Atomwise

Uncertainty quantification for neural network based molecular property prediction

Aryan Pedawi, **Hossam Ashtawy, Brandon M. Anderson**

Atomwise, Inc.

Disclaimer ⚠

- I’m a statistician by training, not a chemist
- Some of my interests:
  - Applying statistics to interesting/important problems in the sciences
  - Applying statistics to state-of-the-art machine learning systems, especially deep neural networks
- This talk is mainly focused on the statistics, but with a lot of visuals to gently introduce the concepts and their relevance
- At the end, we will look at an application in chemistry to ground things

Motivation

- Neural networks are increasingly being utilized in virtual high throughput screening of large compound libraries
- Premium for reliability
- Point predictors vs. interval predictors
- Growing understanding in the statistics/ML community on how to make interval predictors statistically rigorous
- We apply these ideas to NN-based molecular property prediction and develop some new ideas along the way

Uncertainty quantification

Reliability vs. usefulness

- Uncertainty quantification should be **reliable**
  - If a model predicts that an event will occur with 90% probability, then across all such predictions, the event should occur 90% of the time
  - This property is sometimes called **coverage**
- Uncertainty quantification should be **useful**
  - Overly broad or non-adaptive prediction intervals aren’t helpful
  - Easy (cf. hard) examples → tight (cf. wide) prediction intervals

Interval predictors

Preliminaries and notation

- We consider prediction tasks from an input domain $\mathbf{X}$ to a target domain $\mathbf{Y} \subseteq \mathbb{R}$
- We focus on *set-valued* predictors C_β: $\mathbf{X} \to \Delta\mathbf{Y}$
  - A function that takes x as input and returns a prediction interval $C_\beta(x)$ over plausible values of y
  - $\beta \in (0, 1)$ is the desired confidence level

Reliability desiderata

Notions of coverage

Suppose we have a set-valued function $C_\beta(x)$ which returns a $100\beta\%$ prediction interval. We would like the following:

1. Marginal (or, unconditional) coverage guarantee
   - $\Pr[y \in C_\beta(x)] \approx \beta$

2. Conditional coverage guarantee
   - $\Pr[y \in C_\beta(x) \mid x] \approx \beta$

3. Balanced coverage guarantee (for interval predictors)
   - $\Pr[y > C_\beta(x)] \approx \Pr[y < C_\beta(x)]$
   - $\Pr[y > C_\beta(x) \mid x] \approx \Pr[y < C_\beta(x) \mid x]$

Reliability desiderata

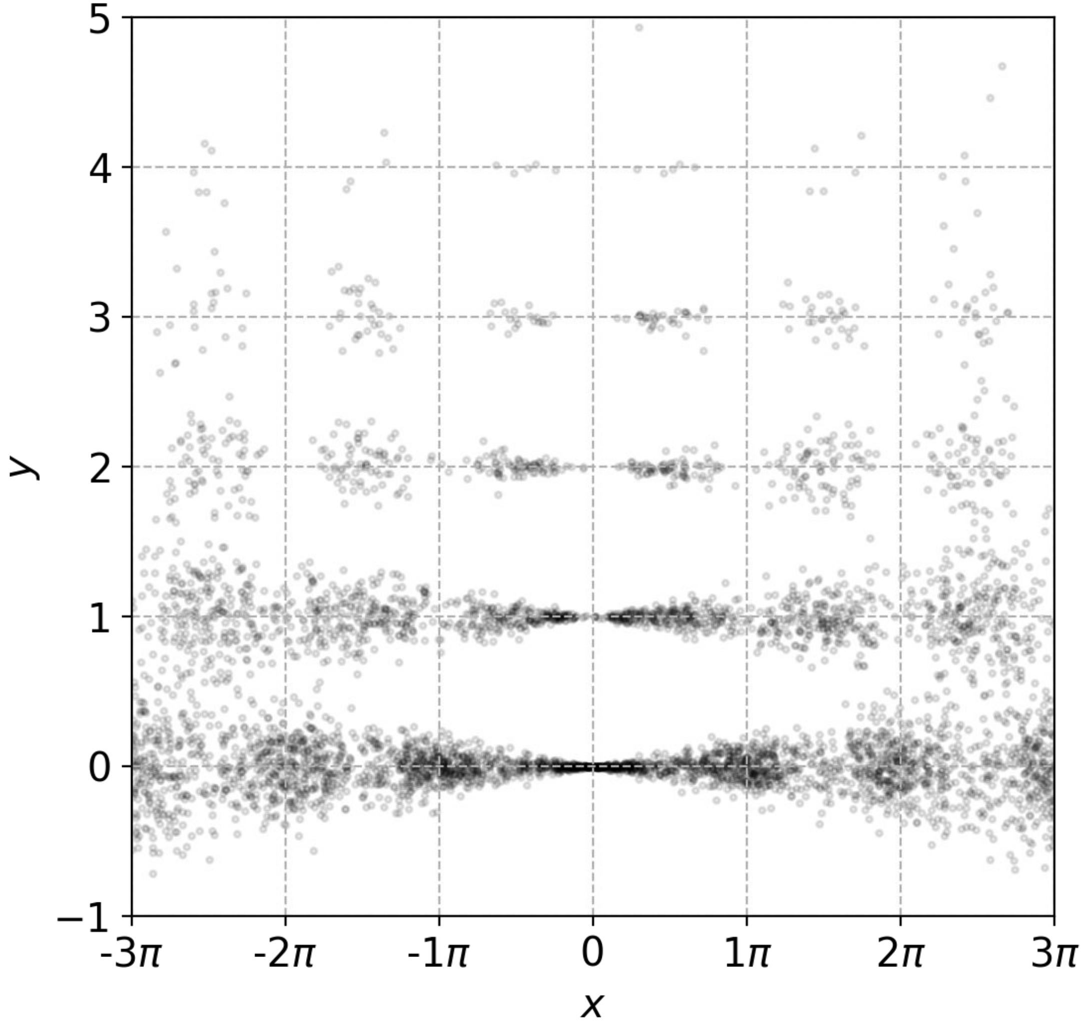
Notions of coverage

Suppose we have a set-valued function $C_\beta(x)$ which returns a $100\beta\%$ prediction interval. We would like the following:

1. Marginal (or, unconditional) coverage guarantee **Easy**
   - $\Pr[y \in C_\beta(x)] \approx \beta$

2. Conditional coverage guarantee **Hard**
   - $\Pr[y \in C_\beta(x) \mid x] \approx \beta$

3. Balanced coverage guarantee (for interval predictors)
   - $\Pr[y > C_\beta(x)] \approx \Pr[y < C_\beta(x)]$
   - $\Pr[y > C_\beta(x) \mid x] \approx \Pr[y < C_\beta(x) \mid x]$



Consider the distribution $p^*(x, y)$, defined as follows[1]:

$$x \sim \mathrm{Uniform}(-3\pi, 3\pi)$$

$$z_1 \sim \mathrm{Normal}(0, 1)$$

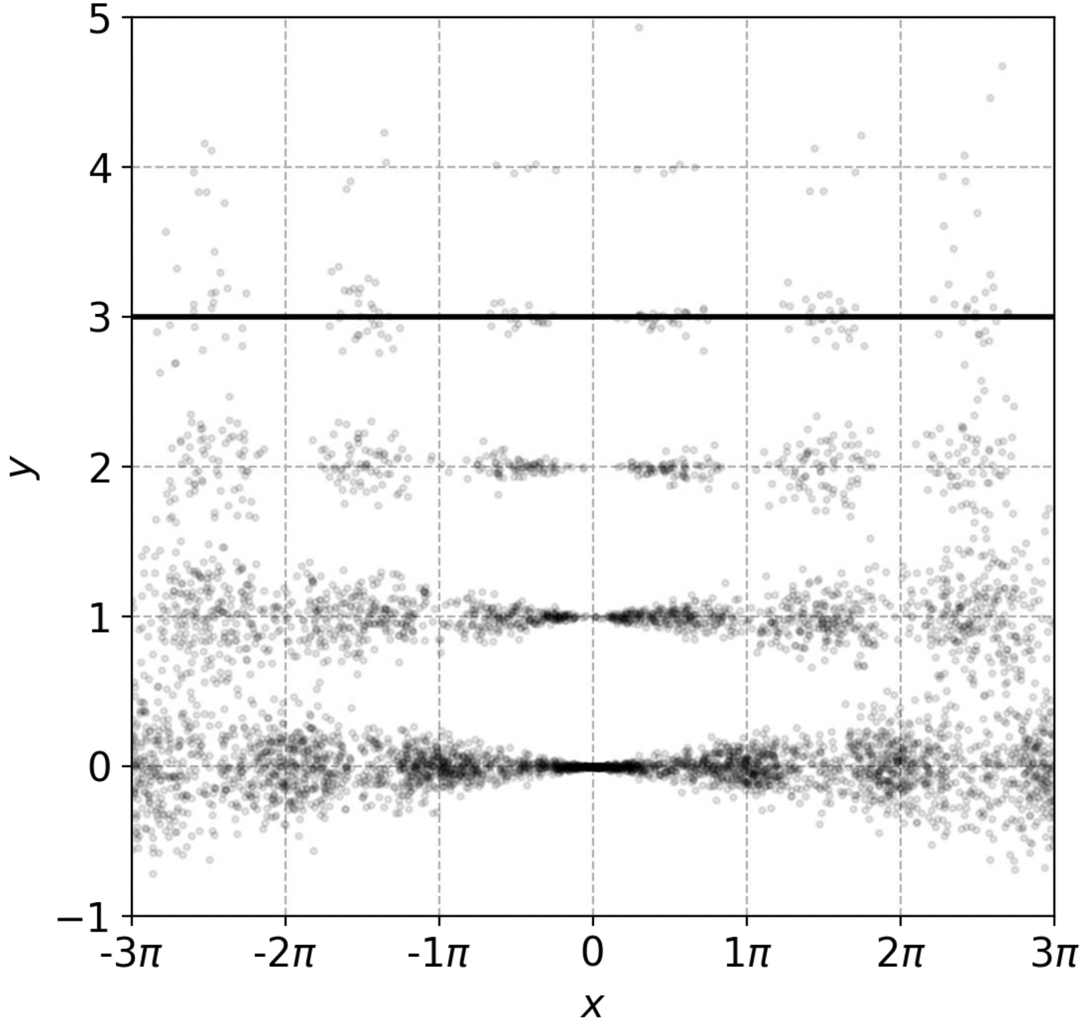
$$z_2 \sim \mathrm{Normal}(0, 1)$$

$$u \sim \mathrm{Uniform}(0, 1)$$

$$v \sim \mathrm{Poisson}(\sin^2(x) + 0.1)$$

$$y = v + 0.03\, x\, z_1 + 25\, \mathrm{I}[u < 0.01]\, z_2$$

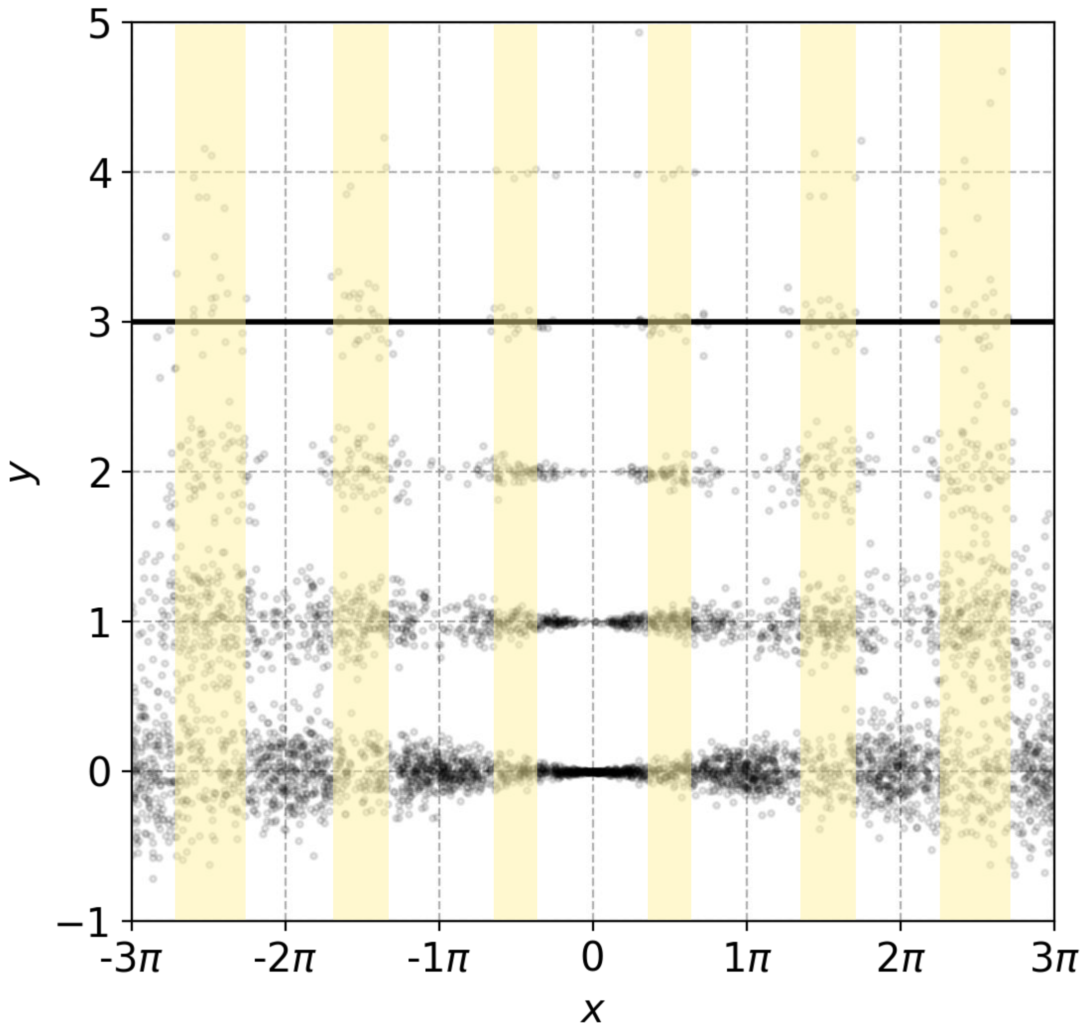
[1] Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, *32*.



Imagine the following game:

- We are given a dataset *D* of (*x*, *y*) pairs from *p**(*x*, *y*), as shown to the left
- A new pair is sampled from *p**(*x*, *y*)
- We observe *x*, but *y* is hidden
- We can pay 5¢ to reveal *y*
- If *y* > 3, we get $1; otherwise, we get $0
- When should we take the gamble*?

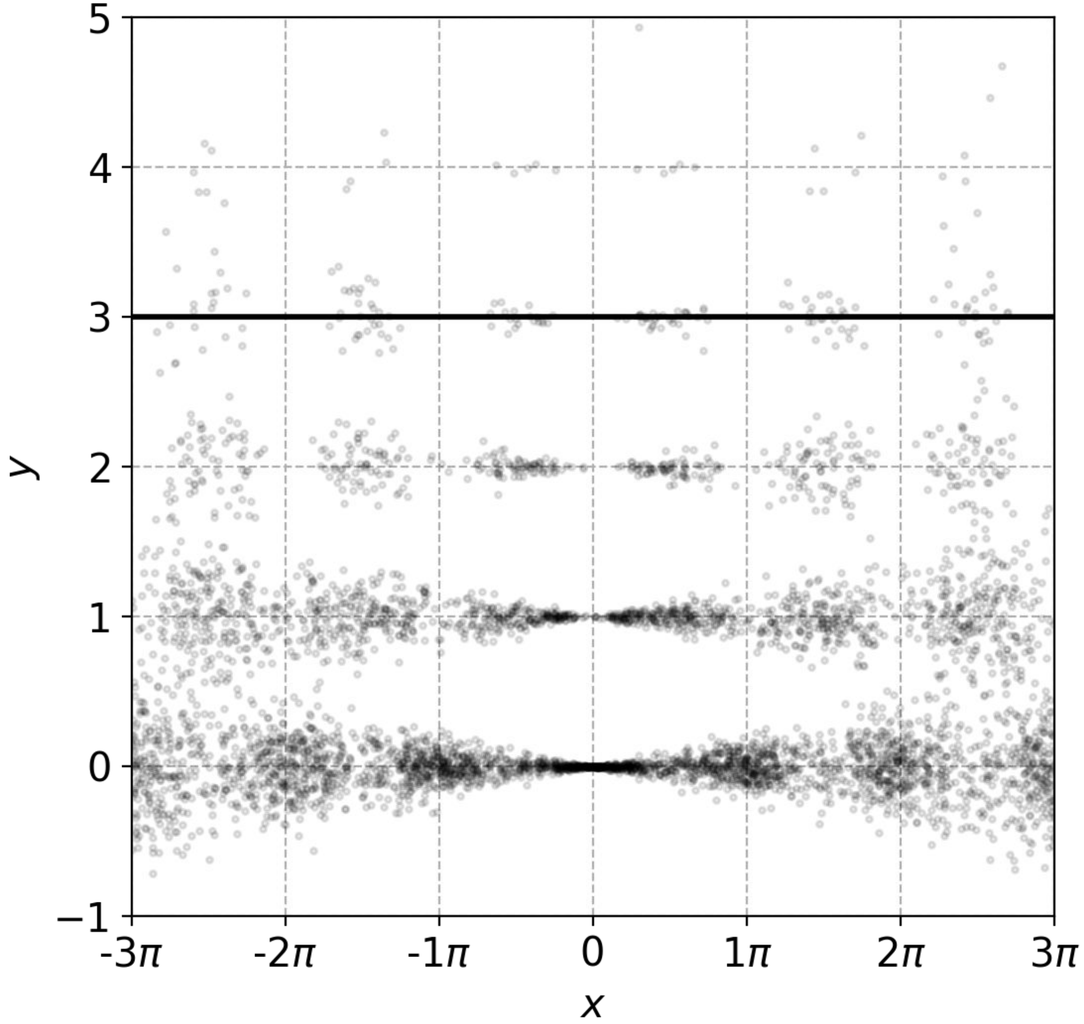
* Assuming objective is to maximize expected profit.



Imagine the following game:

- We are given a dataset D of (x, y) pairs from $p^*(x, y)$, as shown to the left
- A new pair is sampled from $p^*(x, y)$
- We observe x, but y is hidden
- We can pay 5¢ to reveal y
- If $y > 3$, we get \$1; otherwise, we get \$0
- When should we take the gamble*?
  - When $\Pr[Y > 3 \mid D, X = x] \geq 5\%$

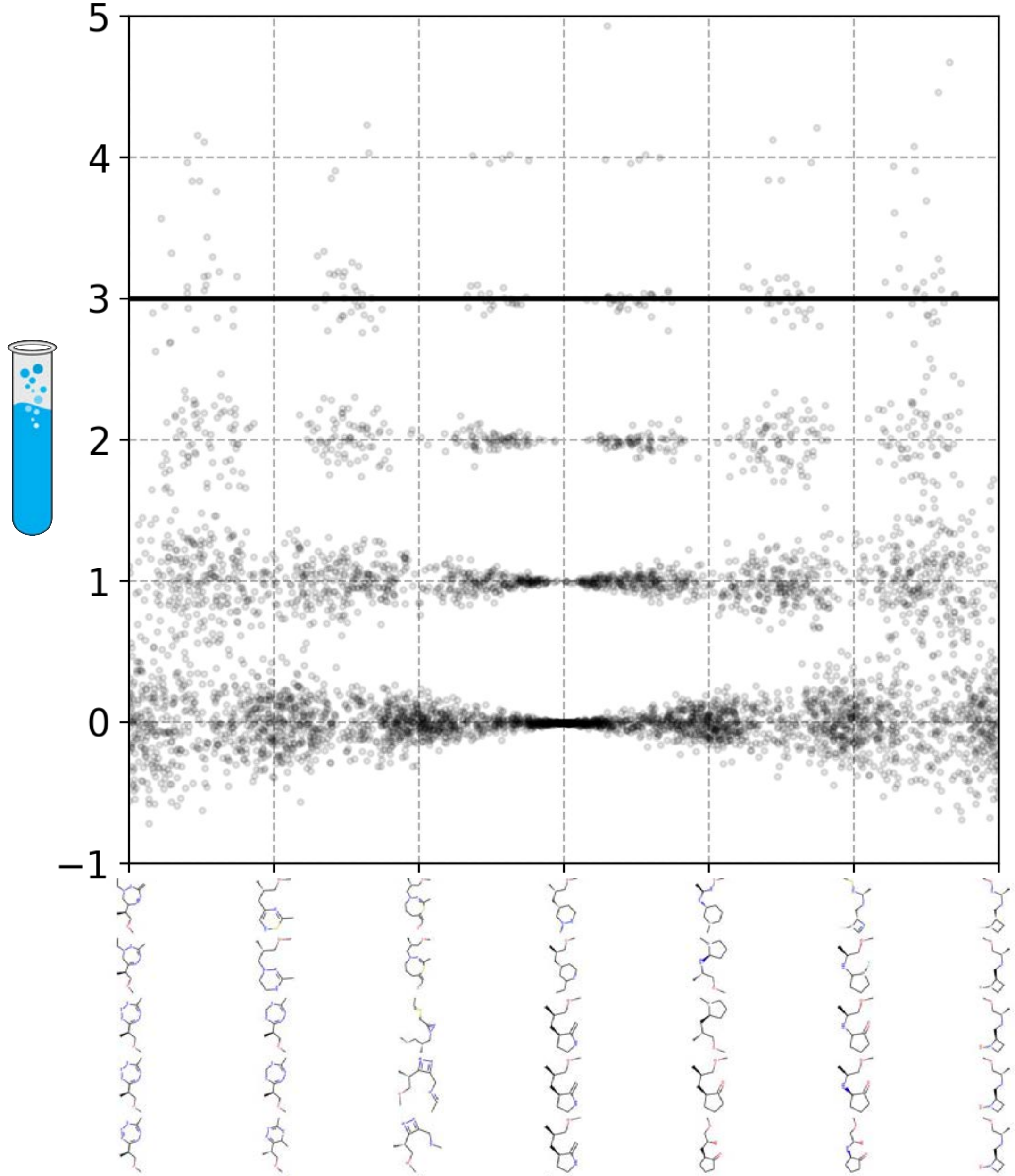
* Assuming objective is to maximize expected profit.



Overly simple stylized 1d example of problems in drug discovery

Given a set of molecule-endpoint pairs, we wish to identify regions of chemical space where the probability of finding a promising candidate is sufficiently high

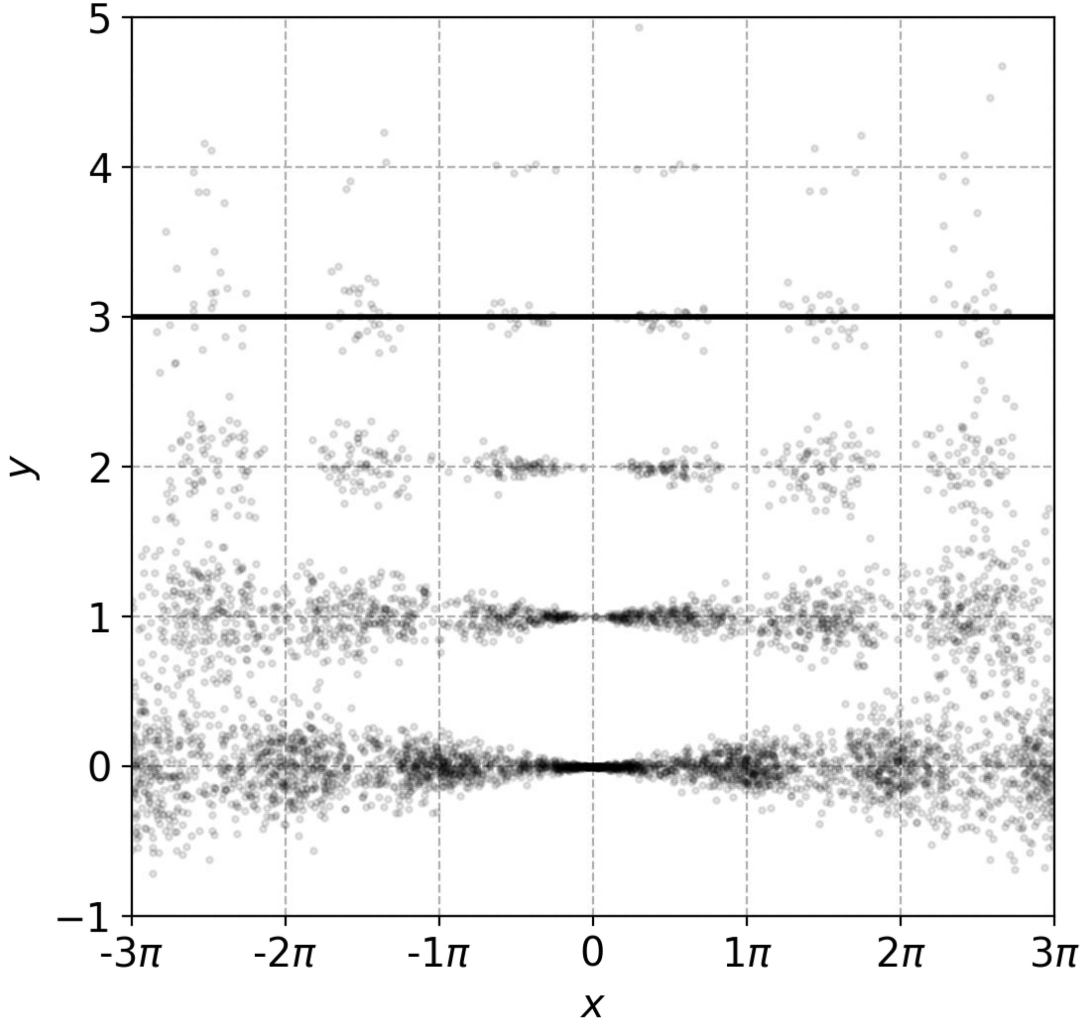
Formally, given a dataset $D_n \in (\mathbf{X}, \mathbf{Y})^n$, we wish to find regions $\mathbf{Z} \subset \mathbf{X}$ such that $\Pr[Y > t \mid D_n, X \in \mathbf{Z}] \geq 1 - \beta$



Overly simple stylized 1d example of problems in drug discovery

Given a set of molecule-endpoint pairs, we wish to identify regions of chemical space where the probability of finding a promising candidate is sufficiently high

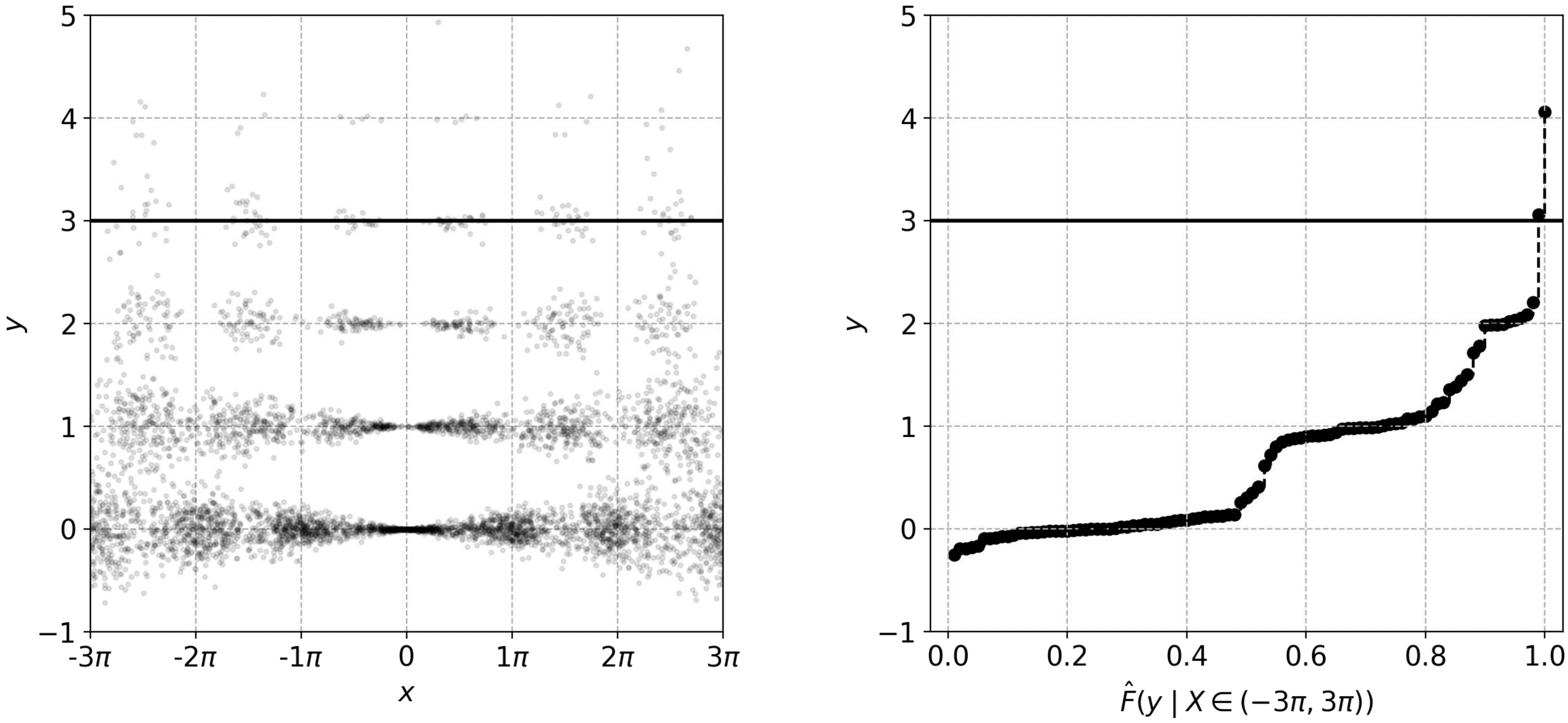
Formally, given a dataset $D_n \in (\mathbf{X}, \mathbf{Y})^n$, we wish to find regions $\mathbf{Z} \subset \mathbf{X}$ such that $\Pr[Y > t \mid D_n, X \in \mathbf{Z}] \geq 1 - \beta$



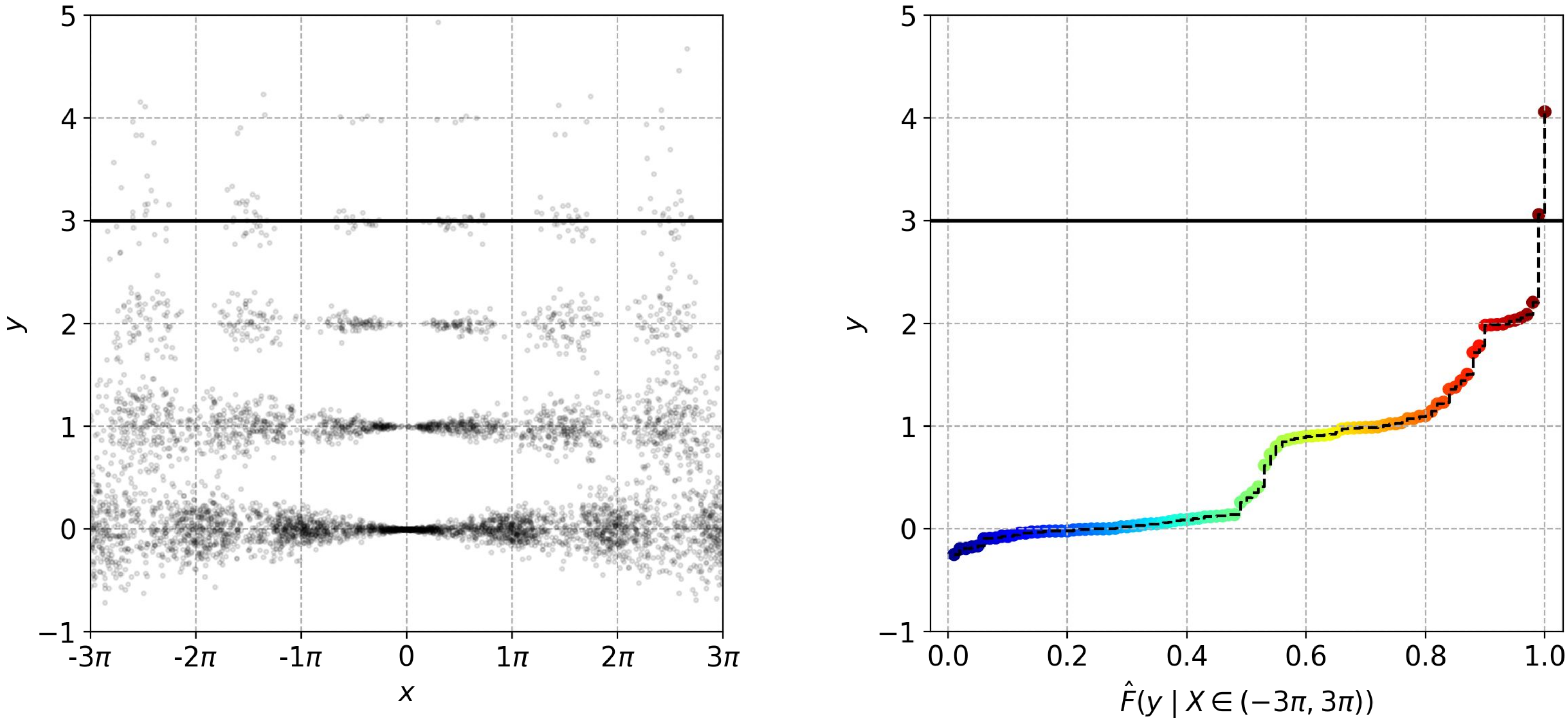
Let's make things simpler: suppose we don't know the value of the new *x*, other than it is in the interval (-3π, 3π)

Can use *D* to estimate the CDF of *y* non-parametrically, i.e., via a counting procedure:

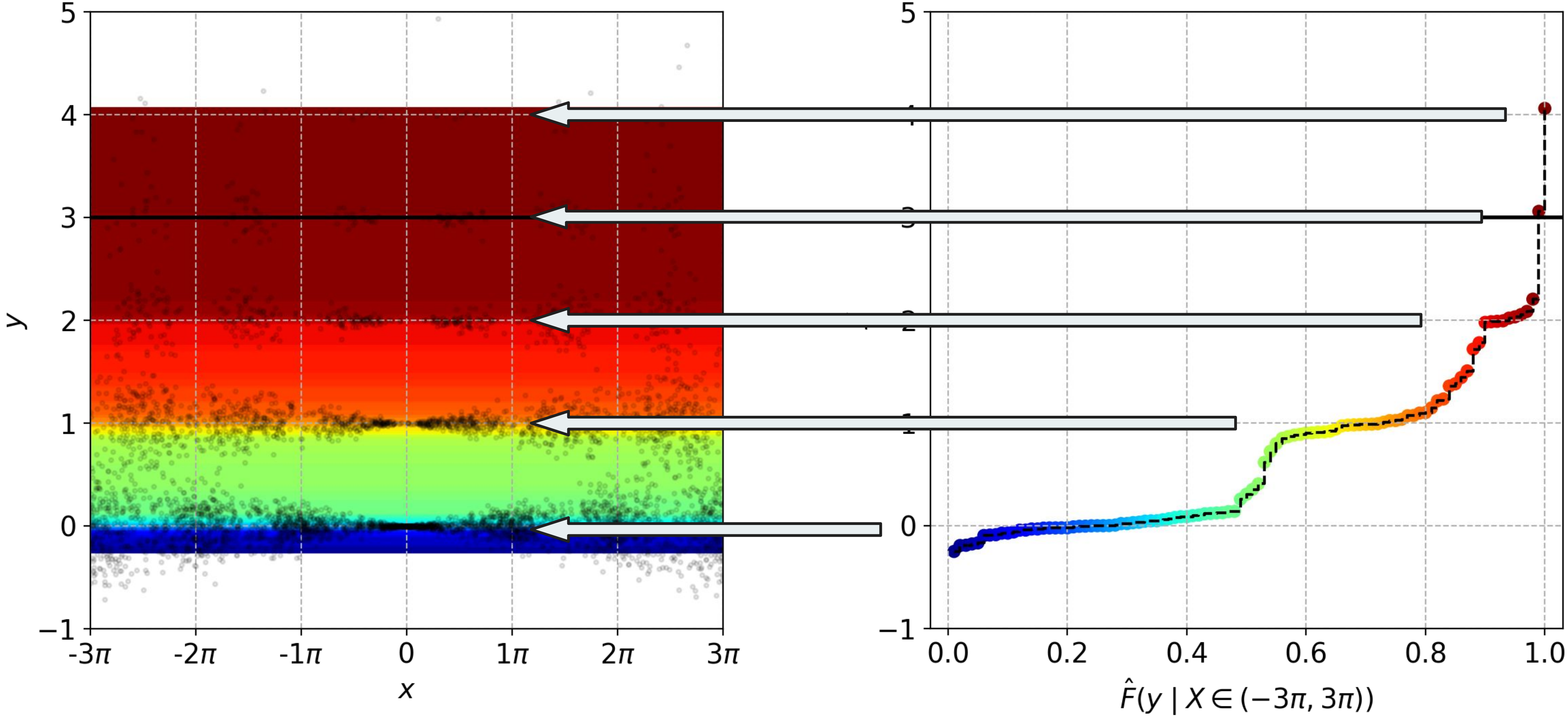
$$F(y \mid D) = \Pr[Y < y \mid D] = n^{-1}\sum_i [y_i < y]$$



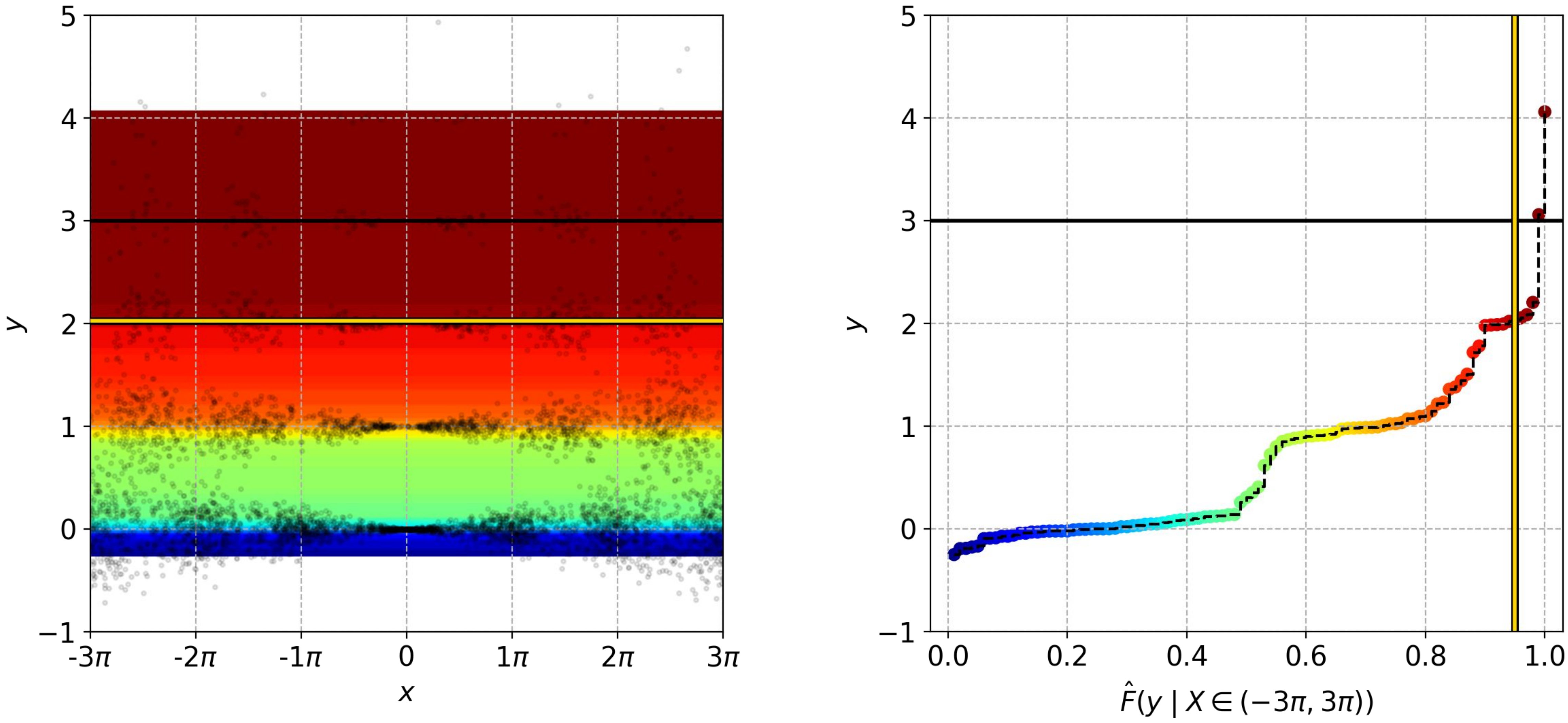
We fit an empirical CDF to 100 observations from *D*



Color indicates the corresponding quantile, with **blue for β = 0** for and **red for β = 1**

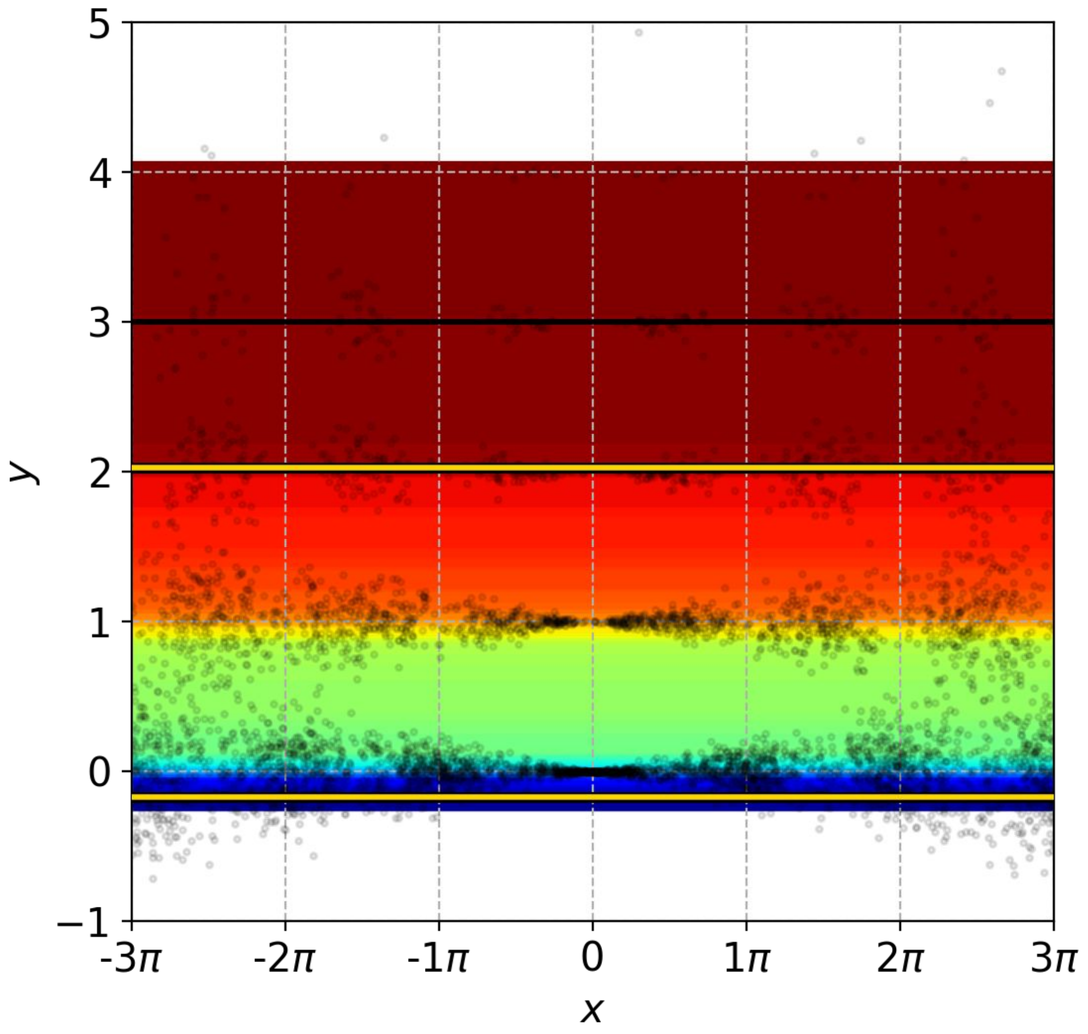


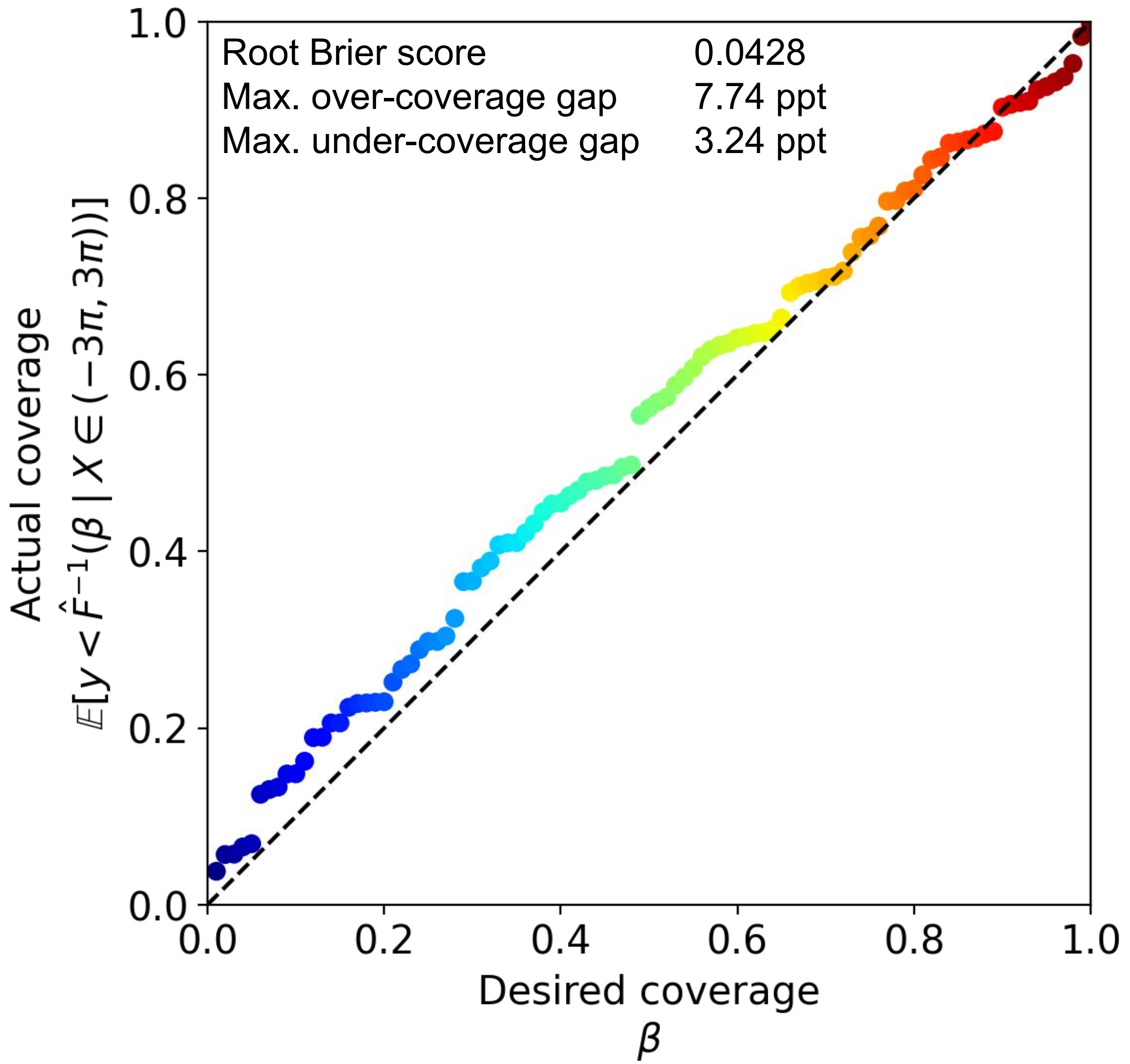
We can color values of *y* based on its associated quantile



The 95% quantile of *y* is ~2, which is less than 3

Hence, in the absence of further information, we reject the gamble





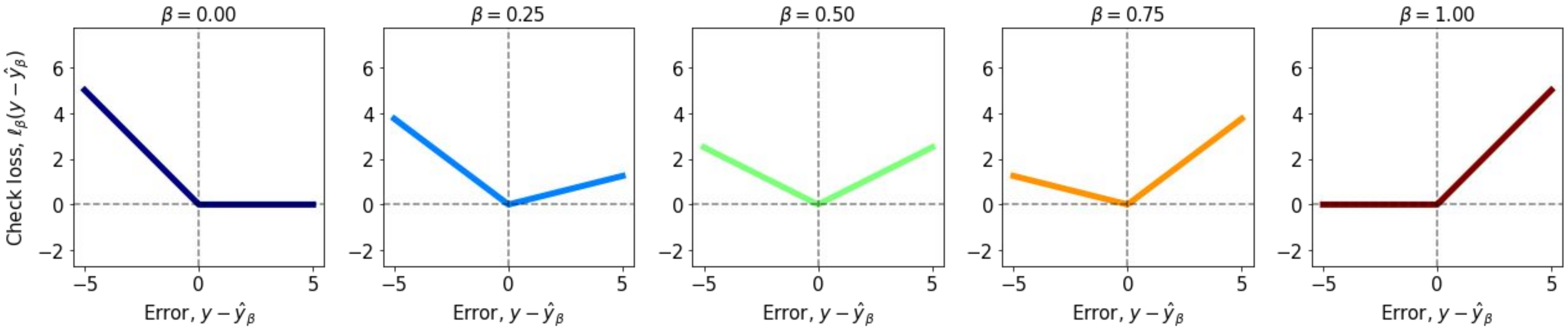
This predictor is reliable (in the marginal sense), but is it useful?

Can be do better by conditioning on *x*?

Quantile regression

Estimating the conditional quantile function

- Regression variant which strives for a consistent estimator of the **conditional quantile function**
  - As opposed to standard least squares regression, which strives for a consistent estimator of the conditional expectation function
- Quantiles are:
  - Robust
  - Fully descriptive of the conditional distribution
  - Equivariant to transformations that often plague likelihood-based inference
    - Scale/shift
    - Monotonic transformations (log, power-law, etc.)



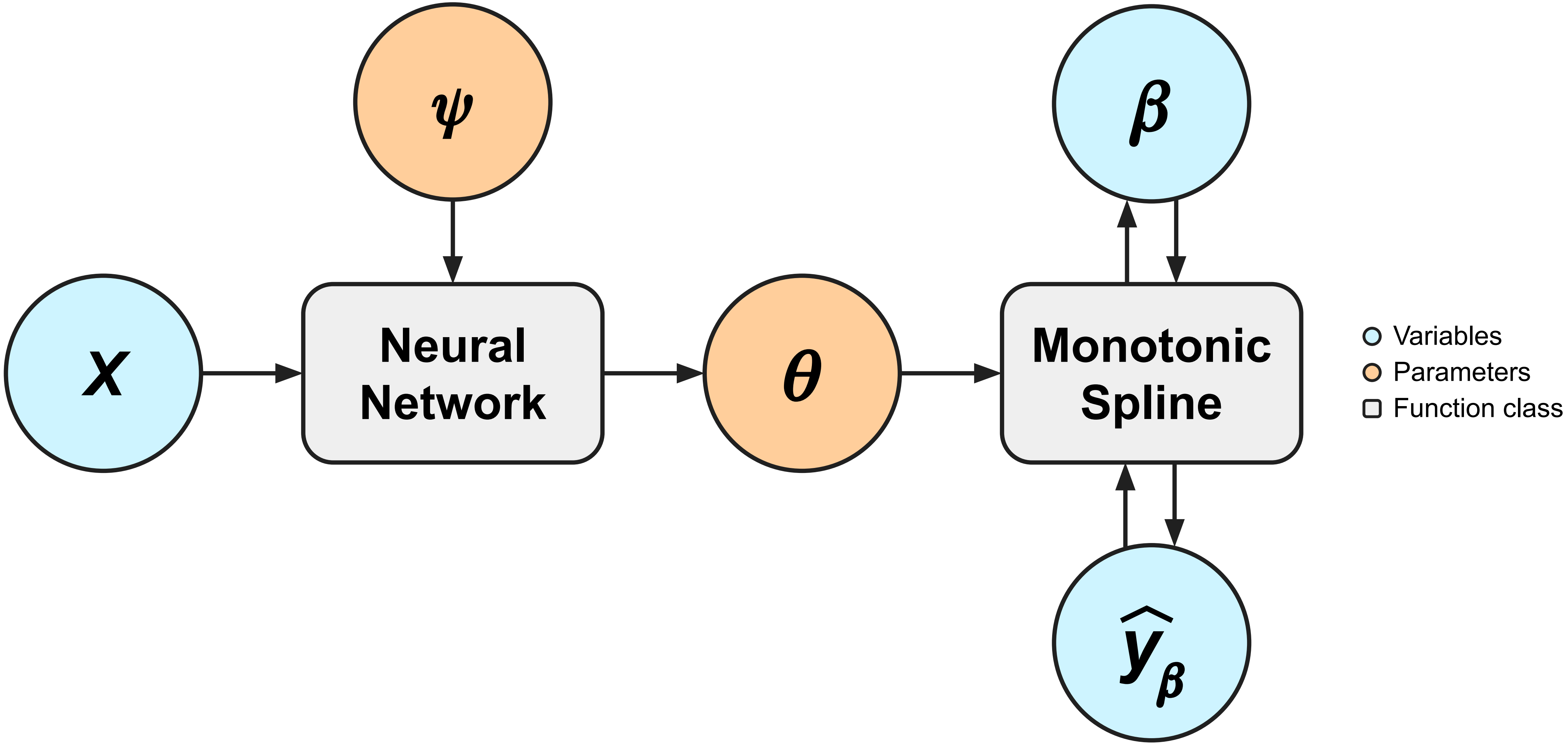
Parametric procedure for estimating the desired conditional quantile is to carry out minimization with the **check function**

Given a set of *n* observations, a consistent estimator for the β-quantile of $y \mid x$ is given by:

$$\widehat{\theta}_\beta = \text{argmin}_\theta \, n^{-1} \sum_i \ell_\beta(y_i - q_\beta(x_i; \theta)),$$

where $\ell_\beta(\varepsilon) = \beta\,|\varepsilon|\,[\varepsilon > 0] + (1 - \beta)\,|\varepsilon|\,[\varepsilon \leq 0]$ is the **check function**

Quantile regression spline neural network



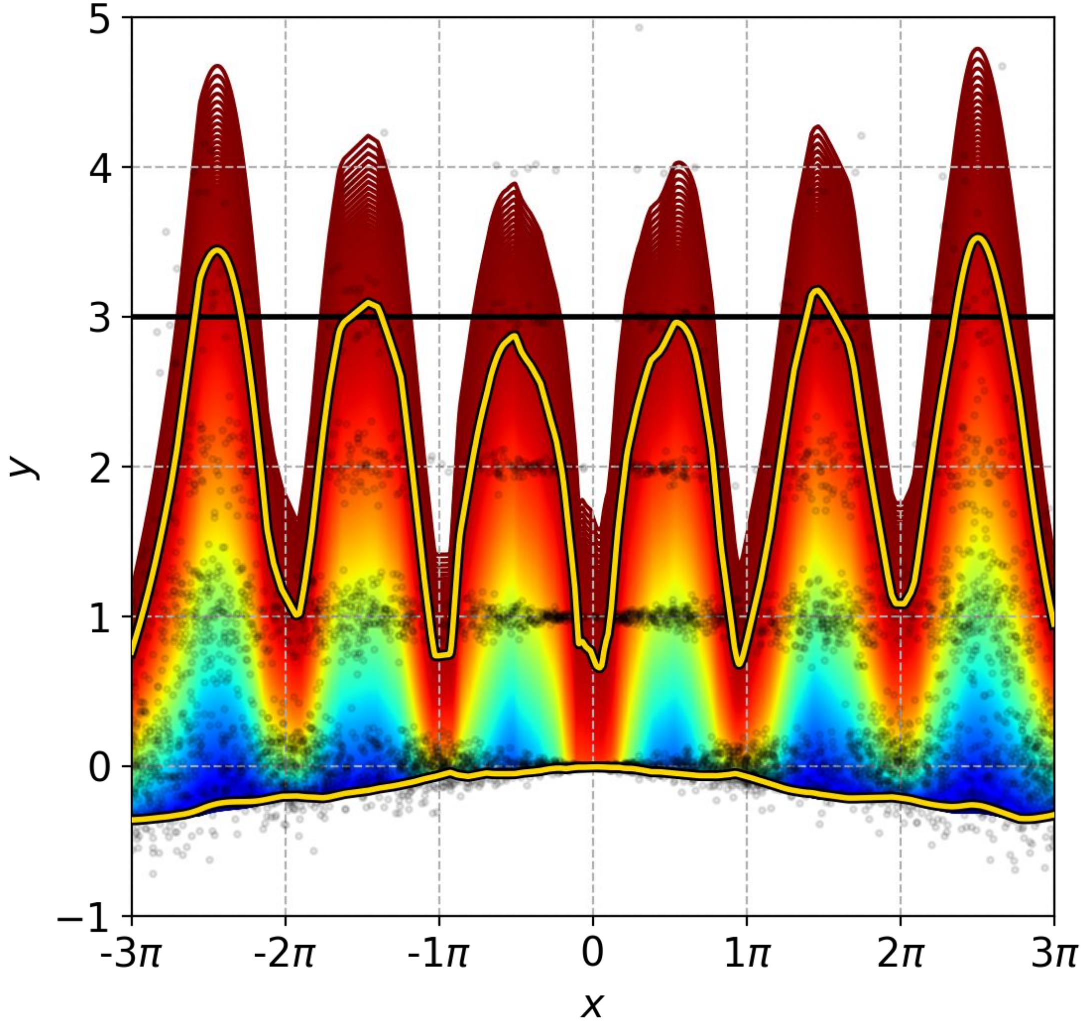
Training

Estimating the full predictive distribution

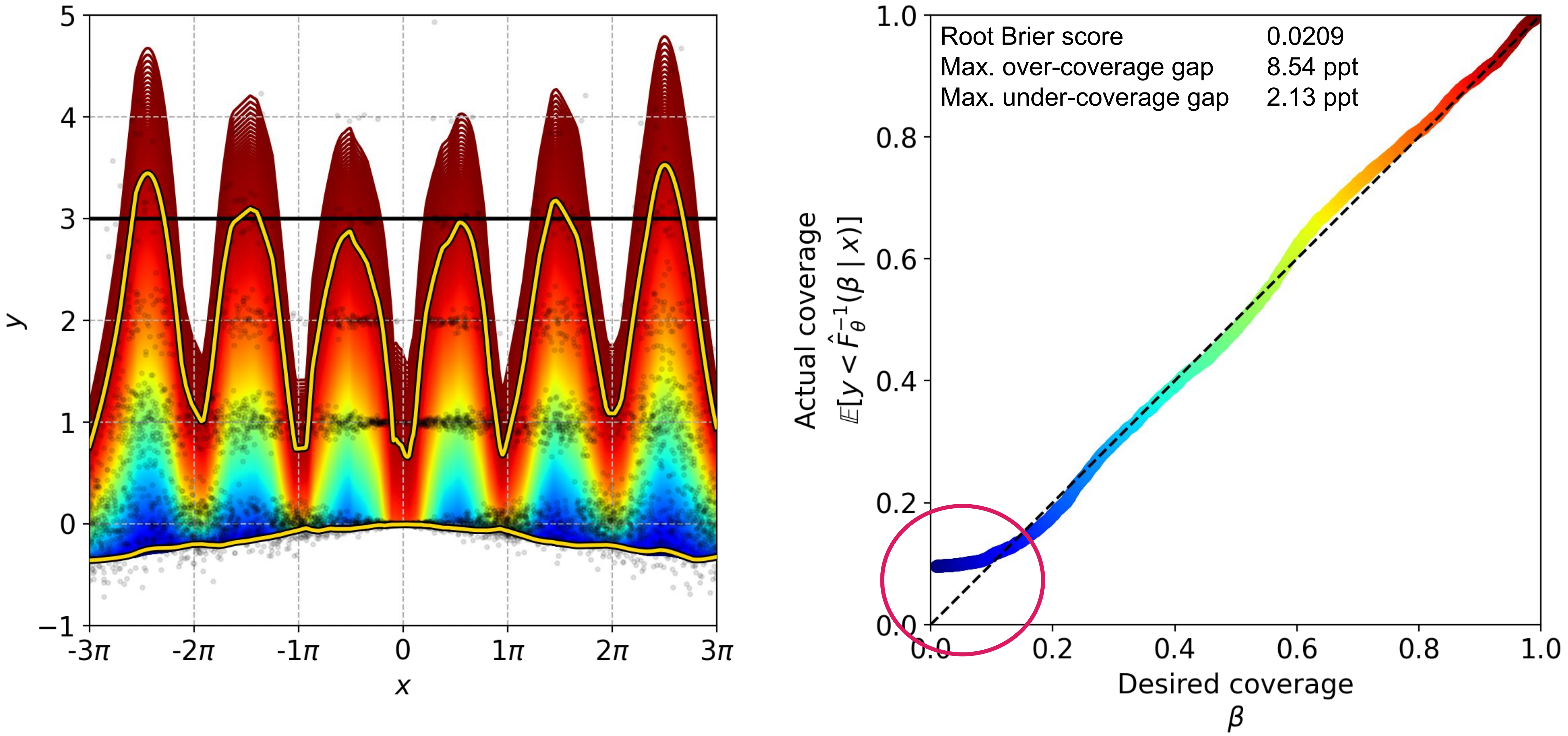
Train end-to-end using the following procedure[2]:

1. Sample a minibatch of (x, y) pairs
2. Sample a minibatch of quantiles $\beta \sim \text{Uniform}(0, 1)$
3. Pass the x's through the neural network to get θ's
4. Pass the β's through the monotonic spline to get $\widehat{y}_{\beta}$'s
5. Compute the check function losses and average
6. Backprop

[2] Tagasovska, N., & Lopez-Paz, D. (2019). Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, *32*.



- Quantile regression neural networks can provide estimated prediction intervals for any *x*
- On this toy problem, the network correctly identifies regions where Pr[*Y* > 3 | *D*, *X* = *x*] ≥ 5%
- Prediction interval length is highly adaptive with respect to *x*
- Training makes no assumptions about the likelihood of *y* given *x*
- Minimization of the check function loss corresponds to a type of *M*-estimator and comes with associated robustness



Root Brier score	0.0209
Max. over-coverage gap	8.54 ppt
Max. under-coverage gap	2.13 ppt

Coverage looks pretty good!

But in general, we don’t have rigorous (finite sample) statistical guarantees

Any function can be made marginally reliable

Using conformalization to confer marginal coverage guarantees

- Suppose that, in addition to a fitted quantile function $\hat{q}_\beta(x)$, we have a held-out *calibration set* $D_{\text{cal}} = \{(x_i, y_i) : i = 1, \ldots, n_{\text{cal}}\}$
- Let $E^\beta_{\text{cal}} = \{y_i - \hat{q}_\beta(x_i) : i = 1, \ldots, n_{\text{cal}}\}$ denote the residuals associated with quantile β on the calibration data
- Consider the adjusted predictor,

$$\check{q}_\beta(x) = \hat{q}_\beta(x) + Q_\beta(E^\beta_{\text{cal}}),$$

where Q_β computes the β-quantile of ECDF associated with E_{cal}

- For (x, y) exchangeable with D_{cal}, the predictor $\check{q}_\beta(x)$ marginally covers

Any function can be made marginally reliable

Using conformalization to confer marginal coverage guarantees

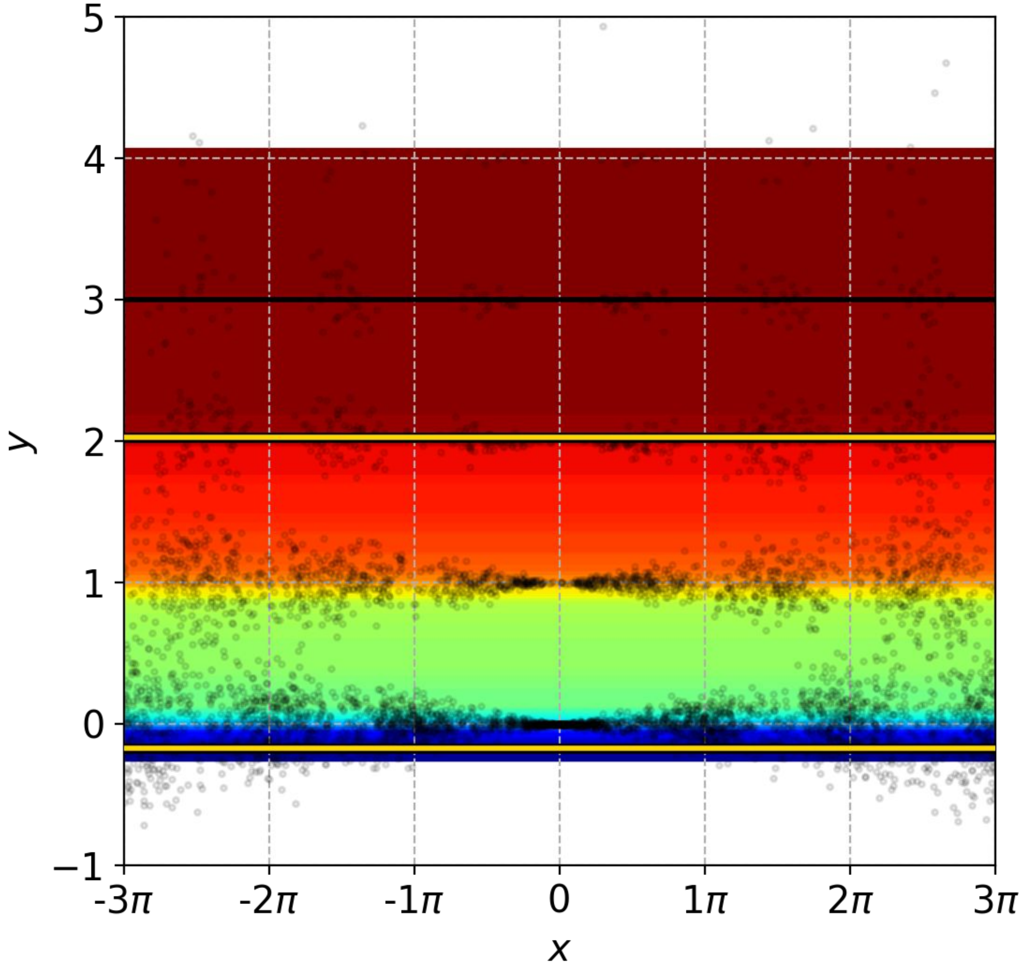
- If $\widehat{q}_\beta(x) = 0$ everywhere…
  - $E^{\beta}_{\text{cal}} = \{y_i - \cancel{\widehat{q}_\beta}(x_i) : i = 1, \ldots, n_{\text{cal}}\}$
  - $\check{q}_\beta(x) = \cancel{\widehat{q}_\beta}(x) + Q_\beta(E^{\beta}_{\text{cal}})$

Any function can be made marginally reliable

Using conformalization to confer marginal coverage guarantees

- If $\widehat{q}_\beta(x) = 0$ everywhere…
  - $E^{\beta}_{\text{cal}} = \{y_i - \xcancel{\widehat{q}_\beta}(x_i) : i = 1, \ldots, n_{\text{cal}}\}$
  - $\check{q}_\beta(x) = \xcancel{\widehat{q}_\beta}(x) + Q_\beta(E^{\beta}_{\text{cal}})$

- **This is exactly the first predictor we looked at!**



Any function can be made marginally reliable

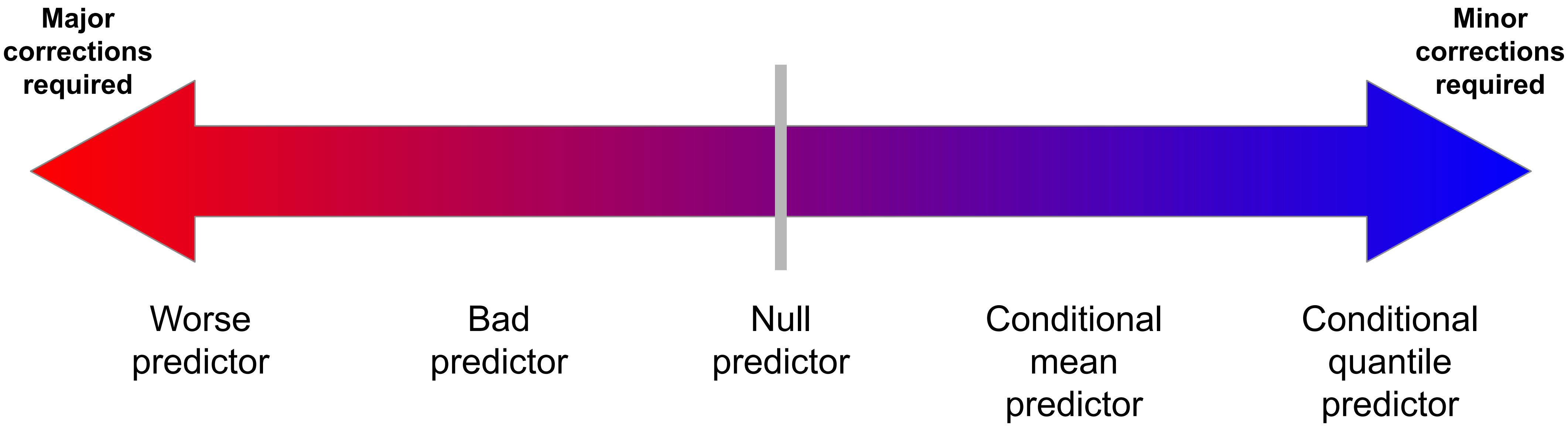
Using conformalization to confer marginal coverage guarantees

- This procedure is an instance of ***conformalization***[3][4], and its associated **marginal coverage guarantees do not require any assumptions about the initial predictor** $\widehat{q}_{\beta}(x)$
    - All that is required is exchangeability of test instances with the calibration set
- The better $\widehat{q}_{\beta}(x)$ approximates the true conditional quantile function, the less correction is required as part of the conformalization step

[3] Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
[4] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Spectrum of heuristic predictors

“All models are wrong, but some are useful”



Major corrections required

Minor corrections required

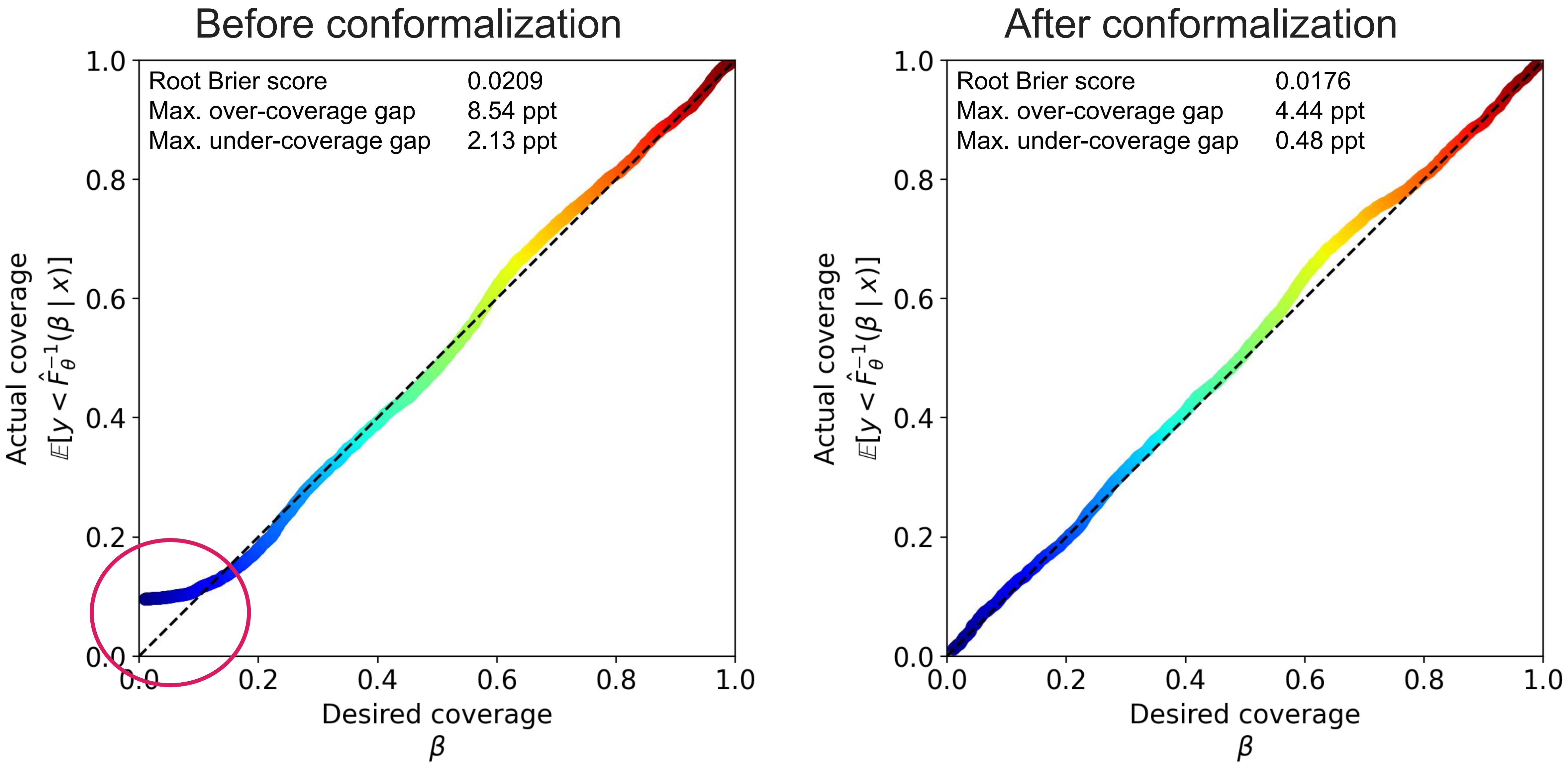
Worse predictor

Bad predictor

Null predictor

Conditional mean predictor

Conditional quantile predictor

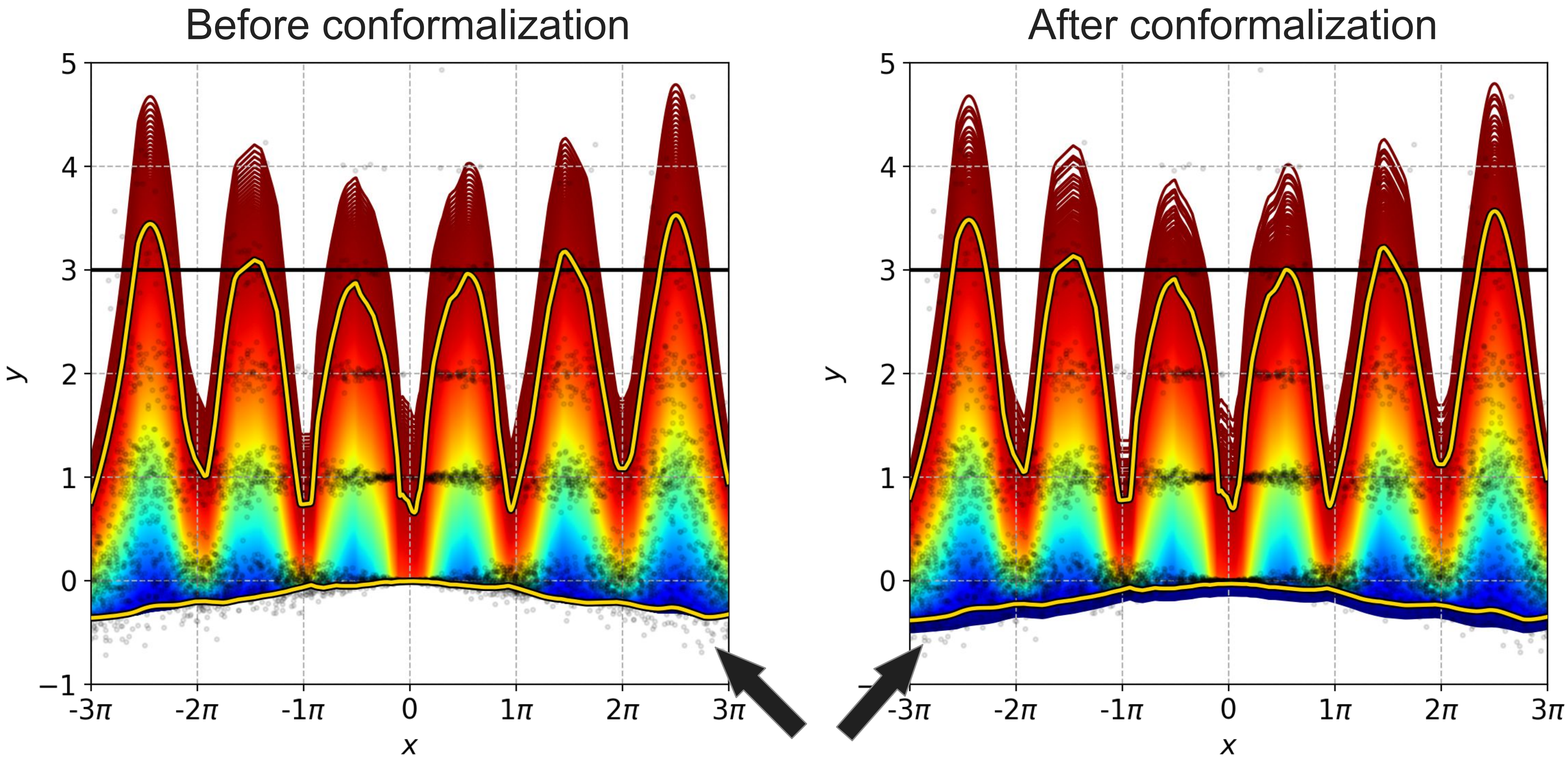


Before conformalization

Root Brier score	0.0209
Max. over-coverage gap	8.54 ppt
Max. under-coverage gap	2.13 ppt

After conformalization

Root Brier score	0.0176
Max. over-coverage gap	4.44 ppt
Max. under-coverage gap	0.48 ppt



Before conformalization

After conformalization

Uncertainty quantification in higher dimensions

Challenges when moving beyond toy problems

- The curse of dimensionality poses challenges for reliable uncertainty estimation
  - With increasing dimensionality, data become more spread out
  - Highly flexible models like NNs can memorize examples that are easily separable (overfitting)
  - As such, quantile regression networks can become very concentrated in their predictions on the training set (tight prediction intervals)
  - As a consequence, such models may fail to leverage their uncertainty capabilities and collapse to point predictors

Uncertainty quantification in higher dimensions

Challenges when moving beyond toy problems

- Motivates greater need for regularization in such regimes
- In addition to usual regularization strategies for NNs, we investigate a number of **consistency-style regularizers** for quantile regression
  - Calibration-based regularizers (own work)
  - Independence of interval length and miscoverage events[5]

[5] Feldman, S., Bates, S., & Romano, Y. (2021). Improving conditional coverage via orthogonal quantile regression. *Advances in Neural Information Processing Systems*, *34*.

Lipophilicity benchmark

Experimental setup

- We apply these ideas to a dataset[6] of 4200 compounds curated from ChEMBL with experimental results of octanol/water distribution coefficient (logD at pH 7.4)
- 80-10-10 split of the compounds into training, validation, and testing sets
- The training set and validation/testing sets are split by scaffold
- The validation and testing sets are split at random
  - We will use the validation set for conformalization
  - Random splitting in this way is sufficient for exchangeability, which guarantees marginal coverage on the test set

[6] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, *9*(2), 513-530.

Lipophilicity benchmark

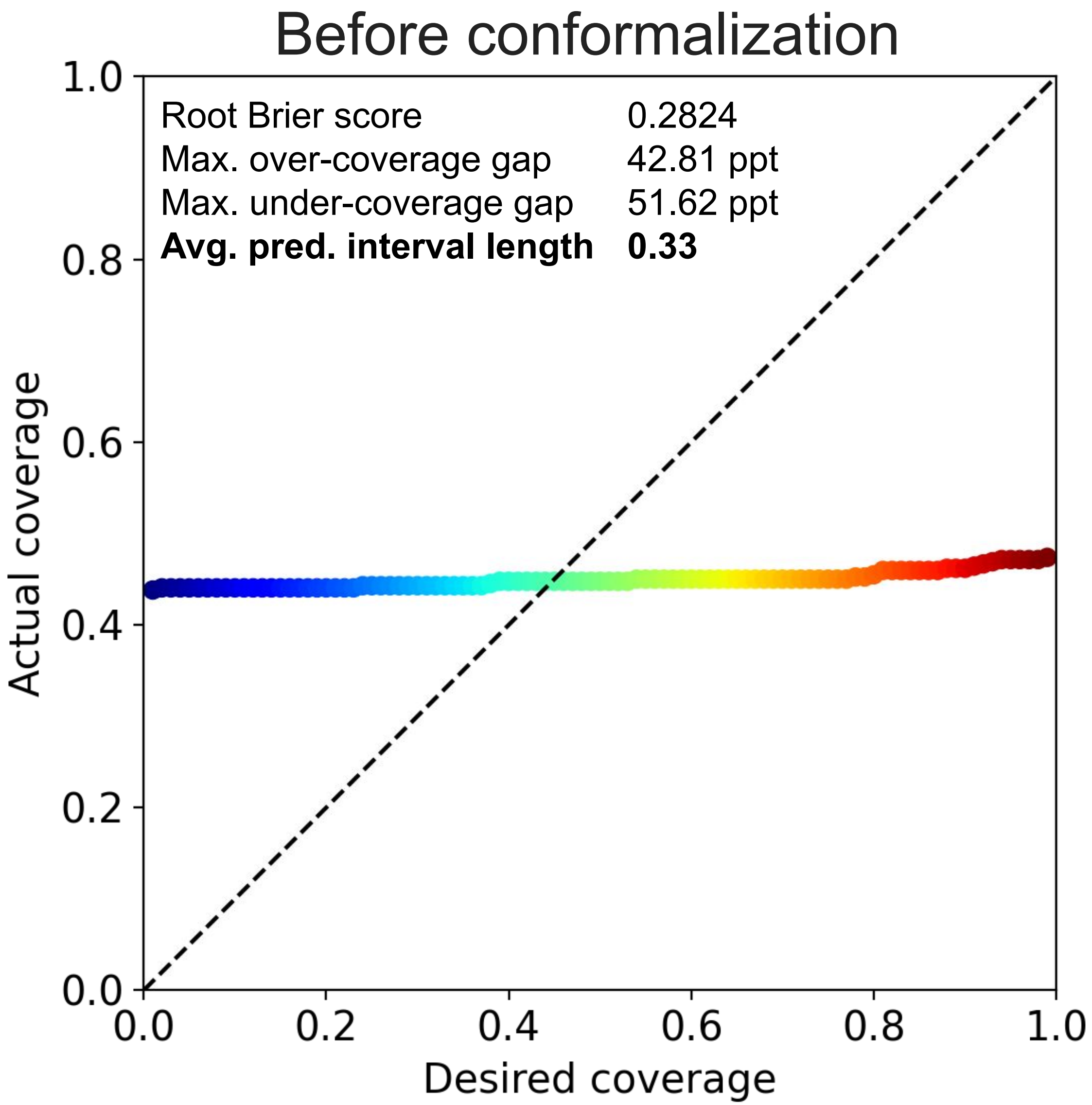
Experimental setup

- Across all experiments, we use the same:
  - Architecture (message passing graph neural network)
  - Optimizer (Adam, lr=1e-4)
  - Checkpoint selection criteria (best validation RMSE)
  - Model regularization (layer normalization, dropout, weight decay)
  - Batch size, max number of iterations, etc.

- After training, we use the validation set to conformalize 90% prediction intervals for each model

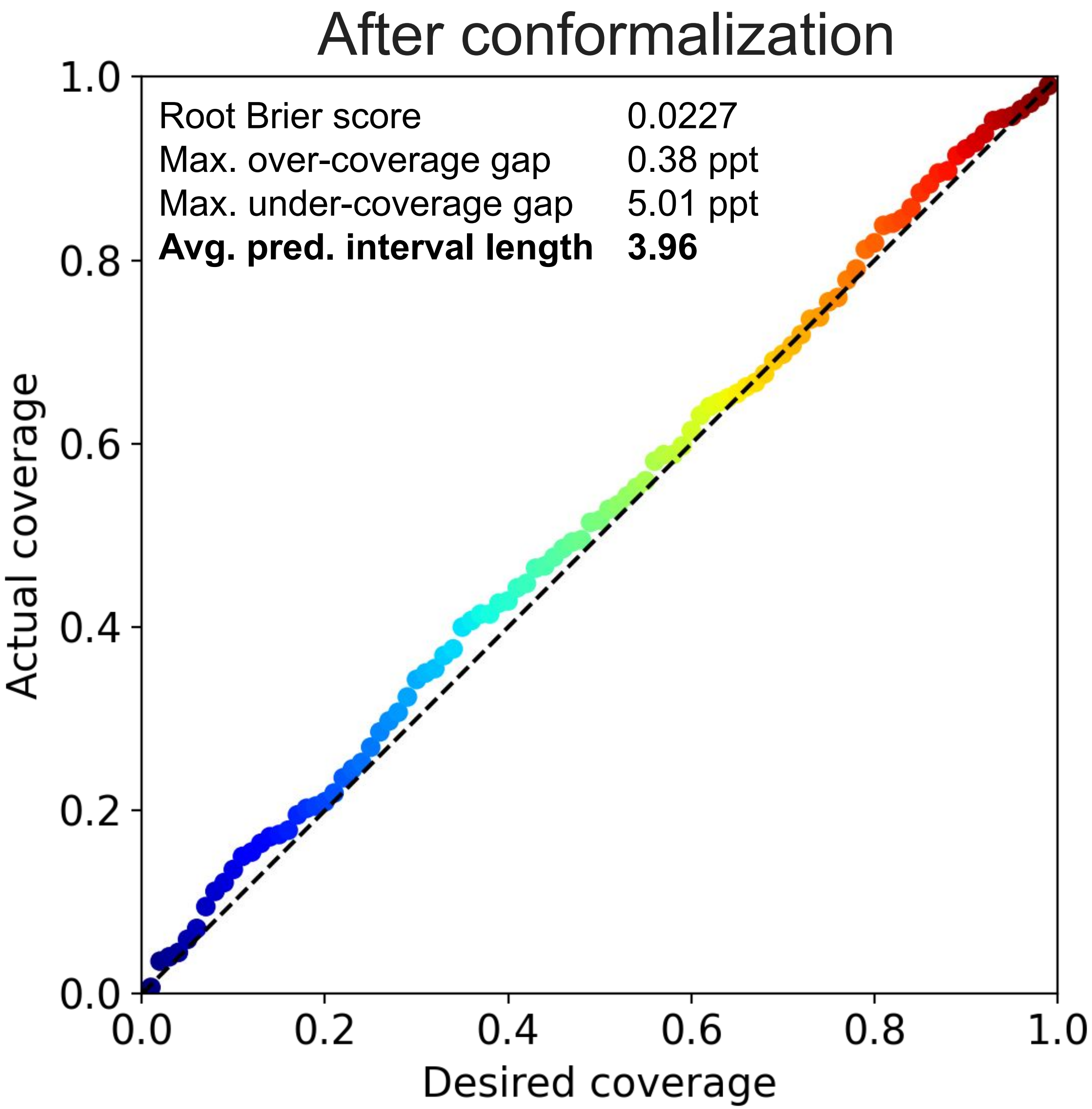
- Marginal coverage on test set after conformalization ranged from 88.7% - 91.6%

Likelihood-based	squared error loss	squared error loss + heteroskedasticity		
Likelihood-free	check function loss	check function loss + calibration reg.	check function loss + orthogonality reg.	check function loss + calibration reg. + orthogonality reg.
	interval score loss	interval score loss + calibration reg.	interval score loss + orthogonality reg.	interval score loss + calibration reg. + orthogonality reg.



Before conformalization

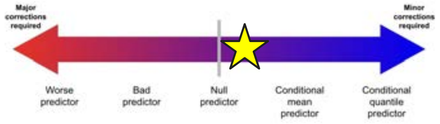
Root Brier score	0.2824
Max. over-coverage gap	42.81 ppt
Max. under-coverage gap	51.62 ppt
Avg. pred. interval length	**0.33**



After conformalization

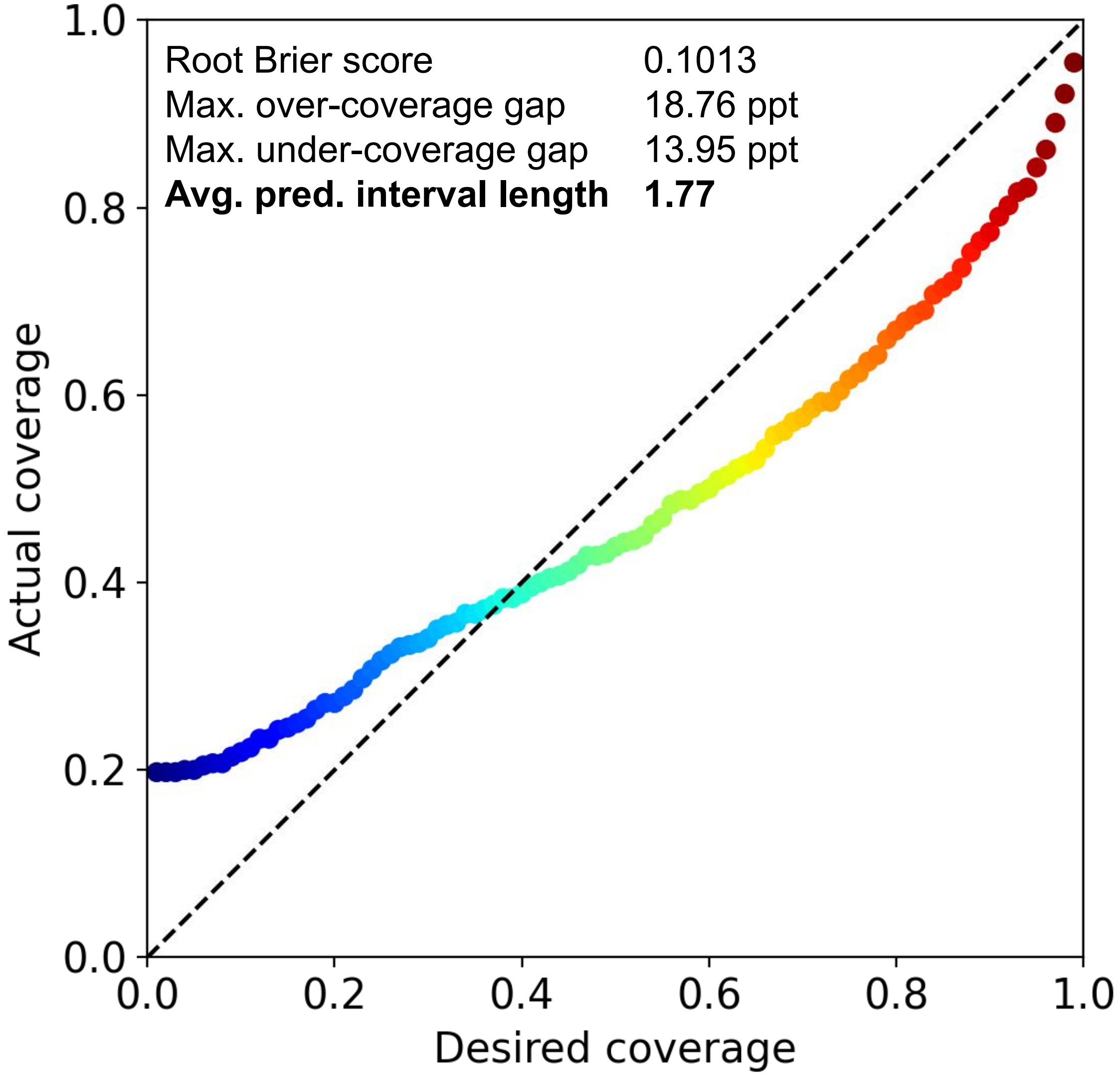
Root Brier score	0.0227
Max. over-coverage gap	0.38 ppt
Max. under-coverage gap	5.01 ppt
Avg. pred. interval length	**3.96**

Inadequate regularization → memorization → unrealistically tight prediction intervals → strong correction required → unusably wide (post-correction) prediction intervals



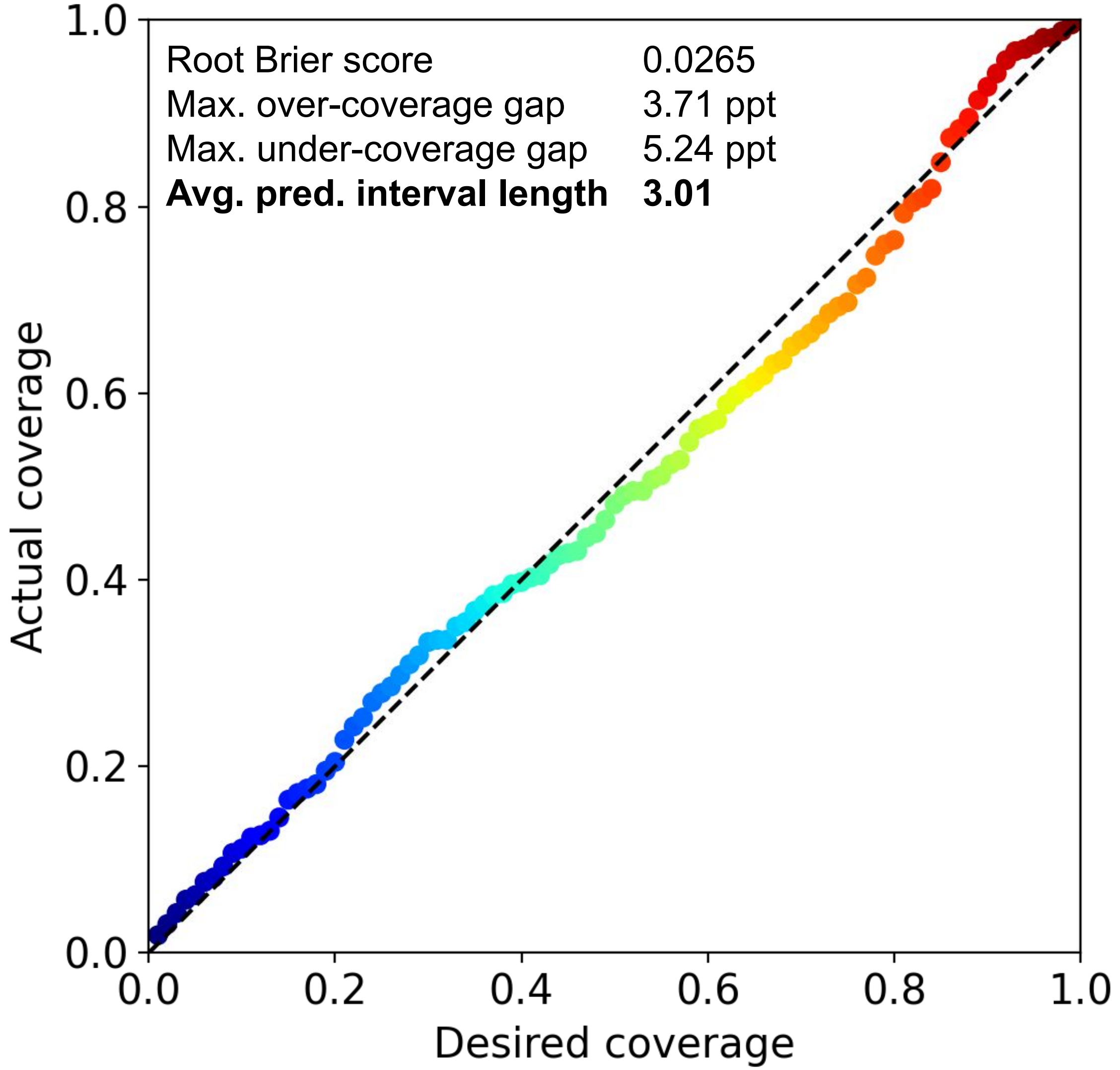
Major corrections required
Minor corrections required
Worse predictor
Bad predictor
Null predictor
Conditional mean predictor
Conditional quantile predictor

Before conformalization



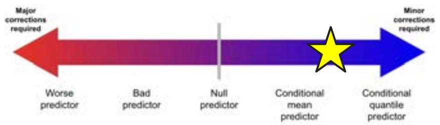
Root Brier score	0.1013
Max. over-coverage gap	18.76 ppt
Max. under-coverage gap	13.95 ppt
Avg. pred. interval length	**1.77**

After conformalization

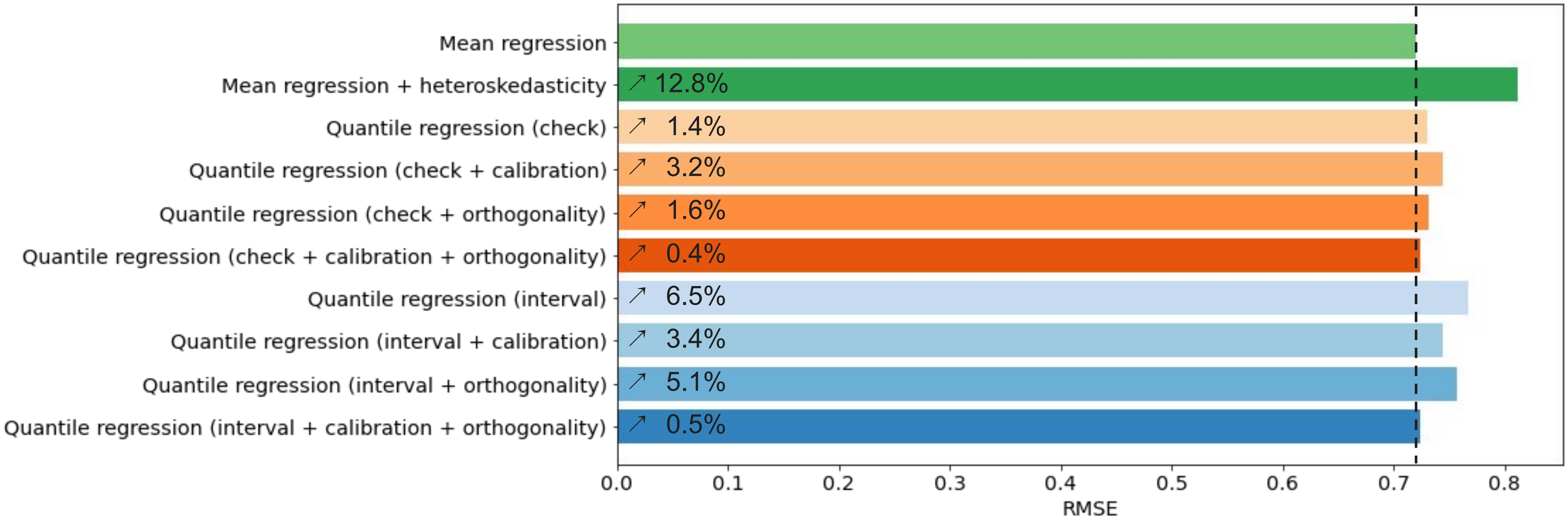


Root Brier score	0.0265
Max. over-coverage gap	3.71 ppt
Max. under-coverage gap	5.24 ppt
Avg. pred. interval length	**3.01**

Appropriate regularization drastically improves generalization of interval predictors, permitting tighter post-correction prediction intervals

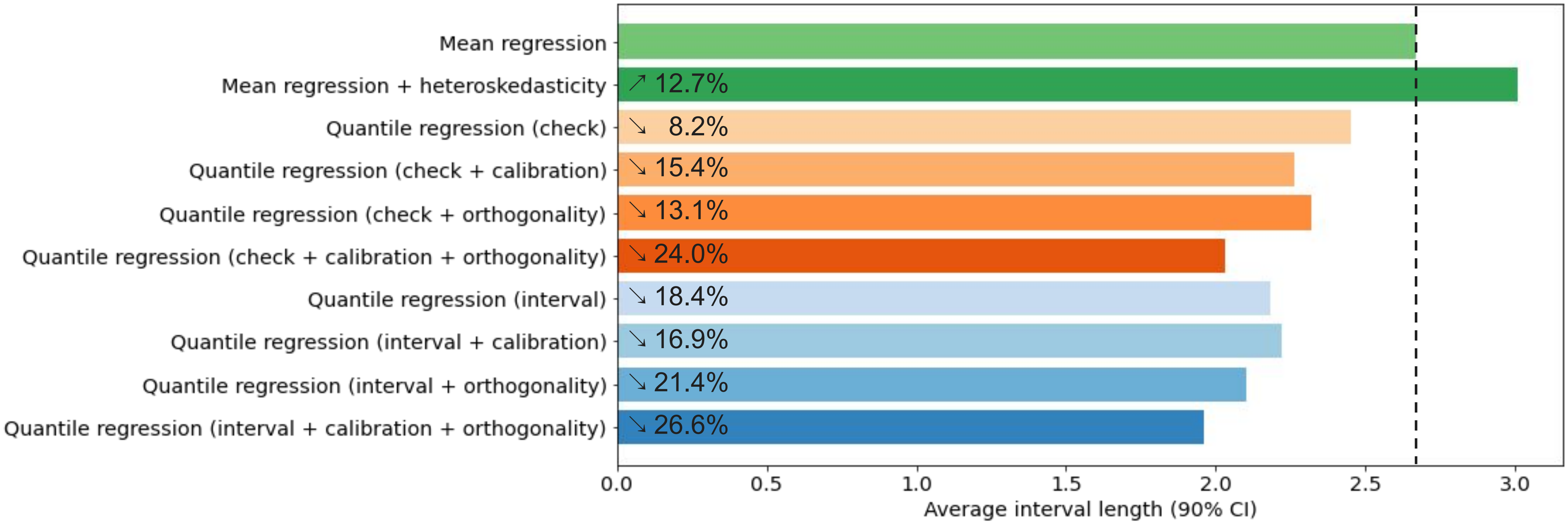

Major corrections required
Minor corrections required
Worse predictor
Bad predictor
Null predictor
Conditional mean predictor
Conditional quantile predictor

Regression variants by RMSE



Regression variants did not differ significantly in test RMSE on lipophilicity benchmark

Regression variants by avg. interval length



Quantile regression yields tighter PIs on lipophilicity benchmark for the same marginal coverage

Consistency-regularized variants offer additional improvements

Explaining RMSE differences

Quantile regression did not improve test RMSE on lipophilicity benchmark

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log(rmse)   R-squared:                       0.310
Model:                            OLS   Adj. R-squared:                 -0.242
Method:                 Least Squares   F-statistic:                    0.5621
No. Observations:                  10   Prob (F-statistic):              0.702
Df Residuals:                       5   Log-Likelihood:                 21.239
Df Model:                           4   AIC:                            -32.48
Covariance Type:            nonrobust   BIC:                            -30.97
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
 Intercept         -0.2682 ***  0.029     -9.272      0.000      -0.343      -0.194
 check             -0.0269      0.041     -0.656      0.541      -0.132       0.078
 interval          -0.0052      0.041     -0.127      0.904      -0.110       0.100
 calibration       -0.0168      0.029     -0.580      0.587      -0.091       0.058
 orthogonality     -0.0166      0.029     -0.573      0.591      -0.091       0.058
==============================================================================
 Omnibus:                        2.486   Durbin-Watson:                   2.819
 Prob(Omnibus):                  0.289   Jarque-Bera (JB):                0.249
 Skew:                          -0.022   Prob(JB):                        0.883
 Kurtosis:                       3.771   Cond. No.                         6.20
==============================================================================
```

No significant differences in RMSE across the regression variants considered

```
***   significance @  1%
**    significance @  5%
*     significance @ 10%
```

Explaining avg. PI length differences

Quantile regression induces statistically significantly tighter PIs on lipophilicity benchmark

```
                            OLS Regression Results
==============================================================================
Dep. Variable:   log(avg_interval_length_90)   R-squared:                0.914
Model:                                   OLS   Adj. R-squared:           0.845
Method:                        Least Squares   F-statistic:              13.29
No. Observations:                         10   Prob (F-statistic):     0.00712
Df Residuals:                              5   Log-Likelihood:          18.992
Df Model:                                  4   AIC:                     -27.98
Covariance Type:                   nonrobust   BIC:                     -26.47
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
 Intercept          1.0420 ***  0.036     28.768      0.000       0.949       1.135
 check             -0.1531 **   0.051     -2.990      0.030      -0.285      -0.021
 interval          -0.2205 ***  0.051     -4.304      0.008      -0.352      -0.089
 calibration       -0.0663      0.036     -1.829      0.127      -0.159       0.027
 orthogonality     -0.0810 *    0.036     -2.235      0.076      -0.174       0.012
==============================================================================
 Omnibus:                      0.254    Durbin-Watson:                   2.472
 Prob(Omnibus):                0.881    Jarque-Bera (JB):                0.402
 Skew:                         0.000    Prob(JB):                        0.818
 Kurtosis:                     2.017    Cond. No.                         6.20
==============================================================================
```

Conformalized quantile regression induces a **~15-22% reduction in average 90% PI length** compared to conformalized likelihood-based regression

Orthogonality regularization induces an additional ~**8% reduction in average 90% PI length**

```
***   significance @  1%
**    significance @  5%
*     significance @ 10%
```

Conclusion

Summary of presentation

- Quantile regression combined with ideas from conformal inference make for reliable and adaptive uncertainty quantification
- In high dimensional settings (e.g., working with molecular graphs or large molecular descriptors), need to think carefully about regularization
- By predicting conditional quantiles directly, we can form adaptive prediction intervals which are tighter on average
- We designed a quantile regression spline neural network which can fully characterize the predictive distribution for a given input
- Applied these ideas to lipophilicity benchmark and observed a 15-22% reduction in average 90% prediction interval length against baselines with matching marginal coverage

Atomwise

Aryan Pedawi
Atomwise, Inc.

aryan@atomwise.com

Thank you!

Team

Hossam Ashtawy
Brandon M. Anderson

Support

Colleagues at Atomwise,
all of you for listening!