CLAPS: Aleatoric-Epistemic Scaling via Last-Layer Laplace for Conformal Regression

Anonymous authors
Paper under double-blind review

Abstract

Conformal regression provides finite-sample marginal coverage, but it does not by itself determine how interval width should adapt across heterogeneous inputs. Existing locally adaptive methods mainly account for aleatoric noise, leaving uncertainty from weak training support less explicit. We propose *Conformal Laplace-Aware Predictive Scaling* (CLAPS), a split conformal regression method that uses heteroscedastic last-layer Laplace uncertainty as the local normalization scale. CLAPS combines learned input-dependent noise with last-layer epistemic uncertainty, while retaining validity through standard conformal calibration. We characterize this aleatoric–epistemic scale, derive its heteroscedastic last-layer precision, and show that it reduces to aleatoric local scaling as epistemic uncertainty contracts. Experiments show nominal-level coverage with competitive interval efficiency.

1 Introduction

Prediction intervals are a standard way to quantify uncertainty in regression, but their usefulness in practice depends on more than attaining a target marginal coverage level. The width of an interval also carries operational meaning. A wide interval caused by intrinsically noisy outcomes suggests irreducible variability, whereas a wide interval caused by limited training support may indicate that additional data, human review, or more cautious downstream decisions are needed. For deployed regression systems, this distinction matters because prediction intervals often function not only as statistical summaries but also as signals for downstream action.

Conformal prediction provides finite-sample marginal coverage under exchangeability by calibrating a nonconformity score on held-out data. In split conformal regression, this makes it possible to wrap a broad class of predictive models with distribution-free coverage guarantees. The guarantee, however, does not determine how interval width should vary across the input space. Standard residual-based conformal intervals use a global residual scale, which can be poorly aligned with heterogeneous regression problems where different regions exhibit different noise levels or different amounts of training support.

Locally adaptive conformal methods address part of this issue by normalizing residuals with an input-dependent scale. In heteroscedastic regression, such scaling is natural: intervals should expand where the response is intrinsically variable and contract where it is stable. Yet local predictive difficulty is not purely aleatoric. In sparse or weakly supported regions, a model may also be uncertain because the training data provide limited evidence about the local regression function. Aleatoric-only conformal scaling can adapt to observation noise, but it does not explicitly account for this epistemic component.

This paper proposes *Conformal Laplace-Aware Predictive Scaling* (CLAPS), a conformal regression method that uses heteroscedastic last-layer Laplace uncertainty as an adaptive local scale. CLAPS combines a learned input-dependent noise estimate with a last-layer posterior uncertainty estimate, and then uses the resulting scale inside a standard split conformal procedure. The method is therefore designed to preserve the validity mechanism of conformal calibration while giving the interval scale a more structured uncertainty interpretation.

The central idea is to separate the source of validity from the source of local adaptivity. Conformal calibration provides the finite-sample marginal coverage guarantee once the score function has been fixed using the training data. The last-layer Laplace approximation does not replace this guarantee; it shapes how the calibrated interval width is allocated across inputs. This separation allows approximate Bayesian uncertainty to guide local interval scaling without requiring the approximation itself to serve as the basis for distribution-free validity.

This paper makes the following contributions:

- **Aleatoric–epistemic conformal scaling.** CLAPS introduces a split conformal regression scale that combines input-dependent aleatoric uncertainty with last-layer epistemic uncertainty.
- **A separation between validity and uncertainty modeling.** The method assigns finite-sample marginal validity to conformal calibration and local interval allocation to the uncertainty model.
- **A heteroscedastic characterization of last-layer uncertainty.** The analysis shows how the learned noise variance affects both the test-time aleatoric scale and the posterior geometry of the last-layer epistemic term.
- **Special-case connections and graceful reduction.** The proposed scale connects global residual scaling, aleatoric locally adaptive scaling, and posterior-uncertainty scaling, while reducing to aleatoric local scaling as epistemic uncertainty contracts.

2 Related Work

2.1 Conformal Prediction for Regression

Conformal prediction provides a general framework for constructing prediction sets with finite-sample marginal coverage under exchangeability. Early work introduced transductive and inductive conformal prediction, nonconformity scores, and rank-based calibration as model-agnostic tools for predictive inference (Saunders et al., 1999; Papadopoulos et al., 2002; Gammerman & Vovk, 2007; Shafer & Vovk, 2008; Papadopoulos et al., 2011). In regression, these ideas lead naturally to residual-based prediction intervals: a model is fitted on training data, residual scores are computed on held-out or resampled data, and an empirical quantile of those scores determines the interval radius.

Regression conformal methods have developed several ways to trade statistical efficiency, computational cost, and data usage. Full conformal inference evaluates candidate labels jointly with the training sample, split conformal inference separates fitting and calibration, and jackknife or cross-validation variants reuse data across multiple fits (Johansson et al., 2014; Lei et al., 2018; Ndiaye & Takeuchi, 2019; Barber et al., 2021; Steinberger & Leeb, 2023). These methods differ in how the conformity scores are computed, but they share the same rank-calibration principle.

Recent work has extended conformal regression to conditional quantile models, covariate shift, online or drifting distributions, non-exchangeable data, and training-conditional or conditional-coverage analyses (Romano et al., 2019; Tibshirani et al., 2019; Foygel Barber et al., 2021; Gibbs & Candes, 2021; Barber et al., 2023; Bian & Barber, 2023; Gibbs & Candès, 2024; Oliveira et al., 2024). This literature establishes conformal regression as a flexible framework for valid prediction intervals. It also foregrounds the central design choice relevant to this paper: the nonconformity score determines how calibrated uncertainty is expressed across the input space.

2.2 Locally Adaptive and Distributional Conformal Methods

Locally adaptive conformal methods modify the nonconformity score so that interval width can vary with input-dependent predictive difficulty. A common strategy is to normalize residuals by an estimated scale or difficulty function, yielding a calibrated threshold that is applied on a local rather than global scale. This idea appears in normalized nonconformity measures and has been extended through local weighting,

learned score transformations, feature-dependent calibration, and group-conditional or multivalid calibration criteria (Papadopoulos et al., 2008; Seedat et al., 2023; Guan, 2023; Hore & Barber, 2025; Gibbs et al., 2025; Jung et al., 2022; Bastani et al., 2022; Kiyani et al., 2024a;b).

Distributional conformal methods use richer estimates of the conditional response distribution to construct more efficient scores and sets. Instead of relying only on absolute residuals, these methods calibrate estimated conditional distributions, histograms, conditional random samples, discretized response models, normalizing flows, or shape templates for multimodal regions (Chernozhukov et al., 2021; Sesia & Romano, 2021; Wang et al., 2023; Plassier et al., 2024; Guha et al., 2024; Colombo, 2024; Tumu et al., 2024; Plassier et al., 2025; van der Laan & Alaa, 2024). Their common premise is that conformal efficiency improves when the score reflects the local structure of the conditional response.

Adaptive score design has also been studied in classification, where set size should grow with input ambiguity. Adaptive prediction sets, regularized classification scores, conformal training, and label-ranking scores show how the choice of conformity score affects set efficiency and conditional behavior (Romano et al., 2020; Angelopoulos et al., 2020; Stutz et al., 2021; Huang et al., 2024). In regression, the analogous question is how to choose a local scale that captures not only heteroscedastic noise, but also uncertainty from limited training support.

2.3 Bayesian and Epistemic Uncertainty in Prediction Intervals

Bayesian predictive modeling offers a principled language for distinguishing aleatoric uncertainty, which reflects irreducible response variability, from epistemic uncertainty, which reflects uncertainty about the learned predictor. Bayesian neural networks represent this uncertainty through distributions over weights, functions, or predictive laws, usually with approximate inference procedures that remain tractable for modern architectures (Graves, 2011; Hernández-Lobato & Adams, 2015; Blundell et al., 2015; Kingma et al., 2015; Gal & Ghahramani, 2016).

Scalable Bayesian deep learning methods approximate posterior uncertainty through variational objectives, dropout interpretations, normalizing-flow posteriors, Bayesian hypernetworks, function-space approximations, and modular Bayesian layers (Louizos & Welling, 2017; Gal et al., 2017; Krueger et al., 2017; Sun et al., 2019; Tran et al., 2019). These approaches make epistemic uncertainty accessible in neural prediction, but their empirical behavior depends on the likelihood, approximation family, optimization procedure, and calibration of the predictive distribution.

Other uncertainty estimators avoid full posterior inference. Deep ensembles use variation across independently trained models, evidential methods parameterize uncertainty over predictive distributions, and deterministic uncertainty methods seek distance-sensitive uncertainty estimates from a single network (Lakshminarayanan et al., 2017; Sensoy et al., 2018; Amini et al., 2020; Van Amersfoort et al., 2020; Charpentier et al., 2020). Complementary work on uncertainty calibration and dataset shift studies when these probabilistic estimates remain reliable under misspecification or changing input distributions (Kendall & Gal, 2017; Kuleshov et al., 2018; Ovadia et al., 2019; Maddox et al., 2019; Wilson & Izmailov, 2020). This line of work motivates using epistemic uncertainty as a source of local information for prediction intervals, while leaving open how it should be combined with distribution-free calibration.

2.4 Last-Layer Laplace Approximation for Neural Uncertainty

Laplace approximation turns a trained neural network into an approximate Bayesian predictor by fitting a Gaussian posterior around a point estimate, with covariance determined by a curvature approximation. Recent work has revisited this classical approximation for modern neural networks, developing scalable implementations, post-hoc uncertainty estimates, marginal-likelihood objectives, and function-space prior variants (Ritter et al., 2018; Daxberger et al., 2021a; Kristiadi et al., 2021; Immer et al., 2021a; Cinquin et al., 2024).

Linearized and subnetwork Laplace methods reduce the complexity of posterior prediction by approximating the network locally or restricting Bayesian inference to selected parameter directions. These approaches connect neural uncertainty to generalized linear models, Gaussian processes, neural tangent kernels, and

efficient adaptation procedures (Immer et al., 2021b; Antorán et al., 2022; Daxberger et al., 2021b; Deng et al., 2022; Ortega et al., 2023; Khan et al., 2019; Jacot et al., 2018; Maddox et al., 2021). They provide a practical middle ground between full-network Bayesian inference and purely deterministic prediction.

Last-layer Bayesian methods are a particularly simple version of this idea. They keep the learned representation fixed and place a posterior on the final predictive layer, yielding a neural-linear model in which representation learning and epistemic uncertainty estimation are separated (Kristiadi et al., 2020; Watson et al., 2021; Harrison et al., 2024; Brunzema et al., 2024). Their efficiency depends on the quality of curvature and likelihood information, which has motivated work on Kronecker-factored, Gauss–Newton, and automatic-differentiation-based curvature approximations, as well as studies of heteroscedastic neural regression (Martens & Grosse, 2015; Botev et al., 2017; Dangel et al., 2019; Eschenhagen et al., 2023; Seitzer et al., 2022; Stirn et al., 2023).

3 Method

We consider regression with inputs $X \in \mathcal{X}$ and responses $Y \in \mathbb{R}$. Given a training set, a calibration set, and a target miscoverage level $\alpha \in (0, 1)$, CLAPS constructs a split conformal prediction interval using a heteroscedastic last-layer Laplace predictive scale.

The base model maps each input x to a representation $\phi(x) \in \mathbb{R}^d$, and the final linear layer defines the predictive mean

$$\mu(x) = \phi(x)^\top w.$$

The model is trained with the heteroscedastic Gaussian likelihood

$$Y \mid x, w \sim \mathcal{N}\big(\phi(x)^\top w, h^2(x)\big),$$

where $h^2(x)$ is an input-dependent variance produced by a separate variance head. After training, $\phi(\cdot)$ and $h^2(\cdot)$ are fixed, and a Laplace approximation is applied only to the final-layer weights.

Let $H \in \mathbb{R}^{n_{\text{tr}} \times d}$ be the training feature matrix with rows $\phi(x_i)^\top$, and define

$$W = \operatorname{diag}\left(\frac{1}{h^2(x_1)}, \ldots, \frac{1}{h^2(x_{n_{\text{tr}}})}\right).$$

With prior precision $\lambda > 0$, the last-layer posterior covariance is

$$\Sigma = \left(\lambda I + H^\top W H\right)^{-1}.$$

CLAPS uses the predictive variance

$$v(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x)$$

as the local scale, combining the heteroscedastic aleatoric variance with the last-layer epistemic variance.

For each calibration example (x_i, y_i), define the normalized score

$$A(x_i, y_i) = \frac{|y_i - \mu(x_i)|}{\sqrt{v(x_i)}}.$$

Let q_α be the $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$-th order statistic of the calibration scores, with the standard conservative convention when this index exceeds n_{cal}. The prediction interval for a test input x is

$$C_\alpha(x) = \left[\mu(x) - q_\alpha \sqrt{v(x)},\ \mu(x) + q_\alpha \sqrt{v(x)}\right].$$

The resulting procedure is summarized in Algorithm 1.

Algorithm 1 Conformal Laplace-Aware Predictive Scaling (CLAPS)

Require: Training set $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n_{\mathrm{tr}}}$, calibration set $\mathcal{D}_{\mathrm{cal}} = \{(x_i, y_i)\}_{i=1}^{n_{\mathrm{cal}}}$, target miscoverage level α, prior precision $\lambda > 0$ or candidate grid Λ

Ensure: Prediction interval $C_\alpha(x)$ for a test input x

Fit a heteroscedastic neural regression model on $\mathcal{D}_{\mathrm{tr}}$ with representation $\phi(x)$, predictive mean

$$\mu(x) = \phi(x)^\top \hat{w},$$

and variance head $h^2(x)$.

if a prior-precision grid Λ is used **then**

  Select $\lambda \in \Lambda$ using a held-out split of $\mathcal{D}_{\mathrm{tr}}$ by minimizing the heteroscedastic predictive negative log-likelihood based on $h^2(x) + \phi(x)^\top \Sigma_\lambda \phi(x)$.

end if

Form the feature matrix $H \in \mathbb{R}^{n_{\mathrm{tr}} \times d}$ whose i-th row is $\phi(x_i)^\top$.

Form the heteroscedastic weight matrix

$$W = \mathrm{diag}\left(\frac{1}{h^2(x_1)}, \ldots, \frac{1}{h^2(x_{n_{\mathrm{tr}}})}\right).$$

Compute the last-layer Laplace covariance

$$\Sigma = \left(\lambda I + H^\top W H\right)^{-1}.$$

for each calibration example $(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}$ **do**

  Compute the CLAPS predictive variance

$$v(x_i) = h^2(x_i) + \phi(x_i)^\top \Sigma \phi(x_i).$$

  Compute the normalized conformal score

$$A_i = \frac{|y_i - \mu(x_i)|}{\sqrt{v(x_i)}}.$$

end for

Let

$$k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil.$$

Let q_α be the k-th order statistic of $\{A_i\}_{i=1}^{n_{\mathrm{cal}}}$, with the standard conservative convention if $k > n_{\mathrm{cal}}$.

for a test input x **do**

  Compute

$$v(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x).$$

  Return

$$C_\alpha(x) = \left[\mu(x) - q_\alpha \sqrt{v(x)}, \mu(x) + q_\alpha \sqrt{v(x)}\right].$$

end for

4 Theory

4.1 Finite-Sample Validity under Adaptive Scaling

Let $\mathcal{D}_{\mathrm{tr}} = \{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{tr}}}$ and $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=n_{\mathrm{tr}}+1}^{n_{\mathrm{tr}}+n_{\mathrm{cal}}}$, and write $n = n_{\mathrm{tr}} + n_{\mathrm{cal}}$. After fitting $\mu(\cdot)$, $h^2(\cdot)$, $\phi(\cdot)$, and Σ on $\mathcal{D}_{\mathrm{tr}}$, define

$$v(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x), \qquad A(x, y) = \frac{|y - \mu(x)|}{\sqrt{v(x)}}.$$

Assume $v(x) > 0$ on the input domain. For calibration scores $A_i = A(X_i, Y_i)$, $i = n_{\text{tr}} + 1, \ldots, n$, let q_α be the $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$-th order statistic, with $q_\alpha = \infty$ if this index exceeds n_{cal}. The resulting interval is

$$C_\alpha(x) = \left[\mu(x) - q_\alpha \sqrt{v(x)},\ \mu(x) + q_\alpha \sqrt{v(x)} \right].$$

Proposition 4.1 (Finite-sample validity)**.** *Suppose that the calibration examples and the test point (X_{n+1}, Y_{n+1}) are exchangeable conditional on $\mathcal{D}_{\text{tr}}$. Then*

$$\mathbb{P}\left\{ Y_{n+1} \in C_\alpha(X_{n+1}) \right\} \geq 1 - \alpha.$$

Proof. Condition on $\mathcal{D}_{\text{tr}}$, so that $A(\cdot, \cdot)$ is fixed. Exchangeability of the calibration examples and the test point implies exchangeability of

$$A_{n_{\text{tr}}+1}, \ldots, A_n, A_{n+1}, \qquad A_{n+1} = A(X_{n+1}, Y_{n+1}).$$

Hence the conformal rank argument gives

$$\mathbb{P}\left\{ A(X_{n+1}, Y_{n+1}) \leq q_\alpha \mid \mathcal{D}_{\text{tr}} \right\} \geq 1 - \alpha.$$

The event $A(X_{n+1}, Y_{n+1}) \leq q_\alpha$ is equivalent to

$$|Y_{n+1} - \mu(X_{n+1})| \leq q_\alpha \sqrt{v(X_{n+1})},$$

which is $Y_{n+1} \in C_\alpha(X_{n+1})$. Taking expectation over $\mathcal{D}_{\text{tr}}$ completes the proof. □

4.2 Aleatoric–Epistemic Predictive Scale

After training, CLAPS treats the representation $\phi(\cdot)$, variance function $h^2(\cdot)$, and final-layer point estimate $\widehat{w}$ as fixed, and uses the last-layer Gaussian approximation

$$w \mid \mathcal{D}_{\text{tr}} \approx \mathcal{N}(\widehat{w}, \Sigma).$$

Together with

$$Y \mid x, w \sim \mathcal{N}(\phi(x)^\top w, h^2(x)),$$

this yields the local predictive scale used in the conformal score.

Proposition 4.2 (Aleatoric–epistemic variance decomposition)**.** *Under the fitted representation, heteroscedastic variance function, and last-layer Gaussian posterior above,*

$$\mathrm{Var}(Y \mid x, \mathcal{D}_{\text{tr}}) = h^2(x) + \phi(x)^\top \Sigma \phi(x).$$

Proof. The law of total variance gives

$$\mathrm{Var}(Y \mid x, \mathcal{D}_{\text{tr}}) = \mathbb{E}\left[\mathrm{Var}(Y \mid x, w, \mathcal{D}_{\text{tr}}) \mid x, \mathcal{D}_{\text{tr}} \right] + \mathrm{Var}\left(\mathbb{E}[Y \mid x, w, \mathcal{D}_{\text{tr}}] \mid x, \mathcal{D}_{\text{tr}} \right).$$

The first term equals $h^2(x)$, and the second term equals

$$\mathrm{Var}\left(\phi(x)^\top w \mid \mathcal{D}_{\text{tr}} \right) = \phi(x)^\top \Sigma \phi(x).$$

Combining the two terms gives the claim. □

The term $h^2(x)$ captures input-dependent observation noise, while $\phi(x)^\top \Sigma \phi(x)$ captures final-layer posterior uncertainty under the learned representation. The resulting scale increases in noisy regions through the aleatoric term and in weakly supported feature regions through the epistemic term.

4.3 Heteroscedastic Last-Layer Laplace Covariance

For fixed $\phi(\cdot)$ and $h^2(\cdot)$, the final-layer negative log-likelihood is

$$\ell(w) = \frac{1}{2}\sum_{i=1}^{n_{\text{tr}}} \frac{\left(y_i - \phi(x_i)^\top w\right)^2}{h^2(x_i)}.$$

With prior $w \sim \mathcal{N}(0, \lambda^{-1}I)$, the negative log-posterior is

$$\mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{n_{\text{tr}}} \frac{\left(y_i - \phi(x_i)^\top w\right)^2}{h^2(x_i)} + \frac{\lambda}{2}\|w\|_2^2.$$

Let $H \in \mathbb{R}^{n_{\text{tr}} \times d}$ have rows $\phi(x_i)^\top$, and define

$$W = \operatorname{diag}\left(\frac{1}{h^2(x_1)}, \ldots, \frac{1}{h^2(x_{n_{\text{tr}}})}\right).$$

Then

$$\nabla_w^2 \mathcal{L}(w) = H^\top W H + \lambda I, \qquad \Sigma = \left(\lambda I + H^\top W H\right)^{-1}.$$

Proposition 4.3 (Heteroscedastic precision weighting)**.** *The last-layer posterior precision satisfies*

$$\Sigma^{-1} = \lambda I + \sum_{i=1}^{n_{\text{tr}}} \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

Proof. By the definitions of H and W,

$$H^\top W H = \sum_{i=1}^{n_{\text{tr}}} \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

Adding λI gives the stated precision matrix. □

Thus, $h^2(x)$ enters CLAPS both as the test-time aleatoric component of $v(x)$ and through the training weights $1/h^2(x_i)$ in the last-layer precision. Noisier training examples contribute less curvature, while lower-noise examples contribute more. The epistemic term is therefore shaped jointly by feature geometry and the heteroscedastic noise pattern.

4.4 Special Cases and Graceful Reduction

The CLAPS scale

$$v_{\text{CLAPS}}(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x)$$

contains several standard scale choices as limiting cases. A constant scale $v(x) = \sigma^2$ recovers residual-based split conformal prediction up to rescaling of the conformal quantile. Removing the epistemic term gives

$$v_{\text{LACP}}(x) = h^2(x),$$

the locally adaptive aleatoric scale. Taking the aleatoric variance to be constant while retaining posterior uncertainty gives

$$v_{\text{post}}(x) = \sigma^2 + \phi(x)^\top \Sigma \phi(x).$$

CLAPS retains both the input-dependent aleatoric term and the last-layer epistemic term.

Now consider a sequence of training sets with

$$\Sigma_n = \left(\lambda I + H_n^\top W_n H_n\right)^{-1}, \qquad v_n(x) = h^2(x) + \phi(x)^\top \Sigma_n \phi(x).$$

Proposition 4.4 (Graceful reduction to aleatoric scaling)**.** *Suppose* $h^2(x) > 0$ *and*

$$\phi(x)^\top \Sigma_n \phi(x) \to 0.$$

Then

$$v_n(x) \to h^2(x), \qquad \frac{\sqrt{v_n(x)}}{h(x)} \to 1.$$

Proof. Since

$$v_n(x) = h^2(x) + \phi(x)^\top \Sigma_n \phi(x),$$

the first convergence follows immediately. Because $h^2(x) > 0$,

$$\frac{\sqrt{v_n(x)}}{h(x)} = \sqrt{1 + \frac{\phi(x)^\top \Sigma_n \phi(x)}{h^2(x)}} \to 1.$$

□

A sufficient condition is

$$\lambda_{\min}\left(\lambda I + H_n^\top W_n H_n\right) \to \infty \qquad \text{and} \qquad \|\phi(x)\|_2 < \infty.$$

Indeed,

$$0 \le \phi(x)^\top \Sigma_n \phi(x) \le \frac{\|\phi(x)\|_2^2}{\lambda_{\min}\left(\lambda I + H_n^\top W_n H_n\right)} \to 0.$$

Therefore, as weighted feature information grows, the last-layer epistemic contribution vanishes and the CLAPS scale reduces to the heteroscedastic aleatoric scale.

5 Experiments

5.1 Common Experimental Design

All experiments evaluate prediction intervals at the nominal coverage level $1-\alpha = 0.90$. We compare CLAPS with five conformal regression baselines: Split CP, CV+, Locally Adaptive Split CP (LACP), Conformalized Quantile Regression (CQR), and Distributional Conformal Prediction (DCP). Split CP and CV+ use mean-regression residuals, LACP uses a learned heteroscedastic scale, CQR uses learned conditional quantiles, and DCP uses a heteroscedastic Gaussian predictive distribution. CLAPS uses the scale $h^2(x) + \phi(x)^\top \Sigma \phi(x)$. In this protocol, DCP uses symmetric standardized-residual calibration under a heteroscedastic Gaussian predictive distribution, making its score algebraically equivalent to the LACP normalized residual score.

All neural methods share the same backbone architecture and training protocol within each experiment. The synthetic studies use controlled one-dimensional regression settings designed to isolate aleatoric and epistemic sources of difficulty. The real-data benchmark uses eight tabular regression datasets: Concrete Compressive Strength, Energy Efficiency, Yacht Hydrodynamics, Airfoil Self-Noise, Wine Quality Red, Naval Propulsion, Bike Sharing, and California Housing. Each real dataset is split into training, calibration, and test sets using a 60/20/20 split, with at most 5000 samples per dataset.

We report empirical marginal coverage, coverage error, coverage violation rate, average width, and interval score. For the real-data benchmark, aggregate efficiency is reported using relative width reduction against Split CP and rank-based summaries of average width and interval score, since raw widths and interval scores are dataset-scale dependent.

5.2 Experiment 1: Controlled Synthetic Mechanism Study

Experimental setup. The first experiment uses a one-dimensional synthetic regression problem with a nonlinear mean function, an input-dependent high-noise region, and a separate sparse region created by low-probability subsampling of training inputs. Calibration and test inputs are sampled uniformly over the full domain, so evaluation covers both difficult regions. We compare all six methods using marginal coverage, sparse-region coverage, noisy-region coverage, region coverage gap, average interval width, and interval score.

Results and analysis. Table 1 shows that all methods attain marginal coverage near the nominal level, while their regional behavior differs. Split CP and CV+ under-cover the noisy region, whereas CQR improves regional coverage at the cost of much wider intervals and higher interval scores. LACP and DCP improve noisy-region coverage using the same heteroscedastic Gaussian scale in this implementation. CLAPS gives the smallest region coverage gap, average width, and interval score.

Table 1: Controlled synthetic mechanism study.

Method	Marg. Cov.	Sparse Cov.	Noisy Cov.	Region Gap	Avg. Width	Int. Score
Split CP	0.9084 ± 0.0093	0.9282 ± 0.0382	0.7590 ± 0.0182	0.1410 ± 0.0182	1.0553 ± 0.0360	1.3669 ± 0.0809
CV+	0.9058 ± 0.0083	0.9754 ± 0.0071	0.7358 ± 0.0162	0.1642 ± 0.0162	1.0131 ± 0.0272	1.3318 ± 0.0548
LACP	0.9126 ± 0.0204	0.8621 ± 0.0312	0.8923 ± 0.0354	0.0520 ± 0.0248	0.9867 ± 0.0597	1.1863 ± 0.0302
CQR	0.9138 ± 0.0226	0.9926 ± 0.0090	0.8959 ± 0.0616	0.0939 ± 0.0066	1.7655 ± 0.2875	1.9934 ± 0.2520
DCP	0.9126 ± 0.0204	0.8621 ± 0.0312	0.8923 ± 0.0354	0.0520 ± 0.0248	0.9867 ± 0.0597	1.1863 ± 0.0302
CLAPS	0.9126 ± 0.0196	0.8715 ± 0.0254	0.8850 ± 0.0355	0.0467 ± 0.0183	0.9831 ± 0.0579	1.1817 ± 0.0303

5.3 Experiment 2: Posterior Contraction and Graceful Reduction

Experimental setup. The second experiment varies the number of training examples in the same synthetic regression family while keeping the calibration and test set sizes fixed. For each training size, we compare LACP, which uses only $h^2(x)$, with CLAPS, which adds the last-layer epistemic term $\phi(x)^\top \Sigma \phi(x)$. We report the epistemic fraction, scale ratio, width ratio, interval-score difference, and marginal coverage of both methods.

Results and analysis. Table 2 shows that the epistemic fraction decreases as the training size grows. The scale ratio moves toward one, the width ratio stays close to one, and the interval-score difference remains near zero. LACP and CLAPS have nearly identical marginal coverage across the training-size sweep.

Table 2: Posterior contraction and graceful reduction.

Train n	Epi. Frac.	Scale Ratio	Width Ratio	Score Δ	Cov. LACP	Cov. CLAPS
100	0.0943 ± 0.0617	1.0578 ± 0.0409	0.9965 ± 0.0116	-0.0019 ± 0.0155	0.8990 ± 0.0119	0.9001 ± 0.0144
250	0.0529 ± 0.0312	1.0309 ± 0.0187	0.9848 ± 0.0120	-0.0283 ± 0.0336	0.8967 ± 0.0129	0.8977 ± 0.0094
500	0.0310 ± 0.0119	1.0177 ± 0.0064	0.9947 ± 0.0097	-0.0076 ± 0.0052	0.9021 ± 0.0214	0.9043 ± 0.0210
1000	0.0179 ± 0.0013	1.0107 ± 0.0008	0.9926 ± 0.0049	-0.0040 ± 0.0033	0.8973 ± 0.0088	0.8959 ± 0.0096
2000	0.0083 ± 0.0021	1.0051 ± 0.0013	0.9977 ± 0.0049	-0.0016 ± 0.0009	0.8871 ± 0.0098	0.8873 ± 0.0092

5.4 Experiment 3: Real Regression Benchmark

Experimental setup. The third experiment evaluates all six methods on the eight real tabular regression datasets. Each dataset and random seed uses the same train/calibration/test protocol. The aggregate table averages dataset-level summaries and uses relative width reduction and rank-based efficiency metrics to avoid domination by datasets with larger target scales.

Results and analysis. Table 3 reports the aggregate benchmark. All methods achieve coverage near the nominal level. CV+ has the lowest coverage violation rate, but its relative width reduction is slightly negative and its efficiency ranks are weak. CQR has reasonable coverage but the worst relative width reduction and width rank. LACP and DCP improve efficiency over Split CP and CV+. CLAPS obtains the largest relative width reduction and the best width and score ranks while maintaining nominal-level coverage.

Table 3: Real regression benchmark summary across datasets.

Method	Coverage	Cov. Err.	Viol. Rate	Rel. Width Red.	Width Rank	Score Rank
Split CP	0.8956	0.0192	0.575	0.0000	3.850	4.4500
CV+	0.9083	0.0207	0.300	−0.0051	4.100	4.4000
LACP	0.8995	0.0252	0.500	0.1299	2.925	2.8625
CQR	0.9024	0.0208	0.450	−0.2240	5.150	4.5750
DCP	0.8995	0.0252	0.500	0.1299	2.825	2.7375
CLAPS	0.8995	0.0245	0.500	0.1367	2.150	1.9750

6 Discussion

6.1 Validity and Uncertainty Modeling Play Different Roles

CLAPS separates two roles that are often conflated in uncertainty-aware prediction intervals. Split conformal calibration provides the finite-sample marginal coverage guarantee once the score function has been fixed using the training data. The last-layer Laplace approximation does not serve as the source of validity; it provides the normalization scale used by the nonconformity score. This separation allows an approximate Bayesian predictive variance to guide local interval shape without requiring the approximation itself to be a fully calibrated posterior.

The resulting view is that conformal calibration sets the global quantile, while the uncertainty model determines how interval width is distributed across the input space. The aleatoric term $h^2(x)$ increases the scale in regions with high estimated observation noise, and the last-layer epistemic term $\phi(x)^\top \Sigma \phi(x)$ increases the scale in regions with weak feature support. CLAPS therefore uses Bayesian last-layer uncertainty not as a replacement for conformal calibration, but as a structured local scale inside a conformal prediction procedure.

6.2 Aleatoric–Epistemic Scaling as Local Adaptivity

Local adaptivity in conformal regression is usually associated with heteroscedasticity: intervals should be wider where the response is intrinsically more variable and narrower where it is more stable. CLAPS broadens this view by treating local predictive difficulty as a combination of aleatoric noise and epistemic uncertainty. The scale

$$v(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x)$$

therefore adapts both to the estimated variability of the response and to the degree of support provided by the weighted training features.

This distinction matters because noisy regions and sparse regions call for different forms of adaptation. An aleatoric-only scale can respond to input-dependent noise, but it need not account for limited training support. A purely epistemic scale, on the other hand, would miss irreducible response variability. By combining the two terms in a single predictive scale, CLAPS extends locally adaptive conformal prediction from noise-adaptive scaling to aleatoric–epistemic scaling.

6.3 When the Epistemic Correction Matters

The epistemic correction is most relevant when feature support is uneven. In regions that are weakly represented in the training data, $\phi(x)^\top \Sigma \phi(x)$ can enlarge the conformal scale beyond what is explained by the aleatoric variance alone. This is the setting where CLAPS differs most clearly from purely heteroscedastic conformal methods: the interval can reflect not only that the response is noisy, but also that the model has limited evidence around the input.

The same mechanism also explains why CLAPS approaches locally adaptive conformal prediction in data-rich regimes. As weighted feature information grows, the last-layer posterior covariance contracts and the

epistemic contribution vanishes. The CLAPS scale then reduces toward $h^2(x)$, leaving the aleatoric component as the dominant source of local adaptation. The epistemic term is therefore selective rather than uniformly conservative: it matters when support is limited and fades when the learned representation is sufficiently informed by the training data.

6.4 Practical Implications for Deployed Regression Intervals

In deployed regression systems, marginal coverage is only one requirement for a useful prediction interval. Practitioners also need intervals whose widths reflect why a prediction is uncertain. CLAPS gives a simple operational decomposition: wide intervals may arise from high estimated observation noise, weak training support, or both. This makes the interval scale more interpretable than a global residual quantile and more informative than an aleatoric-only normalization.

The decomposition can also be used diagnostically. Large aleatoric components indicate regions where irreducible variability limits prediction accuracy, whereas large epistemic components point to regions where additional data, human review, or more cautious decisions may be warranted. Because the conformal quantile calibrates the final interval after the scale has been chosen, the method remains compatible with standard finite-sample conformal validity while providing a more informative account of where uncertainty appears at deployment time.

6.5 Limitations and Future Work

CLAPS shares the standard scope of split conformal prediction. Its guarantee is marginal and relies on exchangeability between calibration and test examples; it does not imply conditional coverage for every input value or subgroup. The method can improve the allocation of interval width across heterogeneous regions, but subgroup and regional coverage behavior remain empirical properties. Extensions to distribution shift, temporal dependence, and stronger subgroup-level guarantees are important directions for future work.

The epistemic component is also limited by the last-layer approximation. CLAPS holds the learned representation and heteroscedastic variance function fixed, so it does not capture posterior uncertainty over the full neural network or uncertainty induced by representation learning. Its efficiency also depends on the quality of the learned variance head and the suitability of the final-layer Gaussian approximation. Future work could study richer posterior approximations, representation-aware uncertainty estimates, calibration under covariate shift, and extensions to higher-dimensional or structured prediction tasks.

7 Conclusion

We presented CLAPS, a conformal regression method that uses heteroscedastic last-layer Laplace uncertainty to shape adaptive prediction intervals. The method keeps the validity mechanism of split conformal prediction intact while using the predictive scale to distinguish regions dominated by observation noise from regions with limited feature support. The analysis shows how the aleatoric and epistemic terms enter the scale, how heteroscedasticity affects the last-layer posterior geometry, and why the method naturally approaches aleatoric local scaling when epistemic uncertainty contracts. Across controlled and real-data experiments, CLAPS maintained nominal-level coverage and yielded competitive interval efficiency. These results suggest that last-layer Bayesian uncertainty can serve as a useful local scaling mechanism for conformal regression without replacing conformal calibration as the source of validity.

Broader Impact Statement

This work studies prediction intervals for regression and is primarily methodological. Its potential positive impact is to make conformal prediction intervals more informative in heterogeneous settings by distinguishing interval width due to estimated observation noise from width due to limited training support. Such information may help practitioners identify regions where uncertainty is driven by irreducible variability and regions where additional data collection, human review, or more cautious decisions may be appropriate.

At the same time, the guarantees considered in this work are marginal and rely on exchangeability between calibration and test examples. They should not be interpreted as conditional coverage guarantees for every input, subgroup, or deployment environment. In high-stakes domains such as healthcare, finance, public policy, or employment, prediction intervals produced by CLAPS should not be used as the sole basis for automated decisions. Practical deployment would require domain-specific validation, subgroup and distribution-shift audits, and careful consideration of how uncertainty estimates affect downstream users and affected individuals.

Reproducibility

To support reproducibility, we provide a supplementary notebook, `CLAPS.ipynb`, containing the implementation of CLAPS and all experiments reported in the paper. The notebook includes the synthetic studies, real-data benchmark, baseline comparisons, evaluation metrics, and table-generation code used to produce the experimental results.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Javier Antorán, David Janz, James U Allingham, Erik Daxberger, Riccardo Rb Barbano, Eric Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised laplace model evidence for modern deep learning. In *International Conference on Machine Learning*, pp. 796–821. PMLR, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in neural information processing systems*, 35:29362–29373, 2022.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pp. 557–565. PMLR, 2017.

Paul Brunzema, Mikkel Jordahn, John Willes, Sebastian Trimpe, Jasper Snoek, and James Harrison. Bayesian optimization via continual variational last layer training. *arXiv preprint arXiv:2412.09477*, 2024.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Tristan Cinquin, Marvin Pförtner, Vincent Fortuin, Philipp Hennig, and Robert Bamler. Fsp-laplace: Function-space priors for the laplace approximation in bayesian deep learning. *Advances in Neural Information Processing Systems*, 37:13897–13926, 2024.

Nicolo Colombo. Normalizing flows for conformai regression. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pp. 881–893, 2024.

Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. *arXiv preprint arXiv:1912.10985*, 2019.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021a.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pp. 2510–2521. PMLR, 2021b.

Zhijie Deng, Feng Zhou, and Jun Zhu. Accelerated linearized laplace approximation for bayesian deep learning. *Advances in Neural Information Processing Systems*, 35:2695–2708, 2022.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. *Advances in Neural Information Processing Systems*, 36:33624–33655, 2023.

Rina Foygel Barber, Emmanuel J Candes, 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.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 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: Statistical Methodology*, 87(4):1100–1126, 2025.

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

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

Etash Guha, Shlok Natarajan, Thomas Möllenhoff, Mohammad Emtiyaz Khan, and Eugene Ndiaye. Conformal prediction via regression-as-classification. *arXiv preprint arXiv:2404.08168*, 2024.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. *arXiv preprint arXiv:2404.11599*, 2024.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *International Conference on Machine Learning*, pp. 20331–20347. PMLR, 2024.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, pp. 4563–4573. PMLR, 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021b.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine learning*, 97(1):155–176, 2014.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. *Advances in neural information processing systems*, 32, 2019.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Shayan Kiyani, George Pappas, and Hamed Hassani. Conformal prediction with learned features. *arXiv preprint arXiv:2404.17487*, 2024a.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024b.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under laplace approximations. In *Uncertainty in Artificial Intelligence*, pp. 344–353. PMLR, 2021.

David Krueger, Chin-Wei Huang, Riashat Islam, Ryan Turner, Alexandre Lacoste, and Aaron Courville. Bayesian hypernetworks. *arXiv preprint arXiv:1710.04759*, 2017.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

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.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International conference on machine learning*, pp. 2218–2227. PMLR, 2017.

Wesley Maddox, Shuai Tang, Pablo Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 2737–2745. PMLR, 2021.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32, 2019.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and Joao Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.

Luis A Ortega, Simón Rodríguez Santana, and Daniel Hernández-Lobato. Variational linearized laplace approximation for bayesian deep learning. *arXiv preprint arXiv:2302.12565*, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, 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 Alex Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*, 2024.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *The 42nd International Conference on Machine Learning*. PMLR, 2025.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International conference on learning representations*, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

C Saunders, A Gammerman, and V Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, pp. 722–726, 1999.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 10160–10177. PMLR, 2023.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in neural information processing systems*, 34:6304–6315, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for stable algorithms. *The Annals of Statistics*, 51(1):290–311, 2023.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 5593–5613. PMLR, 2023.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Dustin Tran, Mike Dusenberry, Mark Van Der Wilk, and Danijar Hafner. Bayesian layers: A module for neural network uncertainty. *Advances in neural information processing systems*, 32, 2019.

Renukanandan Tumu, Matthew Cleaveland, Rahul Mangharam, George Pappas, and Lars Lindemann. Multimodal conformal prediction regions by optimizing convex shape templates. In *6th Annual Learning for Dynamics & Control Conference*, pp. 1343–1356. PMLR, 2024.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Lars van der Laan and Ahmed M Alaa. Self-calibrating conformal prediction. *Advances in Neural Information Processing Systems*, 37:107138–107170, 2024.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In *International Conference on Artificial Intelligence and Statistics*, pp. 8814–8836. PMLR, 2023.

Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent derivative bayesian last layer networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1198–1206. PMLR, 2021.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

A Proofs

A.1 Proof of Proposition 4.1

Proof. Let

$$D_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{tr}}}$$

denote the training set and let

$$D_{\text{cal}} = \{(X_i, Y_i)\}_{i=n_{\text{tr}}+1}^{n_{\text{tr}}+n_{\text{cal}}}$$

denote the calibration set. Write $n = n_{\text{tr}} + n_{\text{cal}}$. After training, the functions $\mu(\cdot)$, $h^2(\cdot)$, $\phi(\cdot)$, and the last-layer covariance matrix Σ are fixed as functions of D_{tr}. Hence the predictive scale

$$v(x) = h^2(x) + \phi(x)^\top \Sigma \phi(x)$$

and the nonconformity score

$$A(x, y) = \frac{|y - \mu(x)|}{\sqrt{v(x)}}$$

are also fixed conditional on D_{tr}.

For each calibration example, define

$$A_i = A(X_i, Y_i), \qquad i = n_{\text{tr}} + 1, \ldots, n.$$

For the test point (X_{n+1}, Y_{n+1}), define

$$A_{n+1} = A(X_{n+1}, Y_{n+1}).$$

By assumption, the calibration examples and the test point are exchangeable conditional on D_{tr}. Since $A(\cdot, \cdot)$ is fixed conditional on D_{tr}, the transformed scores

$$A_{n_{\text{tr}}+1}, \ldots, A_n, A_{n+1}$$

are also exchangeable conditional on D_{tr}.

Let

$$k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil .$$

The conformal quantile q_α is the k-th order statistic of the calibration scores, with the usual conservative convention $q_\alpha = \infty$ if $k > n_{\text{cal}}$. By the standard split conformal rank argument,

$$\mathbb{P}\{A_{n+1} \leq q_\alpha \mid D_{\text{tr}}\} \geq 1 - \alpha.$$

This argument relies only on the exchangeability of the calibration and test scores, and does not require the score distribution to be continuous.

Now observe that

$$A_{n+1} \leq q_\alpha$$

is equivalent to

$$\frac{|Y_{n+1} - \mu(X_{n+1})|}{\sqrt{v(X_{n+1})}} \leq q_\alpha.$$

Since $v(x) > 0$, this is equivalent to

$$|Y_{n+1} - \mu(X_{n+1})| \leq q_\alpha \sqrt{v(X_{n+1})}.$$

Therefore,

$$Y_{n+1} \in \left[\mu(X_{n+1}) - q_\alpha \sqrt{v(X_{n+1})}, \mu(X_{n+1}) + q_\alpha \sqrt{v(X_{n+1})}\right] = C_\alpha(X_{n+1}).$$

Thus,

$$\mathbb{P}\{Y_{n+1} \in C_\alpha(X_{n+1}) \mid D_{\text{tr}}\} \geq 1 - \alpha.$$

Taking expectation over D_{tr} gives

$$\mathbb{P}\{Y_{n+1} \in C_\alpha(X_{n+1})\} \geq 1 - \alpha.$$

□

A.2 Proof of Proposition 4.2

Proof. After training, the representation $\phi(\cdot)$, the heteroscedastic variance function $h^2(\cdot)$, and the last-layer posterior covariance Σ are fixed conditional on D_{tr}. Under the last-layer Gaussian approximation,

$$w \mid D_{\mathrm{tr}} \approx \mathcal{N}(\widehat{w}, \Sigma),$$

and the conditional response model is

$$Y \mid x, w \sim \mathcal{N}(\phi(x)^\top w, h^2(x)).$$

By the law of total variance,

$$\mathrm{Var}(Y \mid x, D_{\mathrm{tr}}) = \mathbb{E}[\mathrm{Var}(Y \mid x, w, D_{\mathrm{tr}}) \mid x, D_{\mathrm{tr}}] + \mathrm{Var}(\mathbb{E}[Y \mid x, w, D_{\mathrm{tr}}] \mid x, D_{\mathrm{tr}}) .$$

The first term is the conditional observation variance. Since $h^2(x)$ is fixed once x and D_{tr} are given,

$$\mathbb{E}[\mathrm{Var}(Y \mid x, w, D_{\mathrm{tr}}) \mid x, D_{\mathrm{tr}}] = \mathbb{E}\big[h^2(x) \mid x, D_{\mathrm{tr}}\big] = h^2(x).$$

The second term is the posterior variance of the last-layer predictive mean. Because

$$\mathbb{E}[Y \mid x, w, D_{\mathrm{tr}}] = \phi(x)^\top w,$$

we have

$$\mathrm{Var}(\mathbb{E}[Y \mid x, w, D_{\mathrm{tr}}] \mid x, D_{\mathrm{tr}}) = \mathrm{Var}(\phi(x)^\top w \mid D_{\mathrm{tr}}).$$

Using $w \mid D_{\mathrm{tr}} \approx \mathcal{N}(\widehat{w}, \Sigma)$, this variance is

$$\mathrm{Var}(\phi(x)^\top w \mid D_{\mathrm{tr}}) = \phi(x)^\top \Sigma \phi(x).$$

Combining the two terms gives

$$\mathrm{Var}(Y \mid x, D_{\mathrm{tr}}) = h^2(x) + \phi(x)^\top \Sigma \phi(x).$$

□

A.3 Proof of Proposition 4.3

Proof. After training, the representation $\phi(\cdot)$ and the heteroscedastic variance function $h^2(\cdot)$ are fixed. For the final-layer weight vector w, the heteroscedastic Gaussian negative log-likelihood, up to constants independent of w, is

$$\ell(w) = \frac{1}{2} \sum_{i=1}^{n_{\mathrm{tr}}} \frac{(y_i - \phi(x_i)^\top w)^2}{h^2(x_i)}.$$

With the Gaussian prior $w \sim \mathcal{N}(0, \lambda^{-1} I)$, the corresponding negative log-posterior is

$$L(w) = \frac{1}{2} \sum_{i=1}^{n_{\mathrm{tr}}} \frac{(y_i - \phi(x_i)^\top w)^2}{h^2(x_i)} + \frac{\lambda}{2} \|w\|_2^2.$$

The last-layer Laplace approximation uses the inverse Hessian of $L(w)$ as the posterior covariance. For each training point,

$$\nabla_w^2 \left[\frac{1}{2} \frac{(y_i - \phi(x_i)^\top w)^2}{h^2(x_i)} \right] = \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

The Hessian of the prior term is

$$\nabla_w^2 \left[\frac{\lambda}{2} \|w\|_2^2 \right] = \lambda I.$$

Therefore,

$$\nabla_w^2 L(w) = \lambda I + \sum_{i=1}^{n_{\text{tr}}} \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

Equivalently, let $H \in \mathbb{R}^{n_{\text{tr}} \times d}$ be the feature matrix whose i-th row is $\phi(x_i)^\top$, and let

$$W = \operatorname{diag}\left(\frac{1}{h^2(x_1)}, \ldots, \frac{1}{h^2(x_{n_{\text{tr}}})}\right).$$

Then

$$H^\top W H = \sum_{i=1}^{n_{\text{tr}}} \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

Thus the last-layer posterior precision is

$$\Sigma^{-1} = \nabla_w^2 L(w) = \lambda I + H^\top W H = \lambda I + \sum_{i=1}^{n_{\text{tr}}} \frac{\phi(x_i)\phi(x_i)^\top}{h^2(x_i)}.$$

□

A.4 Proof of Proposition 4.4

Proof. Fix an input x such that $h^2(x) > 0$, and consider the sequence of CLAPS predictive variances

$$v_n(x) = h^2(x) + \phi(x)^\top \Sigma_n \phi(x).$$

By assumption,

$$\phi(x)^\top \Sigma_n \phi(x) \to 0.$$

Therefore,

$$v_n(x) = h^2(x) + \phi(x)^\top \Sigma_n \phi(x) \to h^2(x).$$

Since $h^2(x) > 0$, we can divide by $h^2(x)$ and obtain

$$\frac{v_n(x)}{h^2(x)} = 1 + \frac{\phi(x)^\top \Sigma_n \phi(x)}{h^2(x)} \to 1.$$

Taking square roots gives

$$\frac{\sqrt{v_n(x)}}{h(x)} = \sqrt{1 + \frac{\phi(x)^\top \Sigma_n \phi(x)}{h^2(x)}} \to 1.$$

This proves that the CLAPS scale converges to the aleatoric locally adaptive scale at x.

It remains to verify the stated sufficient condition. Suppose

$$\lambda_{\min}\big(\lambda I + H_n^\top W_n H_n\big) \to \infty$$

and $\|\phi(x)\|_2 < \infty$. Since

$$\Sigma_n = \big(\lambda I + H_n^\top W_n H_n\big)^{-1},$$

we have

$$0 \le \phi(x)^\top \Sigma_n \phi(x) \le \|\phi(x)\|_2^2 \lambda_{\max}(\Sigma_n).$$

Because Σ_n is the inverse of $\lambda I + H_n^\top W_n H_n$,

$$\lambda_{\max}(\Sigma_n) = \frac{1}{\lambda_{\min}(\lambda I + H_n^\top W_n H_n)}.$$

Hence

$$0 \le \phi(x)^\top \Sigma_n \phi(x) \le \frac{\|\phi(x)\|_2^2}{\lambda_{\min}(\lambda I + H_n^\top W_n H_n)} \to 0.$$

□

B Full Results for Experiment 3

This appendix reports the dataset-level results for Experiment 3. The aggregate results in the main text summarize performance across datasets, while the tables below show the full results for each benchmark dataset. LACP denotes Locally Adaptive Split Conformal Prediction. In this implementation, LACP and DCP use the same heteroscedastic Gaussian scale, so their dataset-level results coincide.

Table 4: Full Experiment 3 results on Concrete.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.9136	0.0268	0.4000	23.7102	29.6977	0.0000
CV+	0.9117	0.0163	0.2000	23.6654	29.7890	-0.0081
LACP	0.9107	0.0375	0.4000	22.2360	27.3074	0.0581
CQR	0.9019	0.0264	0.8000	28.7178	34.6176	-0.2173
DCP	0.9107	0.0375	0.4000	22.2360	27.3074	0.0581
CLAPS	0.9078	0.0384	0.4000	21.9835	26.7948	0.0681

Table 5: Full Experiment 3 results on Energy.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.8915	0.0336	0.6000	11.4294	13.5277	0.0000
CV+	0.8771	0.0349	0.6000	10.8366	13.3069	0.0483
LACP	0.8915	0.0284	0.6000	6.9618	8.3086	0.3887
CQR	0.8980	0.0305	0.4000	8.2924	9.2997	0.2697
DCP	0.8915	0.0284	0.6000	6.9618	8.3086	0.3887
CLAPS	0.8967	0.0310	0.6000	6.9469	8.2805	0.3898

Table 6: Full Experiment 3 results on Yacht.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.8754	0.0252	0.8000	5.6301	9.7653	0.0000
CV+	0.9475	0.0475	0.0000	5.3548	6.5356	0.0273
LACP	0.9049	0.0626	0.4000	2.8589	4.4833	0.5118
CQR	0.9246	0.0502	0.2000	11.9756	15.7035	-1.1811
DCP	0.9049	0.0626	0.4000	2.8589	4.4833	0.5118
CLAPS	0.9049	0.0561	0.4000	2.8003	4.3743	0.5217

Results and analysis. Tables 4–11 provide the dataset-level results underlying the aggregate benchmark in the main text. Across most datasets, CLAPS attains empirical coverage close to the nominal level while reducing average width and interval score relative to Split CP and CV+. The improvement is clearest on Concrete, Energy, Yacht, Airfoil, Bike, and California, where CLAPS yields the narrowest intervals among the conformalized mean-regression and locally adaptive methods, and often obtains the lowest interval score as well.

Compared with LACP and DCP, CLAPS produces small but consistent efficiency gains on several datasets. This pattern is visible in Tables 4, 5, 6, 7, 10, and 11, where the last-layer epistemic correction slightly reduces interval width beyond the aleatoric-only adaptive scale. The gains are modest, but they align with the intended role of CLAPS as a local scaling method: the conformal quantile controls marginal coverage, while the predictive scale reallocates interval width across inputs.

The remaining datasets show the limits of this efficiency pattern. On WineRed, CQR gives the smallest width and interval score, while CLAPS remains close to the locally adaptive baselines. On Naval, the absolute differences between methods are very small, and adaptive scaling does not improve over Split CP.

Table 7: Full Experiment 3 results on Airfoil.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.8833	0.0207	0.8000	12.3212	17.1842	0.0000
CV+	0.8953	0.0087	0.6000	12.3578	16.7552	-0.0058
LACP	0.8860	0.0220	0.8000	11.1616	14.4886	0.0899
CQR	0.8820	0.0180	1.0000	13.8341	17.4969	-0.1284
DCP	0.8860	0.0220	0.8000	11.1616	14.4886	0.0899
CLAPS	0.8853	0.0187	0.8000	10.9975	14.4381	0.1036

Table 8: Full Experiment 3 results on WineRed.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.9012	0.0238	0.4000	2.1148	2.8522	0.0000
CV+	0.9113	0.0212	0.2000	2.1246	2.8122	-0.0060
LACP	0.9044	0.0169	0.4000	2.1567	2.8125	-0.0219
CQR	0.9038	0.0137	0.2000	1.9866	2.7345	0.0584
DCP	0.9044	0.0169	0.4000	2.1567	2.8125	-0.0219
CLAPS	0.9038	0.0188	0.4000	2.1434	2.7959	-0.0156

Table 9: Full Experiment 3 results on Naval.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.8980	0.0092	0.4000	0.0180	0.0216	0.0000
CV+	0.9060	0.0108	0.4000	0.0185	0.0217	-0.0256
LACP	0.9066	0.0098	0.2000	0.0192	0.0219	-0.0659
CQR	0.9096	0.0112	0.2000	0.0233	0.0246	-0.2938
DCP	0.9066	0.0098	0.2000	0.0192	0.0219	-0.0659
CLAPS	0.9072	0.0100	0.2000	0.0192	0.0219	-0.0668

Table 10: Full Experiment 3 results on Bike.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.9018	0.0074	0.4000	360.5115	554.3789	0.0000
CV+	0.9084	0.0100	0.2000	367.0079	552.1318	-0.0185
LACP	0.8986	0.0146	0.6000	341.4513	486.1216	0.0521
CQR	0.9028	0.0064	0.2000	423.8324	509.8446	-0.1770
DCP	0.8986	0.0146	0.6000	341.4513	486.1216	0.0521
CLAPS	0.8976	0.0136	0.6000	338.4354	486.0683	0.0605

Table 11: Full Experiment 3 results on California.

Method	Coverage	Cov. Err.	Viol. Rate	Avg. Width	Int. Score	Rel. Width Red.
Split CP	0.9000	0.0072	0.8000	1.8208	2.9082	0.0000
CV+	0.9088	0.0164	0.2000	1.9129	2.9244	-0.0520
LACP	0.8936	0.0096	0.6000	1.7733	2.7391	0.0264
CQR	0.8966	0.0098	0.6000	2.0406	2.8532	-0.1227
DCP	0.8936	0.0096	0.6000	1.7733	2.7391	0.0264
CLAPS	0.8924	0.0092	0.6000	1.7626	2.7335	0.0324

Taken together, the full results indicate that CLAPS improves interval efficiency on most real regression benchmarks without materially changing the coverage behavior expected from split conformal calibration.

C Sensitivity Analyses

C.1 Common Experimental Design

All sensitivity analyses use a representative subset of four real regression datasets: Concrete, Energy, Bike, and California. Each experiment is repeated over five random seeds, and the target miscoverage level is fixed at $\alpha = 0.10$. Unless a split parameter is explicitly varied, we use the default 60%/20%/20% train/calibration/test split. The base predictor is the same heteroscedastic neural regression model used in the main experiments, with hidden dimension 64, feature dimension 64, and variance floor 10^{-3} unless these quantities are the object of the sensitivity analysis. For each run, CLAPS constructs intervals using the calibrated score based on the combined aleatoric–epistemic scale. When a sensitivity parameter changes the trained model or its training objective, the model is retrained for that setting; when only the last-layer prior precision is varied, the trained model and learned representation are held fixed to isolate the effect of the Laplace ridge term.

We report empirical coverage, absolute coverage error, violation rate, width ratio, interval-score difference, epistemic fraction, and scale ratio. The violation rate is the fraction of dataset–seed runs whose empirical coverage falls below the nominal level. Width ratios and score differences are computed relative to the default setting of the corresponding experiment. The epistemic fraction is the average ratio of the last-layer epistemic variance to the total predictive variance, and the scale ratio is the average multiplicative increase from the aleatoric scale to the combined CLAPS scale. Each table reports averages over the four datasets and five seeds.

C.2 Prior Precision Sensitivity

Experimental setup. We vary the last-layer Laplace prior precision over $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. For each dataset–seed pair, the heteroscedastic model is trained once, and only the posterior covariance $\Sigma = (\lambda I + H^\top W H)^{-1}$ is recomputed. The reference setting is $\lambda = 1$.

Results and analysis. Table 12 shows the expected shrinkage pattern. Larger prior precision reduces the epistemic fraction and moves the scale ratio toward one. Coverage remains close to the nominal level throughout the grid, and the width ratio stays near one despite the change in the epistemic component. The score differences are small, with slightly lower scores for weaker prior precision and slightly higher scores for stronger prior precision. Thus, prior precision changes the epistemic correction in the intended direction without materially changing the calibrated interval-level behavior.

Table 12: Prior precision sensitivity of CLAPS. Width ratio and score delta are computed relative to $\lambda = 1$.

Prior Prec.	Coverage	Cov. Err.	Viol. Rate	Width Ratio	Score Δ	Epi. Frac.	Scale Ratio
10^{-4}	0.9028	0.0195	0.45	0.9963	-0.1264	0.0626	1.0513
10^{-3}	0.9016	0.0207	0.45	0.9932	-0.1205	0.0545	1.0452
10^{-2}	0.9036	0.0203	0.45	0.9975	-0.1073	0.0526	1.0467
10^{-1}	0.9050	0.0217	0.40	1.0003	-0.0627	0.0381	1.0348
1	0.9035	0.0230	0.40	1.0000	0.0000	0.0230	1.0219
10	0.9026	0.0228	0.40	1.0000	0.0666	0.0125	1.0128
100	0.9028	0.0217	0.40	1.0030	0.0880	0.0062	1.0071

C.3 Calibration Size Sensitivity

Experimental setup. We vary the calibration fraction over $\{0.10, 0.15, 0.20, 0.30\}$ while fixing the test fraction at 0.20, which induces training fractions $\{0.70, 0.65, 0.60, 0.50\}$. This experiment therefore measures sensitivity to the train–calibration allocation. The reference setting is the default calibration fraction 0.20.

Results and analysis. Table 13 shows stable coverage across the tested allocations. Width ratios remain close to one, indicating that the final interval size is not strongly tied to the default split. The largest calibration fraction gives a mild increase in width and score, consistent with the loss of training samples when more data are moved into calibration. The default 60%/20%/20% split is therefore a stable middle point rather than a fragile choice.

Table 13: Calibration size sensitivity of CLAPS. Width ratio and score delta are computed relative to calibration fraction 0.20.

Cal. Frac.	Train Frac.	Coverage	Cov. Err.	Viol. Rate	Width Ratio	Score Δ	Epi. Frac.	Scale Ratio
0.10	0.70	0.9093	0.0196	0.50	0.9984	-1.0829	0.0295	1.0269
0.15	0.65	0.9060	0.0228	0.45	0.9879	-0.5222	0.0277	1.0234
0.20	0.60	0.9028	0.0204	0.45	1.0000	0.0000	0.0199	1.0209
0.30	0.50	0.9017	0.0185	0.50	1.0210	0.4816	0.0246	1.0270

C.4 Representation Dimension Sensitivity

Experimental setup. We vary the learned representation dimension over $d_\phi \in \{16, 32, 64, 128\}$ while keeping the hidden dimension fixed at 64. Since d_ϕ changes the architecture and the dimension of the last-layer Laplace approximation, the heteroscedastic model is retrained for each setting. The reference setting is $d_\phi = 64$.

Results and analysis. Table 14 shows that coverage remains near the nominal level across the tested dimensions. The epistemic fraction and scale ratio increase with d_ϕ, reflecting the larger last-layer parameter space. The smallest representation gives wider and less efficient intervals on average, while the largest representation improves the aggregate score but also increases the epistemic contribution. The default $d_\phi = 64$ provides a stable intermediate setting.

Table 14: Representation dimension sensitivity of CLAPS. Width ratio and score delta are computed relative to $d_\phi = 64$.

Feature Dim.	Coverage	Cov. Err.	Viol. Rate	Width Ratio	Score Δ	Epi. Frac.	Scale Ratio
16	0.8986	0.0208	0.45	1.0557	5.9434	0.0063	1.0052
32	0.8956	0.0225	0.45	0.9976	3.7177	0.0133	1.0120
64	0.9028	0.0204	0.45	1.0000	0.0000	0.0199	1.0209
128	0.9008	0.0205	0.55	0.9976	-4.2372	0.0539	1.0563

C.5 Variance Floor Sensitivity

Experimental setup. We vary the lower bound on the aleatoric variance over $\epsilon \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. The floor is used in the heteroscedastic variance head, $h^2(x) = \text{softplus}(\cdot) + \epsilon$, and in the lower-bounded variance terms used for the Laplace and conformal scale computations. Since the floor affects the training objective, the model is retrained for each value. The reference setting is $\epsilon = 10^{-3}$.

Results and analysis. Table 15 shows that interval-level performance is stable across the tested floors. Coverage stays close to the nominal level, and width ratios remain near one. The floor-active rate is essentially zero, and the lower-tail diagnostics of $h^2(x)$ stay above the imposed floor on average, so the default floor does

not drive the results through broad clipping. Smaller floors increase the local scale ratio, but the calibrated interval width and coverage change little.

Table 15: Variance floor sensitivity of CLAPS. Width ratio and score delta are computed relative to $\epsilon = 10^{-3}$.

Var. Floor	Coverage	Cov. Err.	Viol. Rate	Width Ratio	Score Δ	Epi. Frac.	Scale Ratio	Min h^2	P01 h^2	Floor Active
10^{-6}	0.8962	0.0235	0.55	1.0001	0.1646	0.0237	1.2159	0.0147	0.0266	0.0000
10^{-5}	0.8995	0.0212	0.50	1.0036	0.1635	0.0197	1.0788	0.0148	0.0265	0.0000
10^{-4}	0.8994	0.0214	0.50	1.0027	0.0908	0.0184	1.0313	0.0148	0.0267	0.0000
10^{-3}	0.9028	0.0204	0.45	1.0000	0.0000	0.0199	1.0209	0.0154	0.0268	0.0000