

Distribution-free uncertainty quantification for ML in HEP

For a comprehensive intro, see Aishik's talk

Jack Y. Araz

UNIVERSITY COLLEGE LONDON
CITY ST GEORGE'S, UNIVERSITY OF LONDON

Based on arXiv: 2512.17048 & 2606.13780

PARTIQAL 2026

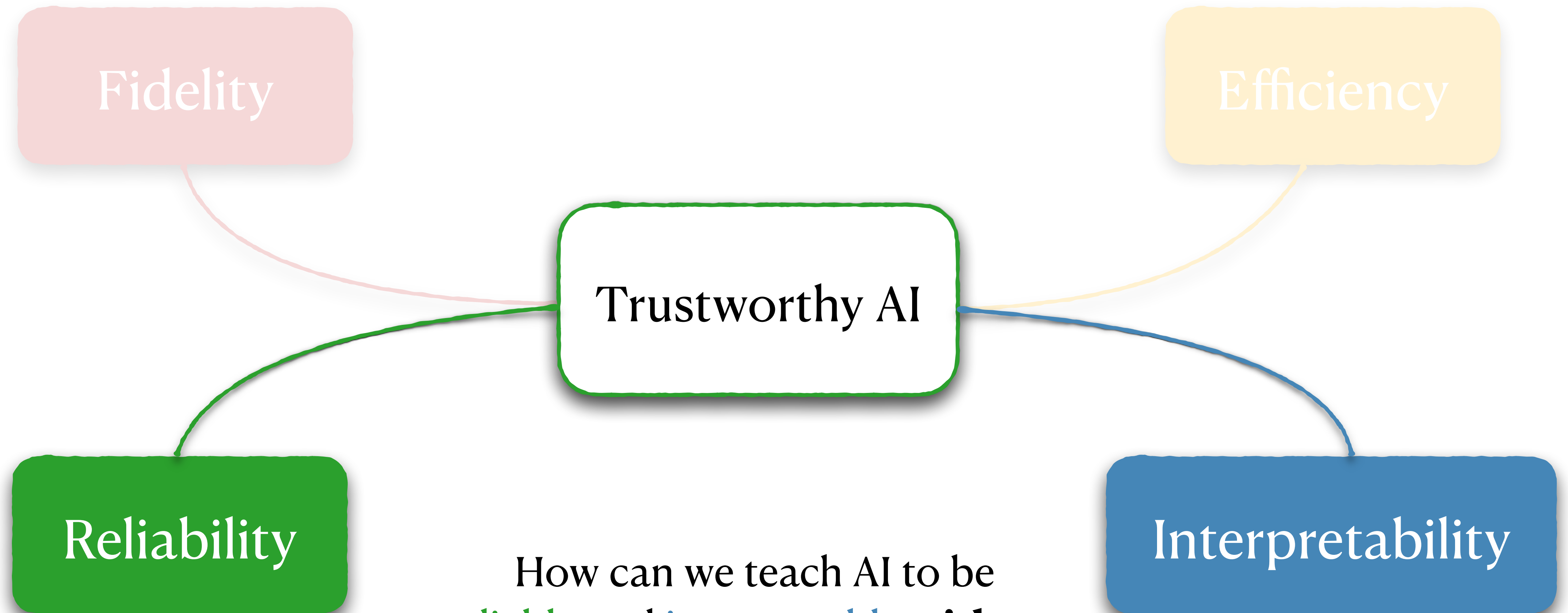
13-17 July, 2026



Towards trustworthy AI for HEP



Towards trustworthy AI for HEP

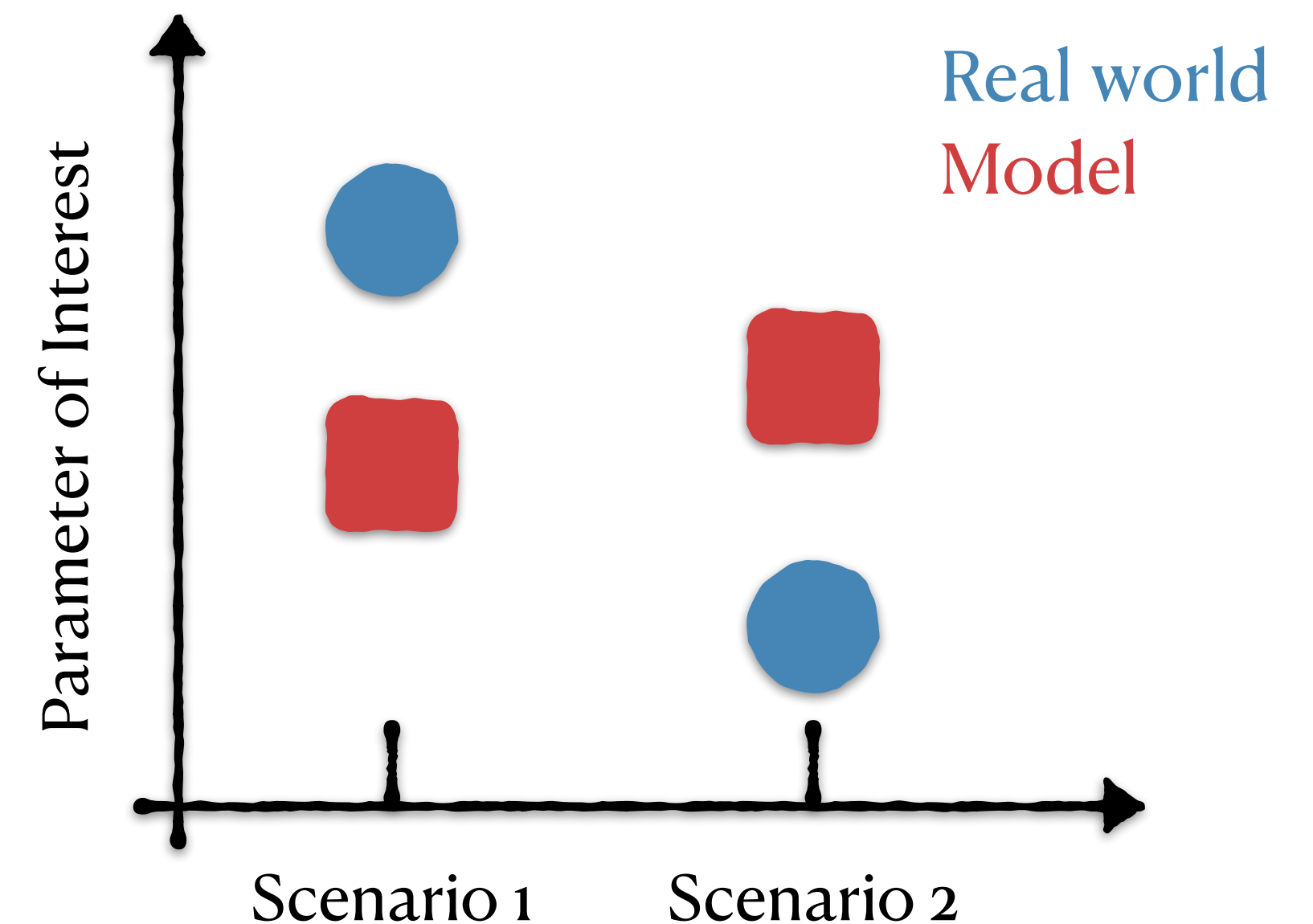


How can we teach AI to be
reliable and **interpretable** **without**
losing fidelity or **efficiency**?

Can we trust AI-driven discovery?

“All models are wrong, but some are useful”
George E. P. Box

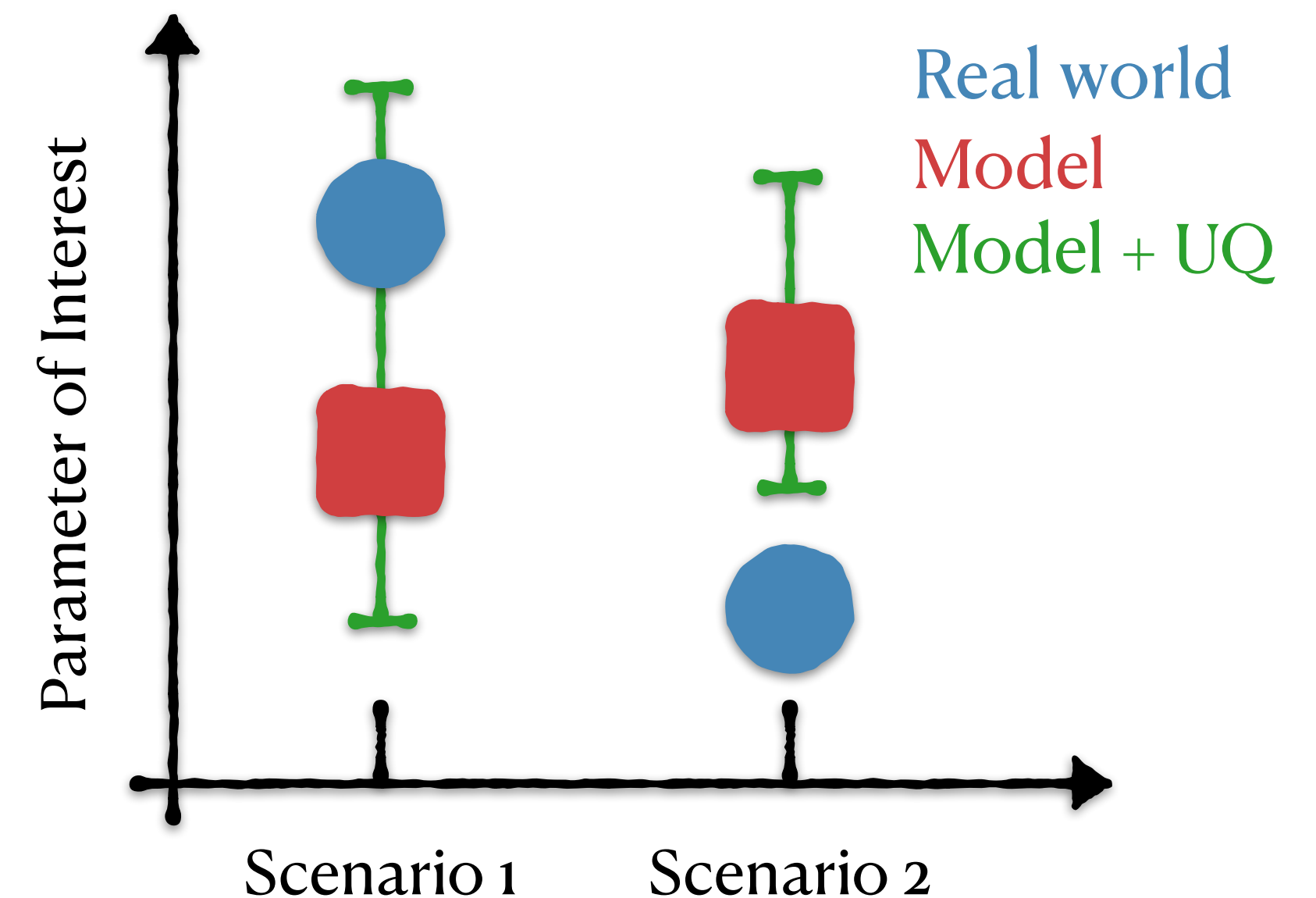
- ❖ **AUC & Accuracy:** Measure average discriminative power, but do not indicate certainty or coverage.
- ❖ **Overconfidence:** Even an AUC ≥ 0.99 can lead to systematic overconfidence.
- ❖ **Distribution shift:** Imperfect simulations, detector effects, and rare signals in sparse feature-space corners.



Can we trust AI-driven discovery?

“All models are wrong, but some are useful”
George E. P. Box

- ❖ **AUC & Accuracy:** Measure average discriminative power, but do not indicate certainty or coverage.
- ❖ **Overconfidence:** Even an $\text{AUC} \geq 0.99$ can lead to systematic overconfidence.
- ❖ **Distribution shift:** Imperfect simulations, detector effects, and rare signals in sparse feature-space corners.

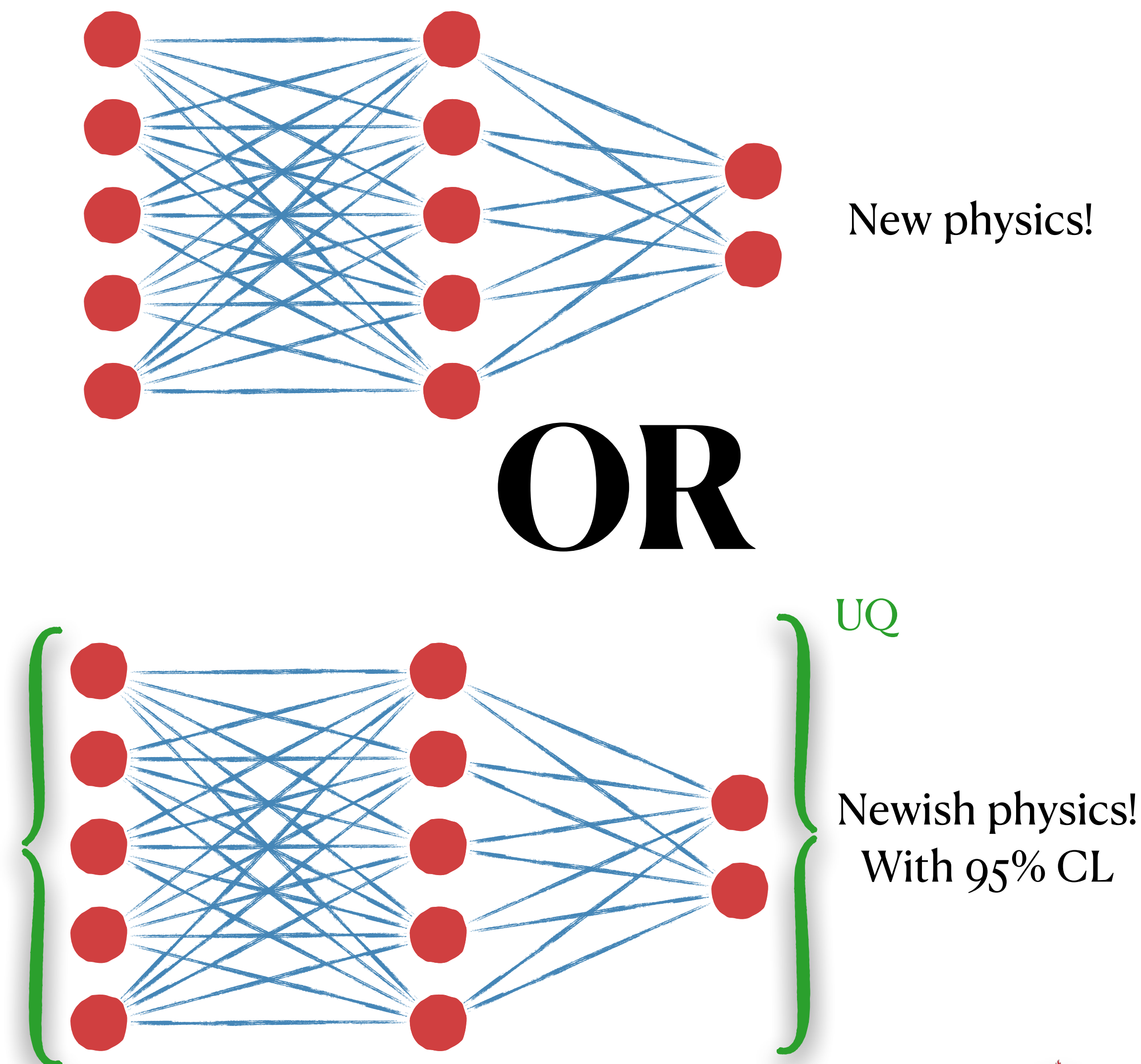


◆ How wrong is our model?

♣ And how confident are we in their predictions?

Outline

- ❖ Widely used methods for UQ
- ❖ Conformal Prediction for HEP
 - ◆ Classification
 - ◆ Generative Modelling
- ❖ Conformal Inference for anomaly detection



What is wrong with what we are doing?

Data-driven methods for UQ

Holdout Data

Reserve a test set; evaluate metrics on unseen data. Provides point estimates but no per-prediction uncertainty.

Cross-validation

Partition into k-folds; train and evaluate k times. This reduces the variance of estimates but assumes stationarity across folds.

Bootstrapping

Resample training data with replacement; develop multiple models. Relies on **asymptotic** arguments; cannot manage distribution shift.

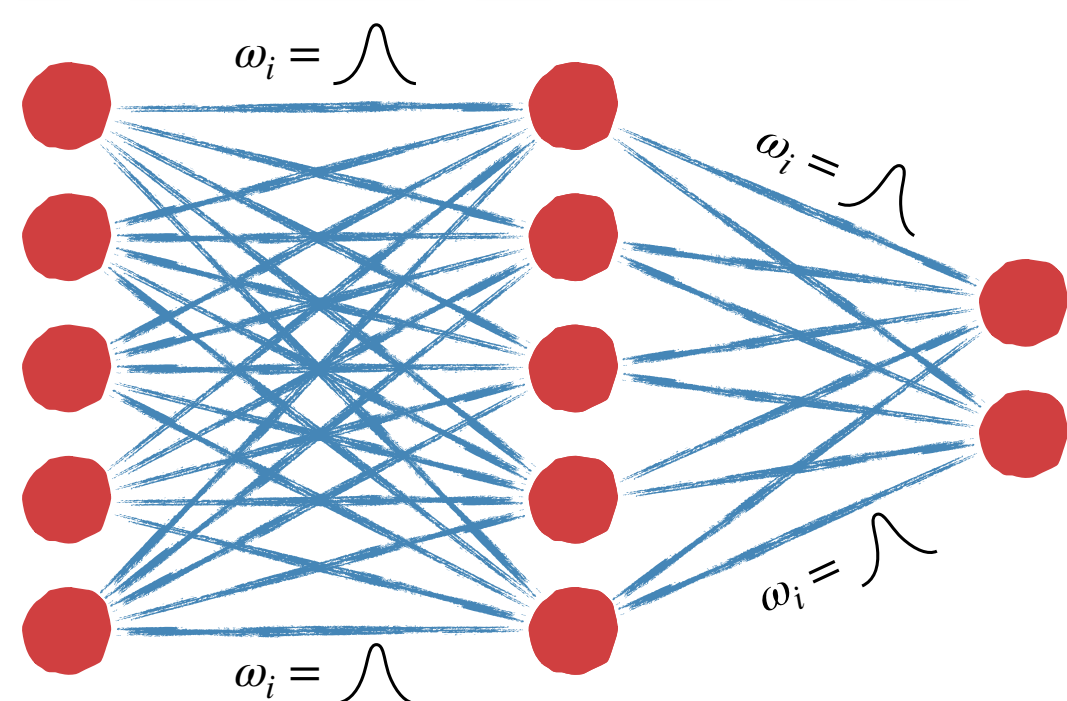
MC Dropout

Uncertainty estimates depend on dropout rate; **no formal coverage guarantee**; tends to underestimate uncertainty in out-of-distribution regions.

Deep Ensembles

Computationally expensive ($N \times$ training cost); ensemble diversity is not guaranteed; provides **no finite-sample statistical guarantee on coverage**.

Bayesian Neural Networks



- ◆ Computational cost
- ◆ Sensitive to prior choice
- ◆ Guarantees are **asymptotic**
- ◆ Architectural modification
- ◆ Posterior approximations introduce uncontrolled bias

What is wrong with what we are doing?

Data-driven methods for UQ

Holdout Data

Reserve a test set; evaluate metrics on unseen data. Provides point estimates but no per-prediction uncertainty.

Cross-validation

Partition into k-folds; train and evaluate k times. This reduces the variance of estimates but assumes stationarity across folds.

Bootstrapping

Resample training data with replacement; develop multiple models. Relies on **asymptotic** arguments; cannot manage distribution shift.

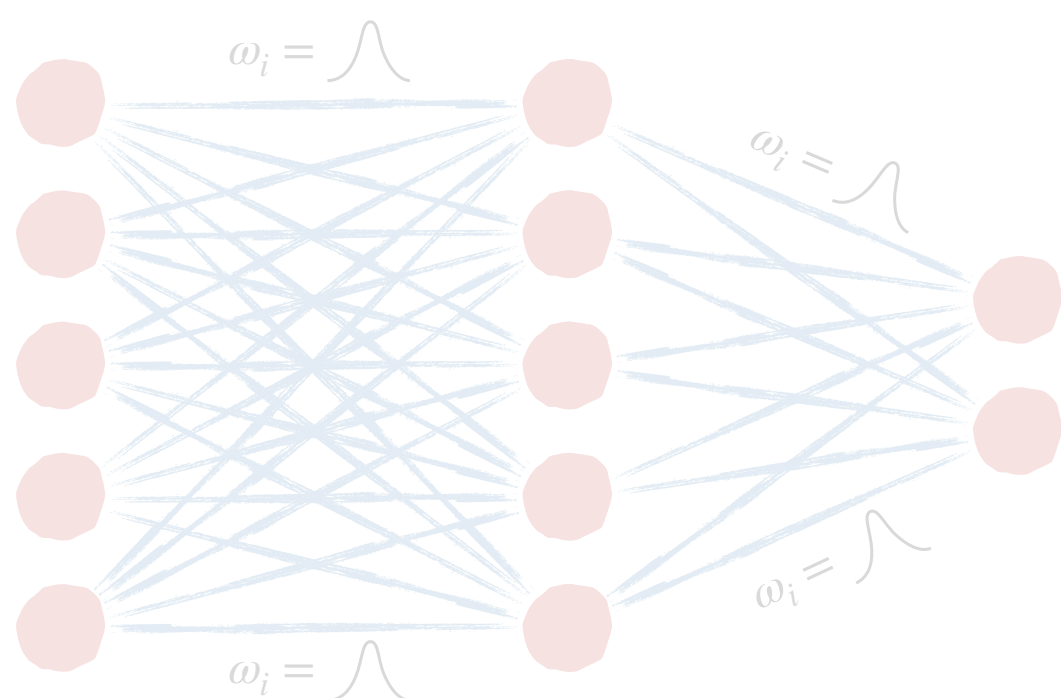
MC Dropout

Uncertainty estimates depend on dropout rate; **no formal coverage guarantee**; tends to underestimate uncertainty in out-of-distribution regions.

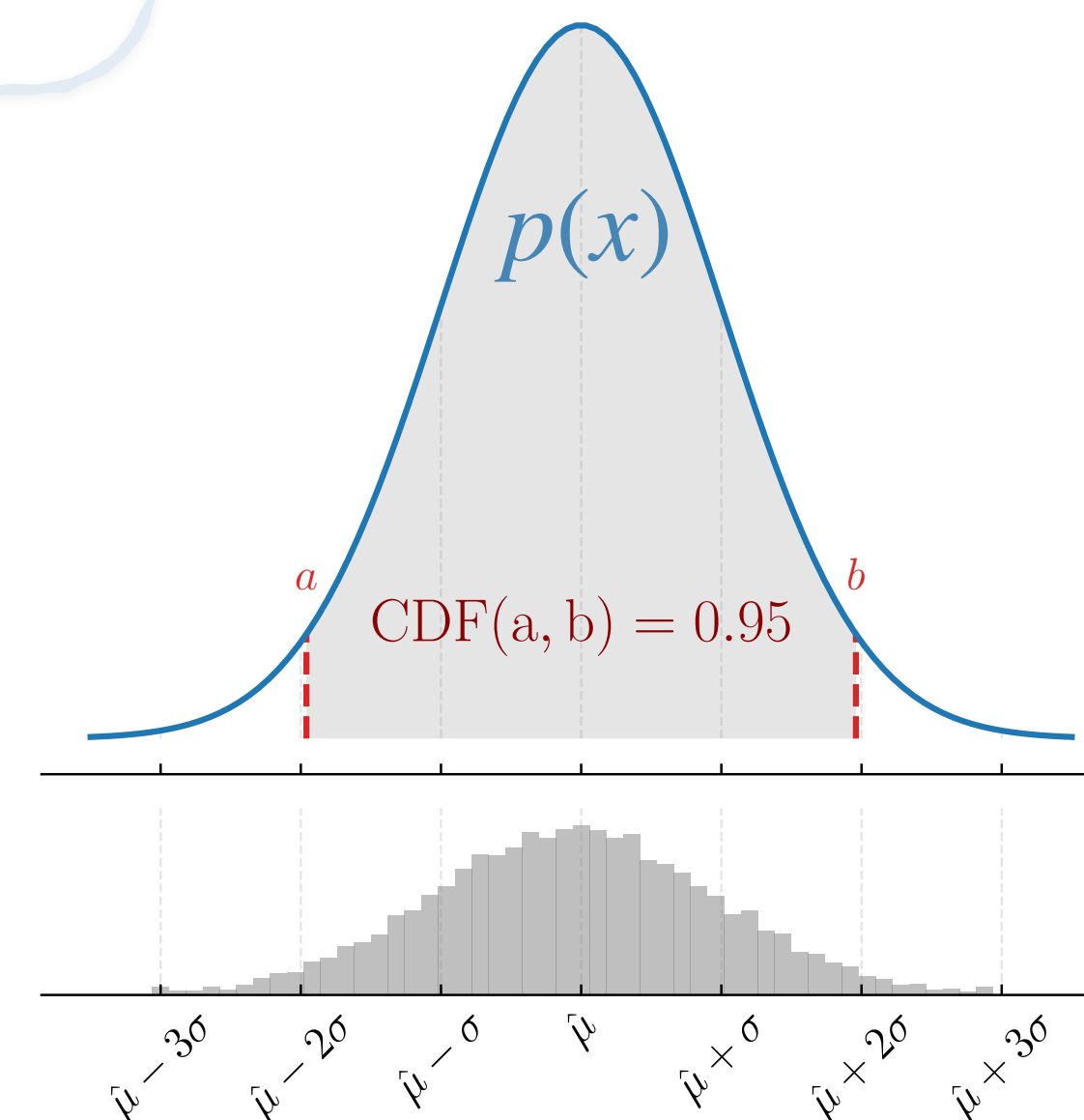
Deep Ensembles

Computationally expensive ($N \times$ training cost); ensemble diversity is not guaranteed; provides **no finite-sample statistical guarantee on coverage**.

Bayesian Neural Networks



- ◆ Computational cost
- ◆ Sensitive to prior choice
- ◆ Guarantees are **asymptotic**
- ◆ Architectural modification
- ◆ Posterior approximations introduce uncontrolled bias



What is wrong with what we are doing?

Data-driven methods for UQ

Holdout Data

Reserve a test set; evaluate metrics on unseen data. Provides point estimates but no per-prediction uncertainty.

Cross-validation

Partition into k-folds; train and evaluate k times. This reduces the variance of estimates but assumes stationarity across folds.

Bootstrapping

Resample training data with replacement; develop multiple models. Relies on **asymptotic** arguments; cannot manage distribution shift.

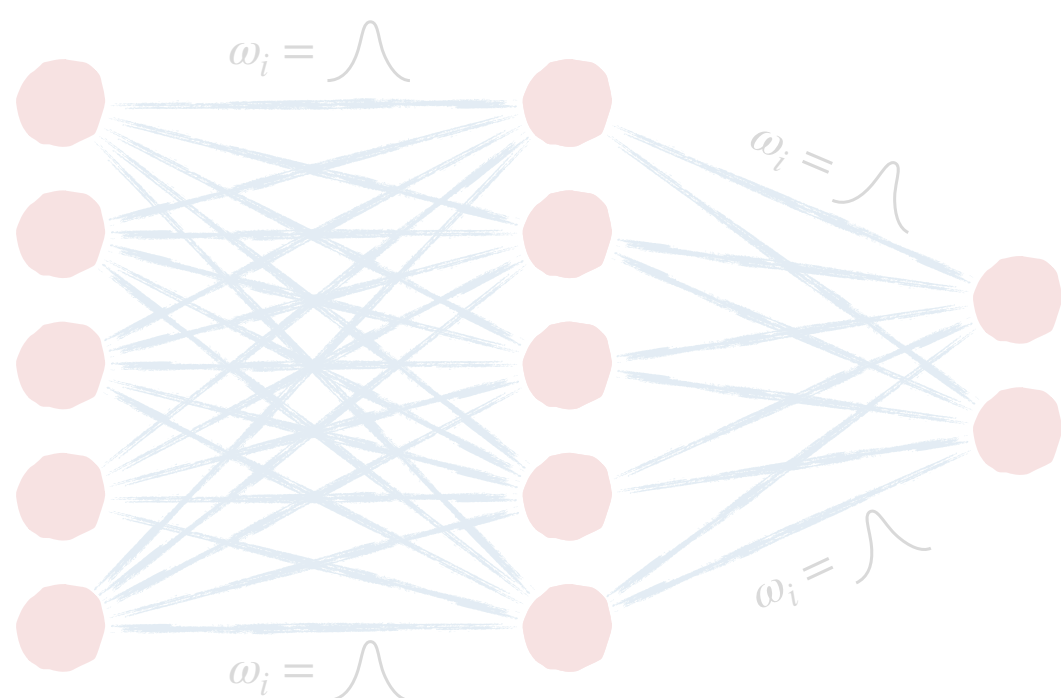
MC Dropout

Uncertainty estimates depend on dropout rate; **no formal coverage guarantee**; tends to underestimate uncertainty in out-of-distribution regions.

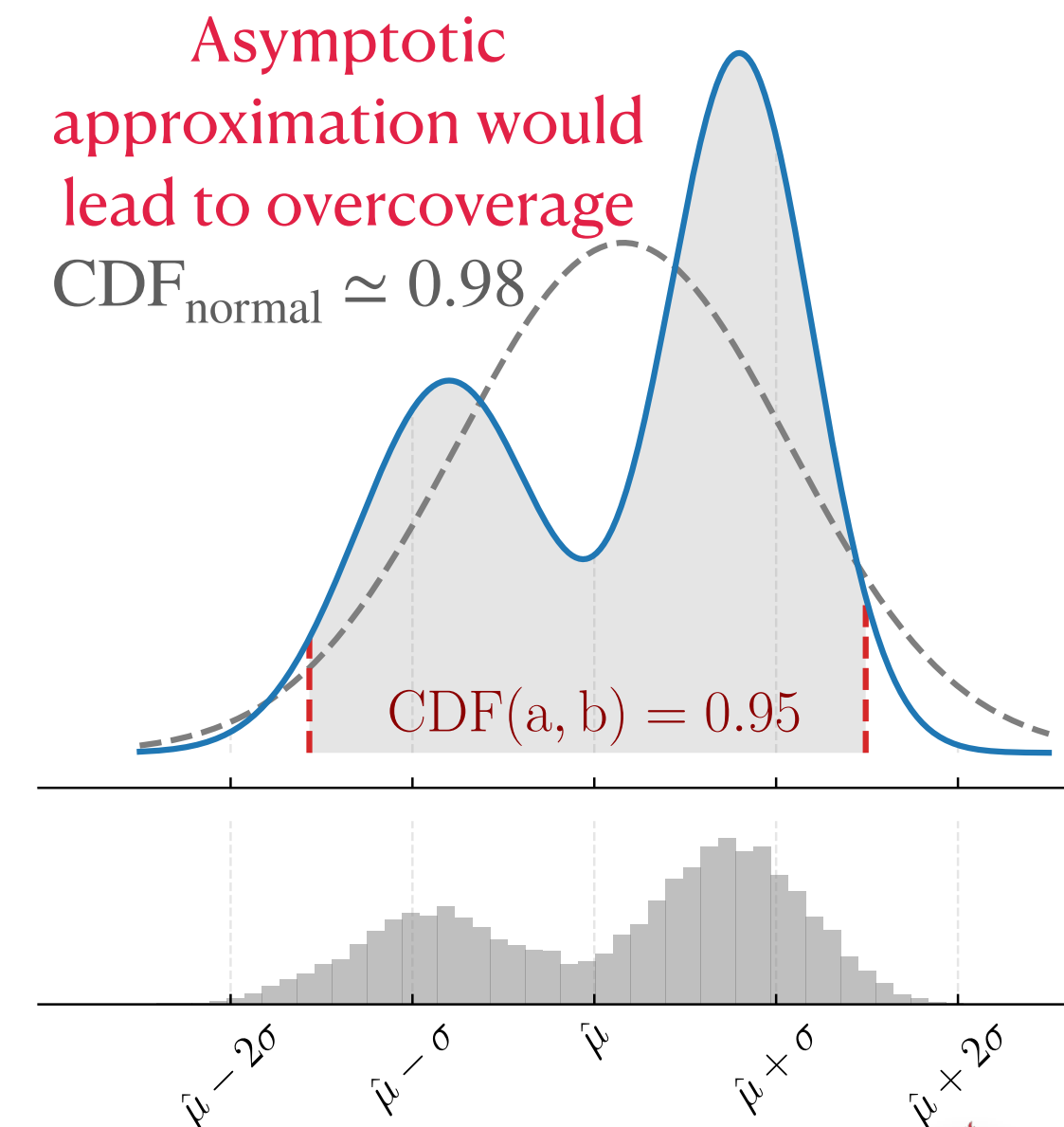
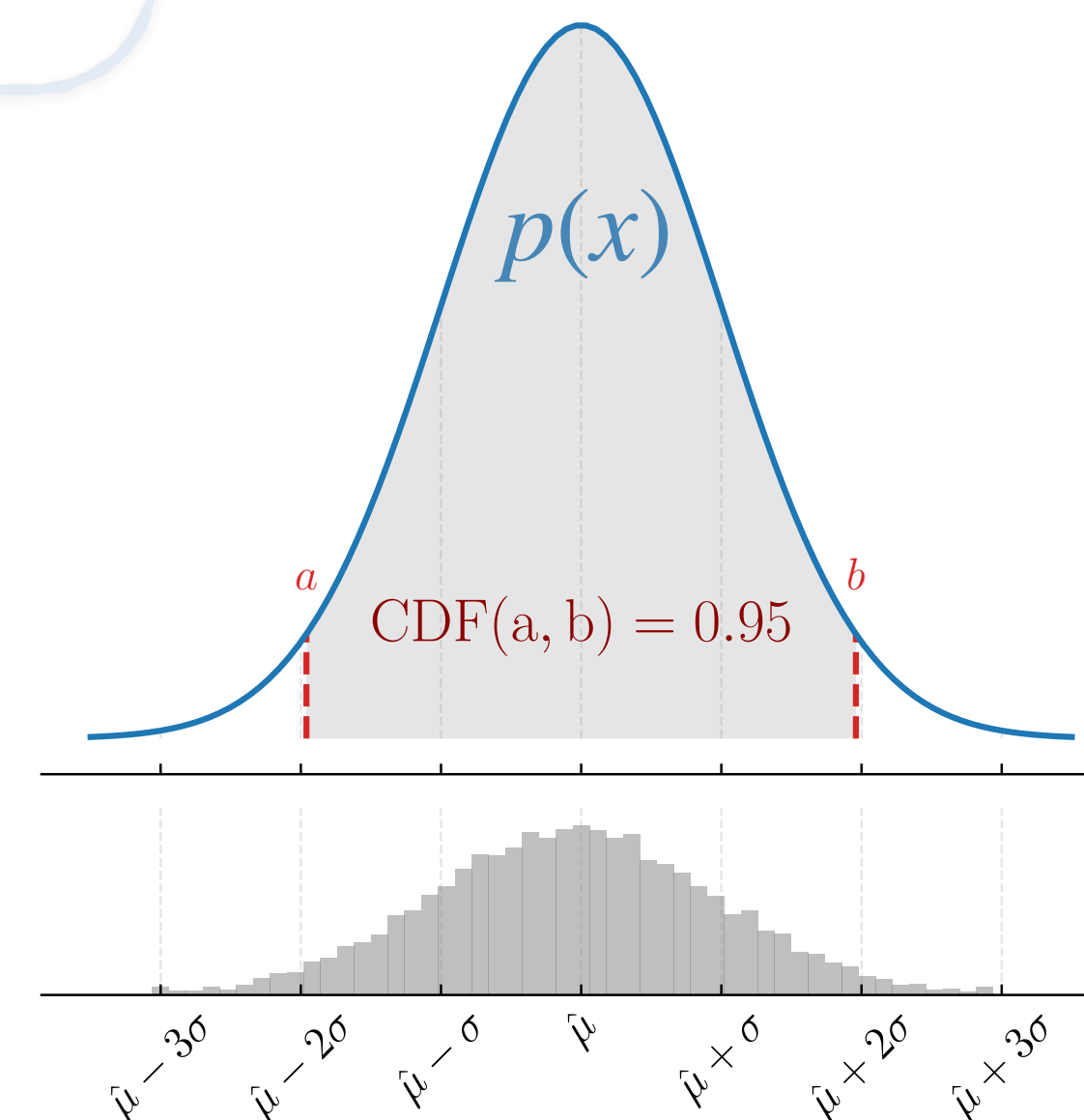
Deep Ensembles

Computationally expensive ($N \times$ training cost); ensemble diversity is not guaranteed; provides **no finite-sample statistical guarantee on coverage**.

Bayesian Neural Networks



- ◆ Computational cost
- ◆ Sensitive to prior choice
- ◆ Guarantees are **asymptotic**
- ◆ Architectural modification
- ◆ Posterior approximations introduce uncontrolled bias



Jack Y. Araz

What is wrong with what we are doing?

Data-driven methods for UQ

Holdout Data

Reserve a test set; evaluate metrics on unseen data. Provides point

How certain is your uncertainty?

- ❖ Rely on assumptions being correct: asymptotic approximation
- ❖ Or test empirically whether our assumptions hold
- ❖ OR use inference methods that don't rely on assumptions (or only rely on weaker assumptions)

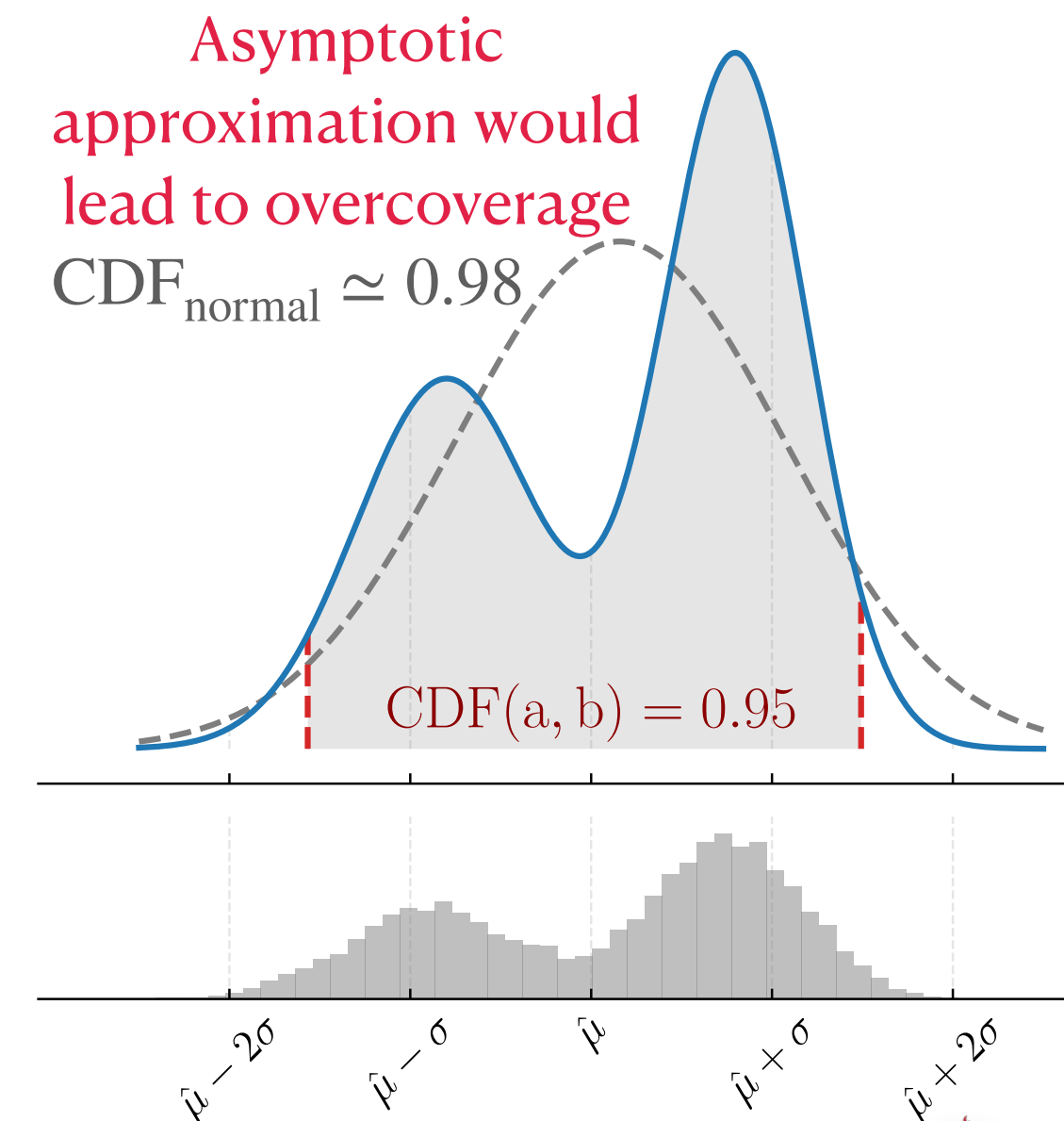
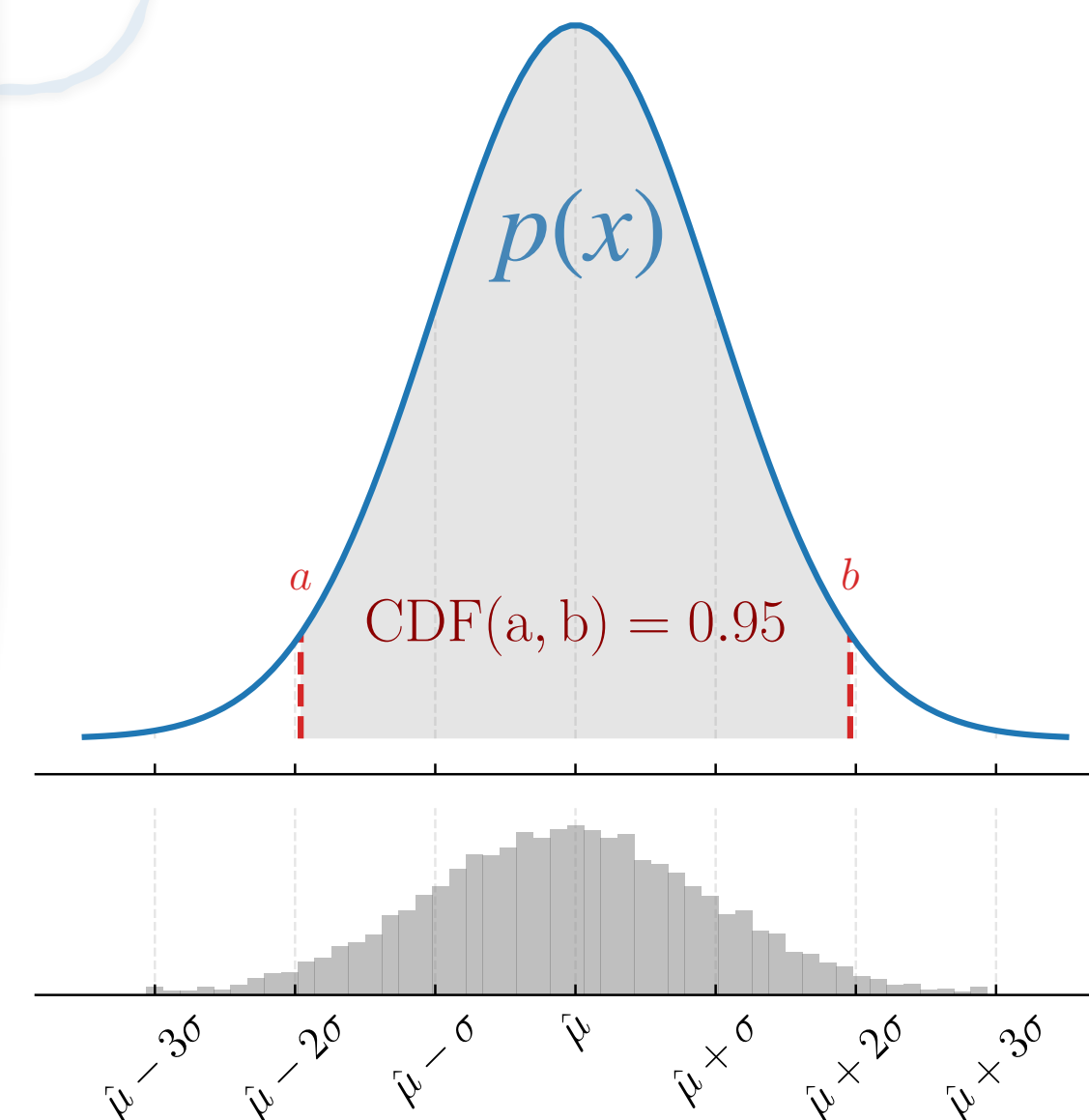
Posterior approximations introduce uncontrolled bias

MC Dropout

Uncertainty estimates depend on dropout rate; **no formal coverage guarantee**; tends to underestimate uncertainty in out-of-distribution regions.

Deep Ensembles

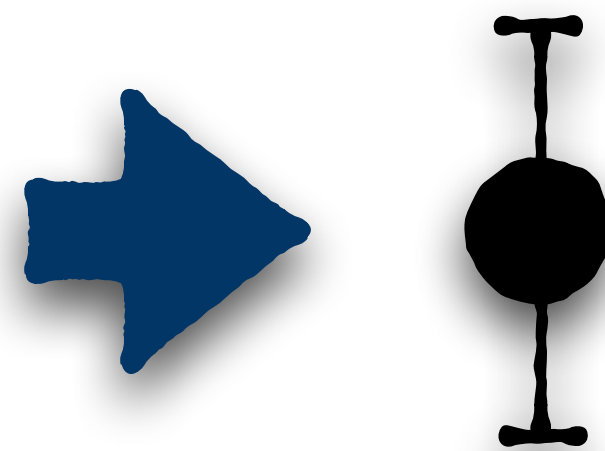
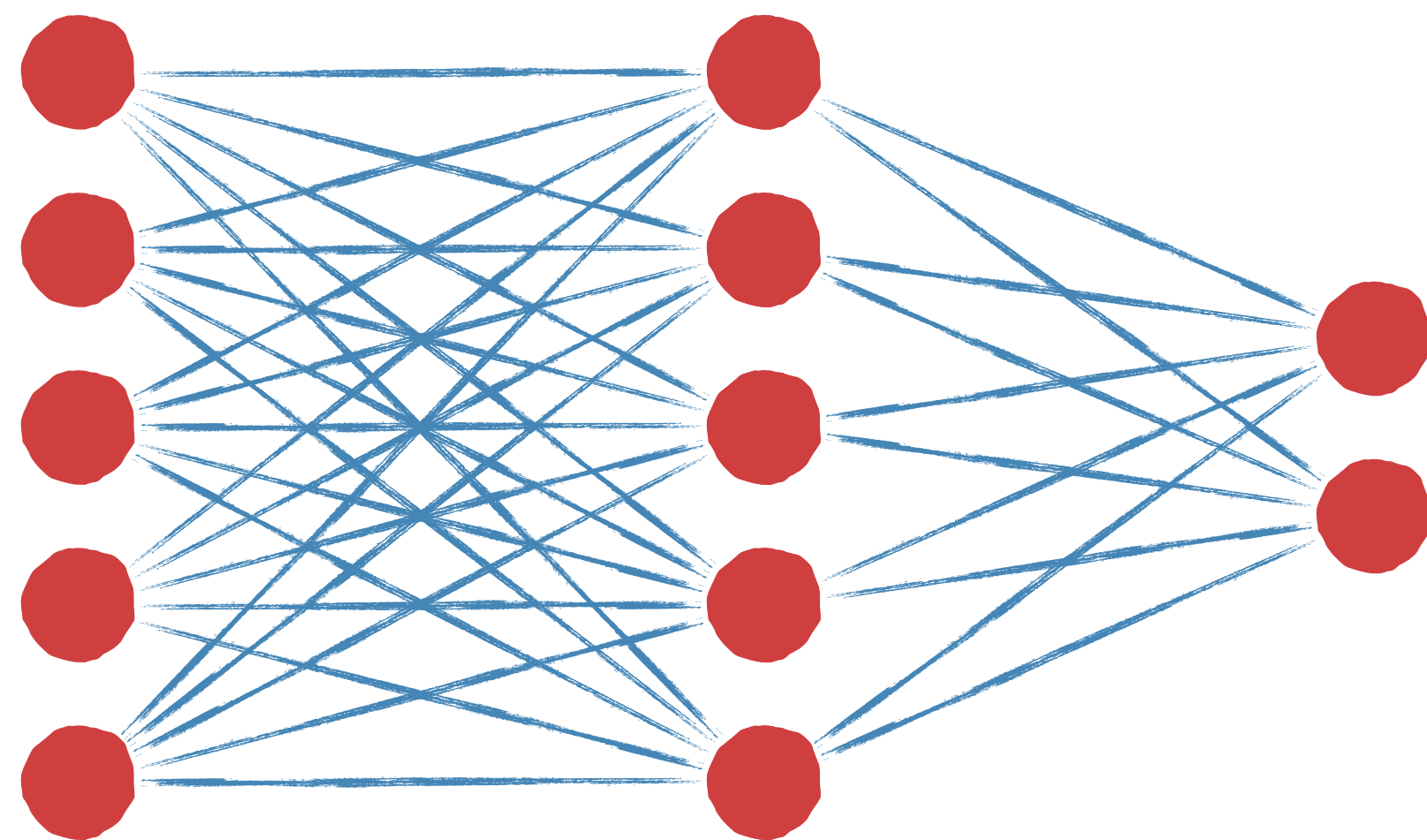
Computationally expensive ($N \times$ training cost); ensemble diversity is not guaranteed; provides **no finite-sample statistical guarantee on coverage**.



Jack Y. Araz

What do we need?

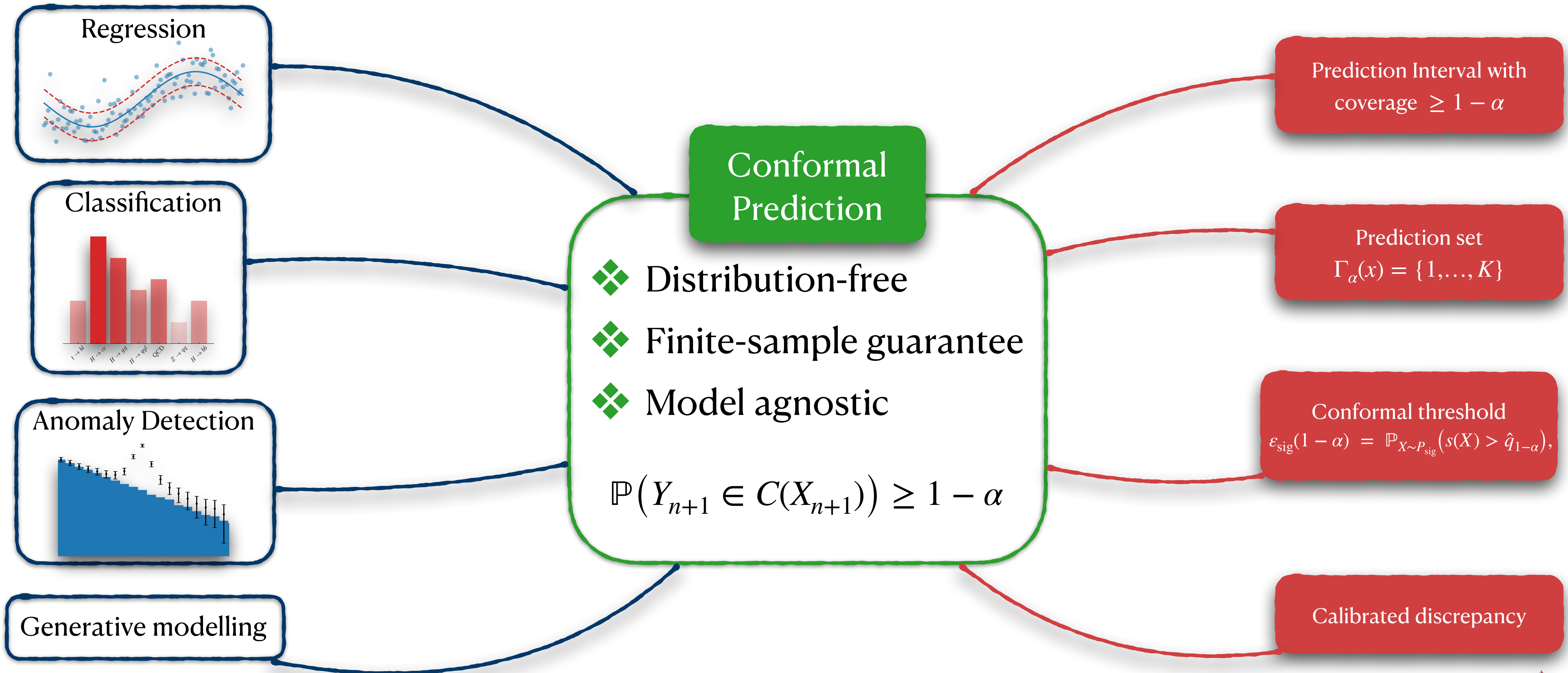
Rigorous, distribution-free, finite-sample CIs for any model and any dataset, for free



Prediction vs Predictive Inference

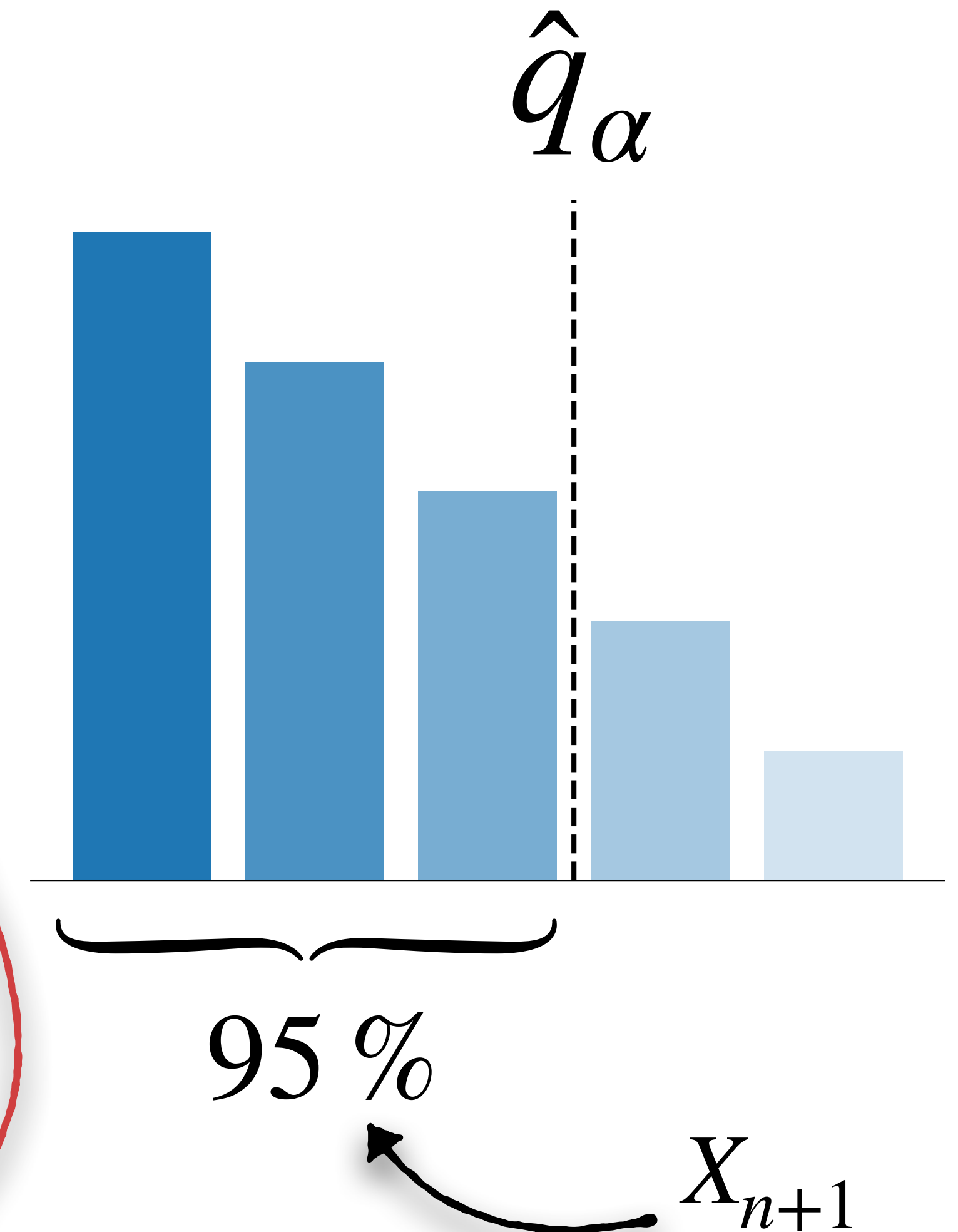


Building trust with confidence intervals



Split Conformal Prediction

- ❖ Using training data $\mathcal{D}_{\text{train}} = \{X_i, Y_i\}_{i=1}^{n_{\text{train}}}$, construct a fitted model $\hat{f}$ using any ML architecture
- ❖ Set aside a calibration set $\mathcal{D}_{\text{cal}} = \{X_i, Y_i\}_{i=1}^{n_{\text{cal}}}$
- ❖ Design a score function which penalises points that are considered out of distribution, e.g. $S_i = |Y_i^{\text{truth}} - \hat{f}(X_i | \theta)|$
- ❖ $\hat{q}_\alpha = \text{Quantile} \left(\{S_1, \dots, S_{n_{\text{cal}}}\}, \left\lceil (1 - \alpha) \left(\frac{n}{2} - 1 \right) \right\rceil \right)$
- ❖ For the test point $n + 1$ return prediction interval
 $C(X_{n+1}) = \hat{f}(X_{n+1} | \theta) \pm \hat{q}$



[Vovk et. al., 2005; Algorithmic learning in a random world]

See [2607.01354](#) for a score function engineering example

Jack Y. Araz

Split Conformal Prediction

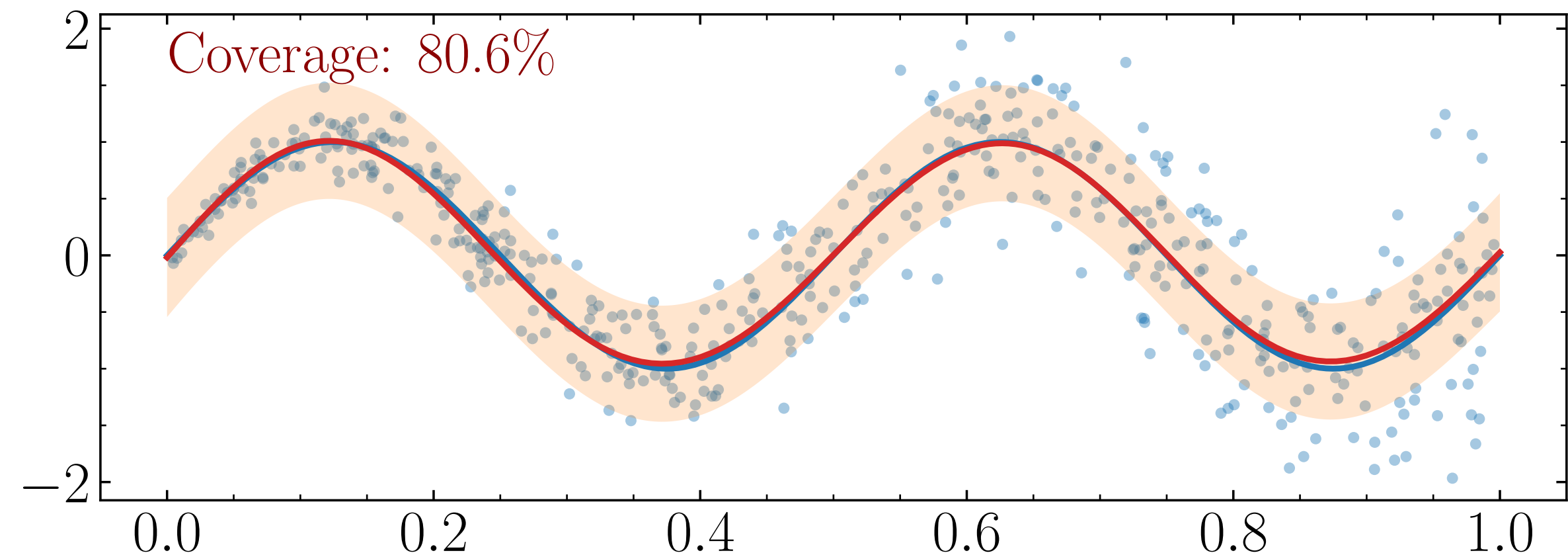
[JYA, Spannowsky; 2512.17048]

Theorem

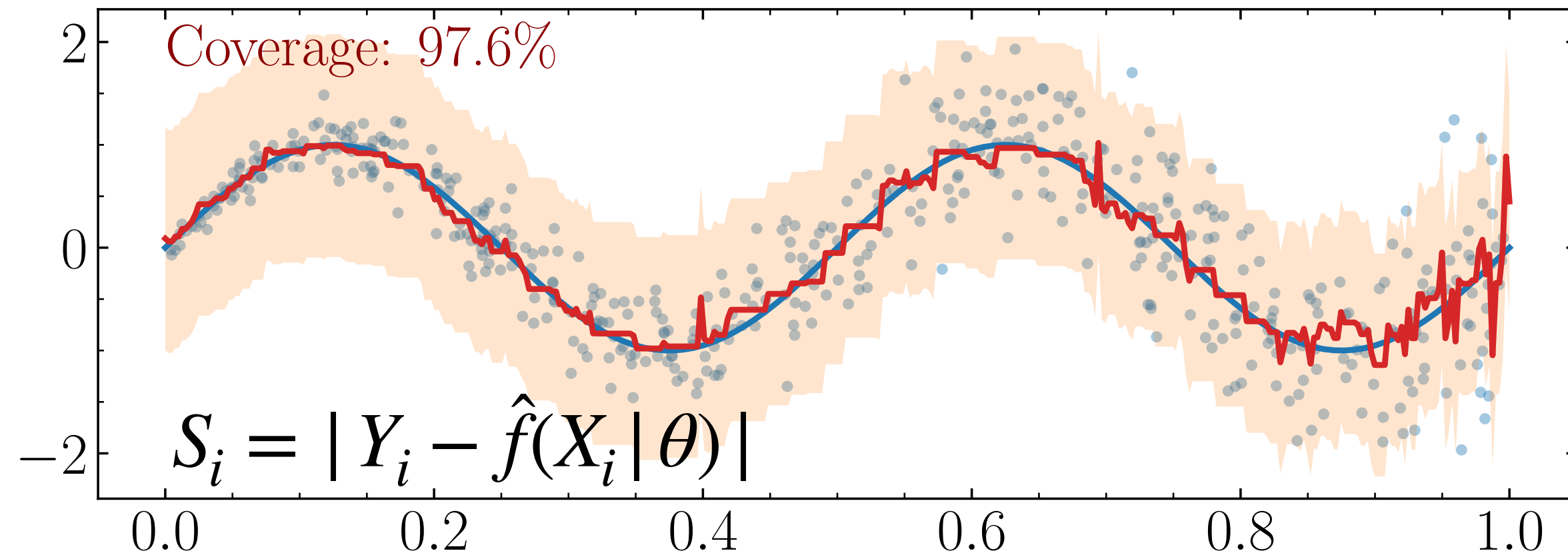
- ❖ If $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{cal}}$ are exchangeable, independent and identically distributed, then split conformal prediction satisfies

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

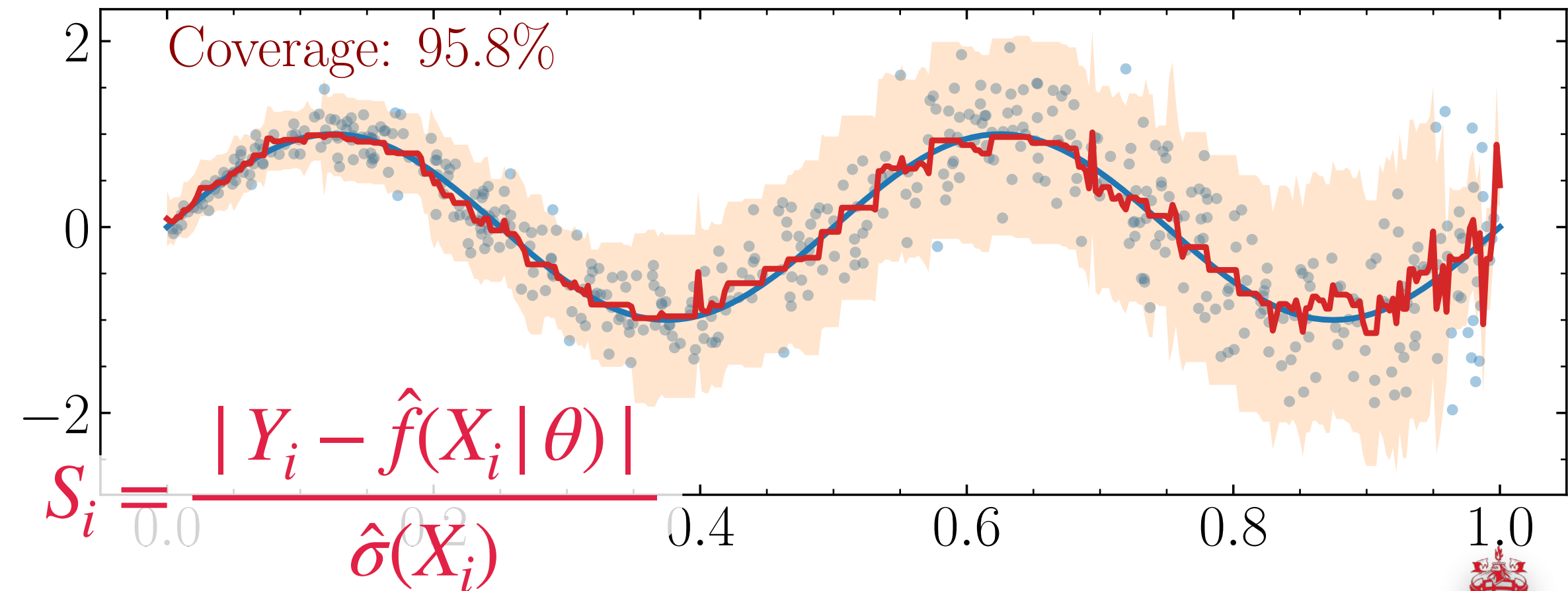
Gaussian Process



Conformal Prediction with an overfitted model



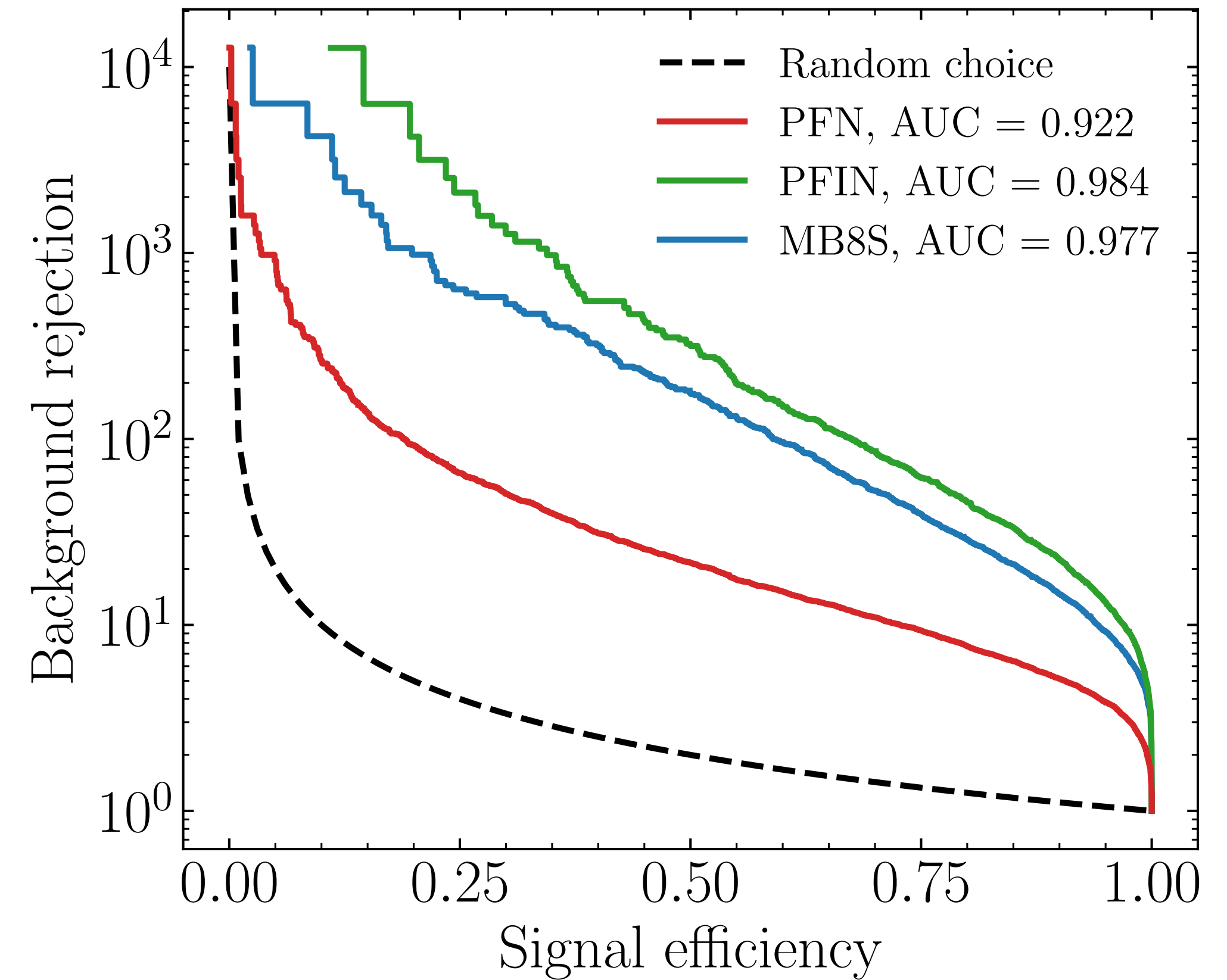
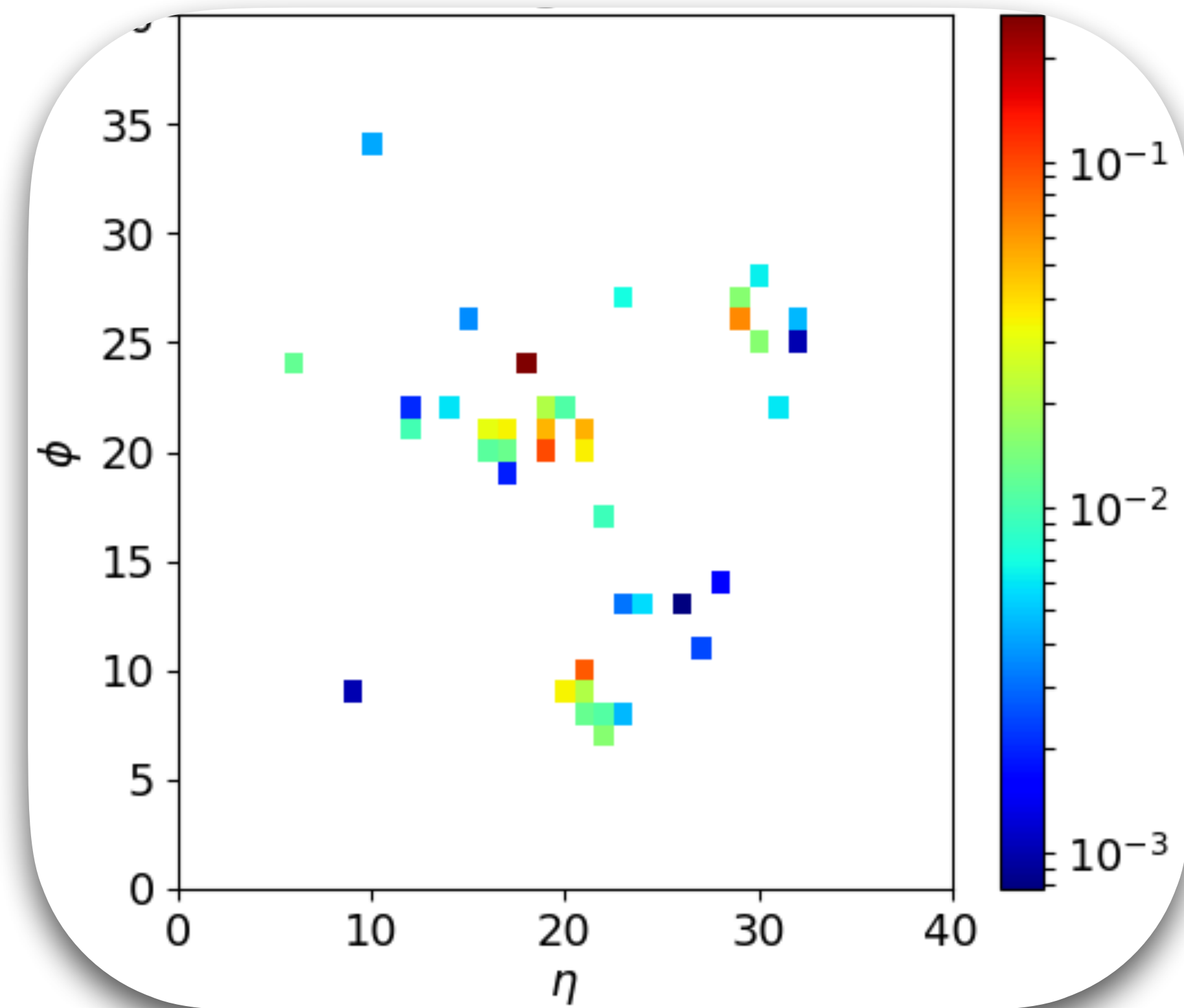
Adaptive Conformal Prediction



Conformal Prediction for Classification

Binary classification

Kasieczka *et. al.* SciPost'19

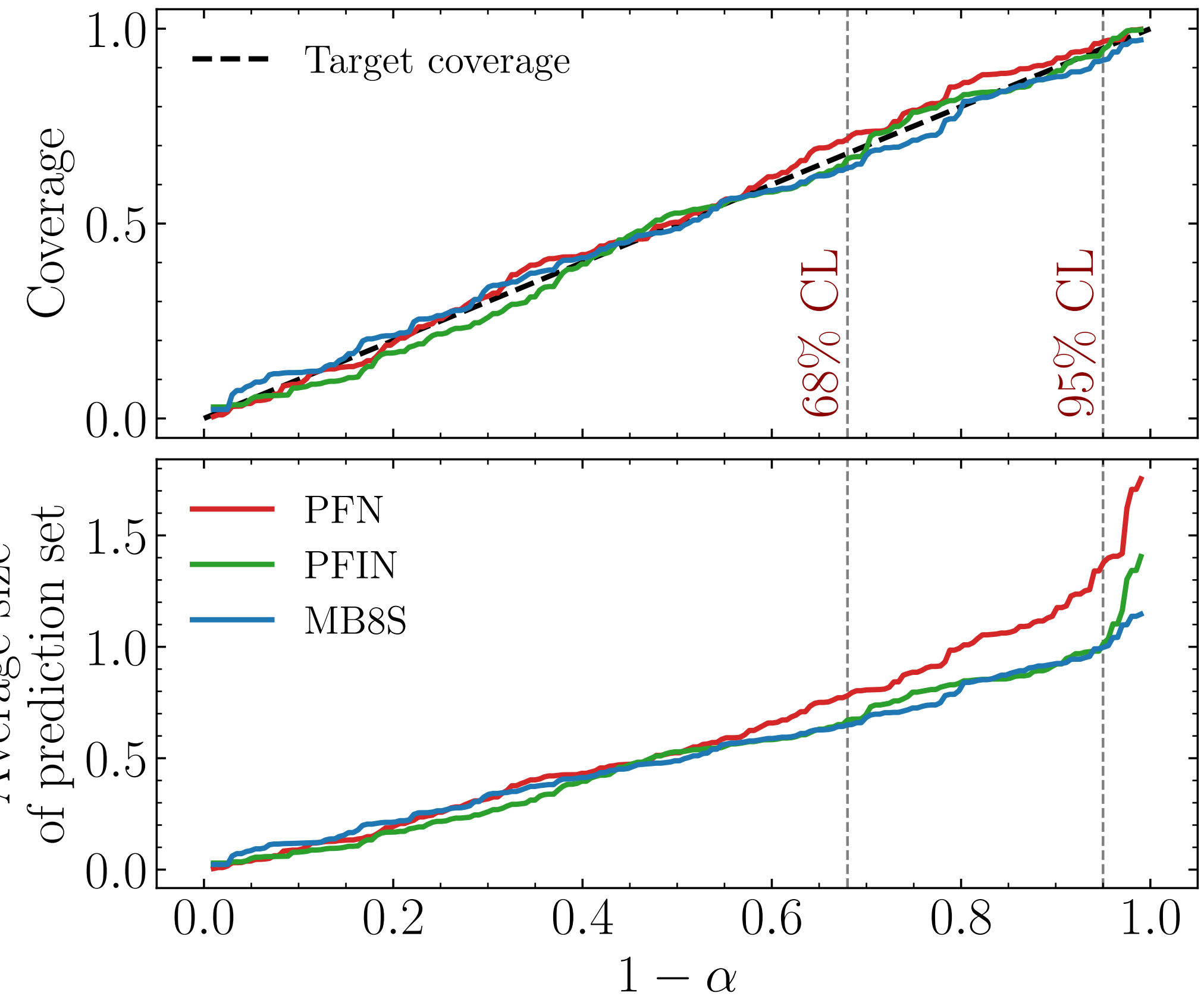
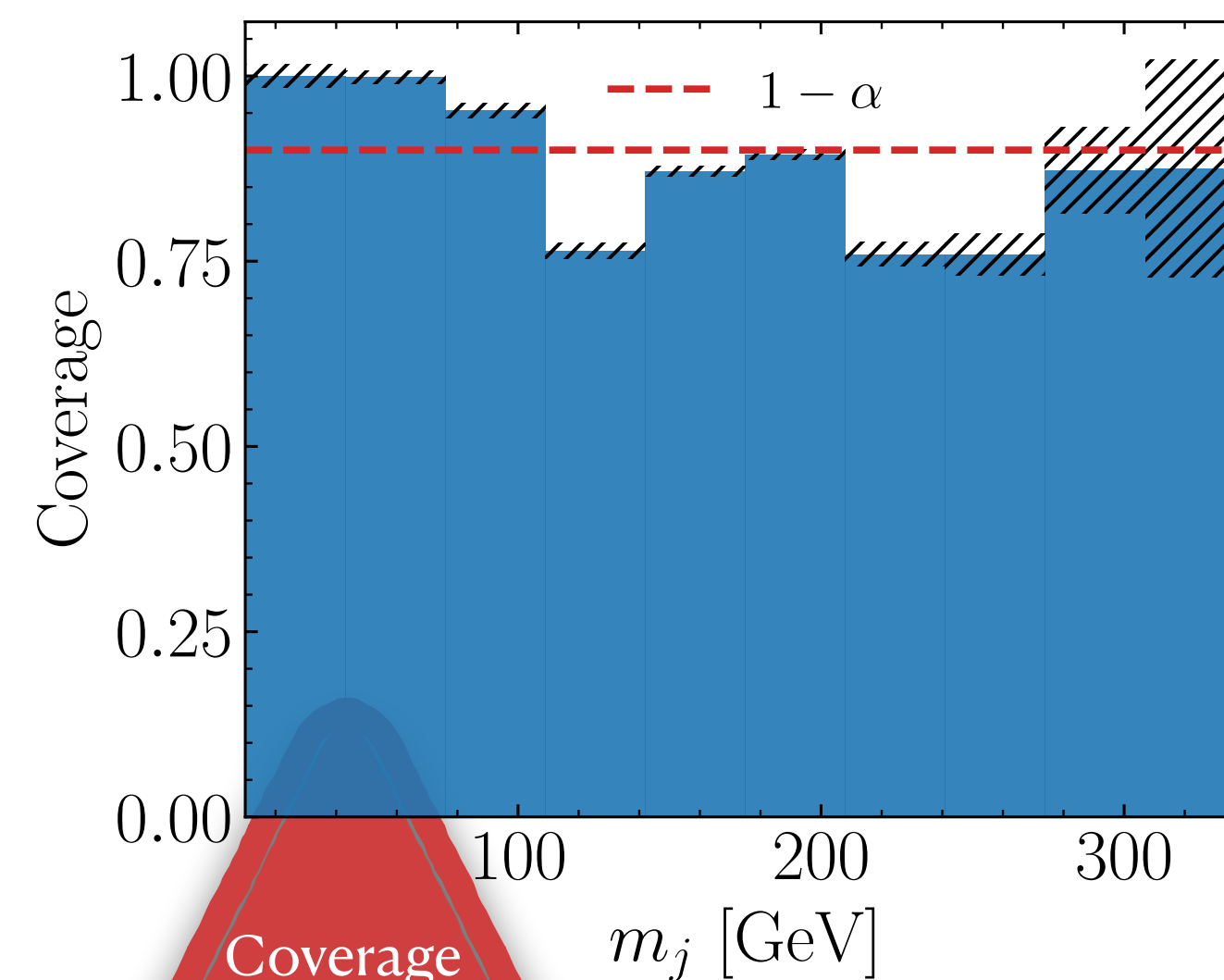
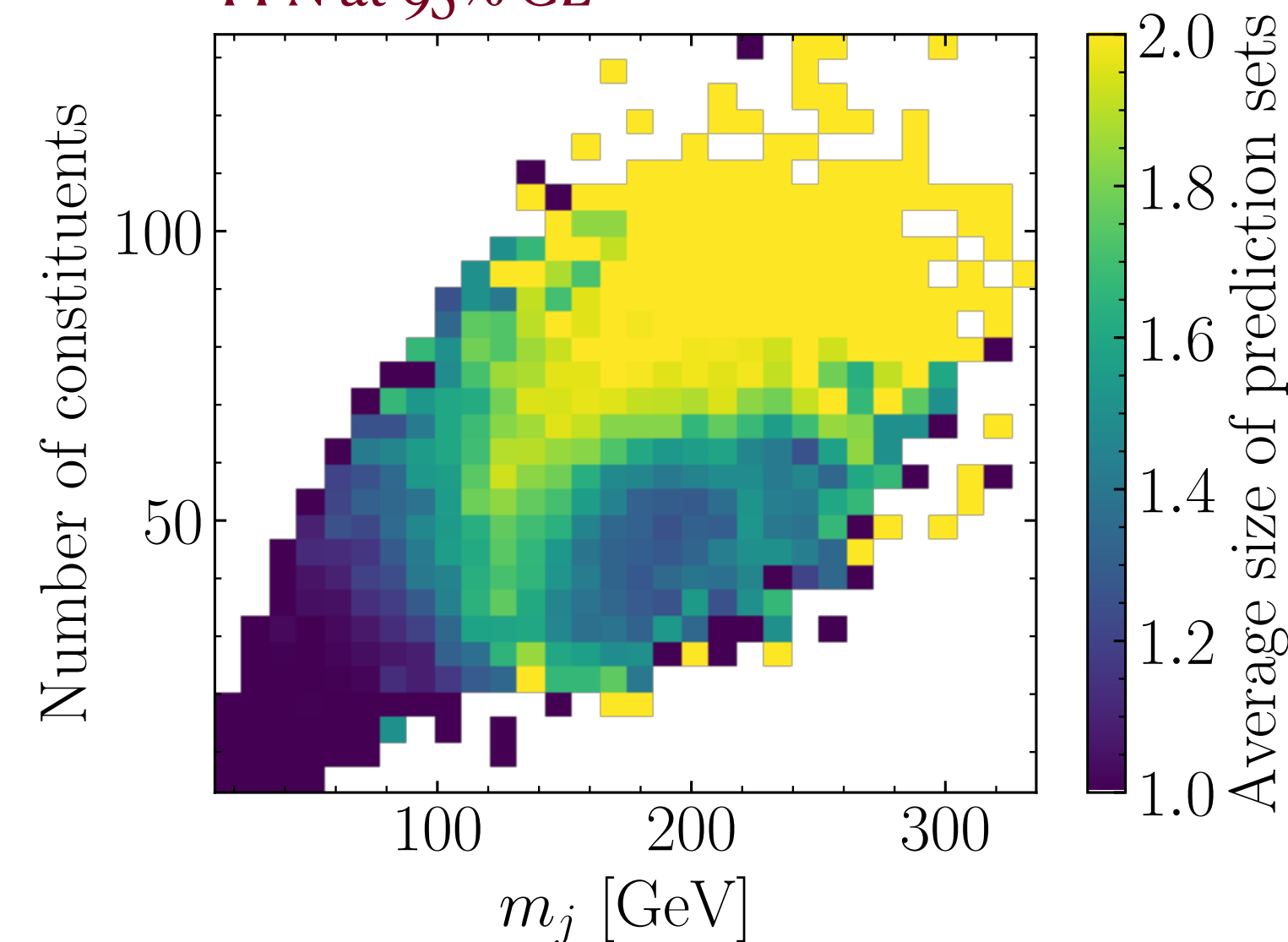


Binary classification

$$s(x, y) = 1 - \hat{f}_y(x)$$

$$\Gamma_\alpha(X_{n+1}) = \{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_{1-\alpha} \}$$

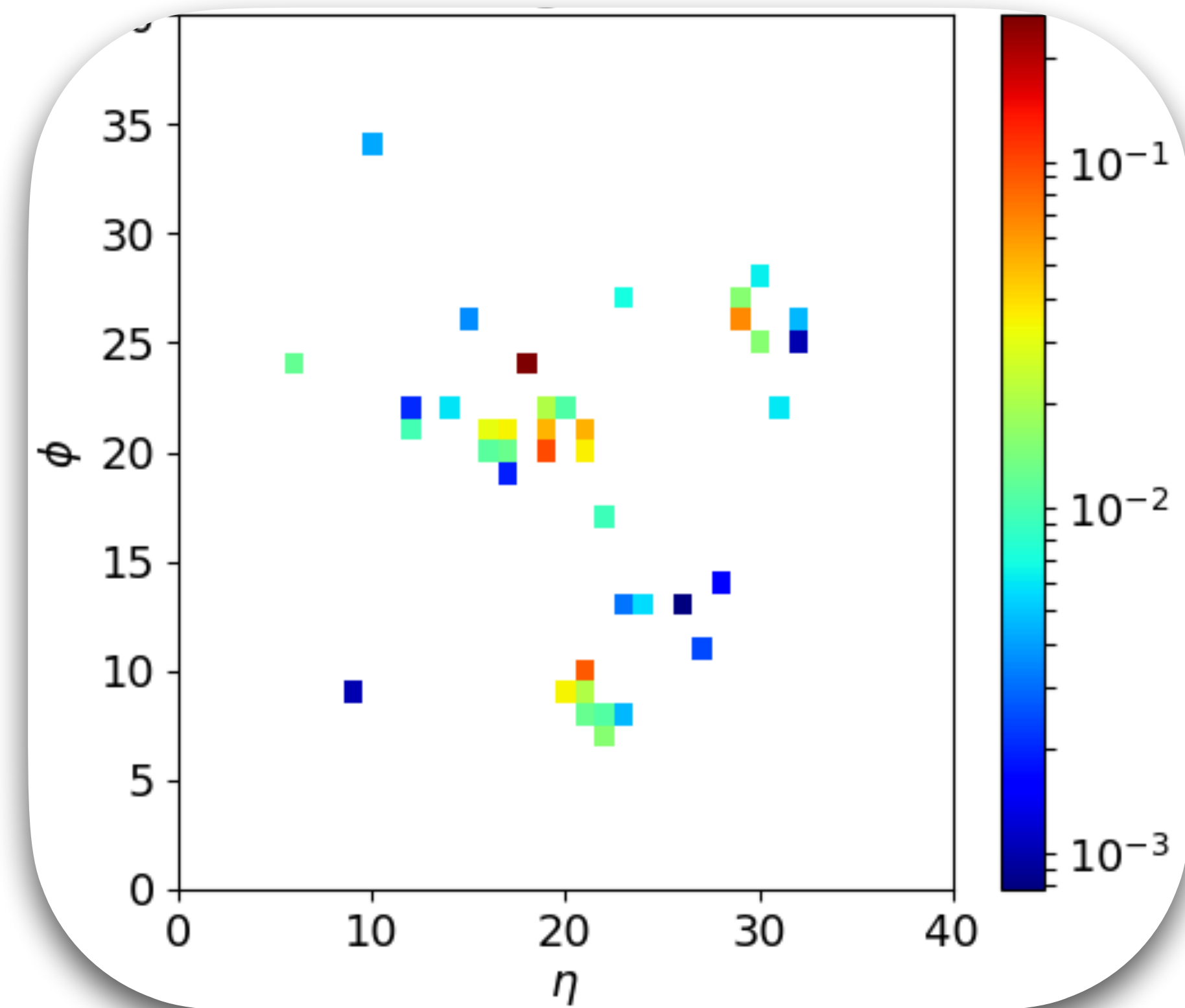
PFN at 95% CL



CP for Classification

[JYA, Spannowsky; [2512.17048](#)]

Kasieczka *et. al.* SciPost'19



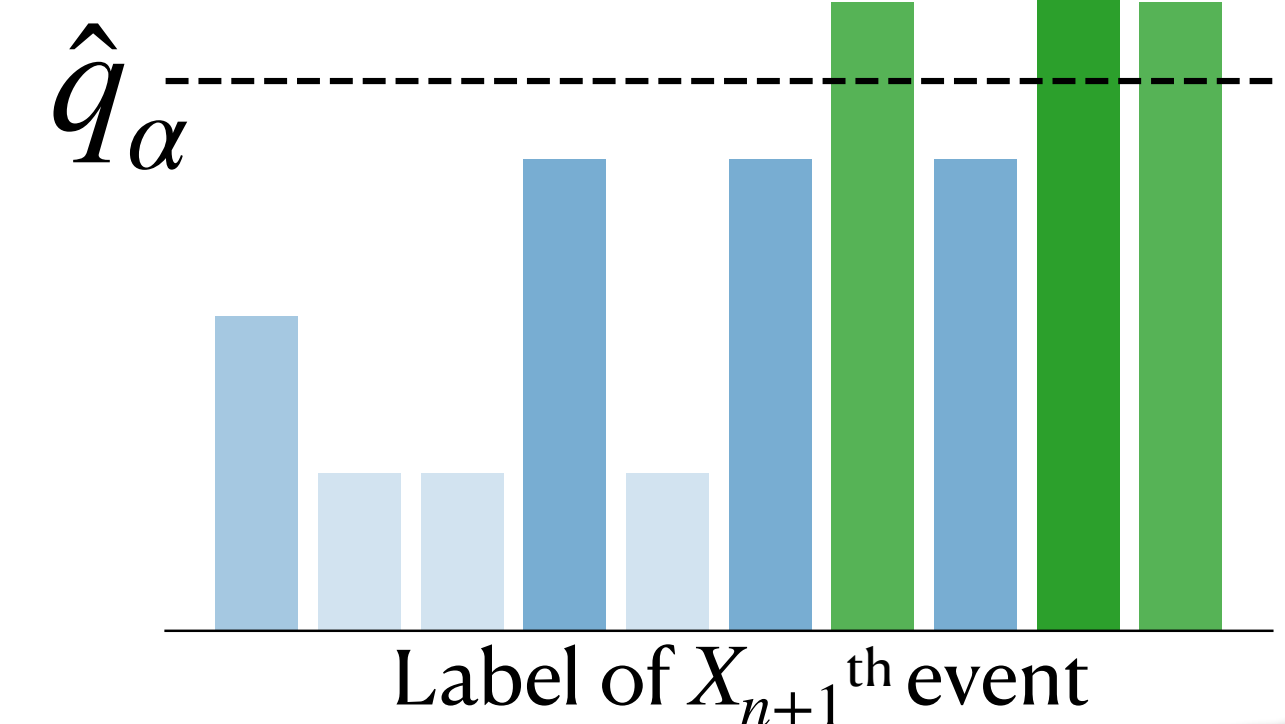
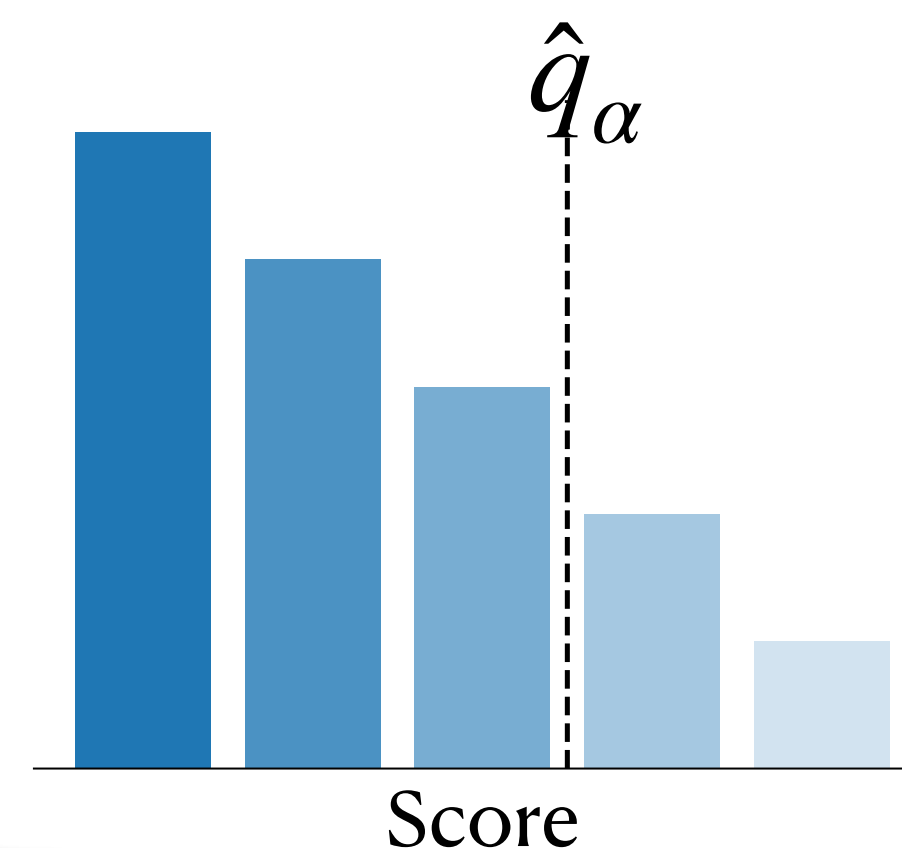
$$\Gamma_\alpha \rightarrow \{\text{Top}, H \rightarrow bb, \text{QCD}\}$$

❖ NN: OmniLearn transformer-based foundation model
OmniLearn: [Mikuni, Nachman; PRD '25]

❖ Dataset: JetClass (10 different classes)

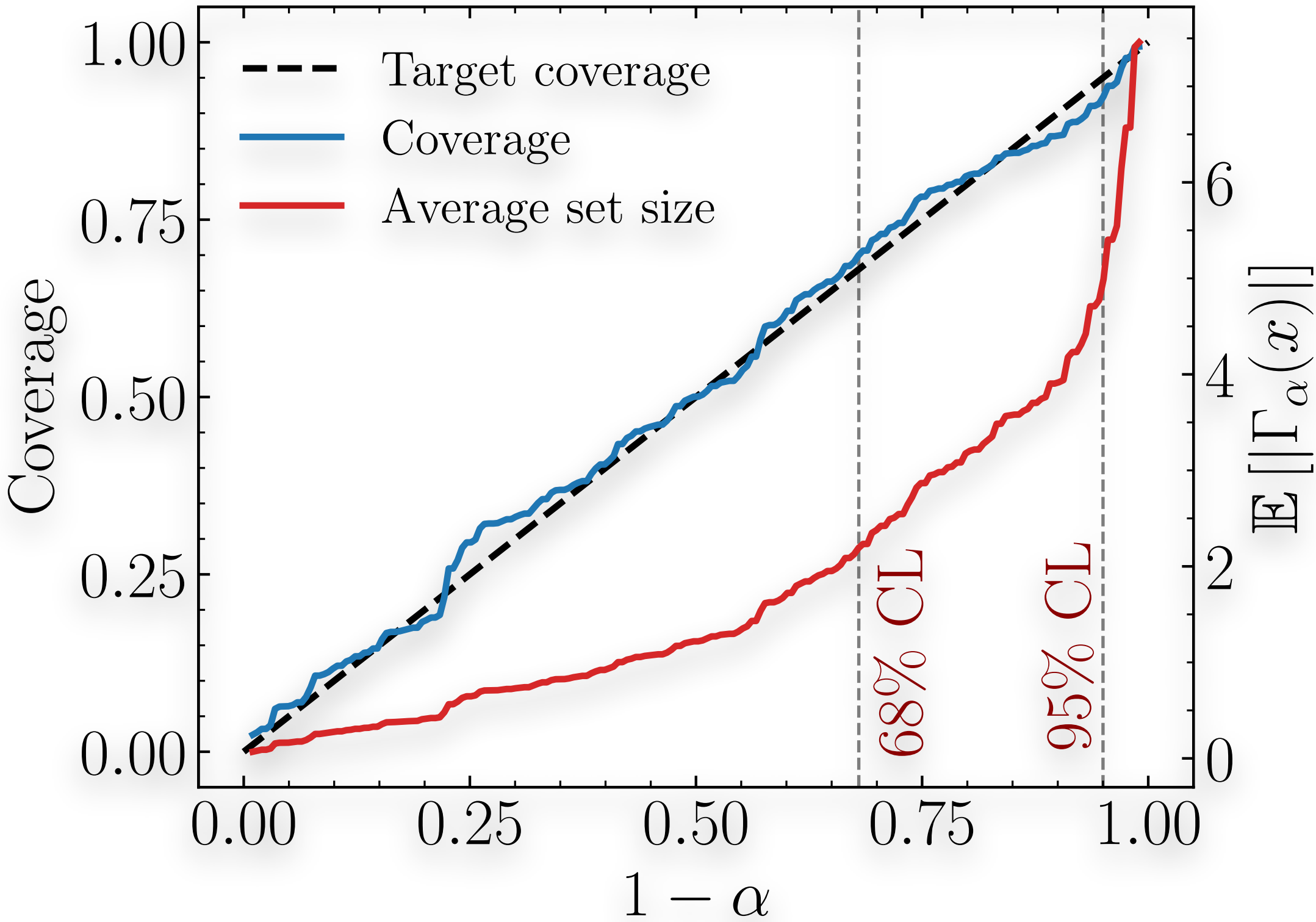
$$\Gamma \in \{QCD, H \rightarrow bb, H \rightarrow cc, H \rightarrow gg, H \rightarrow 4q, H \rightarrow qq\bar{l}, Z \rightarrow qq, W \rightarrow qq, T \rightarrow bqq, T \rightarrow bl\}$$

❖ Score: $\hat{f}(X_i | \theta)$ value of the true class

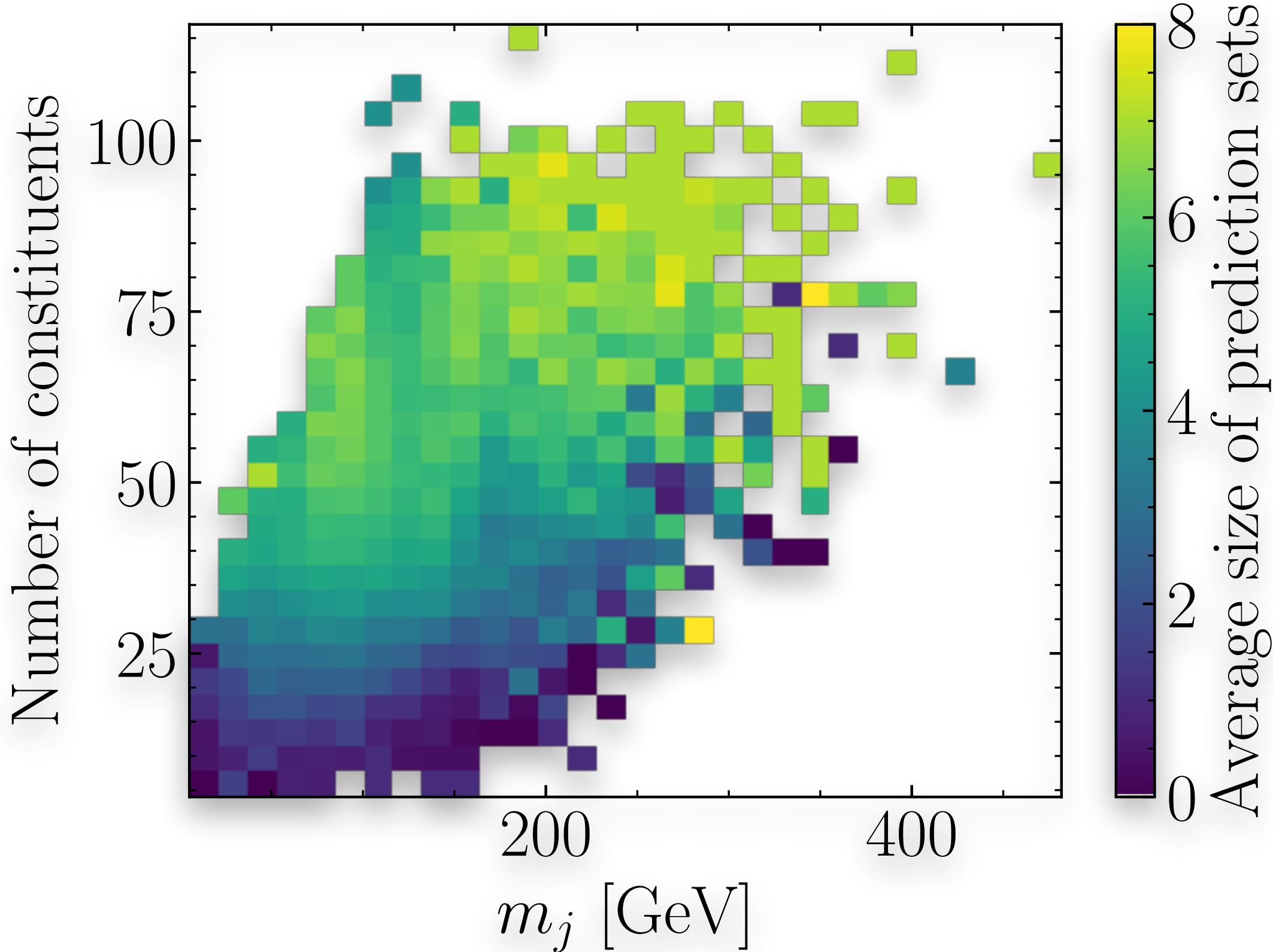


CP for Classification

Reliability



Interpretability

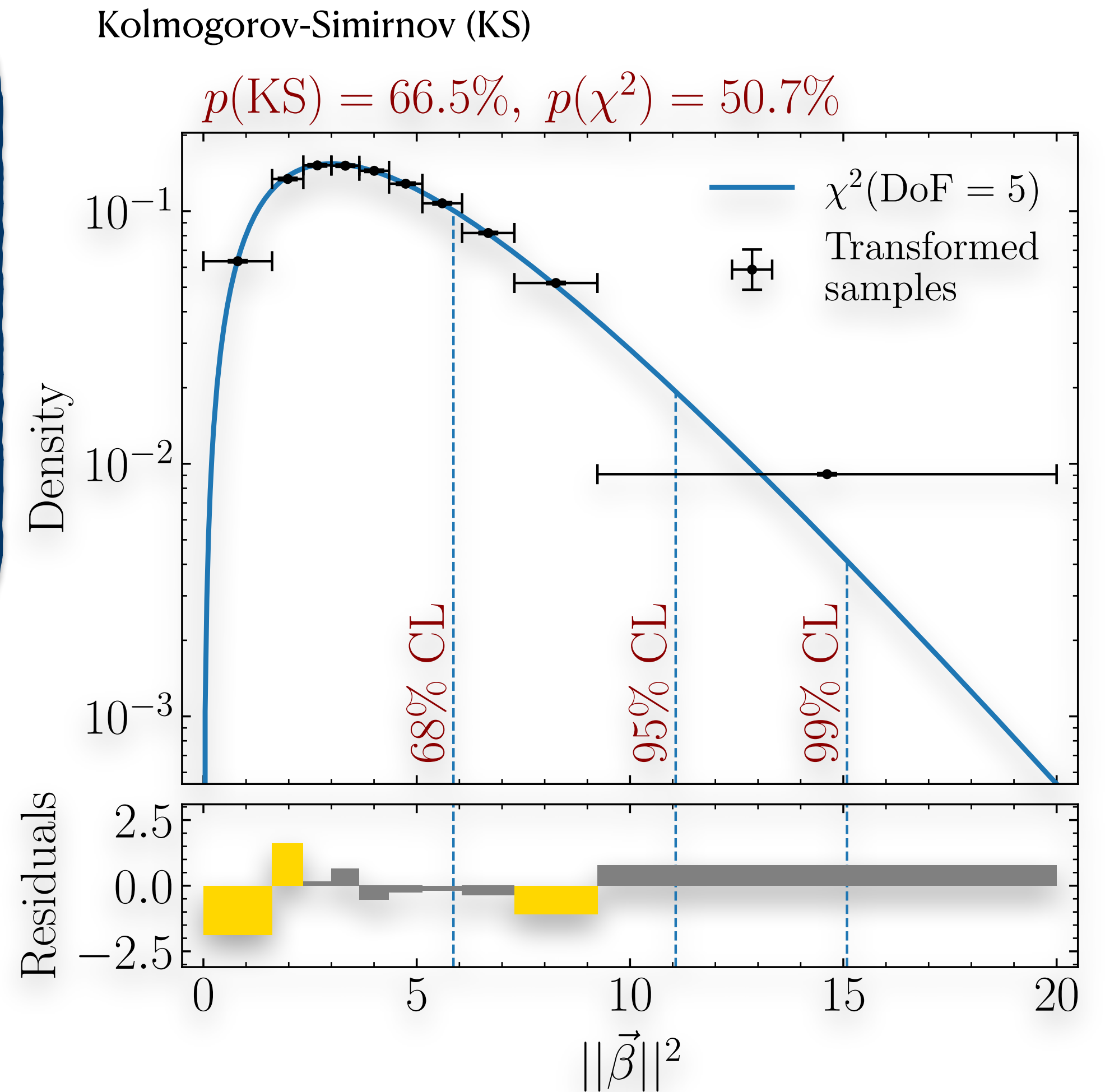
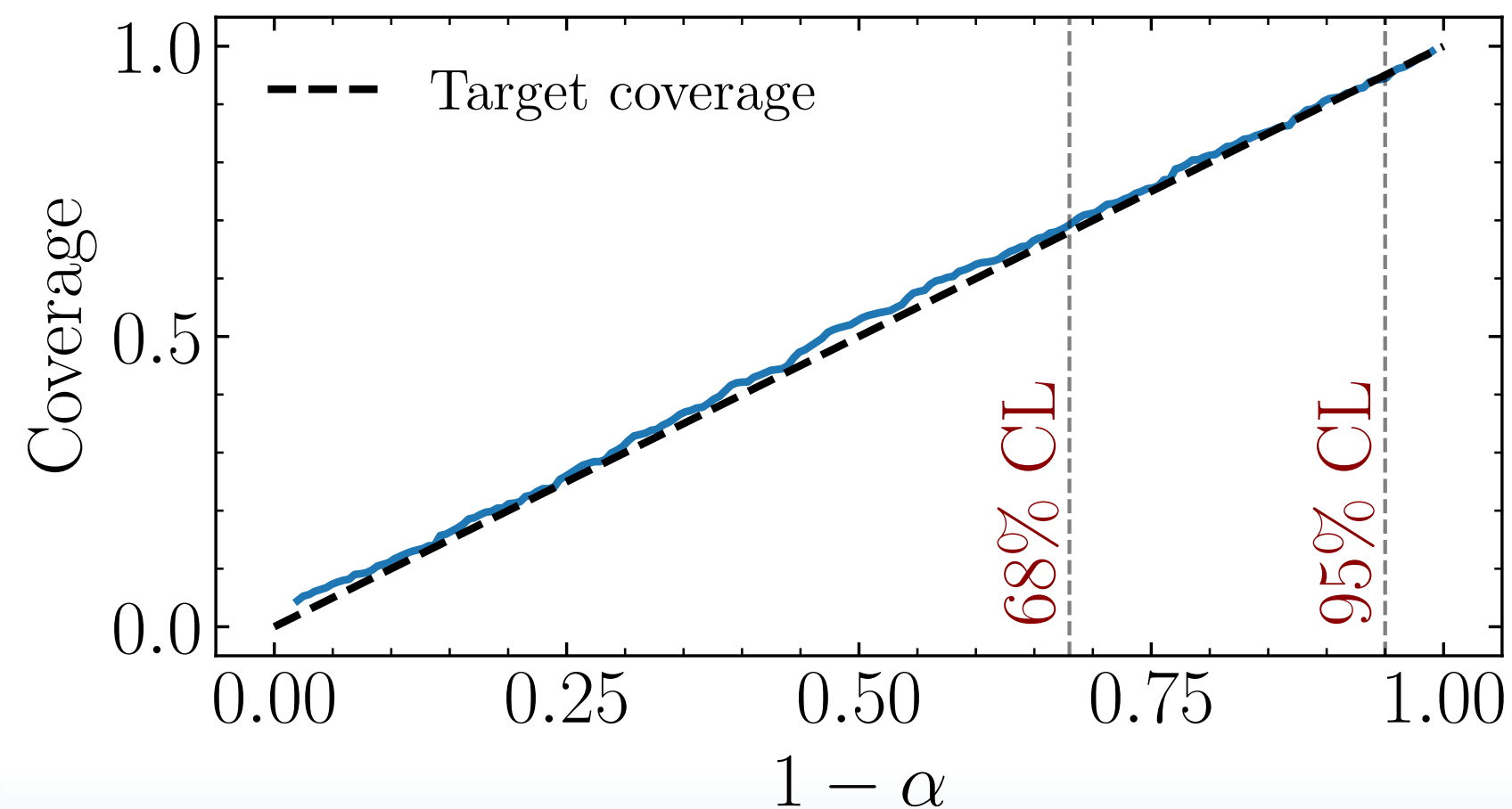


[JYA, Spannowsky; 2512.17048]

Conformal Prediction for Generative Modelling

CP for Generative Modelling

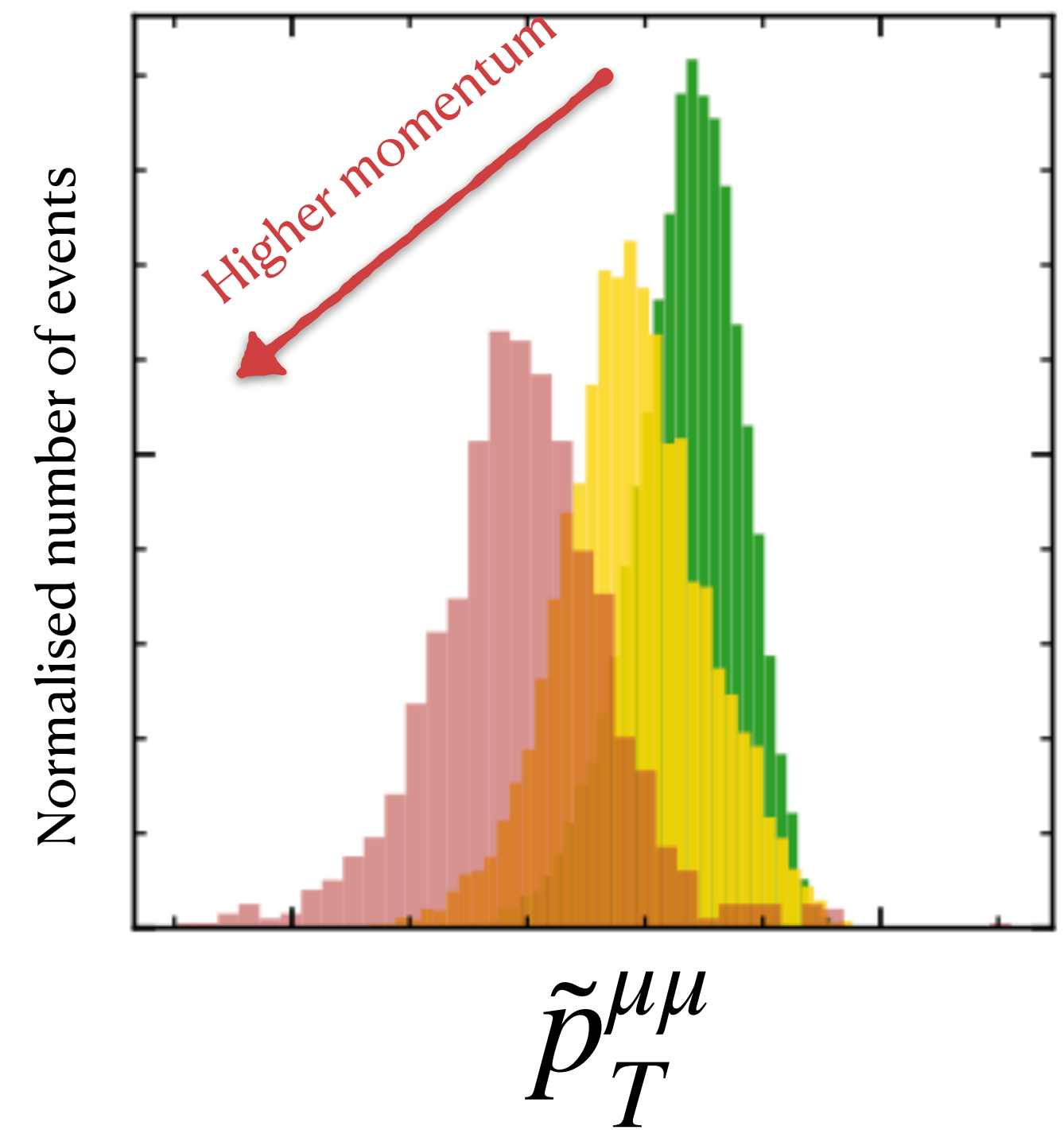
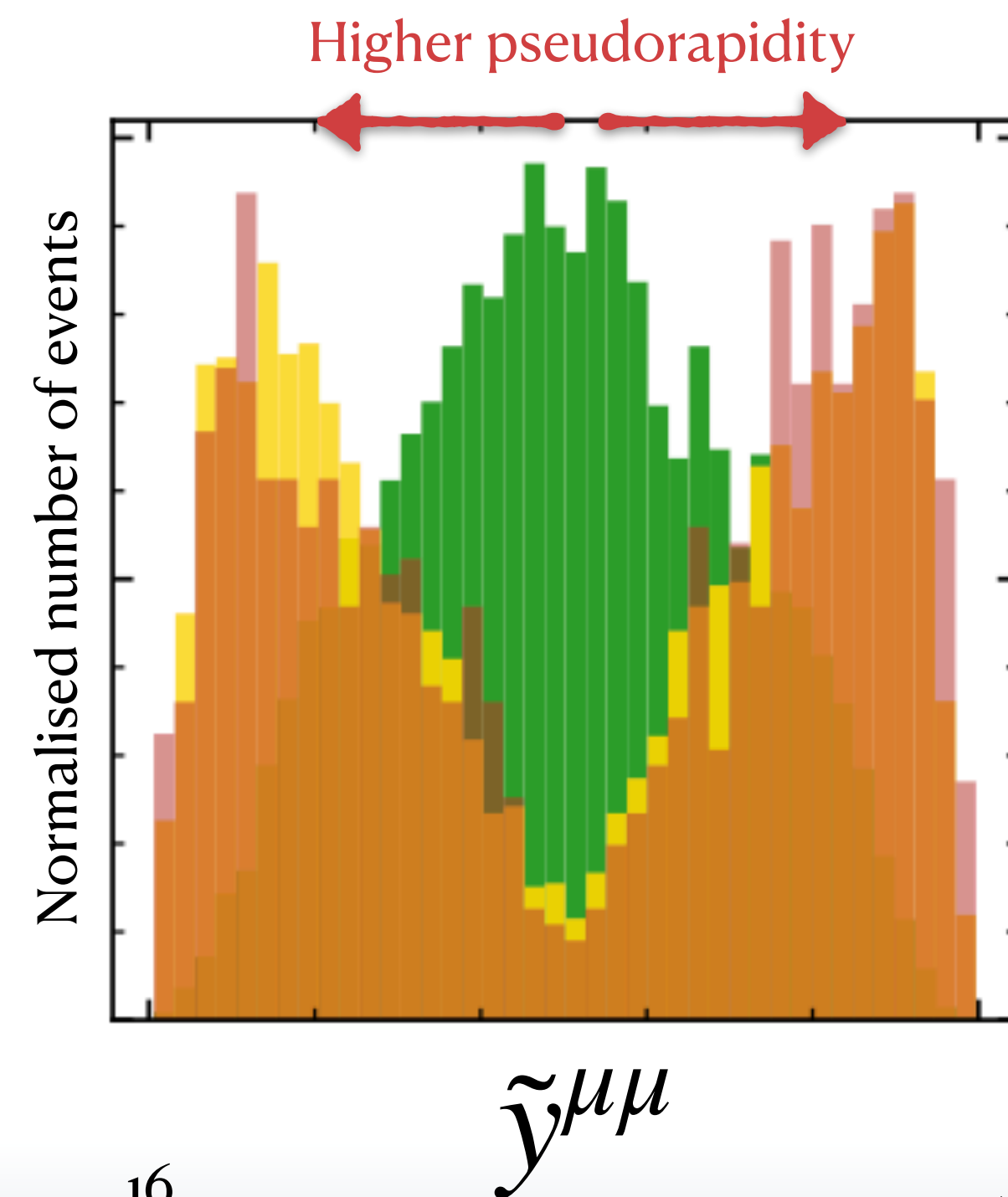
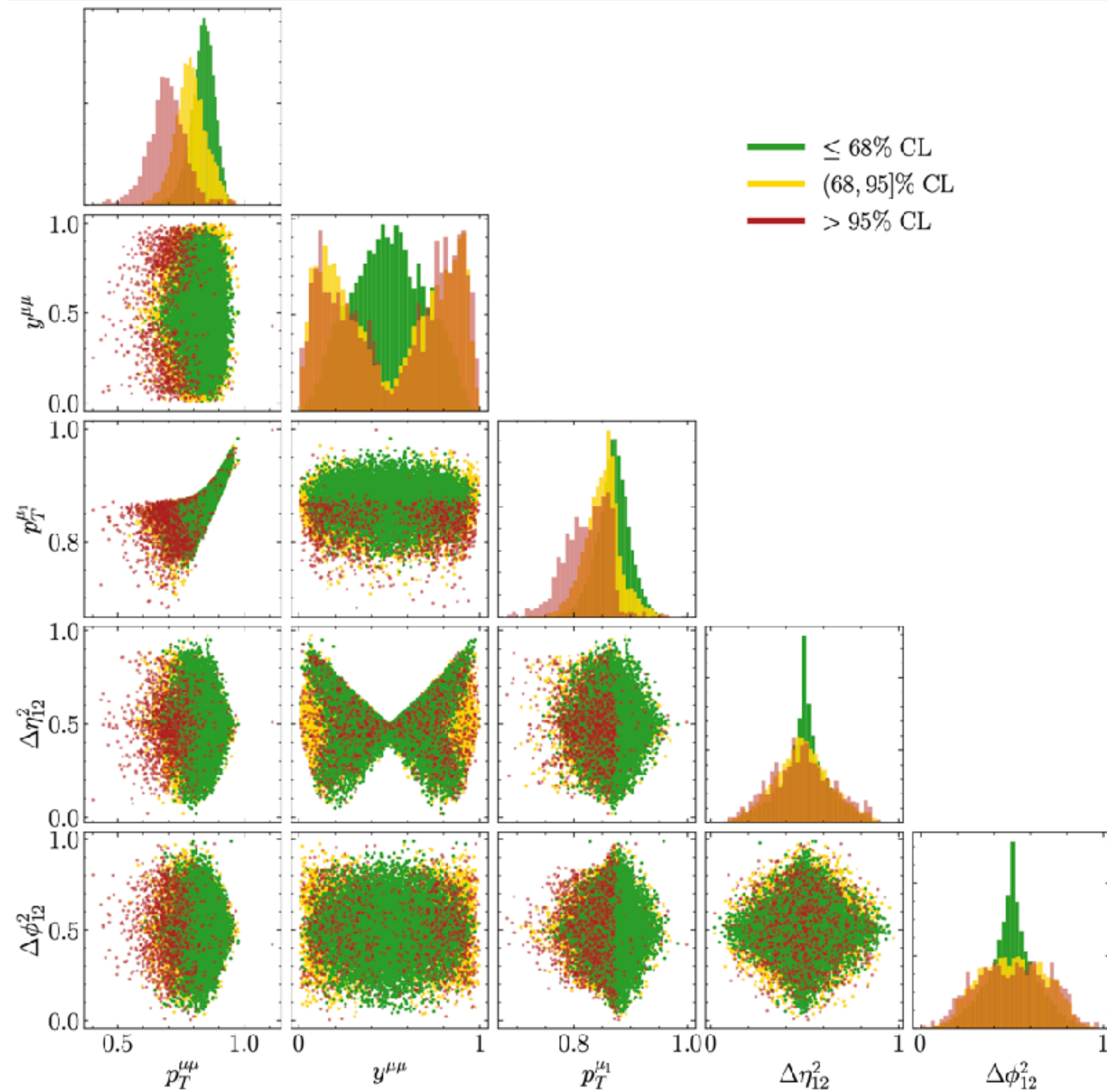
- ❖ ATLAS OmniFold 24-Dimensional $Z \rightarrow \mu\bar{\mu} + \text{jets}$ Open Data
- ❖ Dominant kinematic structure of the underlying $2 \rightarrow 3$ scattering: $\{p_T^{\mu\mu}, y^{\mu\mu}, p_T^{\mu_1}, \Delta\eta_{\mu\mu}, \Delta\phi_{\mu\mu}\}$
- ❖ Score: negative log-probability



[JYA, Beck, Reboud, Dyk, Spannowsky; EPJC '26]

CP for Generative Modelling

[JYA, Spannowsky; 2512.17048]

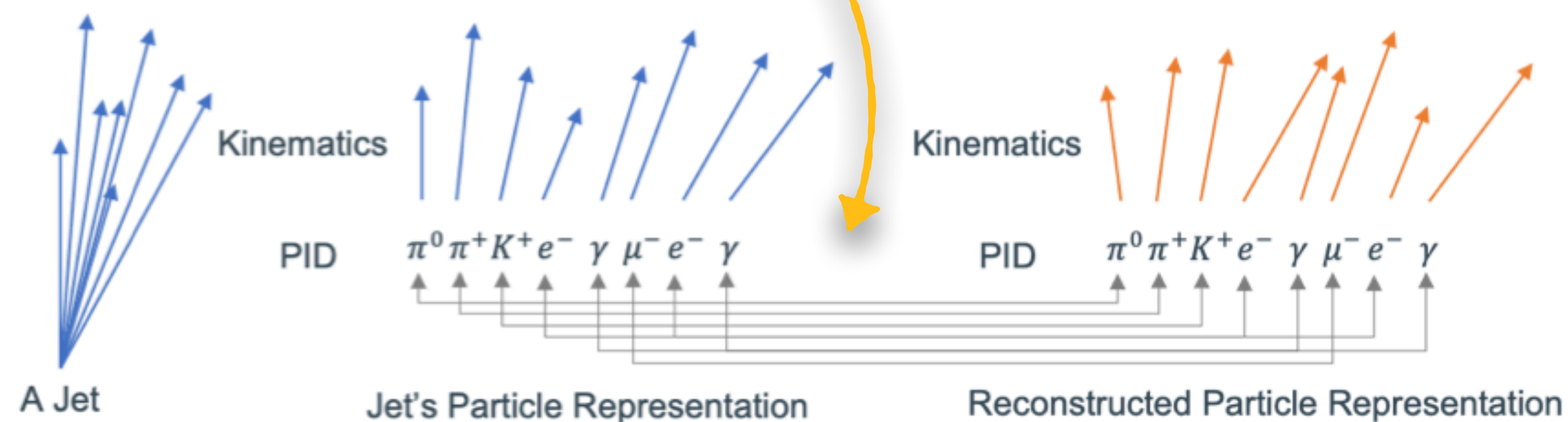


Conformal Prediction for Anomaly Detection

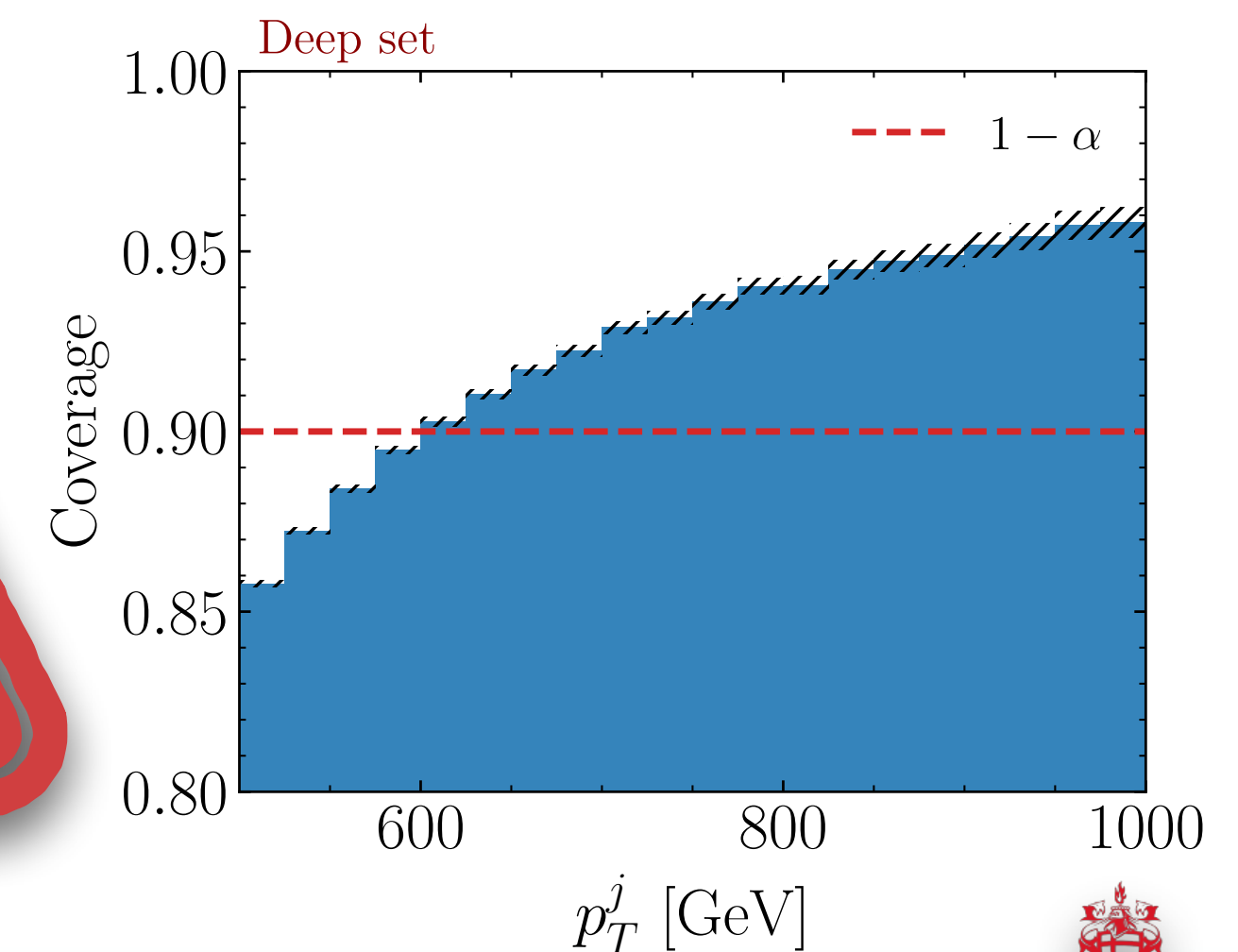
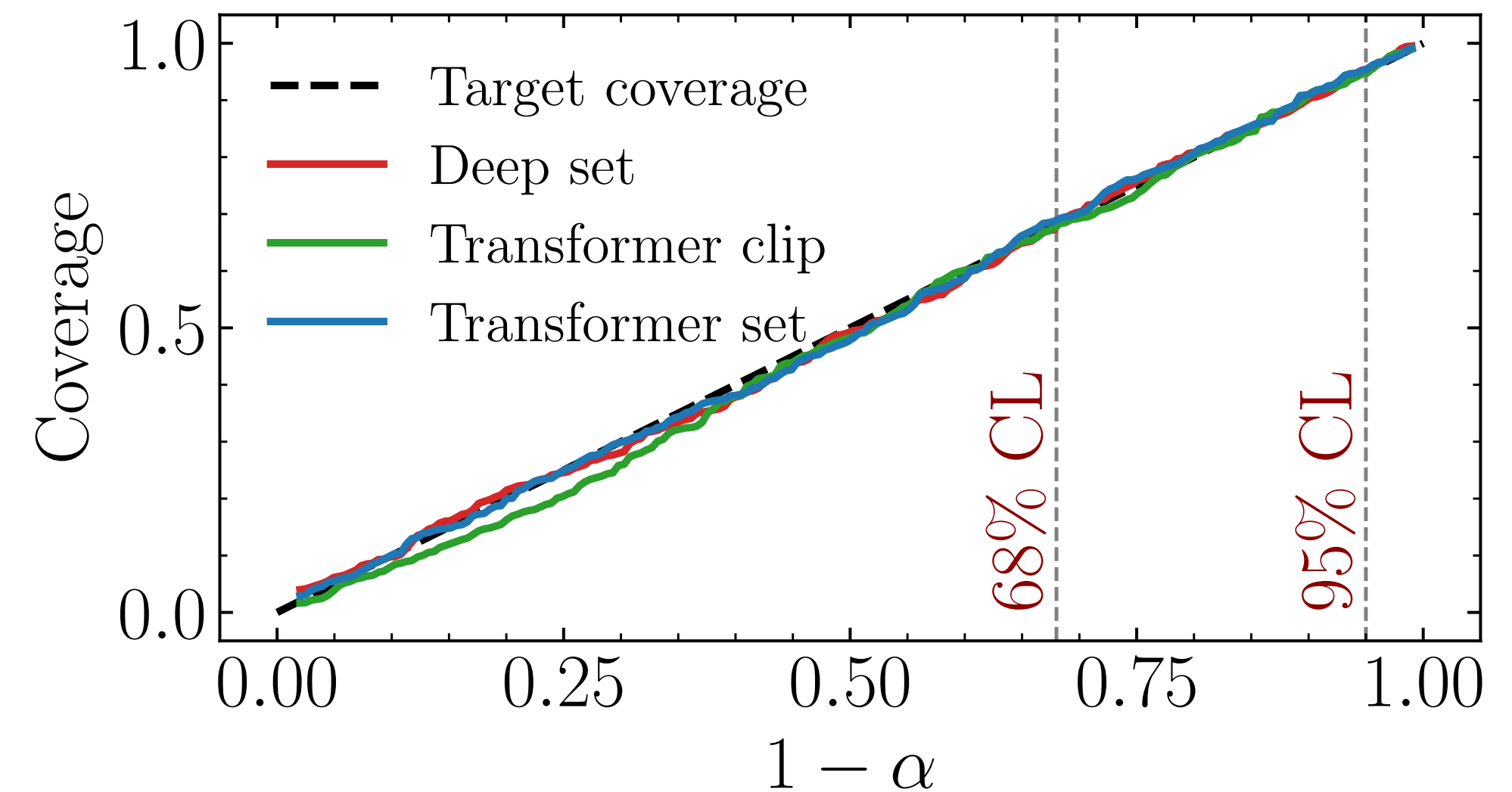
CP for Anomaly Detection

[JYA, Spannowsky; 2512.17048]

- ❖ Particle cloud-based VAE
- ❖ Trained only for the QCD dataset from JetClass
- ❖ Score: Chamfer loss



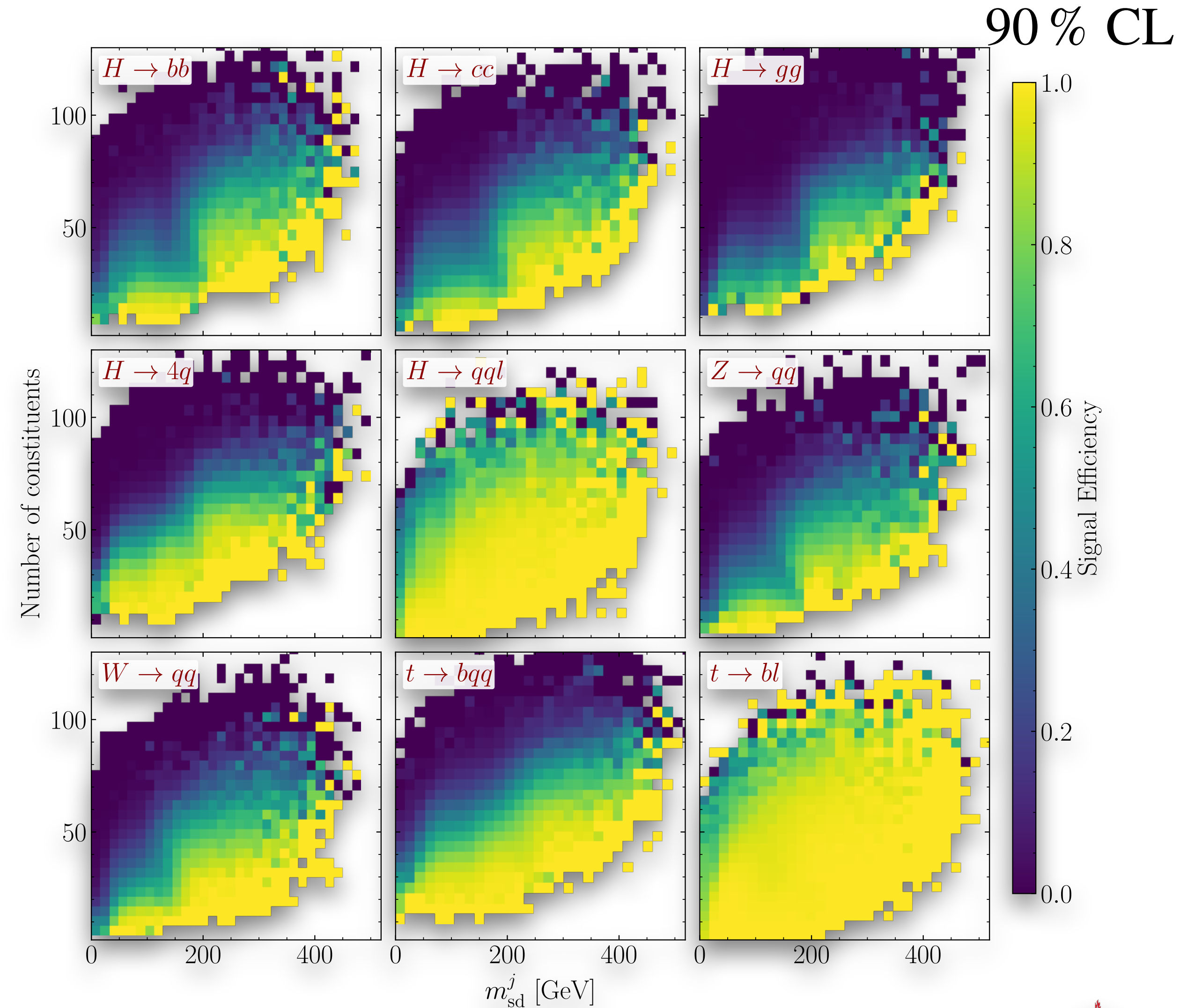
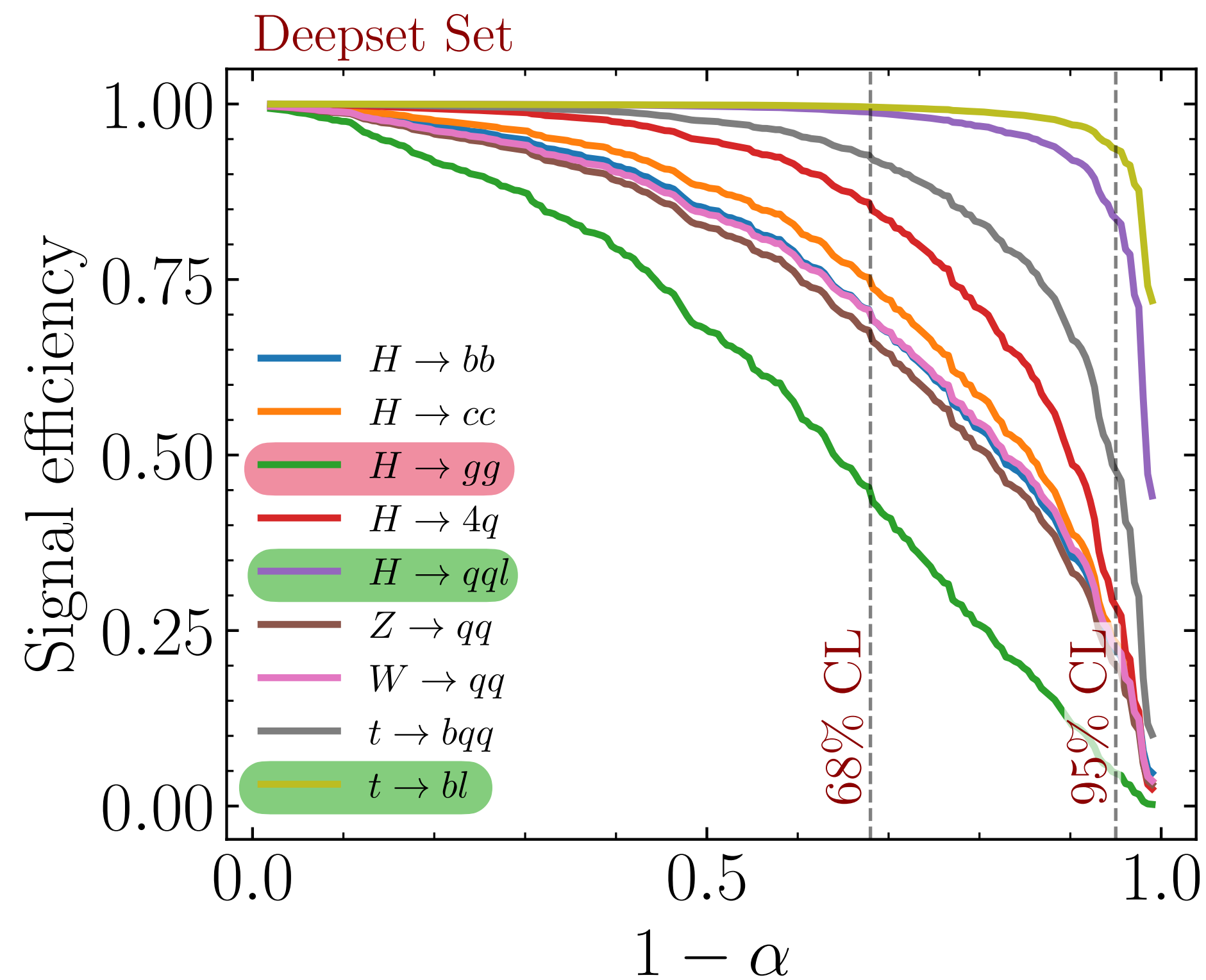
[Liu et. al.; 2311.17162]



CP for Anomaly Detection

[JYA, Spannowsky; 2512.17048]

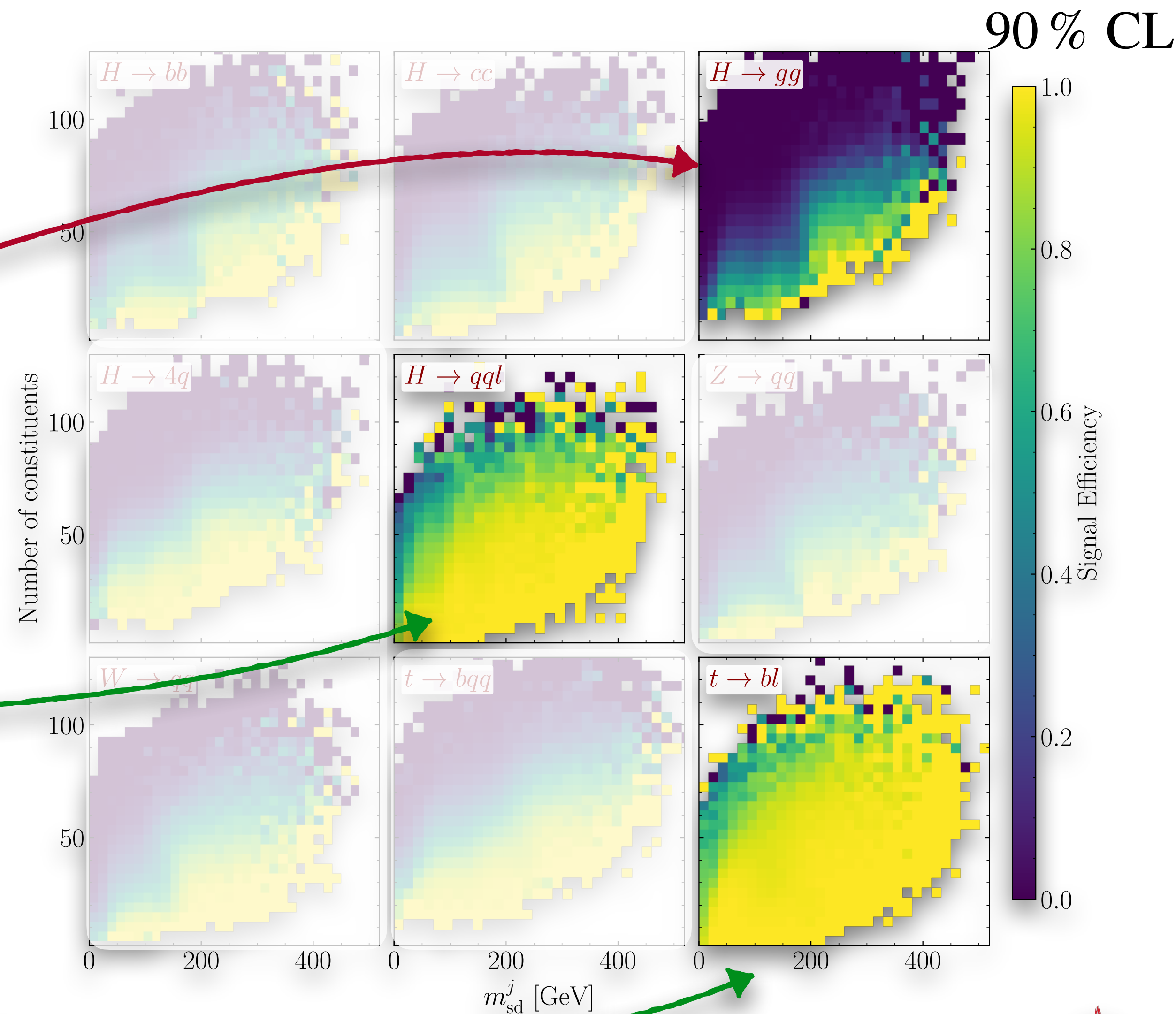
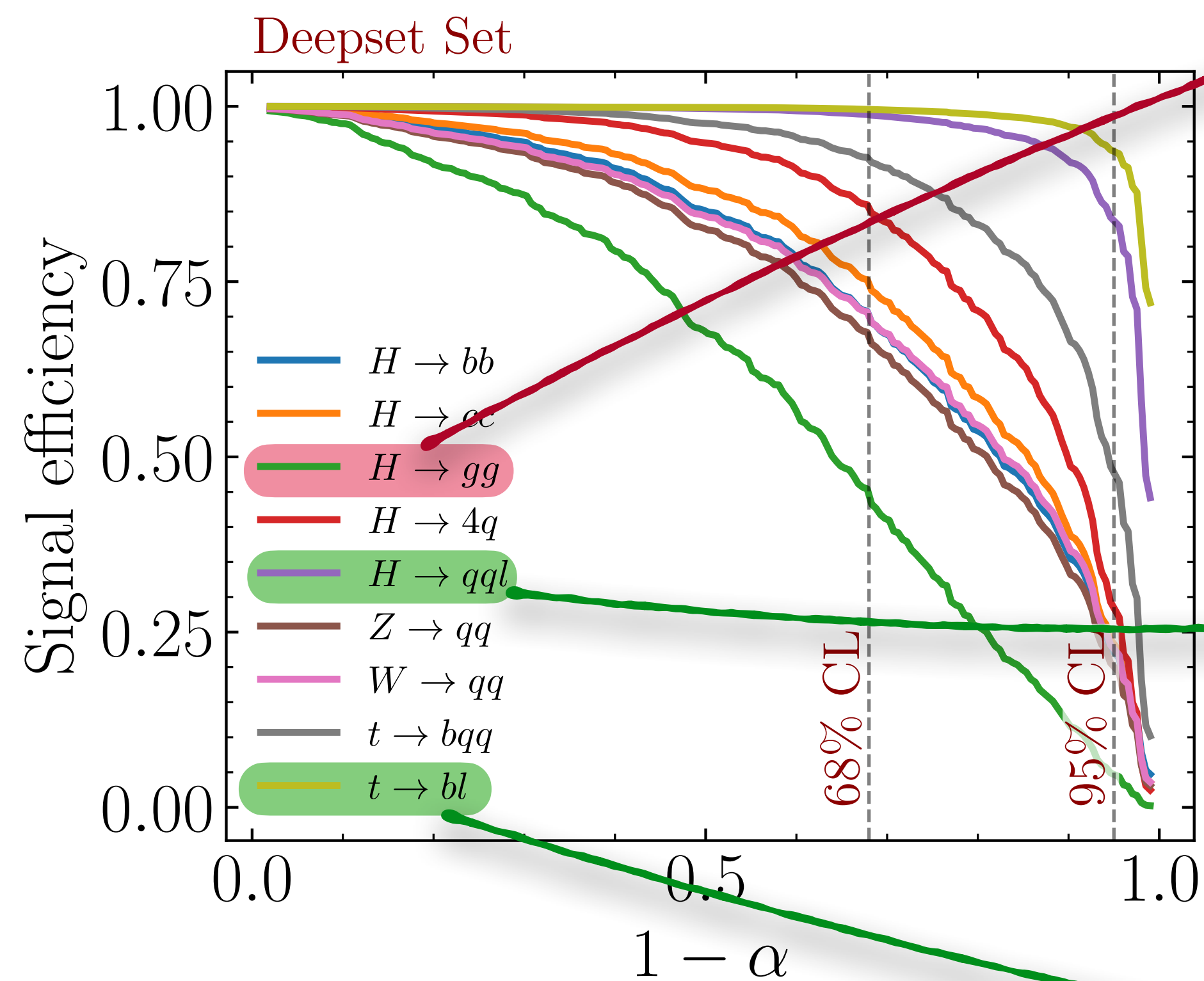
$$\varepsilon_{\text{sig}}(1 - \alpha) = \mathbb{P}_{X \sim P_{\text{sig}}}(s(X) > \hat{q}_{1-\alpha})$$



CP for Anomaly Detection

[JYA, Spannowsky; 2512.17048]

$$\varepsilon_{\text{sig}}(1 - \alpha) = \mathbb{P}_{X \sim P_{\text{sig}}} (s(X) > \hat{q}_{1-\alpha})$$

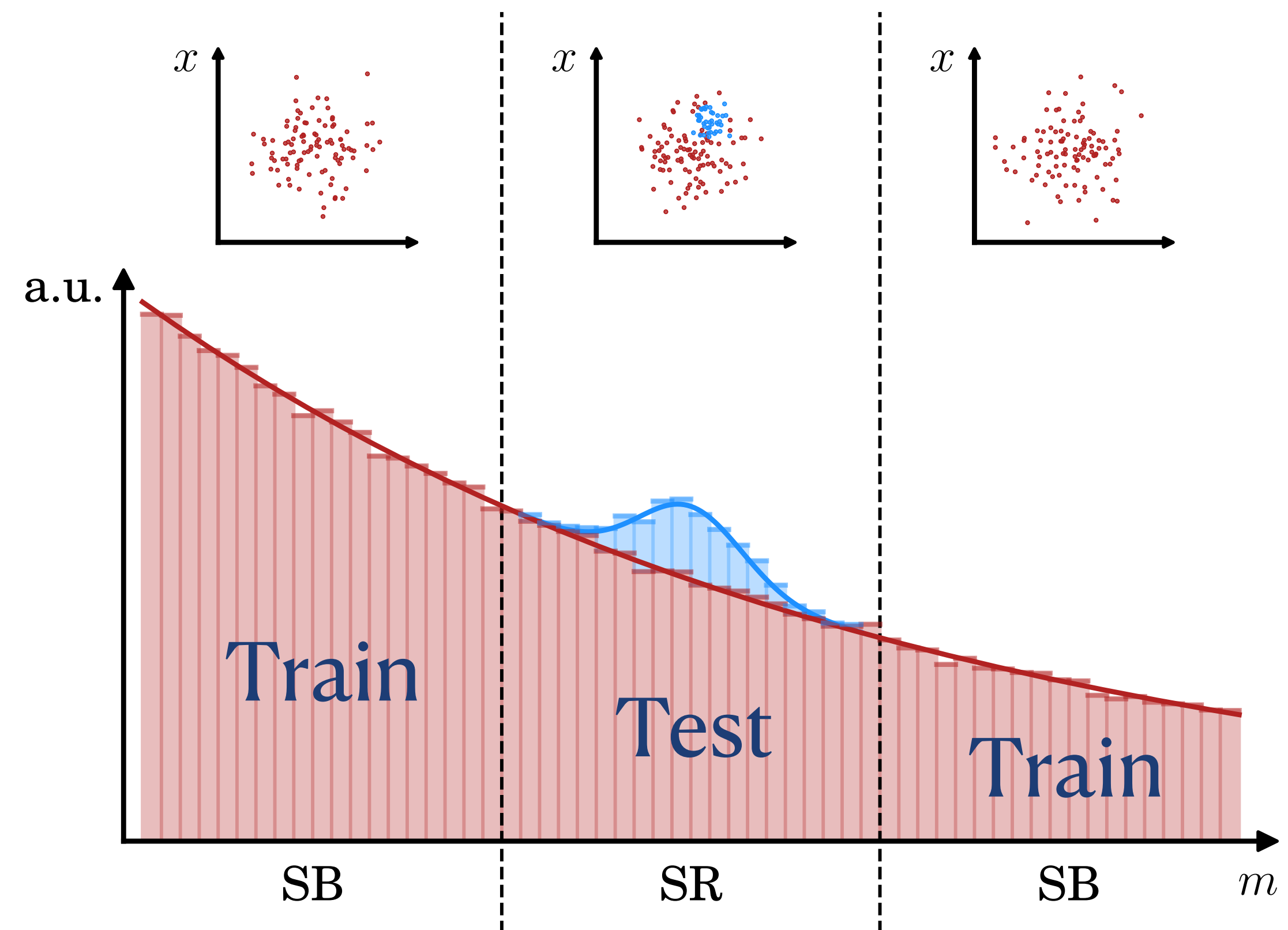


Conformal inference for anomaly detection

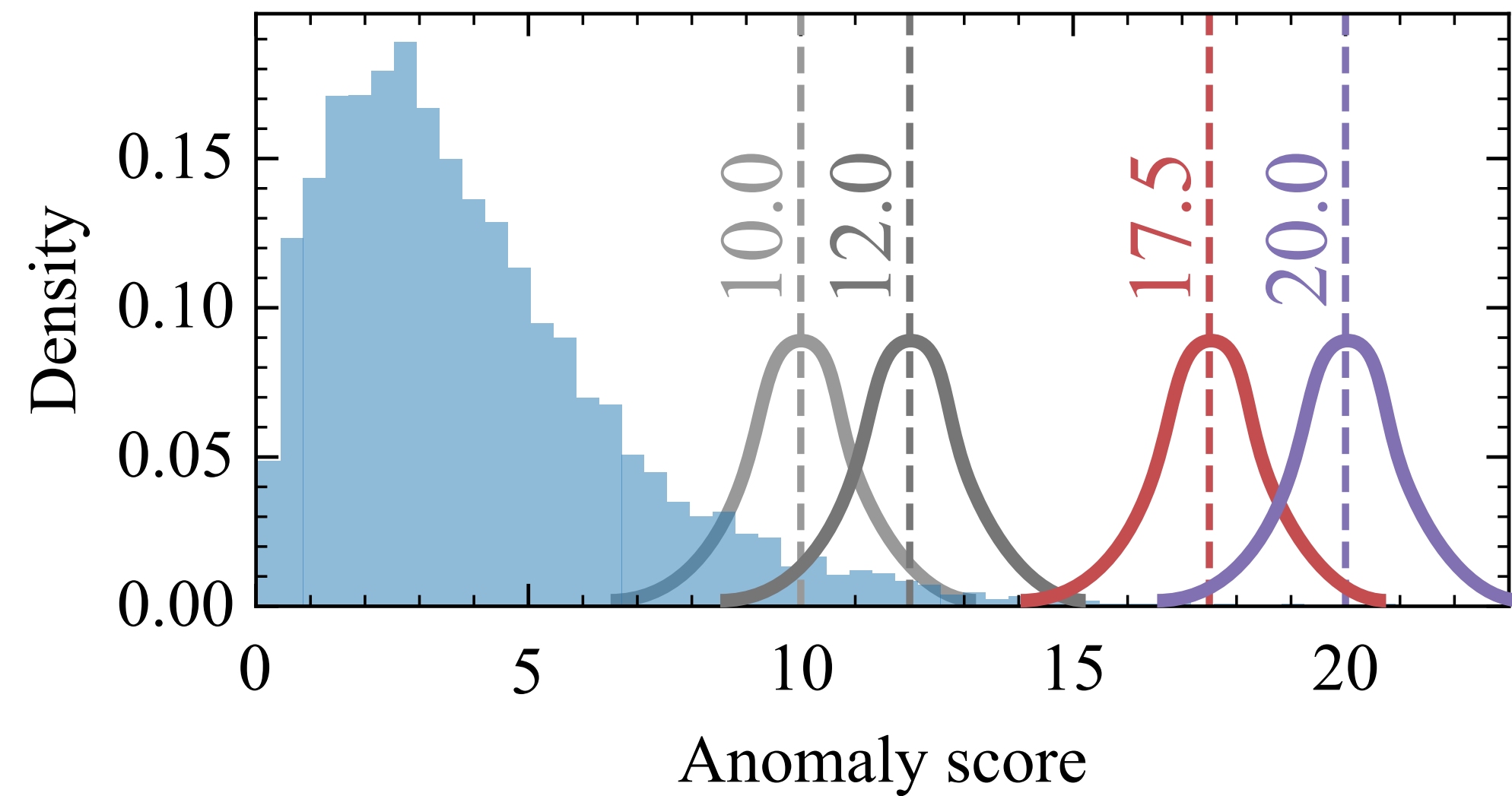
Bump hunt

- ❖ Create control regions that “ideally” do not have any signal
- ❖ Estimate the background template and interpolate to the signal region.
- ❖ Compute the statistical deviation with respect to the background template.
- ❖ Strong dependency on template estimation.
- ❖ Strong dependency on signal leakage in the control region

Image from arXiv: 2109.00546



What if we discover a new particle with AI?

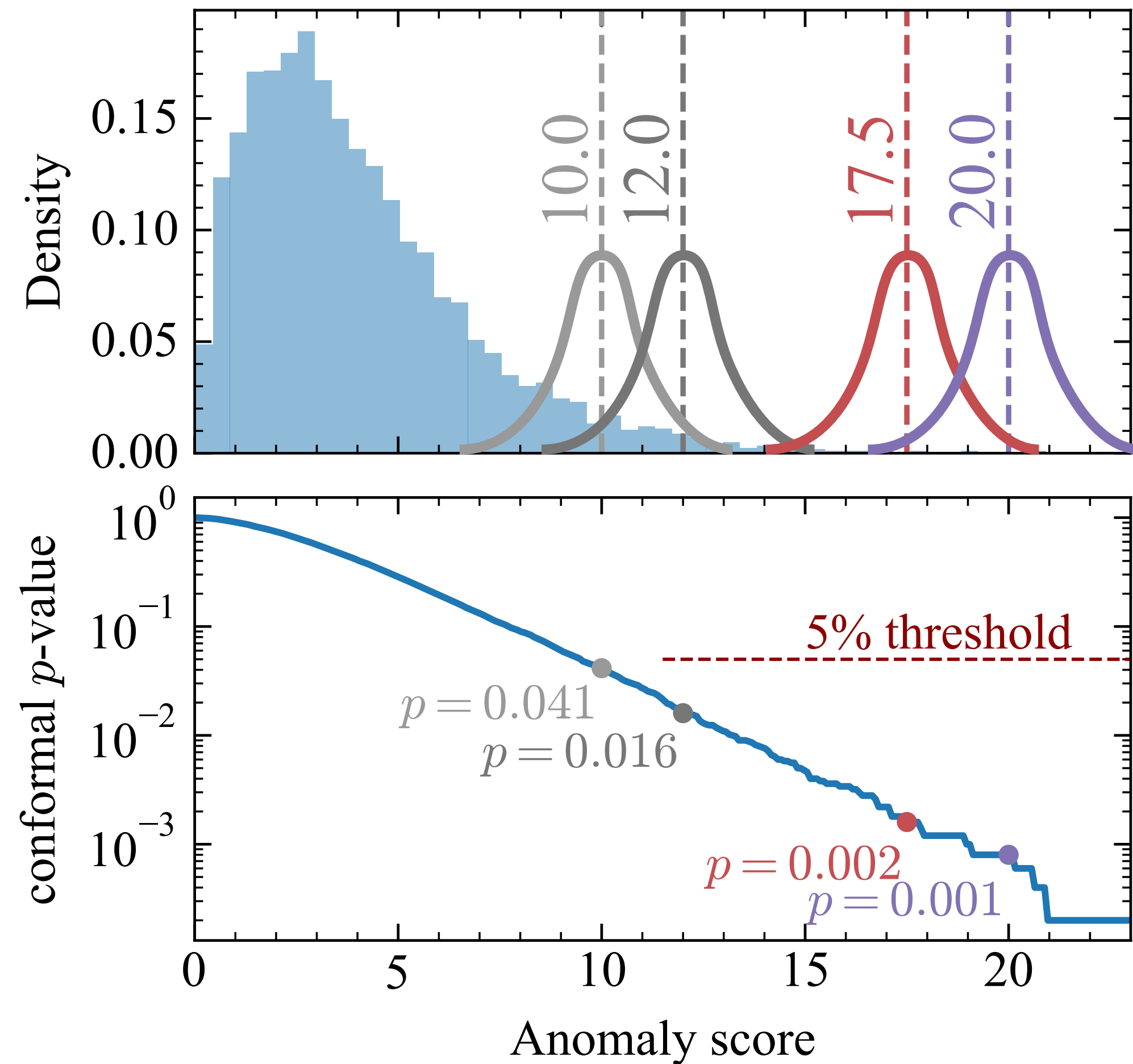


Raw anomaly score
has no meaning!

[JYA, Spannowsky; 2606.13780]

What if we discover a new particle with AI?

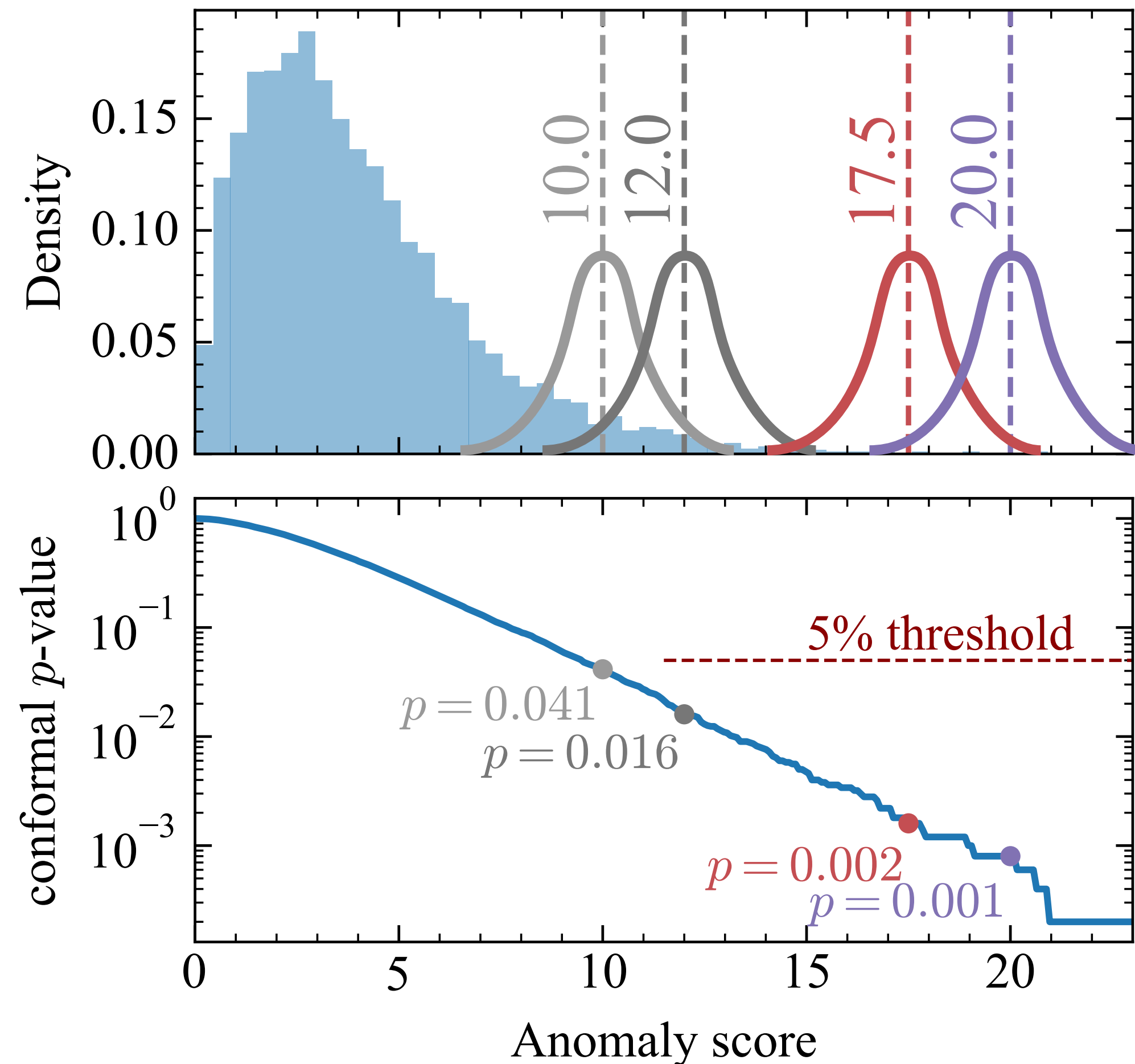
[JYA, Spannowsky; 2606.13780]



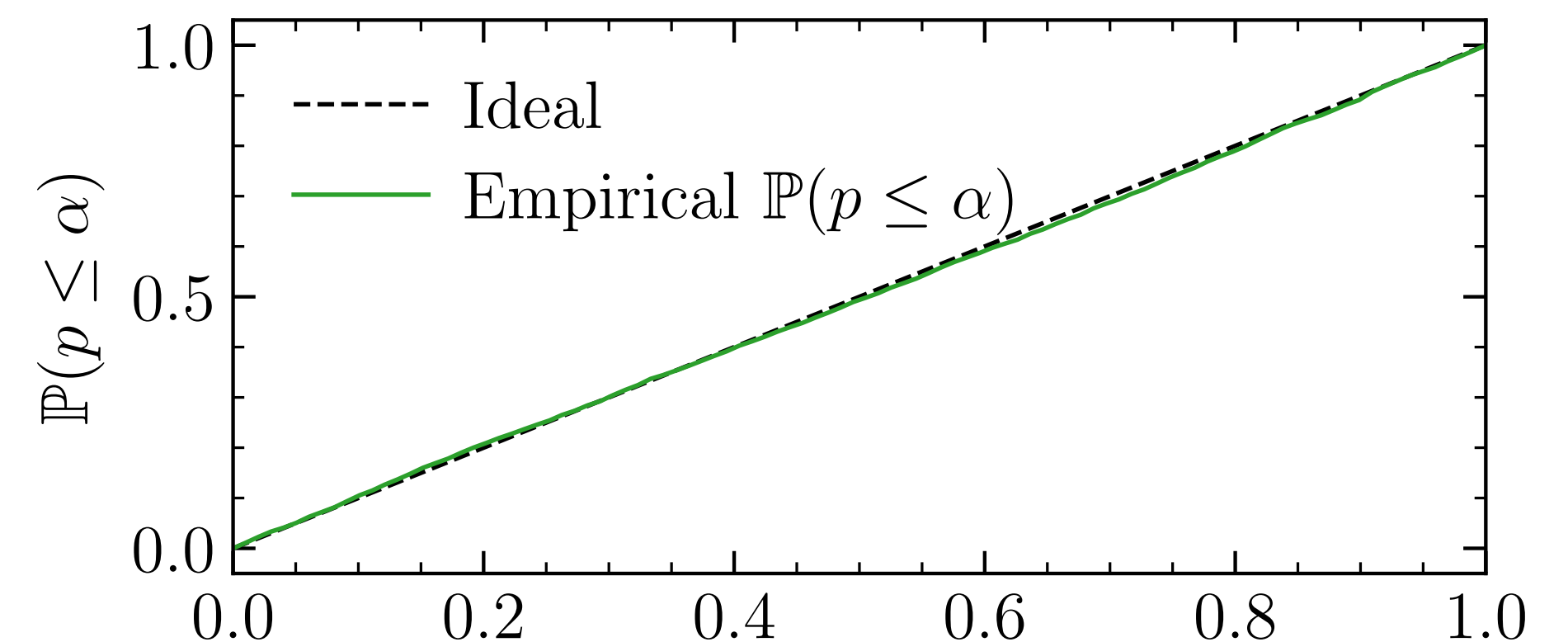
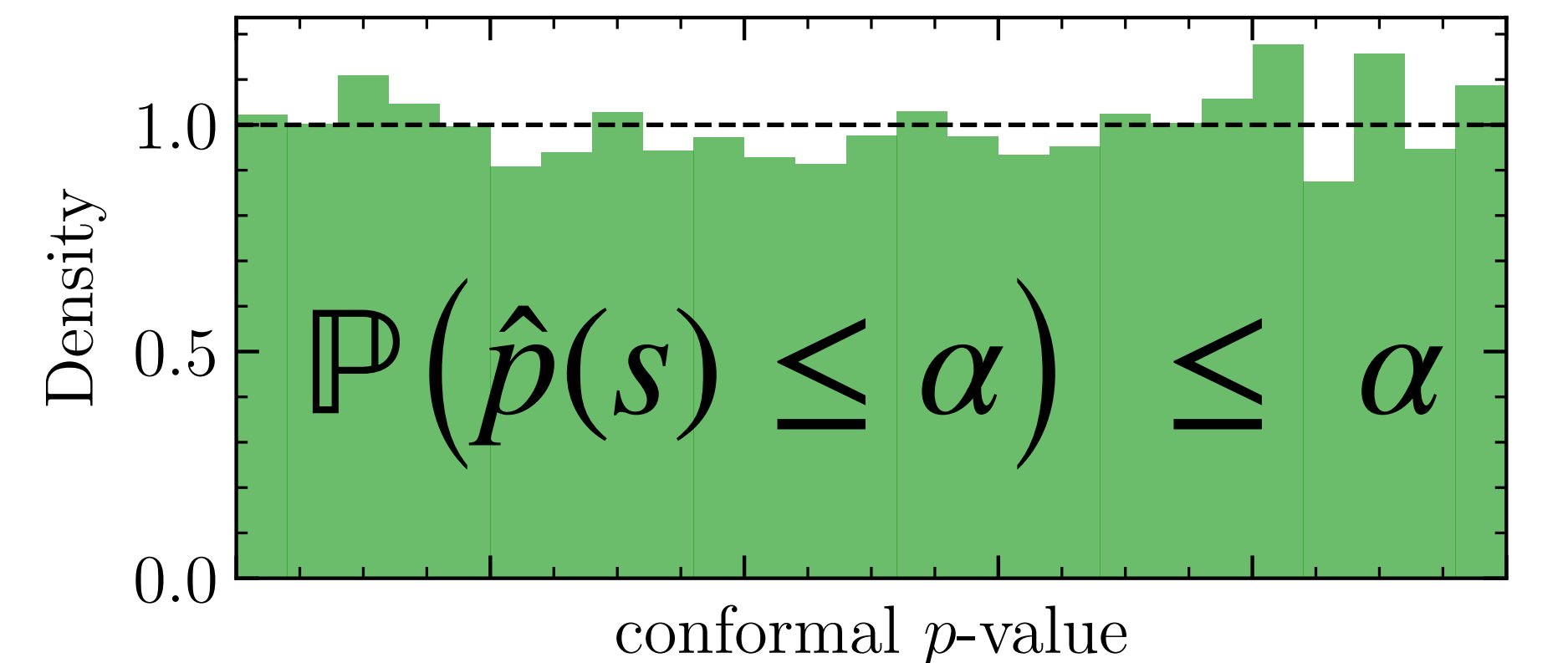
$$\hat{p}(s) = \frac{1 + \left| \{ i : s_i \geq s \} \right|}{n + 1}$$

What if we discover a new particle with AI?

[JYA, Spannowsky; 2606.13780]



$$\hat{p}(s) = \frac{1 + \left| \{ i : s_i \geq s \} \right|}{n + 1}$$



Jack Y. Araz^α

The issue of exchangeability

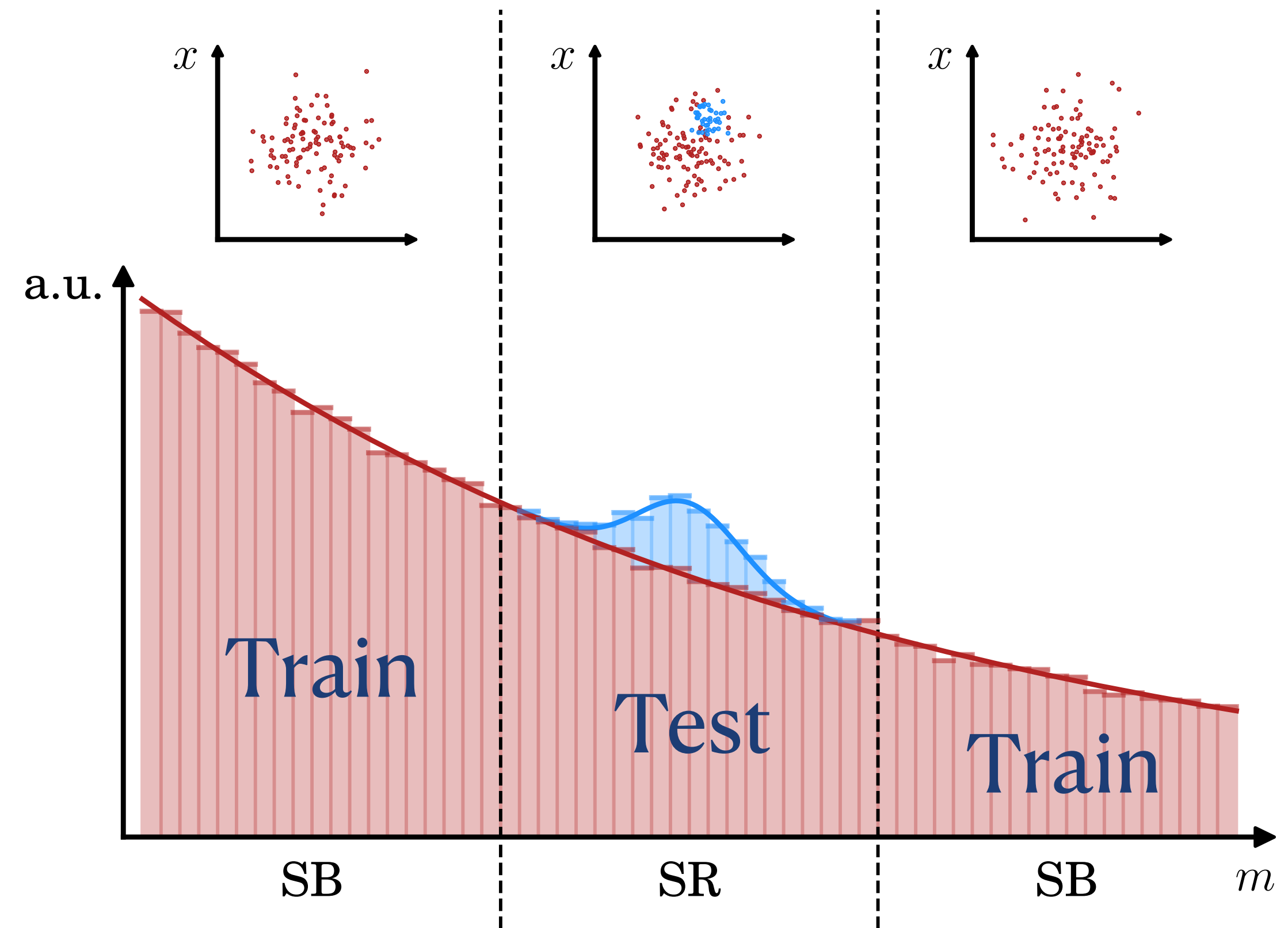
$$\hat{p}(s) = \frac{1 + \left| \{ i : s_i \geq s \} \right|}{n + 1}$$

$p_{\text{SR}}(s) \neq p_{\text{SB}}(s) \rightarrow$ not exchangeable

$$\hat{p}(s) = \frac{w(x^{\text{test}}) + \sum_{i=1}^n w(x_i^{\text{cal}}) \mathbf{1}\{s_i \geq s\}}{w(x^{\text{test}}) + \sum_{i=1}^n w(x_i^{\text{cal}})}$$

$$w(s) = p_{\text{SR}}(s)/p_{\text{SB}}(s)$$

Image from arXiv: 2109.00546



[JYA, Spannowsky; 2606.13780]

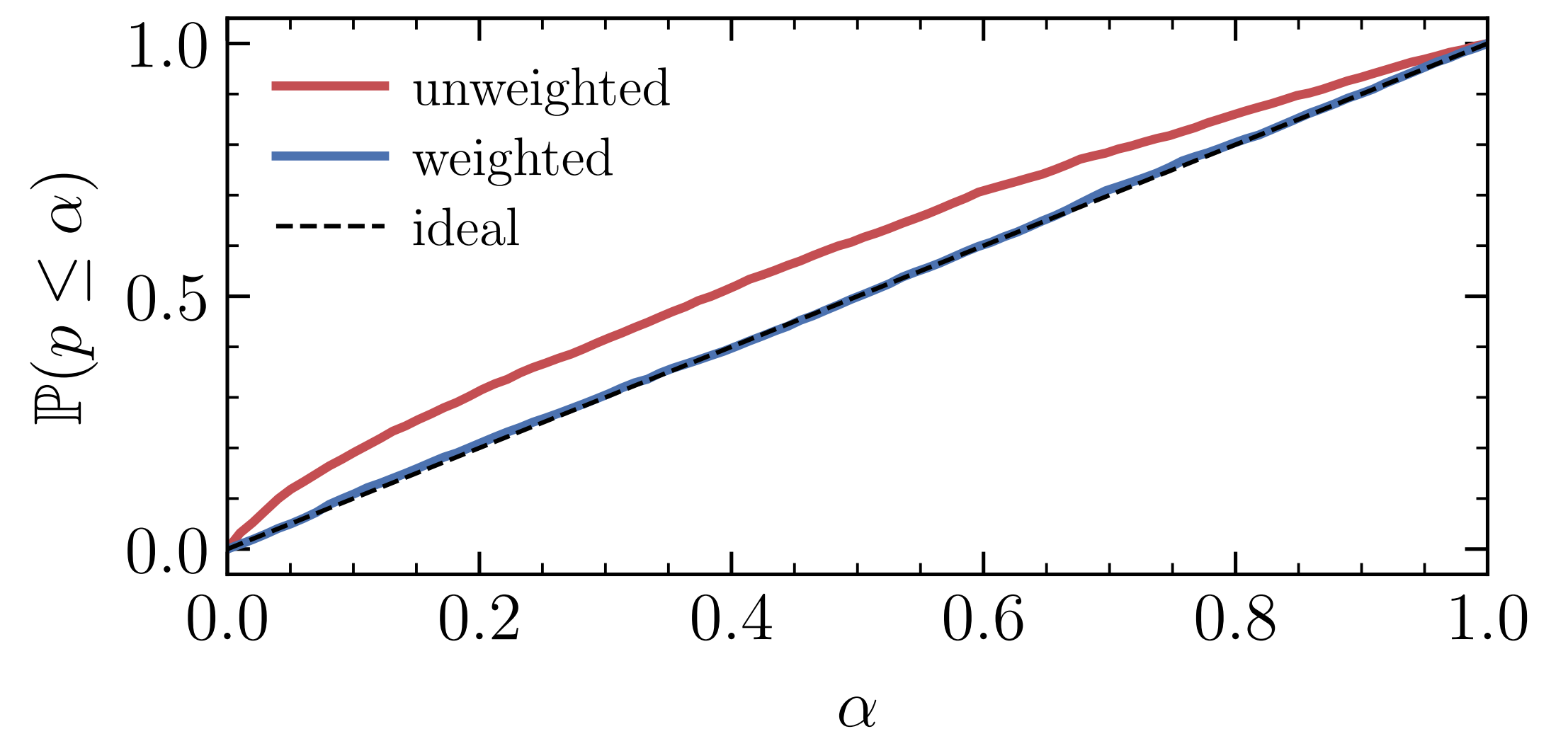
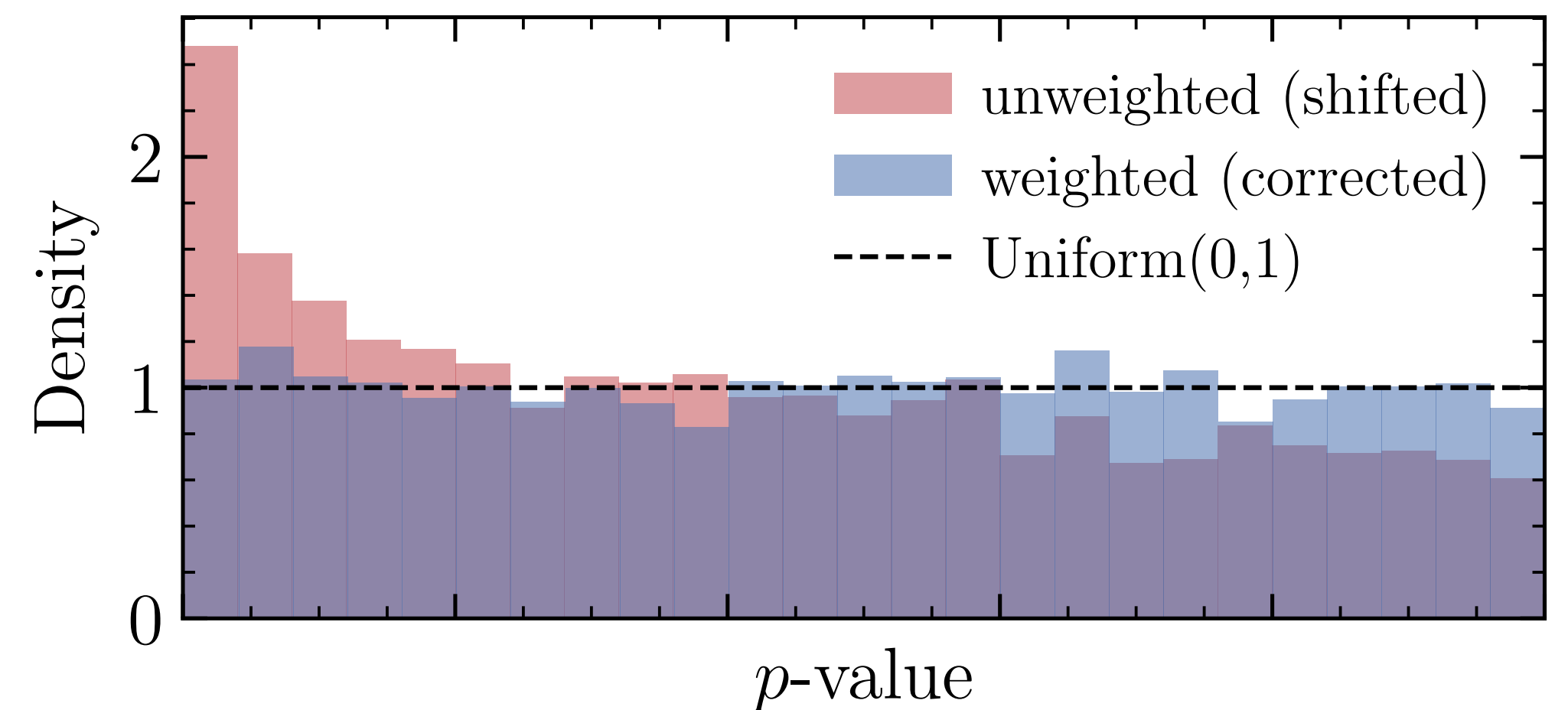
The issue of exchangeability

$$\hat{p}(s) = \frac{1 + \left| \{ i : s_i \geq s \} \right|}{n + 1}$$

$p_{\text{SR}}(s) \neq p_{\text{SB}}(s) \rightarrow$ not exchangeable

$$\hat{p}(s) = \frac{w(x^{\text{test}}) + \sum_{i=1}^n w(x_i^{\text{cal}}) \mathbf{1}\{s_i \geq s\}}{w(x^{\text{test}}) + \sum_{i=1}^n w(x_i^{\text{cal}})}$$

$$w(s) = p_{\text{SR}}(s)/p_{\text{SB}}(s)$$



[JYA, Spannowsky; 2606.13780]

The issue of marginal coverage

$$\mathbb{P}(\hat{p}_b(s) \leq \alpha \mid b) \leq \alpha$$

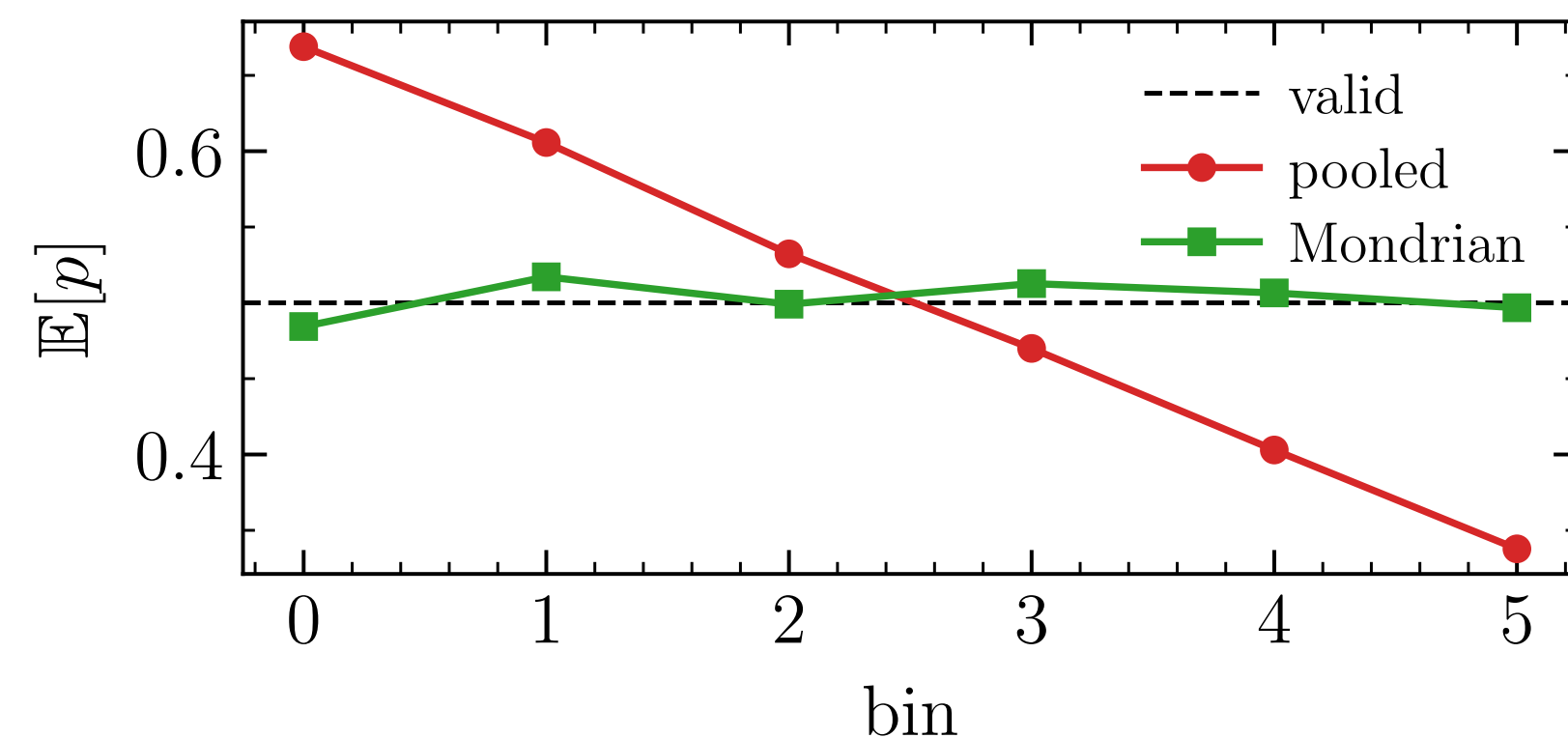
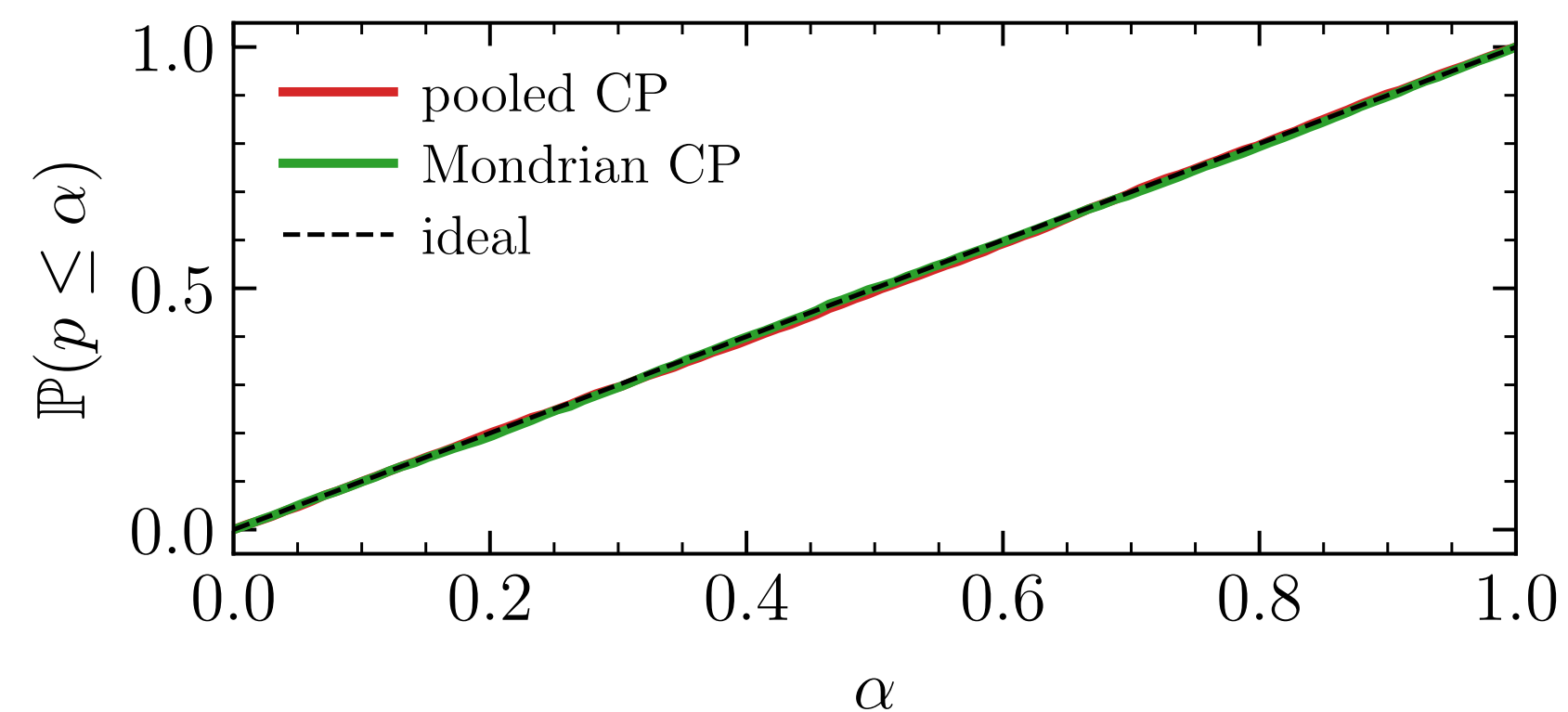
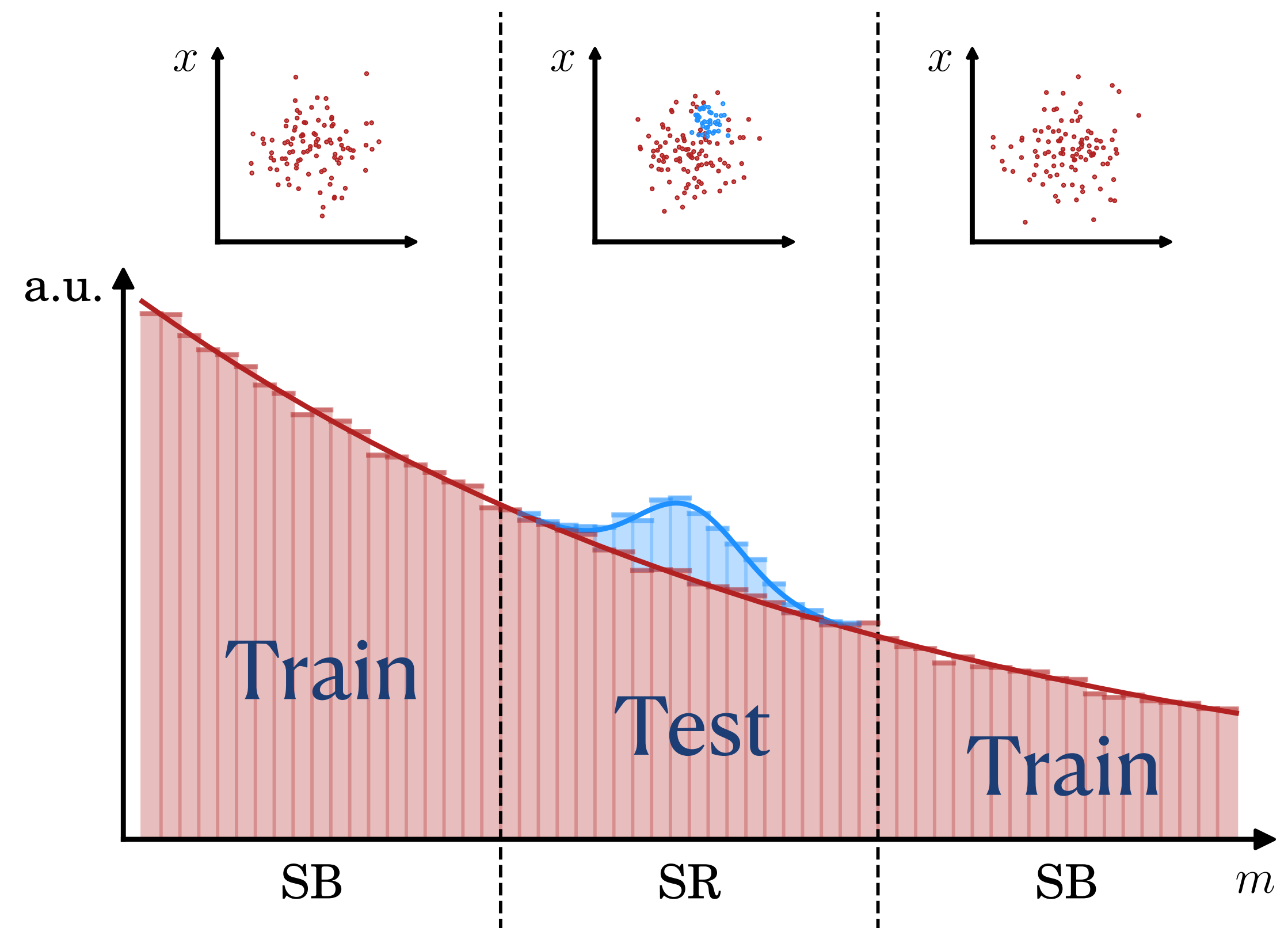
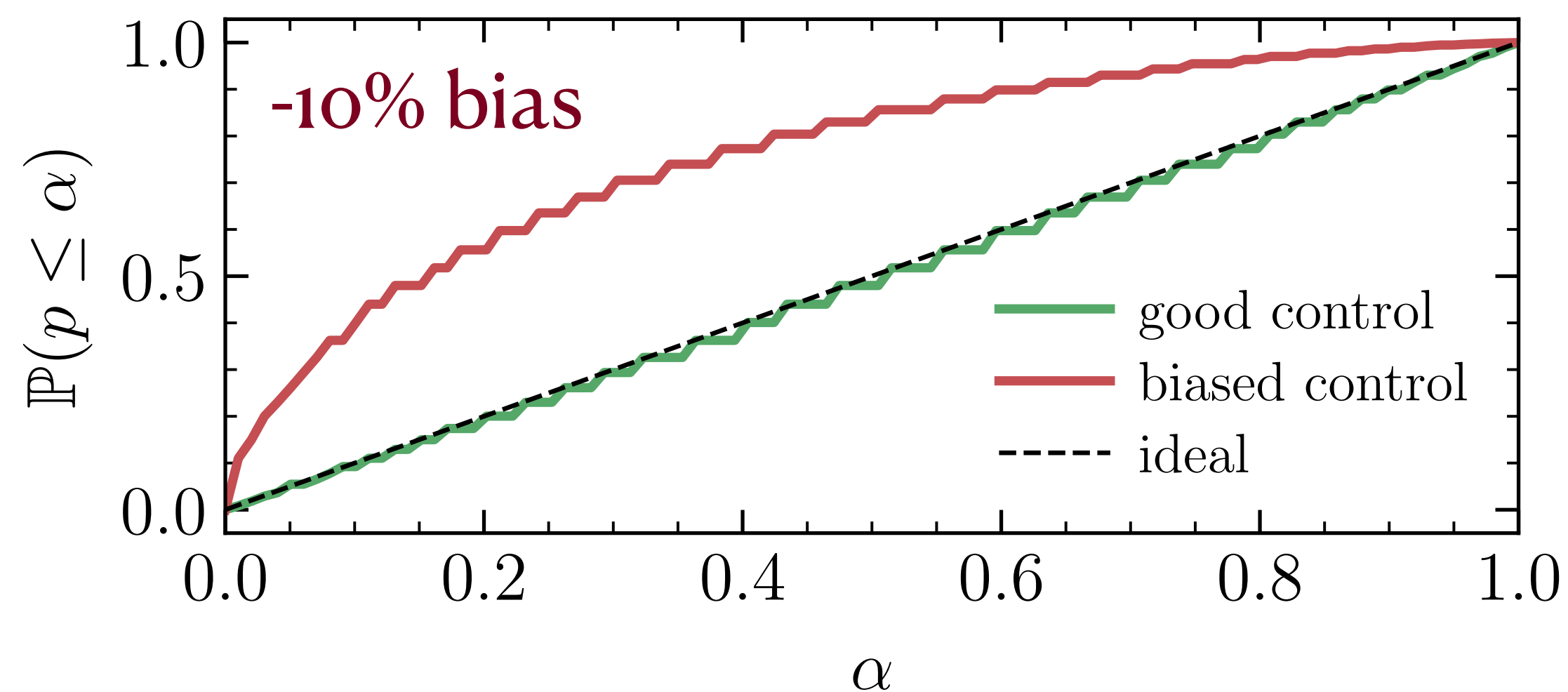


Image from arXiv: 2109.00546



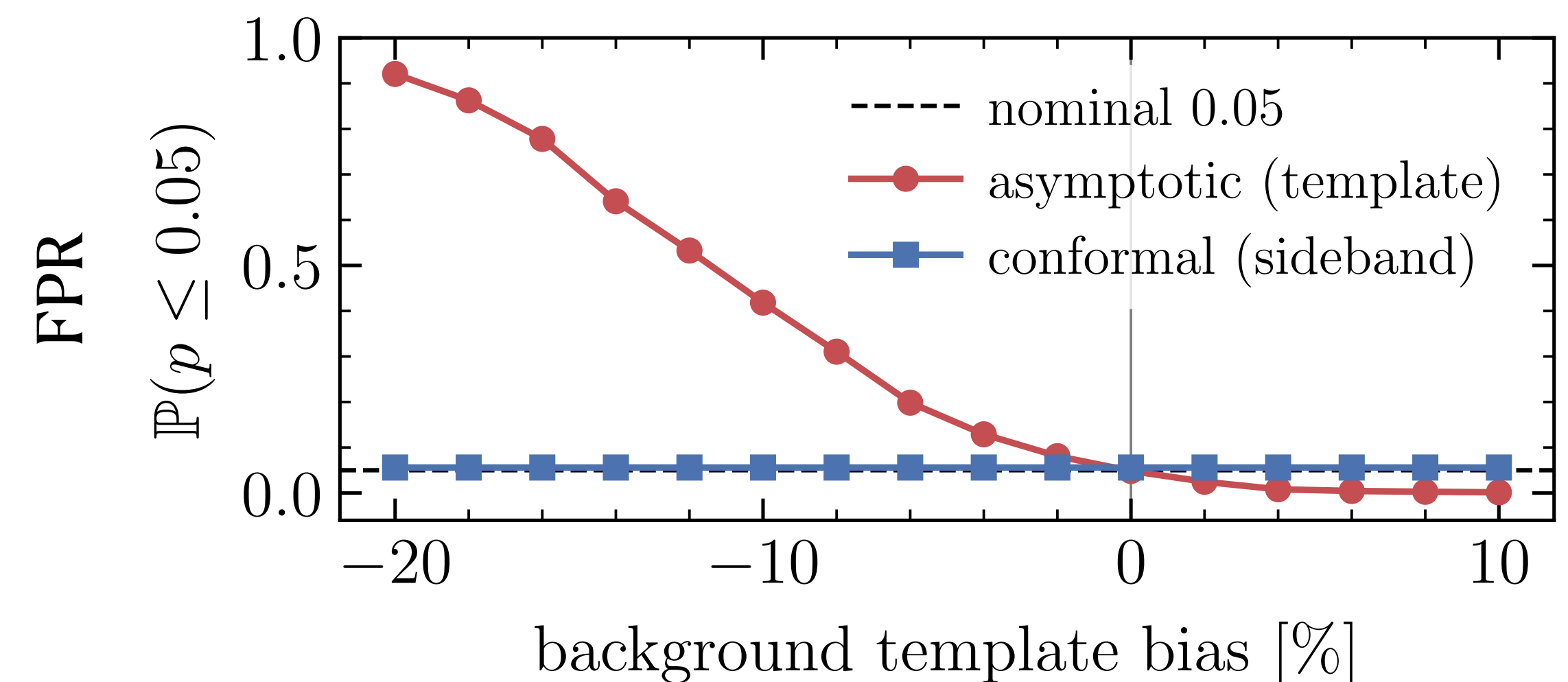
What does finite-sample validity add?

❖ At the LHC luminosities, Wilks' theorem is accurate up to a fraction of a σ , so isn't finite-sample validity a marginal gain?

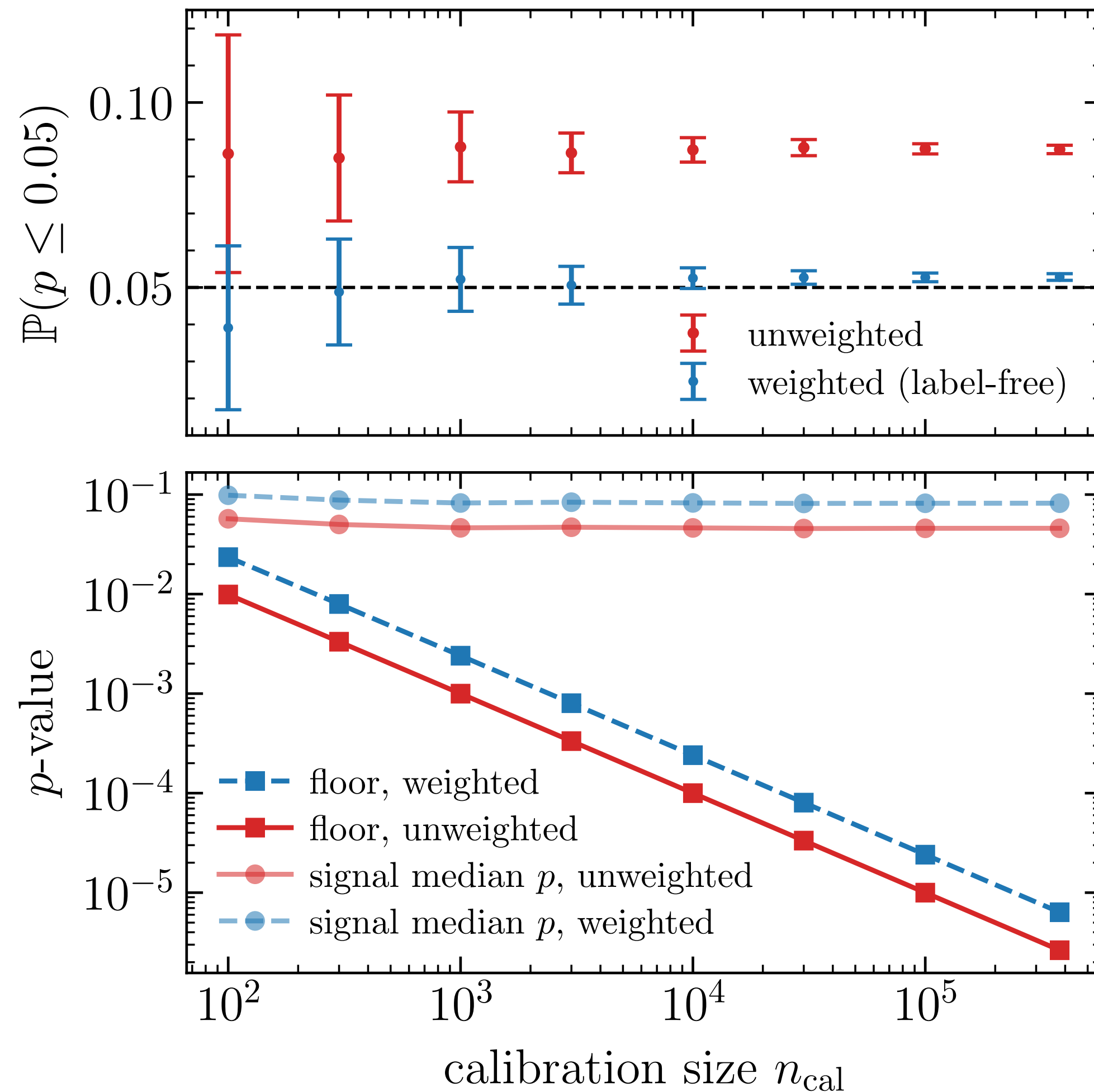


$$Z_{\text{asym}} = \text{sign}(N - b_m) \sqrt{2 [N \ln(N/b_m) - (N - b_m)]}$$

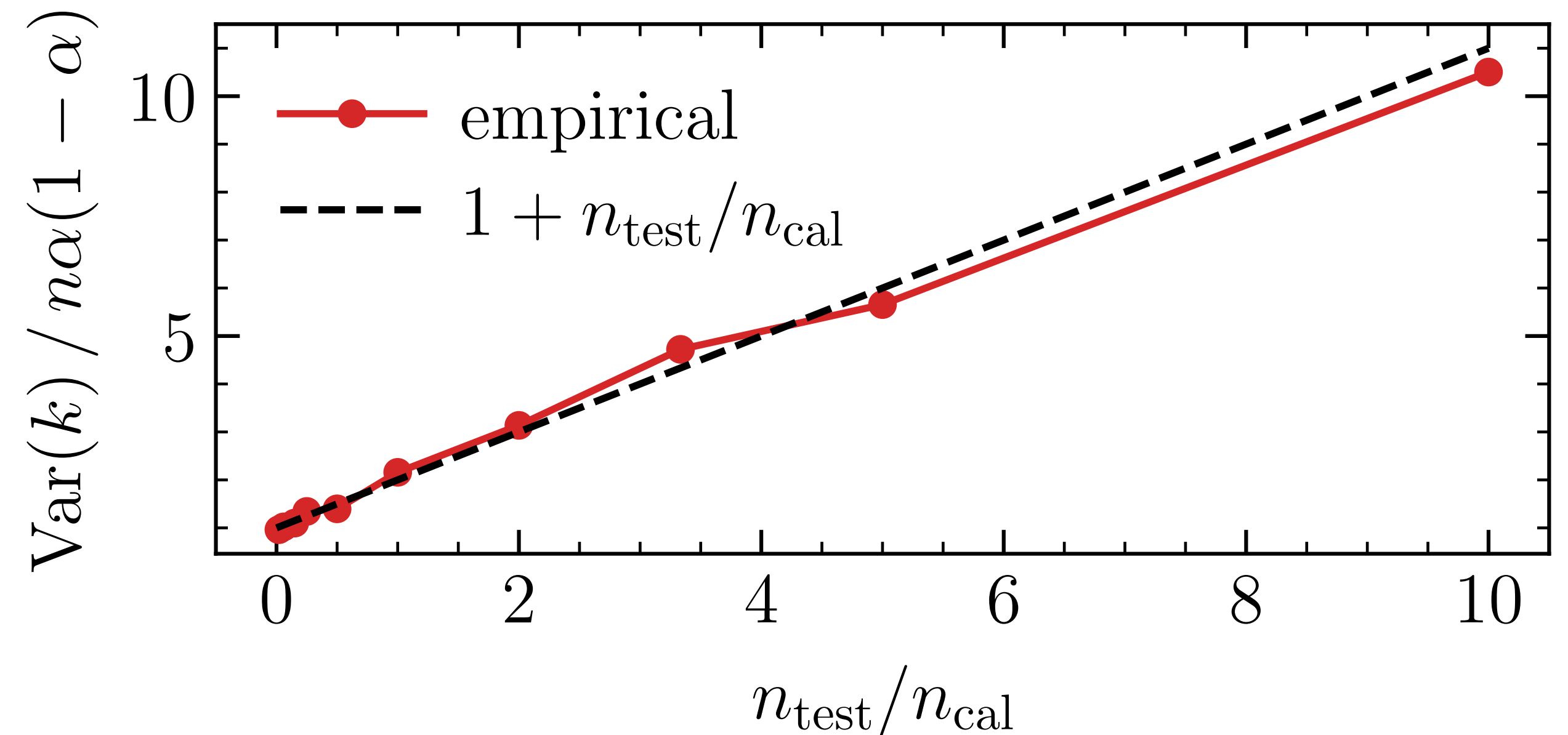
$$Z(m) = \frac{k_m - \alpha n_m}{\sqrt{c_m n_m \alpha(1 - \alpha)}}, \quad k_m = |\{p \leq \alpha\}|$$



How much does it cost?

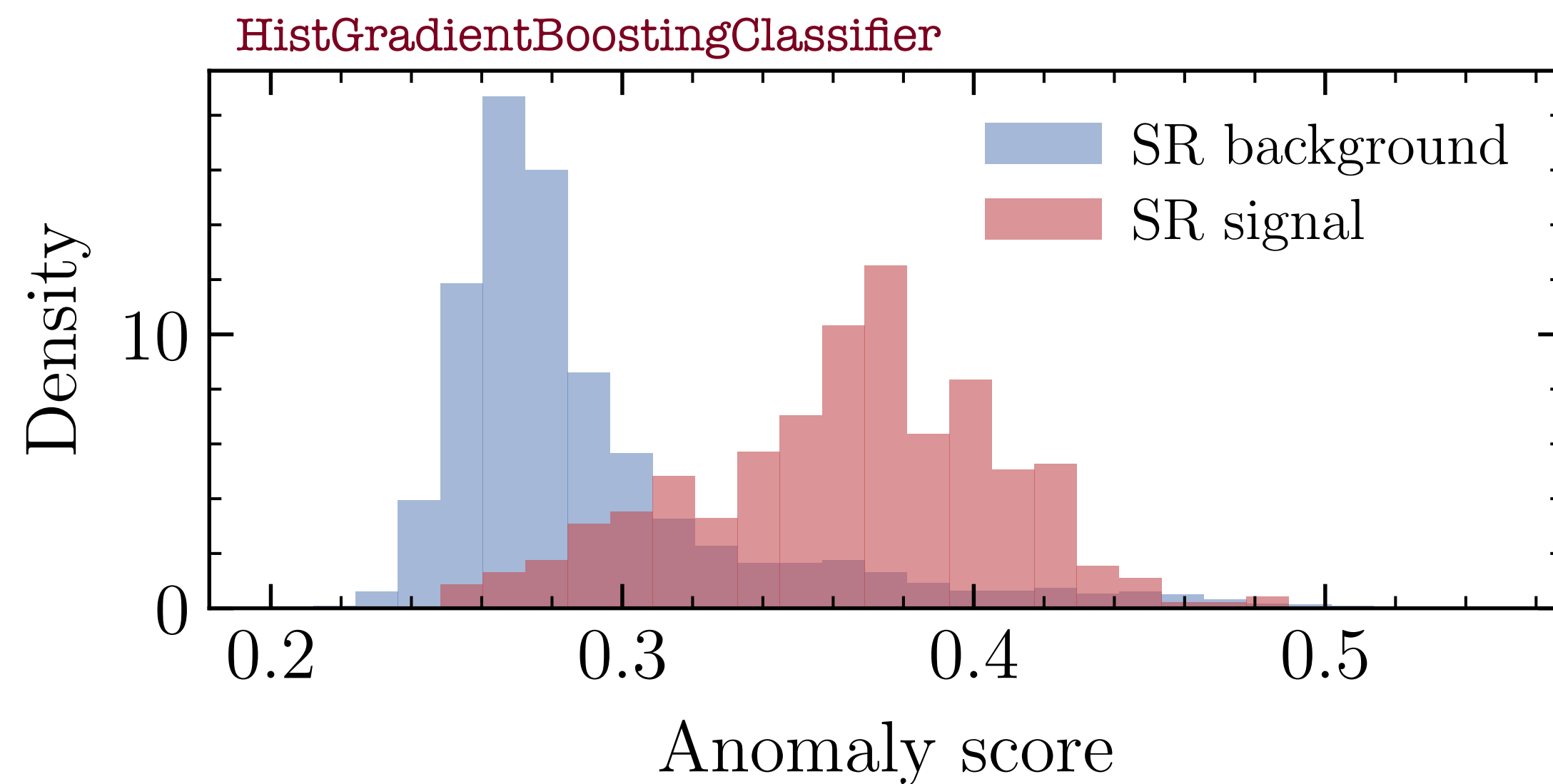


$$Z(m) = \frac{k_m - \alpha n_m}{\sqrt{c_m n_m \alpha(1 - \alpha)}} \quad , \quad k_m = |\{p \leq \alpha\}|$$

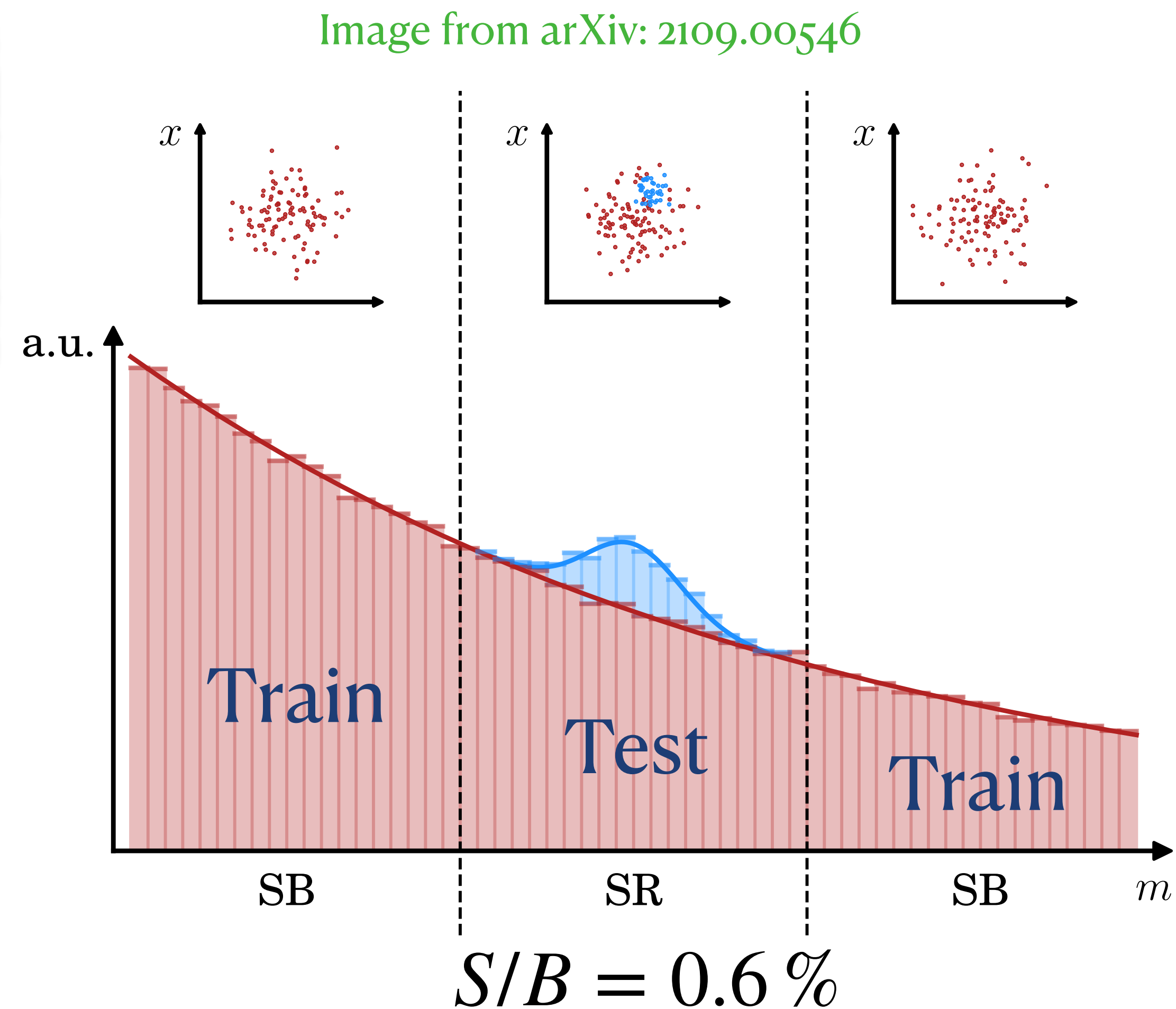


LHC Olympics 2020 R&D dataset

- ❖ QCD dijet + signal $Z' \rightarrow XY \rightarrow 4j$ where $(m_{Z'}, m_X, m_Y) = (3.5, 0.5, 0.1)$ TeV
- ❖ signal is localised as a bump in the dijet invariant mass m_{jj} near 3.5 TeV

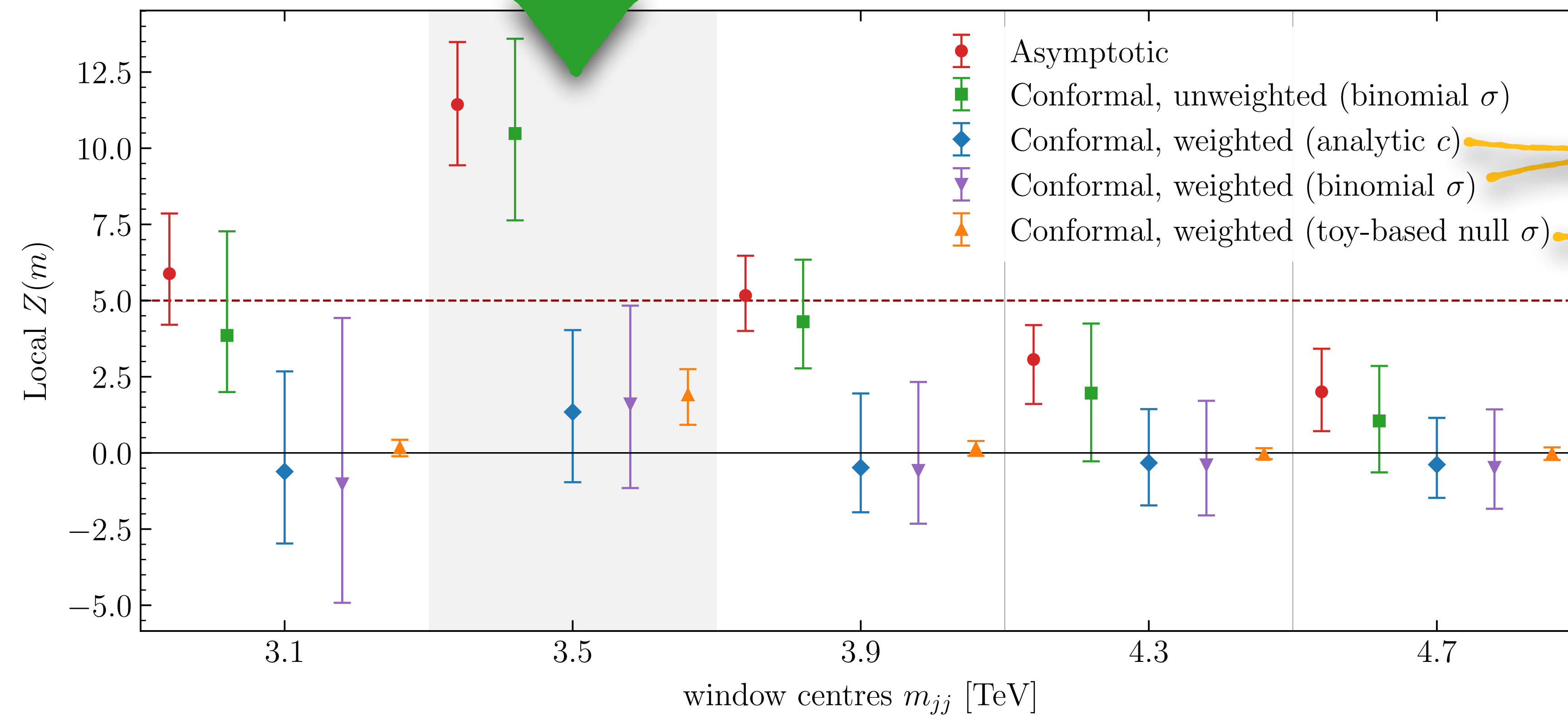
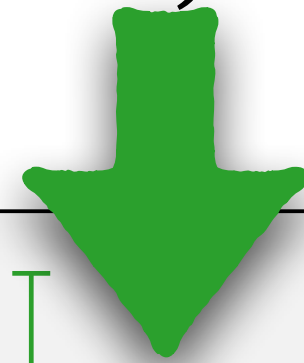


$$\text{SIC } (\epsilon_S / \sqrt{\epsilon_B}) \approx 1.95$$



Local significance

$$S/B = 0.6 \% , \text{ SIC } (\epsilon_S/\sqrt{\epsilon_B}) \approx 1.95$$

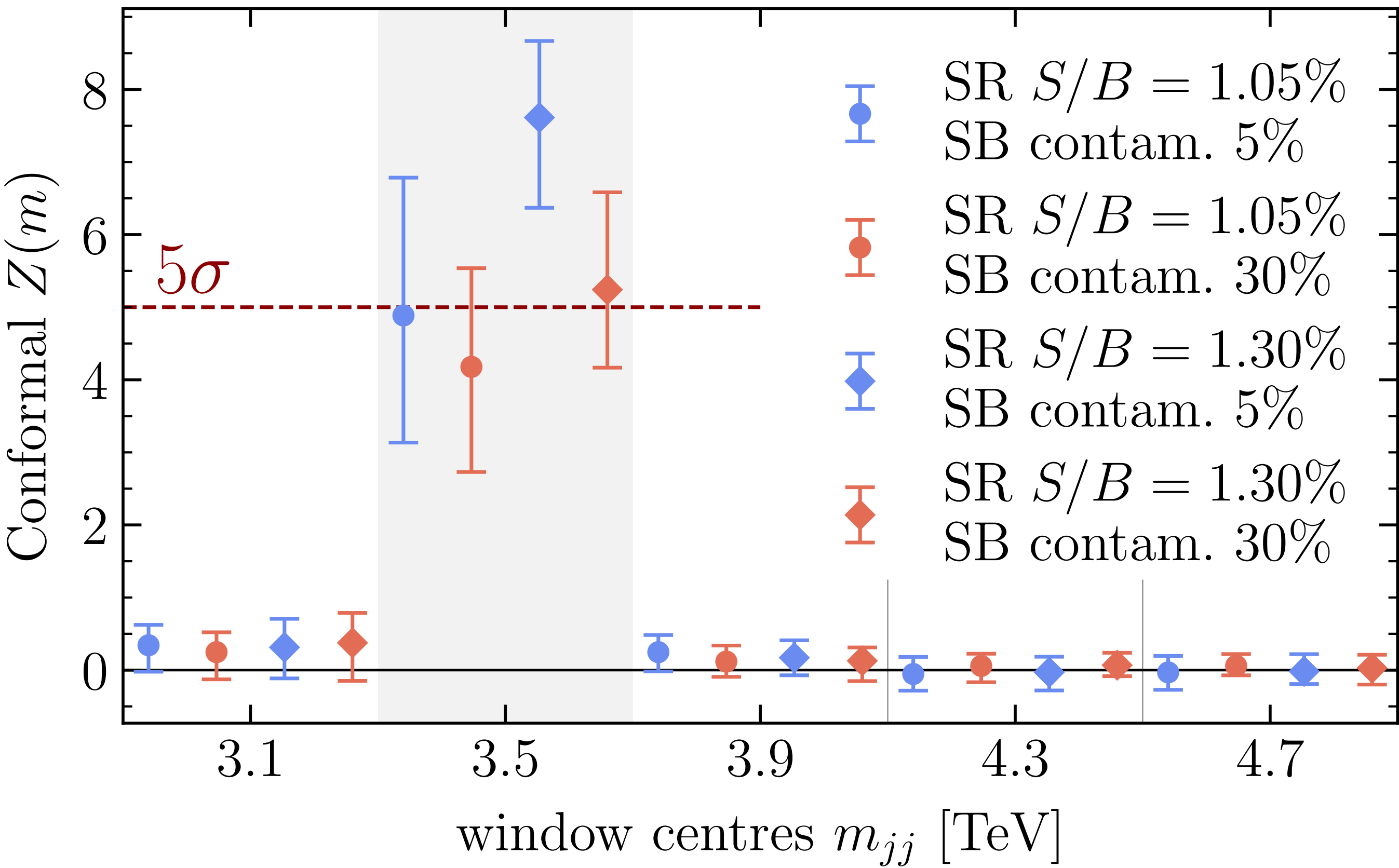


$$c = 1 + n_{\text{test}}/n_{\text{cal}}$$

$$c = 1$$

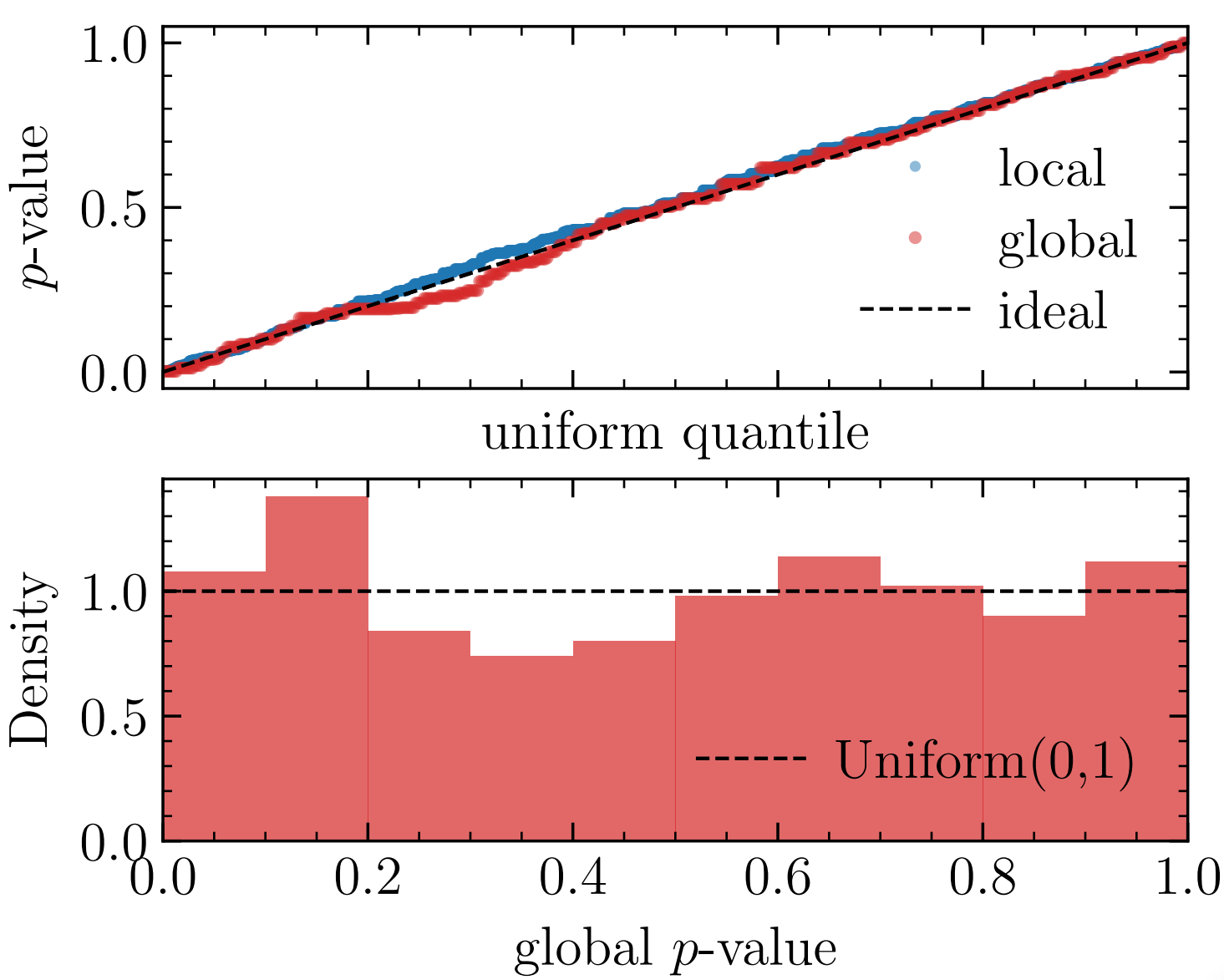
Resample everything
100 times

Signal contamination in the control region



Global Significance

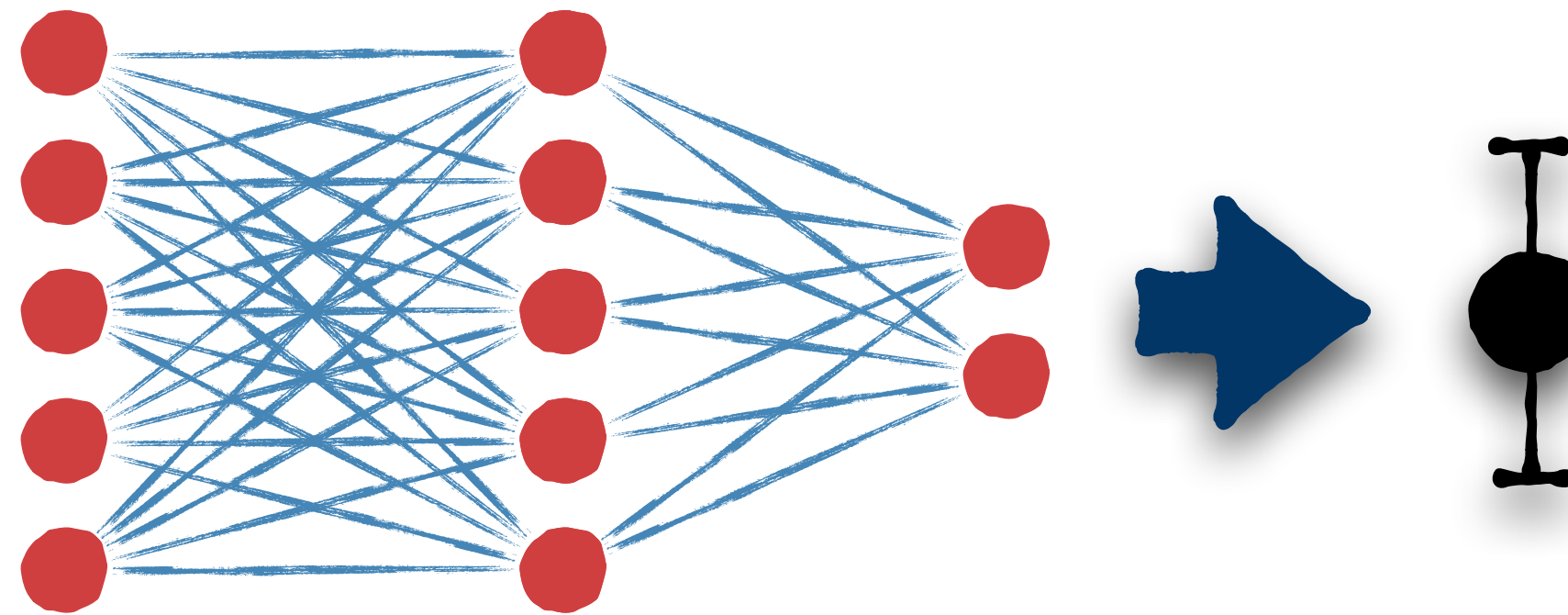
SR S/B	SB contamination	
	5%	30%
1.05%	4.6 σ	3.8 σ
1.30%	7.4 σ	4.9 σ



Conclusion & Outlook

Conclusion & Outlook

- ❖ Conformal prediction
 - ◆ Model agnostic
 - ◆ Distribution free
 - ◆ Coverage guarantee



What's next?

- ❖ Building robust statistical methods for conformal inference
- ❖ Integrating CP into our decision-making tool-chain, e.g. triggers
- ❖ Physics-inspired score function engineering
- ❖ Understanding the effect of a limited calibration dataset

Sponsors

