates the level online and "puts no constraints on the data generating distribution," but guarantees a *long-run time-average*

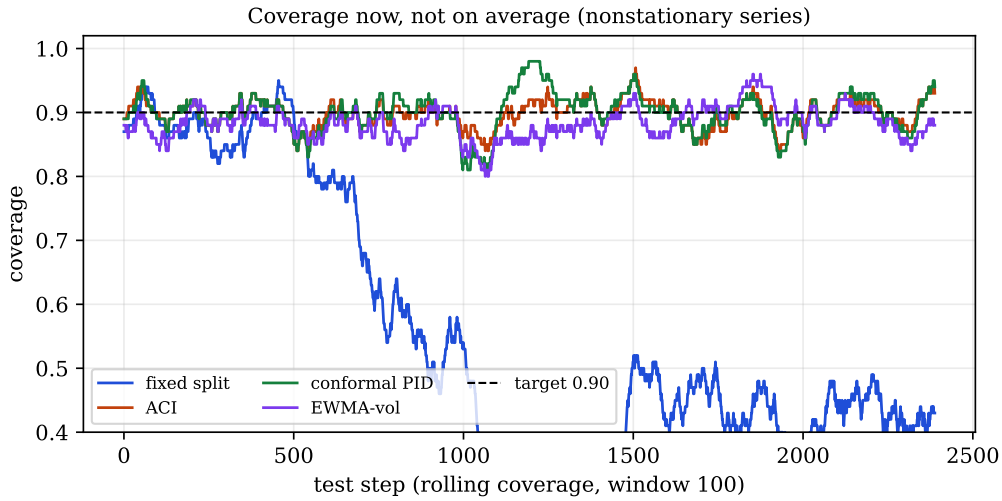

Figure 7: Coverage now, not on average. Rolling coverage on a nonstationary series: fixed split conformal collapses once the volatility regime arrives, while adaptive conformal (ACI, conformal PID) and a probabilistic EWMA-volatility model track the target. Even the recovering methods oscillate around the target rather than holding it pointwise.

miscoverage frequency, $\frac{1}{T}\sum_t \mathrm{err}_t \to \alpha$, not a finite-sample statement about $\mathbb{P}(Y \in C(X))$, and certainly not conditional coverage.

- EnbPI (Xu and Xie, 2021) drops exchangeability for approximate marginal coverage under strong-mixing/stationarity surrogates; later online methods (Zaffran et al., 2022; Gibbs and Candès, 2024) retreat to regret bounds over local intervals. The general beyond-exchangeability analysis of Barber et al. (2023) bounds the coverage gap by a total-variation distance that is exactly what nonstationarity inflates.

It is worth separating three distinct guarantees that the word "conformal" now spans: (i) *split conformal under exchangeability* gives finite-sample marginal coverage; (ii) *adaptive conformal inference* gives long-run, time-average coverage control under distribution shift; (iii) *online conformal under arbitrary shifts* gives regret-style or local-interval performance rather than the original exchangeable guarantee. These are progressively weaker and more diffuse objects; none is the sharp predictive distribution *now*, for *this* step, that a proper score grades. The exchangeability required by guarantee (i) is typically unavailable in deployed time-series settings, though blocking or learned-transform constructions can sometimes recover an exchangeability-like assumption.

An experiment bears this out (Figure 7). On a nonstationary series with known σ_t, fixed split conformal collapses far below target once drift sets in, but the adaptive methods (ACI, conformal PID control, and weighted/NexCP) recover coverage and are competitive on the interval (Winkler) score. They are not straw men. The finding is the one the theory predicts: a simple probabilistic volatility model (an EWMA-variance Gaussian) matches or beats them on the proper score *and* returns a full distribution, and where the conformal methods help they do so by adapting the interval to recently realized residuals, that is, by local conditional modeling, with the conformal step supplying the coverage leveling on top. (Full tables, including MAPIE's EnbPI/ACI and `crepes`, are in the benchmark accompanying the paper.)

The same holds for an off-the-shelf probabilistic forecaster. The author's `skaters` (Cotton, 2021), a streaming forecaster that emits a predictive mean and standard deviation online, has the

best non-oracle interval score (6.57) and CRPS (0.86) here, because its predicted spread tracks the changing volatility. The forecaster is deliberately simple (the `thinking_fast_and_slow` skater composes a fast moving average for the level with a slow moving average of the residuals for the spread), which sharpens the point: even a trivial adaptive model that estimates conditional spread beats every conformal method here, and conformalizing it cannot help. A normalized conformal wrap re-levels its coverage to $1-\alpha$ at an essentially identical interval score (6.58 against 6.57), and a naive split-conformal wrap on the nonstationary series collapses (to 60% coverage, score 11.9). The conformal step re-levels; it does not add sharpness, which is the experience the introduction describes.

8 When coverage *is* the objective

The argument is skeptical, not dismissive, and it comes with a precise converse. The litmus is one question:

> *Is your loss a function of whether the truth lands in a region, or of where it lands?*

If the former, coverage is the native object and conformal prediction is sound and often ideal:

- Selective prediction and risk control. "Return a small set guaranteed to contain the label 95% of the time; a human checks it." The deliverable is a set; risk-controlling prediction sets and conformal risk control formalize this (Angelopoulos et al., 2021; Bates et al., 2021; Angelopoulos et al., 2024).
- Retrieval and shortlisting, where the product is a candidate set and coverage is recall.
- Anomaly and novelty detection via conformal p-values, conformal prediction's most natural home, because here it is not a forecaster but a distribution-free hypothesis test and coverage is Type-I error control (Vovk et al., 2005).
- Compliance and long-run frequency control, where a certificate of the form "contained 95% of the time on average" *is* the product, and the distribution-free, finite-sample nature of (1) is exactly the value added.

Two caveats temper even these. If you trust a calibrated posterior you can form a sharper set yourself (the smallest region of mass $\geq 1-\alpha$) and recover the expectations a bare set discards; conformal's marginal value is then purely the distribution-free guarantee, useful when you distrust the density and exchangeability holds. And even when coverage is the goal you usually want it conditionally, which returns us to the impossibility of Section 3.

9 Scope, and the constructive escapes

Our strong claims target the object that is actually over-sold: *post-hoc, marginal, split conformal prediction with a residual score*. A large and active literature builds conformal methods that do more, and how they relate matters, because in every case they confirm the paper's mechanism rather than contradict it: they buy sharpness or conditional coverage by *conditioning on X or optimizing a proper objective*, which is exactly what Propositions 3 say is required.

Conformal predictive systems that condition on X. Mondrian / binned conformal predictive systems (Boström et al., 2021) fit a separate residual law per region, and Toccaceli (2026) chooses the bins by minimizing a leave-one-out CRPS. These produce sharp, shape-adaptive predictive distributions, and they do so by leaving the single-shape class of Proposition 2 (a per-bin shape is X-dependent) and, in the CRPS case, by making the partition criterion a proper score. This is the most pointed case for the defence, and the concession is real: when the reference class is itself selected by a proper scoring rule, the conformal *construction* is performing conditional density estimation. The category distinction does not vanish; it relocates to "which component estimates the distribution," and the answer is the conditional partition, not the coverage certificate: the value is in the estimation, as throughout. Conformalized quantile regression (Romano et al., 2019) is the same story with a quantile model in place of the bins.

End-to-end conformal training. The orthogonality of Proposition 1 is a statement about *post-hoc* conformalization of a fixed predictor. Conformal training (Stutz et al., 2022; Correia et al., 2024) deliberately differentiates through the conformalizer to minimize expected set size subject to coverage, which couples the coverage and efficiency axes by construction. This does not contradict the proposition; it illustrates its corollary: to obtain sharpness you must optimize for it, rather than read it off the coverage certificate.

The information in set size. Correia et al. (2024) connect conformal set size to information-theoretic uncertainty, with bounds relating expected prediction-set size to the conditional entropy $H(Y \mid X)$. This is complementary to our Proposition 3: both say conformal *geometry* carries information. It also sharpens the two-coordinate plane of Section 6: the *efficiency* (set-size / sharpness) coordinate is the information-bearing one, while the marginal *coverage* coordinate, by Proposition 1, is not. Reporting only the latter discards exactly the coordinate that encodes aleatoric uncertainty.

Approaching conditional coverage. The no-go result of Section 3 concerns *exact, distribution-free, finite-length* conditional coverage. It is not the end of the subject. Diagnostics such as the excess-risk-of-target-coverage family (Braun et al., 2025) quantify and decompose the conditional gap; methods that target conditional or subgroup-wise validity directly interpolate between marginal and conditional coverage (Gibbs et al., 2025); and post-processing via pivotal scores and optimal transport (Laplante, 2026) attains *approximate* conditional coverage by estimating the conditional law of the score, again conditioning on X, and explicitly under the structural assumptions the impossibility theorem rules out for the exact, assumption-free case. These approach the goalpost; they do not move the one the no-go result defends.

Modern time-series conformal. Beyond the adaptive methods benchmarked in Section 7, methods such as HopCPT (Auer et al., 2023) exploit temporal structure to tighten intervals and improve empirical adaptivity. Their guarantees remain of the approximate / long-run kind catalogued there; the improvement, again, comes from modeling the conditional error law.

Beyond the log-score. We used the log-score for the clean information-theoretic identities. The two-coordinate distinction (coverage versus score) carries over to any strictly proper score, but the exact mutual-information identity is special to the log-score; for CRPS or the interval score the analogous gap is the regret of the best X-blind residual law under that score. The accompanying benchmark in fact ranks methods by the interval (Winkler) score and CRPS (Gneiting and Raftery, 2007), on which the same orderings hold.

10 Prediction versus verification

Conformal prediction *verifies* a coverage property; it does not *model*, and a fixed rule cannot substitute for modeling. Conformalization is best viewed as a terminal certification operator: it repairs the marginal coverage coordinate of a set-valued report, but any improvement in a proper score must come from upstream modeling of the conditional location, scale, quantiles, or residual shape. It moves only the coverage coordinate; the residual-information gap is reduced only by conditioning on X (the conditional-shape models of Section 9), never by the conformal step.

Practical remarks. (i) Conformalize *last*: model first, then add the certificate, since the conformal step cannot improve any proper score (estimate first, certify second). (ii) Use a proper score (log-likelihood, CRPS) as the sharpness diagnostic; because conformalization leaves it unchanged, a score that does not move after conformalizing confirms that any gain came from modeling. (iii) To close $I(R;X)$, add conditioning to the model (heteroscedastic spread, quantile or residual-shape estimation), not tightness to the conformal step.

11 Conclusion

Conformal prediction certifies the coverage of a set, finite-sample and distribution-free, under exchangeability. That guarantee is genuine and, for the right objective, exact. The error is reading it as distributional forecast quality. For a fixed location predictor the guarantee you can have (marginal) is rarely the one you want (conditional); the conditional one is unavailable distribution-free without further structure; the output is a set, not a measure; read as a single-shape distribution its log-score loss relative to the oracle is exactly the residual conditional-information term $I(R;X)$ of (4); and the exchangeability the basic guarantee needs is typically unavailable in deployed time-series settings. Within the single-shape residual class it is optimal, and where a guarantee is the product it is the right tool. But coverage answers a question, *is the truth in this region?*, that, in forecasting, is often not the one you were asking. Conformal prediction is not weak uncertainty quantification; it is exact uncertainty quantification for a set-valued coverage objective, and mistaking a set-coverage certificate for evidence of distributional quality is the category error.

Reproducibility. Every figure is generated by the accompanying script `figures.py` (Figures 1–7), and the time-series experiment by the benchmark scripts, all from the synthetic generators described in the text. The identities of Proposition 3 are confirmed numerically by `check_gap.py`, which computes both sides independently (by grid integration) and matches them to machine precision on a heteroscedastic Gaussian and a skewed residual law. All code, figures, and the interactive demonstrations are at `https://conformalprediction.net` (source: `https://github.com/microprediction/conformalprediction`).

References

How to add uncertainty estimation to your models with conformal prediction. Towards Data Science. `https://towardsdatascience.com/how-to-add-uncertainty-estimation-to-your-models-with-conformal-prediction-a5acdb86ea05/`.

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

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.12783.

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.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Anthony Bellotti and Xindi Zhao. Conformal prediction and trustworthy AI. arXiv:2508.06885, 2025.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications (COPA), PMLR*, volume 152, pages 24–38, 2021.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction. *arXiv preprint arXiv:2512.11779*, 2025.

Alvaro H.C. Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.02140.

Peter Cotton. timemachines: continuously evaluated online time-series prediction (skaters). `https://github.com/microprediction/timemachines`, 2021.

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.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B*, 2025. arXiv:2305.12616.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning (ICML), PMLR*, volume 80, pages 2796–2804, 2018.

Félix Laplante. A post-processing conformal prediction approach for conditional coverage via pivotal scores. *arXiv preprint arXiv:2605.25852*, 2026.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.

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.

Valeriy Manokhin. Predicting full probability distributions with conformal prediction. Medium. `https://valeman.medium.com/predicting-full-probability-distributions-with-conformal-prediction-1dd4c1f26973`.

Hendrik Mehrtens, Tabea Bucher, and Titus J. Brinker. Pitfalls of conformal predictions for medical image classification. *arXiv preprint arXiv:2506.18162*, 2025.

Yizhou Min, Yizhou Lu, Lanqi Li, Zhen Zhang, and Jiaye Teng. Questioning the coverage–length metric in conformal prediction: When shorter intervals are not better. *arXiv preprint arXiv:2601.21455*, 2026.

Benjamin Recht. Cover songs; and from intervals to bands. *arg min* blog, 2024. `https://www.argmin.net/`.

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

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2110.09192.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Paolo Toccaceli. CRPS-optimal binning for univariate conformal regression. *arXiv preprint arXiv:2603.22000*, 2026.

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

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108(3):445–474, 2019.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning (ICML), PMLR*, volume 139, pages 11559–11569, 2021.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning (ICML), PMLR*, volume 162, pages 25834–25866, 2022.