# Epistemic Uncertainty in Conformal Scores: A Unified Approach

Luben M. C. Cabezas, Vagner S. Santos, Thiago R. Ramos, Rafael Izbicki

April 18, 2025

Presented by Mian Wei

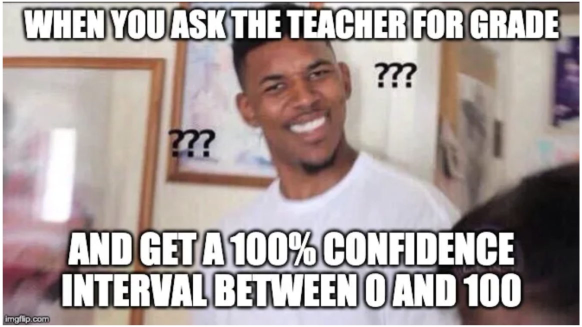
# Uncertainty

What is uncertainty? and why do we care?

[1]source: https://medium.com/data-science/how-confidence-and-prediction-intervals-work-4592019576d8

# Uncertainty

What is uncertainty? and why do we care?

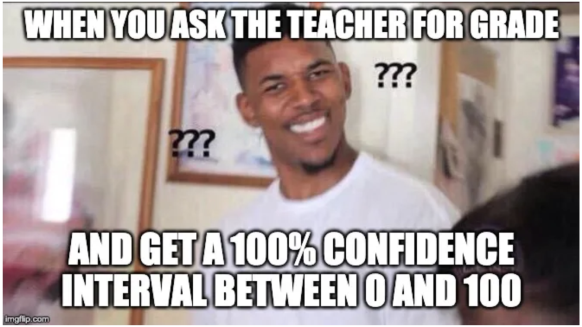


[1]

[1] source: https://medium.com/data-science/how-confidence-and-prediction-intervals-work-4592019576d8

# Uncertainty

What is uncertainty? and why do we care?



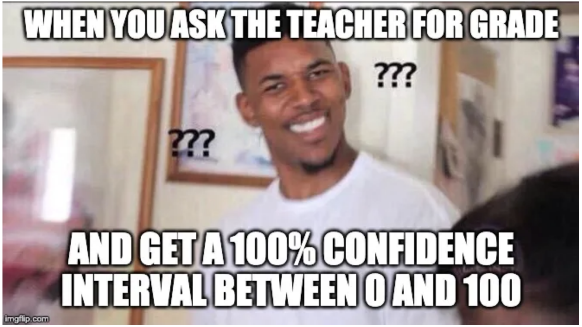
[1]

- 100% confidence: GOOD!
- Between 0 to 100: USELESS!

[1] source: https://medium.com/data-science/how-confidence-and-prediction-intervals-work-4592019576d8

# Uncertainty

What is uncertainty? and why do we care?



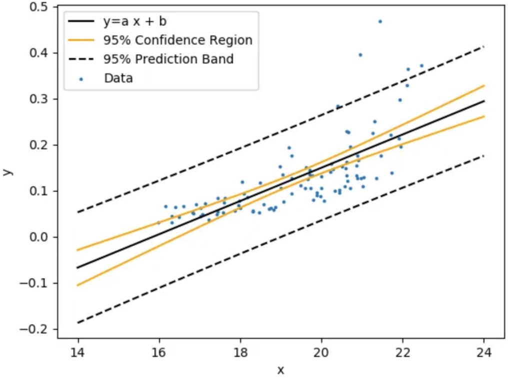
[1]

- 100% confidence: GOOD!
- Between 0 to 100: USELESS!

- $\Longrightarrow$ Less confident but still high
- $\Longrightarrow$ Band as tight as possible

[1] source: https://medium.com/data-science/how-confidence-and-prediction-intervals-work-4592019576d8

# Confidence Interval v.s. Prediction Interval



[2]

"What is the average blood pressure?" v.s. "What might the next patient's blood pressure be?"

[2] source: https://medium.com/data-science/how-confidence-and-prediction-intervals-work-4592019576d8

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)

**Procedure:**

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)

**Procedure:**

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.

**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

- Split data into training set and calibration set;

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \dots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

- Split data into training set and calibration set;
- Fit the model $\hat{f}$ on the training set;

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

- Split data into training set and calibration set;
- Fit the model $\hat{f}$ on the training set;
- Compute $s(x, y)$ on the calibration set;

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \ldots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

- Split data into training set and calibration set;
- Fit the model $\hat{f}$ on the training set;
- Compute $s(x, y)$ on the calibration set;
- Compute the empirical quantile $q_{1-\alpha}$ of the scores;

# Conformal Prediction

**Assumption:** The data points $(x_1, y_1), \dots, (x_n, y_n)$ are **exchangeable**.
**Decide:**

- A predictive model $\hat{f}(x)$ (e.g., regression)
- A nonconformity score function $s(x, y)$ (e.g., $s(x, y) = |y - \hat{f}(x)|$)
- A desired threshold $\alpha$ (e.g., 95%)

**Procedure:**

- Split data into training set and calibration set;
- Fit the model $\hat{f}$ on the training set;
- Compute $s(x, y)$ on the calibration set;
- Compute the empirical quantile $q_{1-\alpha}$ of the scores;
- Construct the prediction interval for a new input $x_{n+1}$ as:

$$\widehat{C}(x_{n+1}) = \left[\hat{f}(x_{n+1}) - q_{1-\alpha},\ \hat{f}(x_{n+1}) + q_{1-\alpha}\right]$$

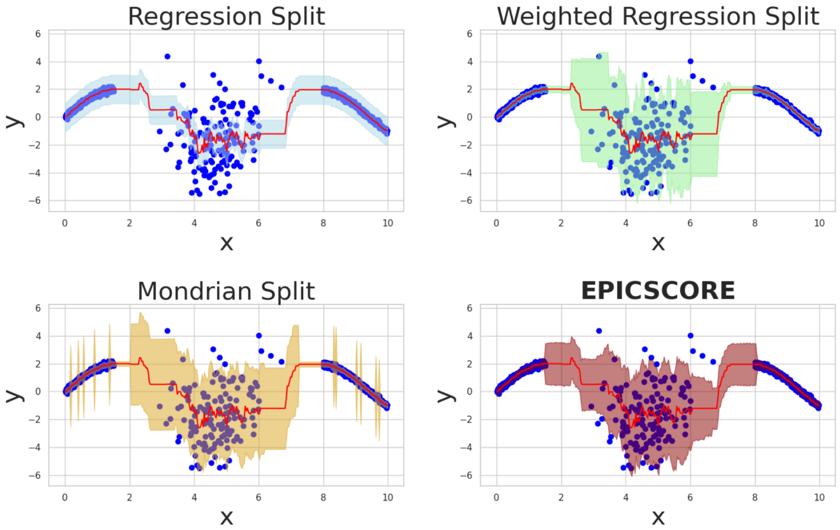
# Aleatoric and Epistemic Uncertainty

**Aleatoric Uncertainty:** inherent in the data, irreducible
**Epistemic Uncertainty:** due to lack of knowledge (model or data), reducible

**For Conformal Prediction**

- Aleatoric Uncertainty: noisy data $\rightarrow$ wider interval, naturally captured
- Epistemic Uncertainty: not captured
  - Model: predefined. ”*Is this model confident here?*”
  - Data: In regions with no training data, may still produce confident-looking intervals

# Aleatoric and Epistemic Uncertainty



# Motivation & Novelty

**Goal:** integrate epistemic uncertainty into the conformal prediction framework.

**Two main directions:**

- Redesign the conformal score, e.g., weighted regression split:

$$s(x, y) = \frac{|y - \hat{y}|}{\hat{\sigma}(x)}$$

- Adapt cutoffs locally, e.g., Mondrian conformal regression, Partition the feature space into bins

**For this paper:**

- Uses Bayesian modeling to capture epistemic uncertainty;
- Model-agnostic — works with any Bayesian model;
- Can be layered on any conformal score;
- Marginal and asymptotic conditional coverage.

# EPICSCORE

**Input:**

- Dataset $D = \{(X_i, Y_i)\}_{i=1}^{n}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$.
- A conformal score function $s(x, y)$ .
- Nominal level $\alpha \in (0, 1)$.
- A new test point $X_{n+1}$.

**Input:**

- Dataset $D = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$.
- A conformal score function $s(x, y)$ .
- Nominal level $\alpha \in (0, 1)$.
- A new test point $X_{n+1}$.

**Step I: Fit Conformal Scores**

1. Split $D$ into:
   - Training set $D_{\text{train}}$
   - Calibration set $D_{\text{cal}}$
2. Use $D_{\text{train}}$ to fit a base predictive model, and construct the initial conformal score function $s(x, y)$.

## Step II: Fit the Predictive Function

1. Split the calibration set $D_{\text{cal}}$ into:
   - $D_{\text{cal},1}$ – for fitting the predictive distribution
   - $D_{\text{cal},2}$ – for computing the quantile cutoff

## Step II: Fit the Predictive Function

1. Split the calibration set $D_{\text{cal}}$ into:
   - $D_{\text{cal},1}$ – for fitting the predictive distribution
   - $D_{\text{cal},2}$ – for computing the quantile cutoff
2. Transform $D_{\text{cal},1}$ into score data:

$$D = \{(X, S) \mid S = s(X, Y),\ (X, Y) \in D_{\text{cal},1}\}$$

**Step II: Fit the Predictive Function**

1. Split the calibration set $D_{\text{cal}}$ into:
   - $D_{\text{cal},1}$ – for fitting the predictive distribution
   - $D_{\text{cal},2}$ – for computing the quantile cutoff
2. Transform $D_{\text{cal},1}$ into score data:

$$D = \{(X, S) \mid S = s(X, Y),\ (X, Y) \in D_{\text{cal},1}\}$$

3. Use Bayesian models to estimate the predictive cumulative distribution function (CDF):

$$F(s \mid x, D) = \int F(s \mid x, \theta) f(\theta \mid D)\, d\theta$$

### Step II: Fit the Predictive Function

1. Split the calibration set $D_{\text{cal}}$ into:
   - $D_{\text{cal},1}$ – for fitting the predictive distribution
   - $D_{\text{cal},2}$ – for computing the quantile cutoff
2. Transform $D_{\text{cal},1}$ into score data:

$$D = \{(X, S) \mid S = s(X, Y),\ (X, Y) \in D_{\text{cal},1}\}$$

3. Use Bayesian models to estimate the predictive cumulative distribution function (CDF):

$$F(s \mid x, D) = \int F(s \mid x, \theta) f(\theta \mid D)\, d\theta$$

### Modeling choices for $F(s \mid x, D)$:

- Gaussian Processes (GPs)
- Bayesian Additive Regression Trees (BART)
- Mixture Density Networks with MC-Dropout

**Step III Procedure:**

1. Compute EPICSCORE for all elements of $D_{\text{cal},2}$ by:

$$s'(x,y) = F(s(x,y) \mid x, D)$$

# EPICSCORE

**Step III Procedure:**

1. Compute EPICSCORE for all elements of $D_{\text{cal},2}$ by:

$$s'(x, y) = F(s(x, y) \mid x, D)$$

2. Compute the $(1 - \alpha)$ empirical quantile $t_{1-\alpha}$ of the conformal scores

**Step III Procedure:**

1. Compute EPICSCORE for all elements of $D_{\text{cal},2}$ by:

$$s'(x, y) = F(s(x, y) \mid x, D)$$

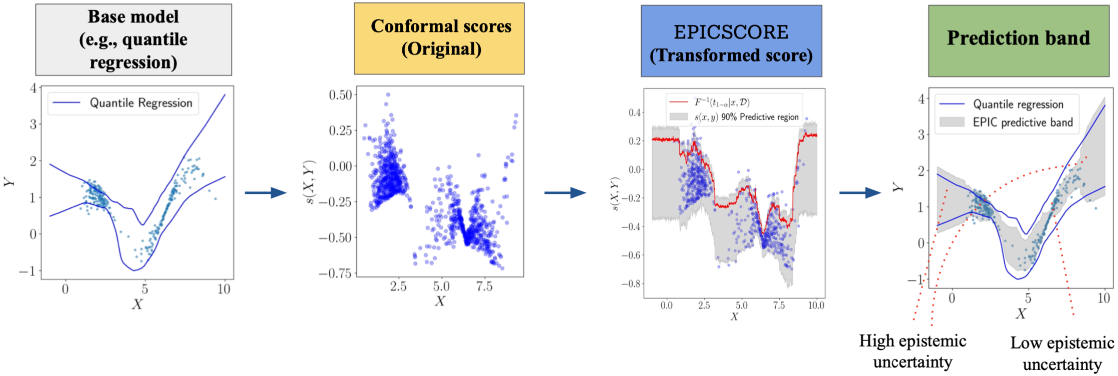
2. Compute the $(1 - \alpha)$ empirical quantile $t_{1-\alpha}$ of the conformal scores
3. Define the prediction region for a new input $x_{n+1}$:

$$\mathcal{R}_{\text{EPIC}}(x_{n+1}) = \{y : s'(x_{n+1}, y) \leq t_{1-\alpha}\}$$

or equivalently, using the original score $s$:

$$\mathcal{R}_{\text{EPIC}}(x_{n+1}) = \{y : s(x_{n+1}, y) \leq F^{-1}(t_{1-\alpha} \mid x_{n+1}, D)\}$$

# Intuition Behind




"Uncertainty of uncertainty"

# Special Cases

**Case 1: Regression**

- Original score

$$s(x, y) = |y - g(x)|$$

- EPICSCORE:

$$s'(x, y) = F(s(x, y) \mid x, D)$$

- Prediction interval becomes:

$$g(x) \pm F^{-1}(t_{1-\alpha} \mid x, D)$$

- **Example:** Suppose $g(10) = 80$, and Monte Carlo Dropout gives:

$$s(x = 10, y) \sim \mathcal{N}(5, 2^2)$$

$$F^{-1}(0.9) \approx 5 + 1.28 \cdot 2 = 7.56$$

$\rightarrow$ Final interval: $\boxed{[72.44,\ 87.56]}$

# Special Cases

**Case 2: Quantile Regression (CQR)**

- Original score:

$$s(x, y) = \max\{q_{\alpha_1}(x) - y,\ y - q_{\alpha_2}(x)\}$$

- EPICSCORE expands both tails using Bayesian uncertainty:

$$[q_{\alpha_1}(x) - F^{-1},\ q_{\alpha_2}(x) + F^{-1}]$$

# Special Cases

**Case 2: Quantile Regression (CQR)**

- Original score:

$$s(x, y) = \max\{q_{\alpha_1}(x) - y,\ y - q_{\alpha_2}(x)\}$$

- EPICSCORE expands both tails using Bayesian uncertainty:

$$[q_{\alpha_1}(x) - F^{-1},\ q_{\alpha_2}(x) + F^{-1}]$$

**Case 3: Classification**

- Examples:

$$s(x, y) = -\hat{P}(y \mid x) \quad \text{or} \quad \sum_{y' : \hat{P}(y'|x) > \hat{P}(y|x)} \hat{P}(y'|x)$$

- EPICSCORE computes:

$$s'(x, y) = \sum_{y' : s(x, y') \leq s(x, y)} P(y' \mid x, D)$$

- Larger prediction sets for uncertain (outlier) inputs.

# Theory

## Theorem (Marginal Coverage)

*Assuming that the data are independent and identically distributed (i.i.d.), the confidence region constructed by* `EPICSCORE` *satisfies marginal coverage, that is,*

$$\mathbb{P}(Y \in R_{EPIC}(\mathbf{X})) \geq 1 - \alpha.$$

*Moreover, if the fitted scores follow a continuous joint distribution, the upper bound also holds:*

$$\mathbb{P}(Y \in R_{EPIC}(\mathbf{X})) \leq 1 - \alpha + \frac{1}{1 + |\mathcal{D}_{cal,2}|}.$$

# Theory

**Assumption 1.** For any $\varepsilon > 0$, we assume uniform convergence in probability over the randomness in $D$:

$$\lim_{|D|\to\infty} \mathbb{P}\left(\sup_{s,\mathbf{x}} |F(s \mid \mathbf{x}, D) - F(s \mid \mathbf{x}, \theta^*)| > \varepsilon\right) = 0.$$

## Theorem (Asymptotic Conditional Coverage)

*Under Assumption 1, and assuming that the data are independent and identically distributed (i.i.d.), the confidence region constructed by* `EPICSCORE` *satisfies the asymptotic conditional coverage condition, that is:*

$$\lim_{|\mathcal{D}_{cal}|\to\infty} \mathbb{P}(Y \in R_{EPIC}(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}) = 1 - \alpha.$$

# Experiments

**Setup**

- 13 datasets
- 40% training, 40% calibration, 20% test, averaged over 50 random splits
- AISL (Average Interval Score Loss)

**Models for the predictive distribution**

- **Bayesian Additive Regression Trees (BART)** [CGM12]: Sum-of-trees model with heteroskedastic noise modeling.
- **Gaussian Processes (GP)** [WR06]: Nonparametric Bayesian model with kernel-defined similarity.
- **Mixture Density Networks with Monte Carlo Dropout (MDN-MC)** [Bis94, GG16]: Neural network modeling score distribution as a mixture of Gaussians with Bayesian approximation.

# Experiments

**Quantile-Regression baselines**

- **CQR** [RPC19]: Conformalized quantile regression with fixed cutoff.
- **CQR-r** [SC20]: Scaled version of CQR to adapt to interval width.
- **UACQR-P, UACQR-S** [RFBW24]: Ensemble-based corrections to capture epistemic uncertainty.

**Regression Baselines**

- **Regression Split** [LW14]: Classic conformal method using residuals.
- **Weighted Regression Split** [LGR$^+$18]: Adjusts cutoff using predicted residual scale.
- **Mondrian Conformal Regression** [BJ20]: Builds local bins to improve conditional coverage.

# Experiments

Table 1: Quantile regression AISL values for each method and dataset. The table reports the mean across 50 runs, with twice the standard deviation in brackets. Bold values indicate the best-performing method within a 95% confidence interval. EPICSCORE demonstrates strong performance across most datasets and consistently ranks among the top methods.

| Dataset | EPIC-BART | EPIC-GP | EPIC-MDN | CQR | CQR-r | UACQR-P | UACQR-S |
|---|---|---|---|---|---|---|---|
| airfoil | 19.361 (0.234) | 19.704 (0.27) | **18.799 (0.29)** | 20.521 (0.234) | 20.535 (0.236) | 23.021 (0.337) | 20.188 (0.3) |
| bike $\times(10^1)$ | 44.722 (0.297) | 47.818 (0.320) | **43.858 (0.326)** | 45.628 (0.256) | 45.638 (0.258) | 53.413 (0.376) | **43.815 (0.385)** |
| concrete | **42.765 (0.723)** | 45.276 (0.764) | 44.442 (0.8) | 46.882 (0.681) | 46.896 (0.683) | 52.789 (1.097) | 47.324 (1.349) |
| cycle | 34.435 (0.142) | 35.054 (0.131) | **34.077 (0.129)** | 39.218 (0.134) | 39.408 (0.136) | 43.775 (0.181) | 35.346 (0.197) |
| electric | 0.099 (< 0.001) | 0.096 (< 0.001) | **0.082 (< 0.001)** | 0.102 (0.001) | 0.102 (0.001) | 0.111 (0.001) | 0.097 (< 0.001) |
| homes $\times(10^5)$ | 7.739 (0.066) | 8.098 (0.072) | **7.225 (0.049)** | 8.360 (0.075) | 8.433 (0.078) | 11.427 (0.131) | 8.544 (0.107) |
| meps19 | **65.085 (1.469)** | **64.907 (1.56)** | **64.3 (1.528)** | **64.239 (1.56)** | **64.239 (1.56)** | 71.015 (1.763) | **63.737 (1.461)** |
| protein | 17.687 (0.019) | 18.096 (0.037) | **17.417 (0.019)** | 17.7 (0.015) | 17.7 (0.016) | 18.149 (0.015) | 17.691 (0.015) |
| star $\times(10^1)$ | **98.466 (0.768)** | **98.033 (0.750)** | **98.725 (0.754)** | **97.770 (0.725)** | **97.791 (0.724)** | 99.782 (0.647) | 99.809 (0.968) |
| superconductivity | 74.37 (0.222) | 80.278 (0.266) | **70.212 (0.196)** | 75.496 (0.219) | 75.508 (0.218) | 87.929 (0.513) | 73.971 (0.404) |
| WEC $\times(10^5)$ | 2.925 (0.009) | 2.665 (0.012) | **2.374 (0.010)** | 3.138 (0.009) | 3.142 (0.009) | 3.517 (0.010) | 3.046 (0.010) |
| winered | **3.007 (0.058)** | **3.009 (0.059)** | **2.977 (0.05)** | **2.979 (0.069)** | **2.978 (0.069)** | 3.059 (0.069) | **2.999 (0.063)** |
| winewhite | 3.334 (0.03) | 3.327 (0.034) | **3.219 (0.03)** | 3.316 (0.036) | 3.315 (0.036) | 3.378 (0.038) | **3.2 (0.036)** |

# Experiments

Table 2: Regression AISL values for each method and dataset. The reported values represent the average across 50 runs, with two times the standard deviation in parentheses. Bolded values highlight the method with superior performance within a 95% confidence interval. EPICSCORE demonstrates competitive or superior performance compared to other methods.

| Dataset | EPIC-BART | EPIC-GP | EPIC-MDN | Mondrian | Reg-split | Weighted |
|---|---|---|---|---|---|---|
| airfoil | **19.747 (0.767)** | **20.287 (0.686)** | **19.823 (0.675)** | 21.532 (0.919) | 21.201 (0.98) | **20.276 (0.819)** |
| bike $\times(10^1)$ | **36.381 (0.463)** | 41.448 (0.575) | **37.041 (0.452)** | 38.190 (0.403) | 43.918 (0.567) | 37.773 (0.468) |
| concrete | **52.098 (2.237)** | **52.998 (2.359)** | **51.648 (2.185)** | 61.915 (2.815) | **54.902 (2.634)** | 58.399 (3.165) |
| cycle | **19.418 (0.211)** | **19.522 (0.221)** | **19.436 (0.213)** | **19.403 (0.226)** | **19.73 (0.208)** | **19.49 (0.207)** |
| electric | **0.048 (<0.001)** | 0.049 (<0.001) | **0.048 (<0.001)** | 0.05 (<0.001) | 0.05 (0.001) | **0.048 (<0.001)** |
| homes $\times(10^5)$ | 5.921 (0.0716) | 6.192 (0.0689) | **5.546 (0.0545)** | 5.710 (0.053) | 7.569 (0.098) | 5.860 (0.056) |
| meps19 | 86.039 (2.421) | 87.086 (2.405) | **75.061 (1.807)** | 79.192 (1.821) | 109.83 (2.695) | 92.433 (3.259) |
| protein | 18.885 (0.054) | 18.772 (0.065) | 17.735 (0.055) | **17.586 (0.051)** | 19.423 (0.055) | 18.314 (0.065) |
| star $\times(10^1)$ | **105.616 (1.255)** | **106.112 (0.998)** | **106.368 (1.173)** | 109.346 (1.119) | **105.250 (1.038)** | 129.492 (1.657) |
| superconductivity | 54.895 (0.364) | 59.16 (0.449) | **53.406 (0.365)** | 58.065 (0.313) | 68.183 (0.418) | 54.981 (0.345) |
| WEC $\times(10^5)$ | 1.437 (0.010) | 1.435 (0.011) | **1.283 (0.009)** | **1.294 (0.009)** | 1.620 (0.009) | 1.410 (0.009) |
| winered | **3.152 (0.07)** | **3.171 (0.064)** | **3.101 (0.062)** | 3.262 (0.069) | **3.214 (0.063)** | 3.415 (0.067) |
| winewhite | **3.104 (0.027)** | 3.187 (0.029) | **3.129 (0.029)** | **3.087 (0.023)** | 3.181 (0.028) | 3.189 (0.033) |

# Accommodation

**Is it worth reading?** Yes.

- Integrates epistemic uncertainty into conformal prediction
- Captures the uncertainty of uncertainty
- Provides theoretical guarantees
- Includes extensive experimental results

Christopher M Bishop.
Mixture density networks.
Technical report, Aston University, 1994.

Henrik Boström and Ulf Johansson.
Mondrian conformal regressors.
In *Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR, 2020.

Hugh A Chipman, Edward I George, and Robert E McCulloch.
Bart: Bayesian additive regression trees.
*The Annals of Applied Statistics*, 6(1):266–298, 2012.

Yarin Gal and Zoubin Ghahramani.
Dropout as a bayesian approximation: Representing model uncertainty in deep learning.
In *ICML*, pages 1050–1059, 2016.

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

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.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett.
Integrating uncertainty awareness into conformalized quantile regression.
In *AISTATS*, pages 1540–1548, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel Candès.
Conformalized quantile regression.
In *NeurIPS*, pages 3543–3553, 2019.

Matteo Sesia and Emmanuel J Candès.
A comparison of some conformal quantile regression methods.
*Stat*, 9(1):e261, 2020.

Christopher KI Williams and Carl Edward Rasmussen.
*Gaussian Processes for Machine Learning*.
MIT Press, 2006.