Beyond 5σ

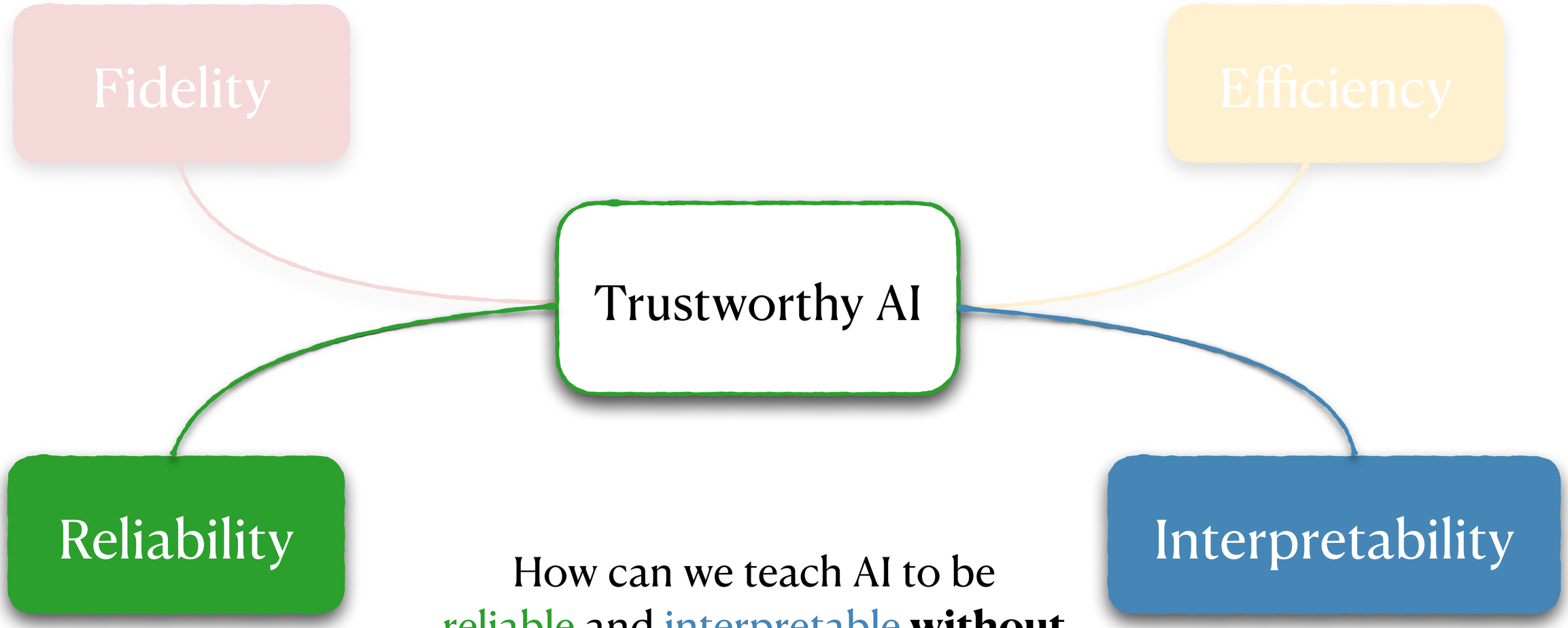
Teaching AI to know what it doesn't know

Jack Y. Araz
University College London
City St George's, University of London

Based on arXiv: 2512.17048
IoP Annual APP and HEPP Conference
10 April, 2026

Towards trustworthy AI for HEP

Towards trustworthy AI for HEP

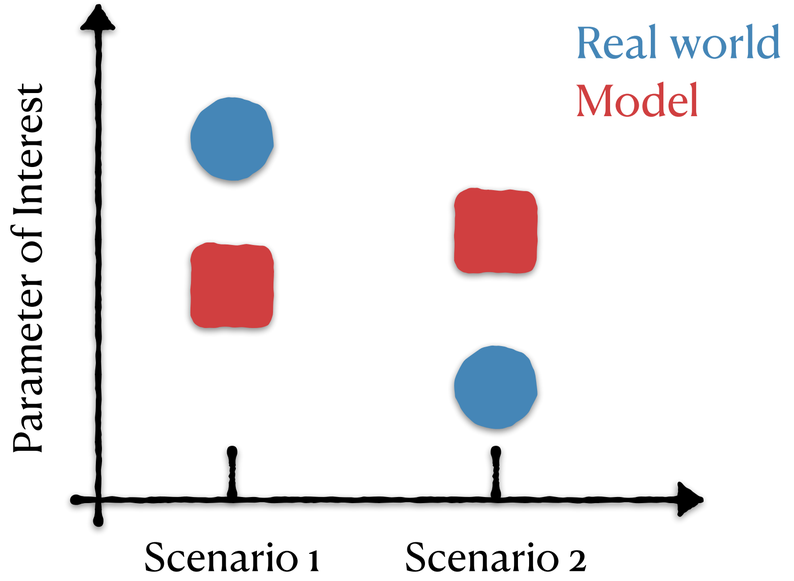


How can we teach AI to be reliable and interpretable **without losing** fidelity or efficiency?

Can we trust AI-driven discovery?

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

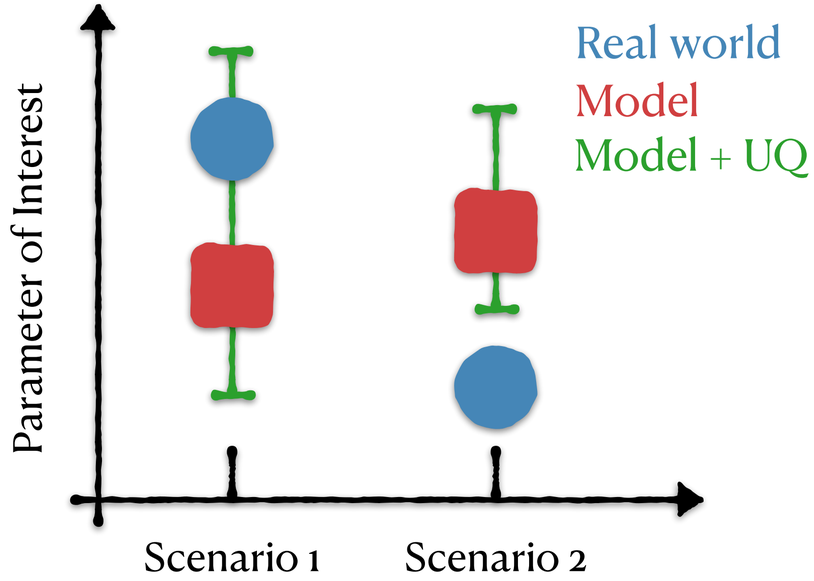
- AUC & Accuracy: Measure average discriminative power, but do not indicate certainty or coverage.
- Overconfidence: Even an AUC ≥ 0.99 can lead to systematic overconfidence.
- Distribution shift: Imperfect simulations, detector effects, and rare signals in sparse feature-space corners.



Can we trust AI-driven discovery?

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

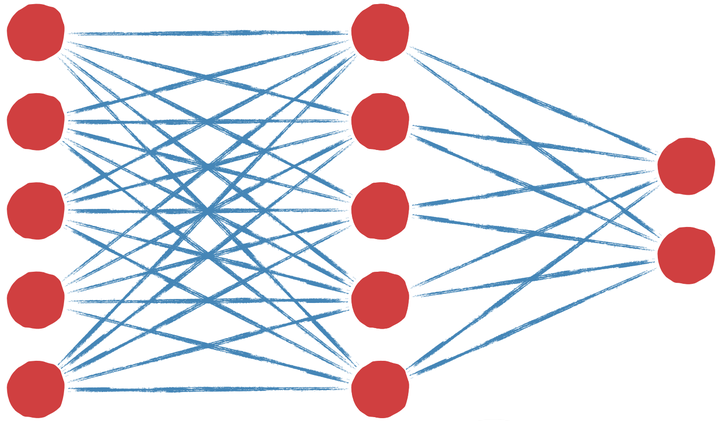
- AUC & Accuracy: Measure average discriminative power, but do not indicate certainty or coverage.
- Overconfidence: Even an AUC ≥ 0.99 can lead to systematic overconfidence.
- Distribution shift: Imperfect simulations, detector effects, and rare signals in sparse feature-space corners.



- How wrong is our model?
- And how confident are we in their predictions?

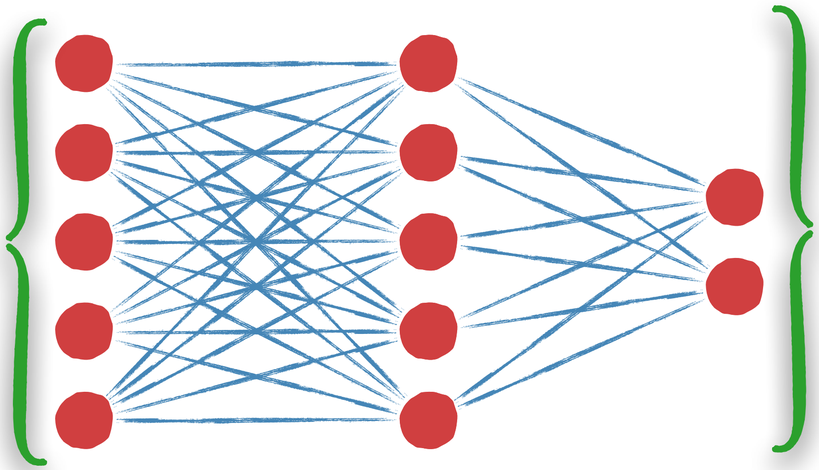
Outline

- Widely used methods for UQ
- Conformal Prediction for HEP
  - Classification
  - Generative Modelling
  - Anomaly detection
- What's next?



New physics!

OR



UQ

Newish physics!
With 95% CL

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.

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× training cost); ensemble diversity is not guaranteed; provides no finite-sample statistical guarantee on coverage.

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.

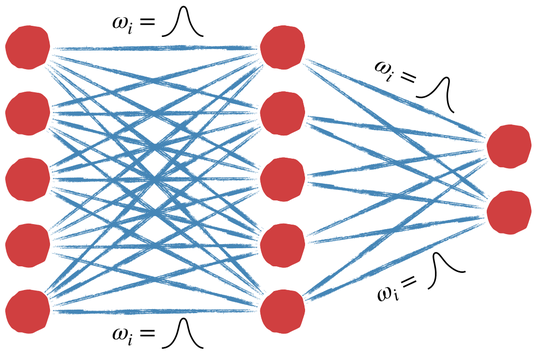
Cross-validation

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.

Bayesian Neural Networks



- Computational cost
- Sensitive to prior choice
- Guarantees are asymptotic
- Architectural modification
- 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× training cost); ensemble diversity is not guaranteed; provides no finite-sample statistical guarantee on coverage.

Ensemble BNN: [JYA, Spannowsky; JHEP '21]
Repulsive Ensembles: [Bahl, Elmer et. al.; SciPost '25]

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

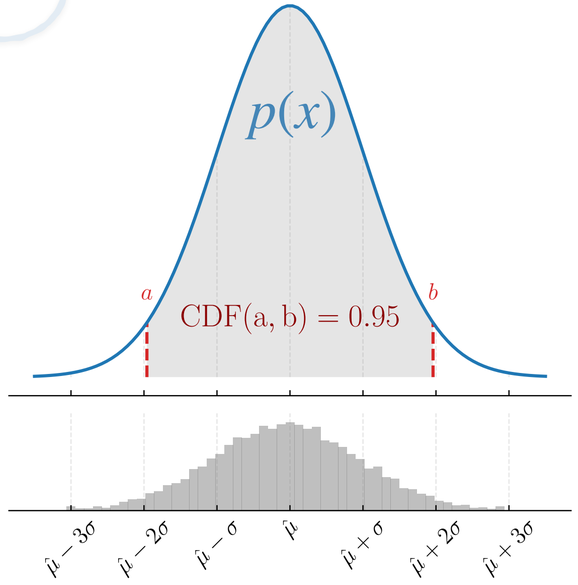
Uncertainty estimates depend on dropout rate; no formal coverage guarantee; tends to underestimate uncertainty in out-of-distribution regions.

Deep Ensembles

Computationally expensive (N× 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



Ensemble BNN: [JYA, Spannowsky; JHEP '21]
Repulsive Ensembles: [Bahl, Elmer et. al.; SciPost '25]

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

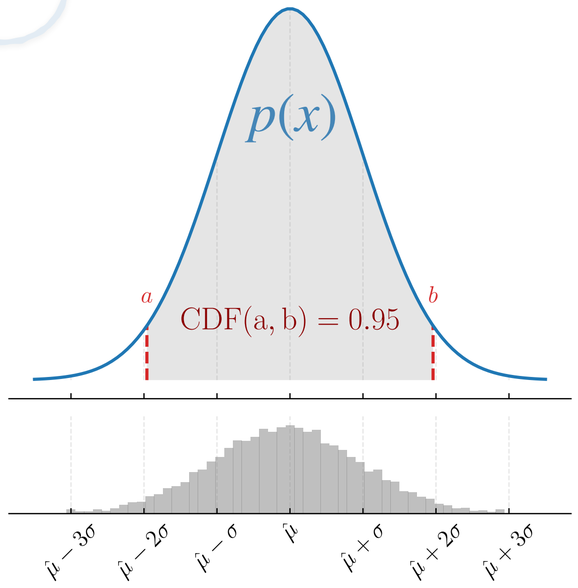
Uncertainty estimates depend on dropout rate; no formal coverage guarantee; tends to underestimate uncertainty in out-of-distribution regions.

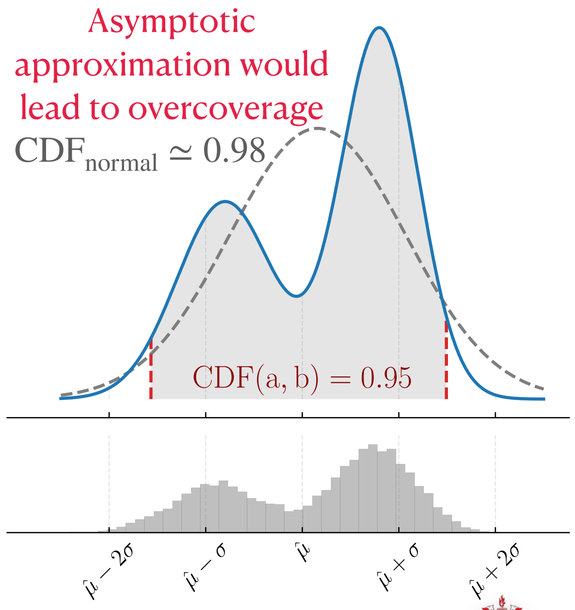
Deep Ensembles

Computationally expensive (N× 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





Ensemble BNN: [JYA, Spannowsky; JHEP '21]
Repulsive Ensembles: [Bahl, Elmer et. al.; SciPost '25]

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?

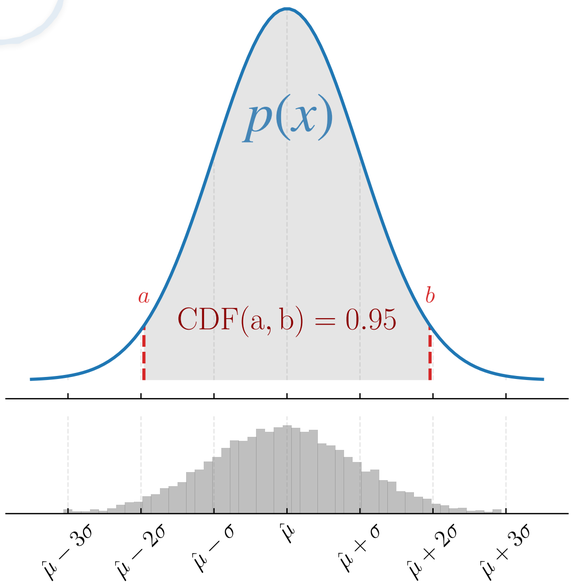
- Rely on assumptions being correct: asymptotic approximation
- Or test emprically weather our assumptions hold
- OR use inference methods that don't rely on assumptions (or only rely on weaker assumptions)

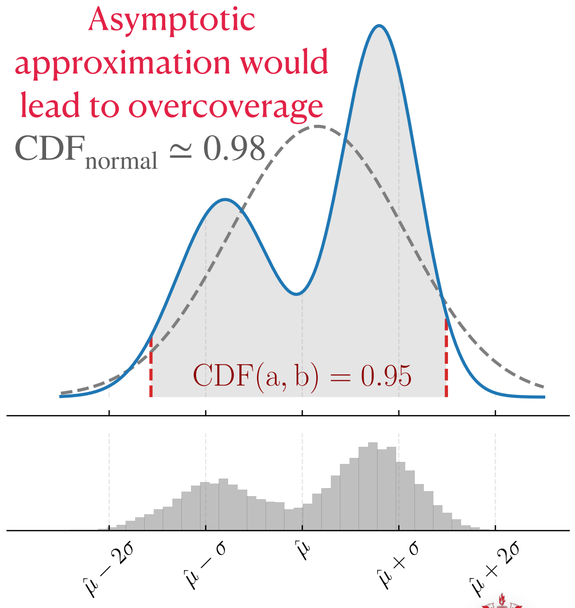
MC Dropout

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

Deep Ensembles

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

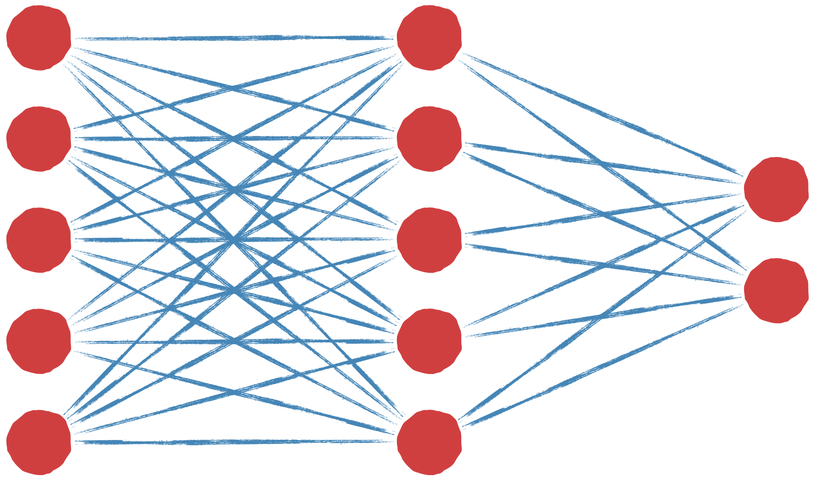




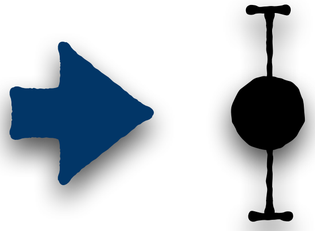
Posterior approximations introduce uncontrolled bias

Ensemble BNN: [JYA, Spannowsky; JHEP '21]
Repulsive Ensembles: [Bahl, Elmer et. al.; SciPost '25]

What do we need?

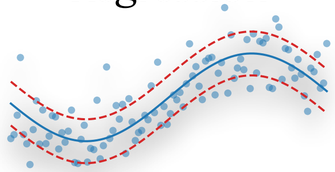


Rigorous, distribution-free, finite-sample CIs for any model and any dataset, for free

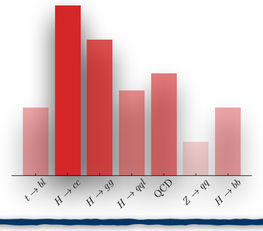


Building trust with confidence intervals

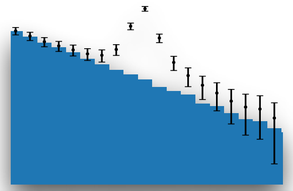
Regression



Classification



Anomaly Detection



Generative modelling

Conformal Prediction

- Distribution-free
- Finite-sample guarantee
- Model agnostic

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

Prediction Interval with coverage $\geq 1-\alpha$

Prediction set $\Gamma_\alpha(x) = \{1,\ldots,K\}$

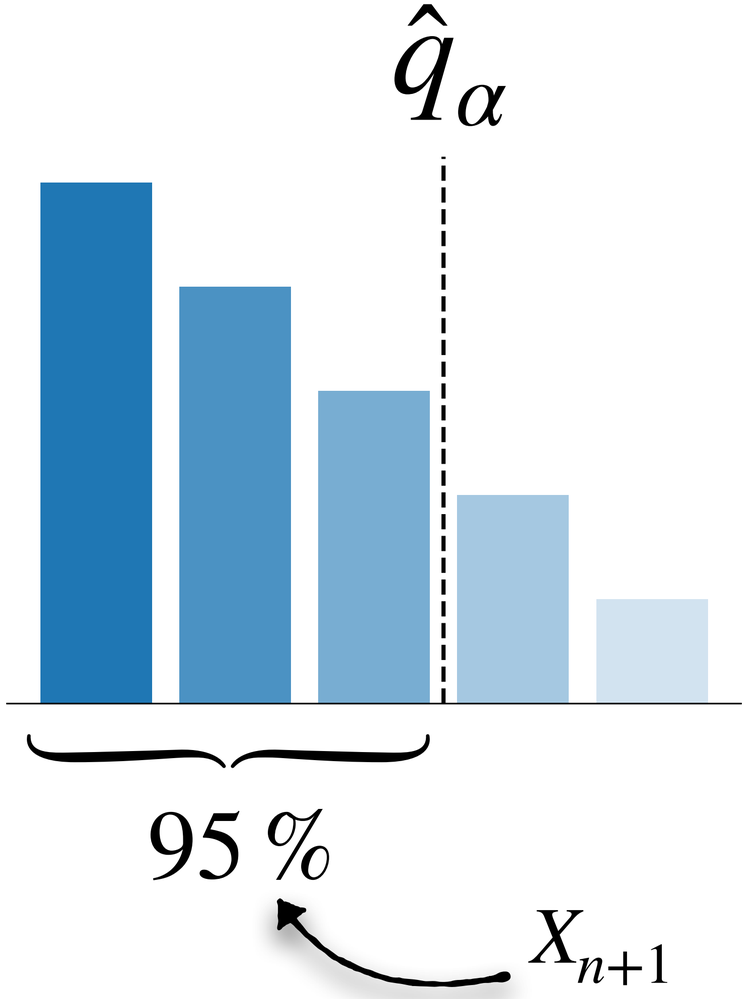
Conformal threshold $\varepsilon_{\text{sig}}(1-\alpha) = \mathbb{P}_{X\sim P_{\text{sig}}}\left(s(X) > \hat{q}_{1-\alpha}\right)$,

Calibrated discrepancy

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

Split Conformal Prediction

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

Split Conformal Prediction

Theorem

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

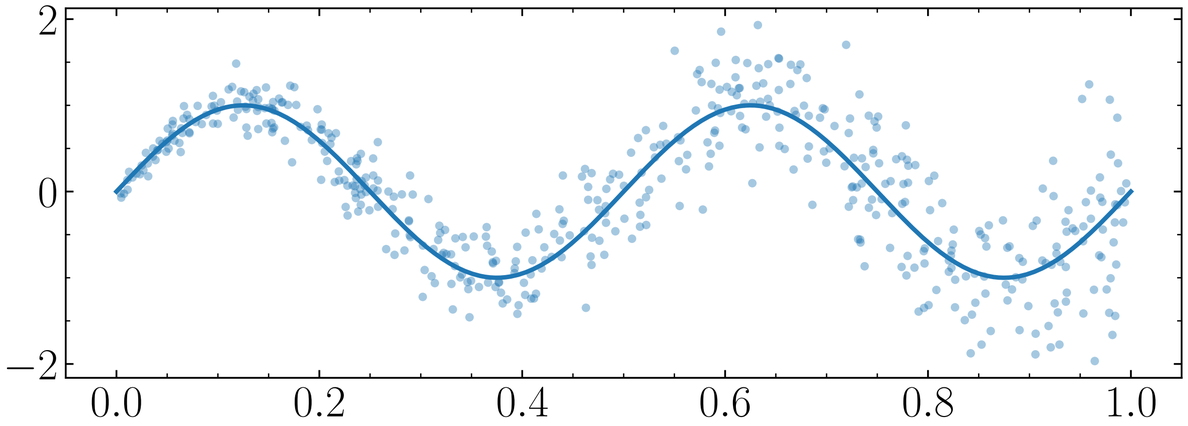
$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha$$

Split Conformal Prediction

Theorem

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

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

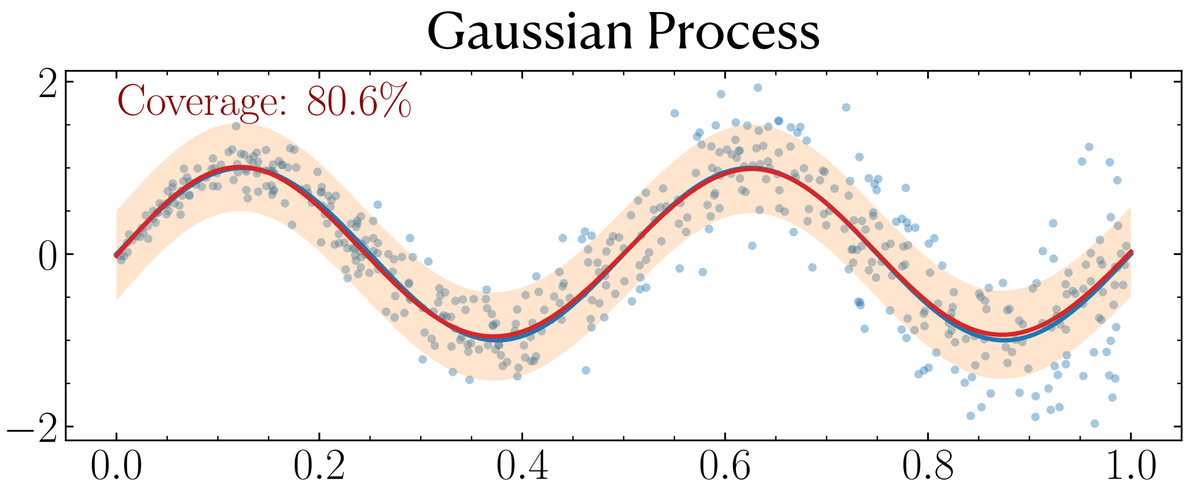


Split Conformal Prediction

Theorem

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

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



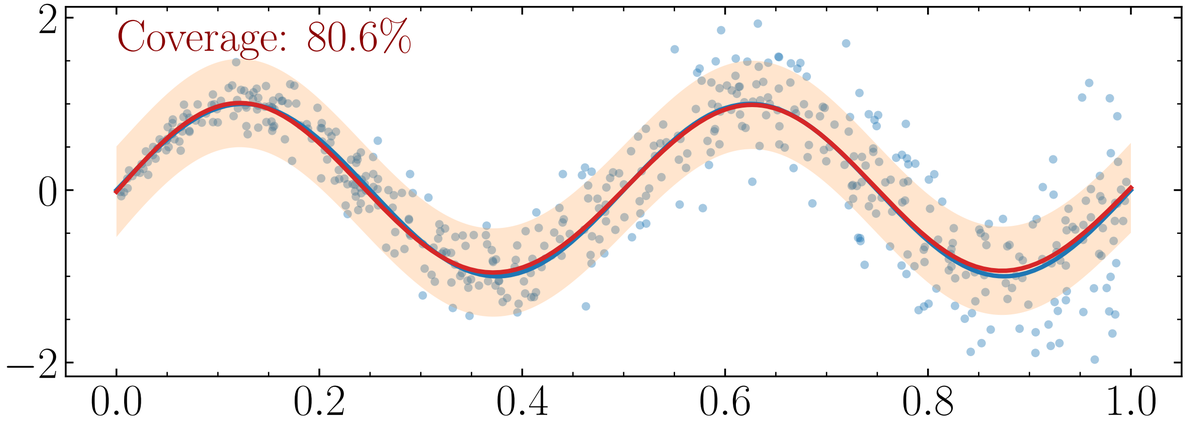
Split Conformal Prediction

Theorem

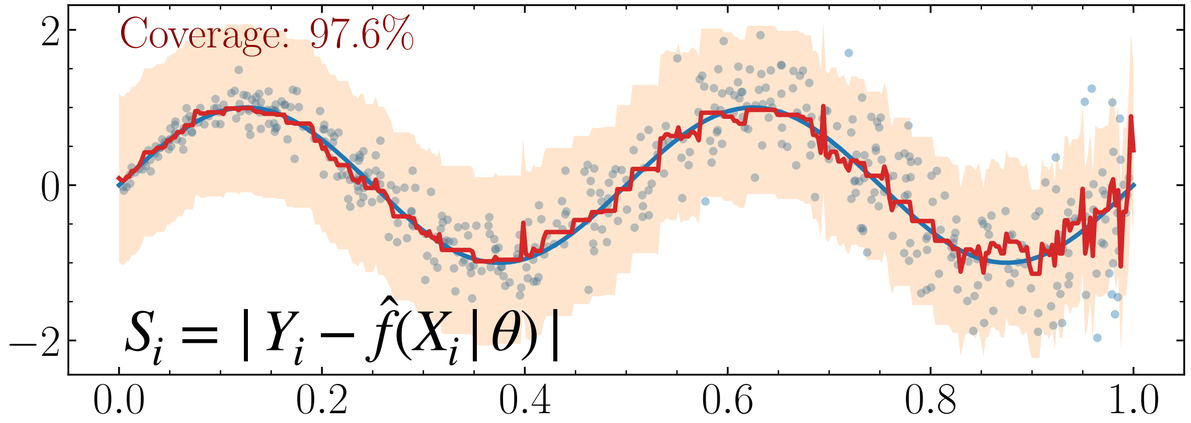
- If $\mathscr{D}^{\text{train}}$ and $\mathscr{D}^{\text{cal}}$ are exchangeable, independent and identically distributed, then split conformal prediction satisfies

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

Gaussian Process



Conformal Prediction with an overfitted model



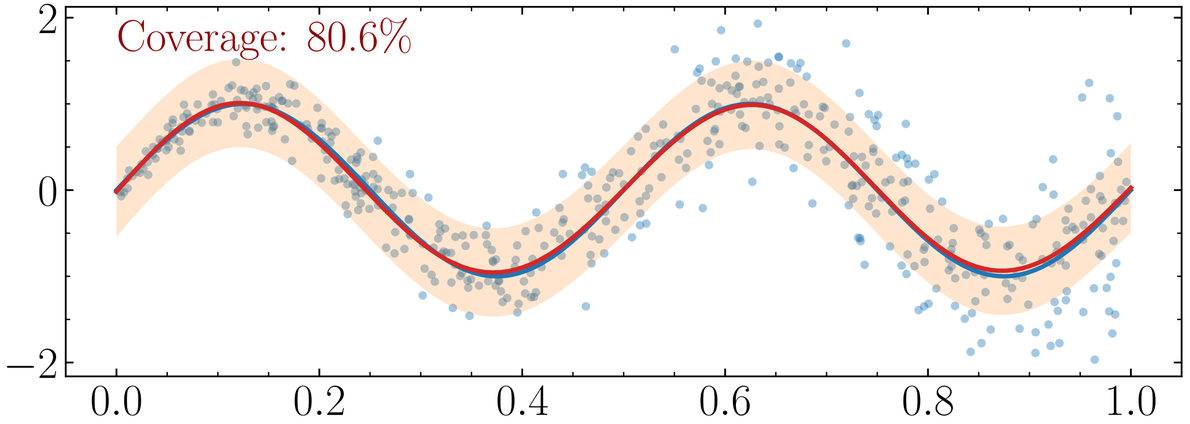
Split Conformal Prediction

Theorem

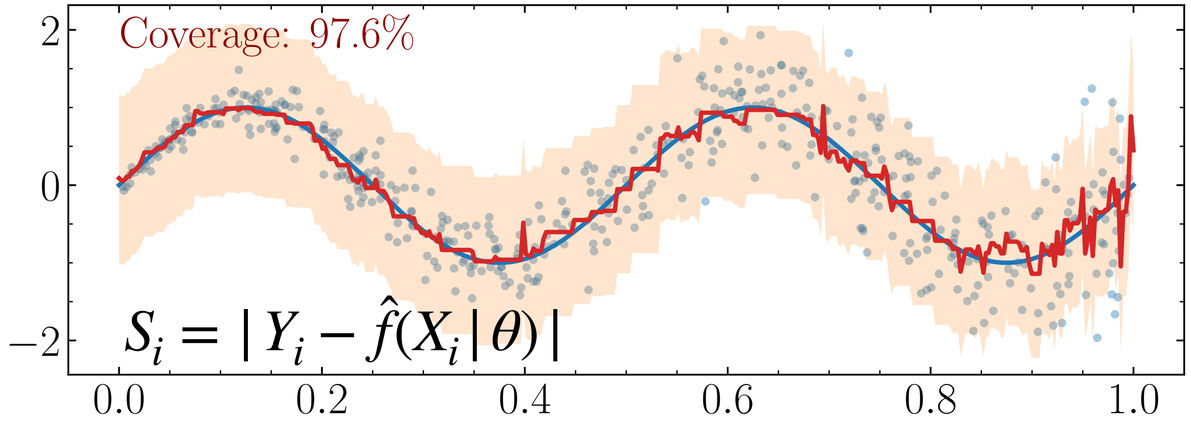
- If $\mathscr{D}^{\text{train}}$ and $\mathscr{D}^{\text{cal}}$ are exchangeable, independent and identically distributed, then split conformal prediction satisfies

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

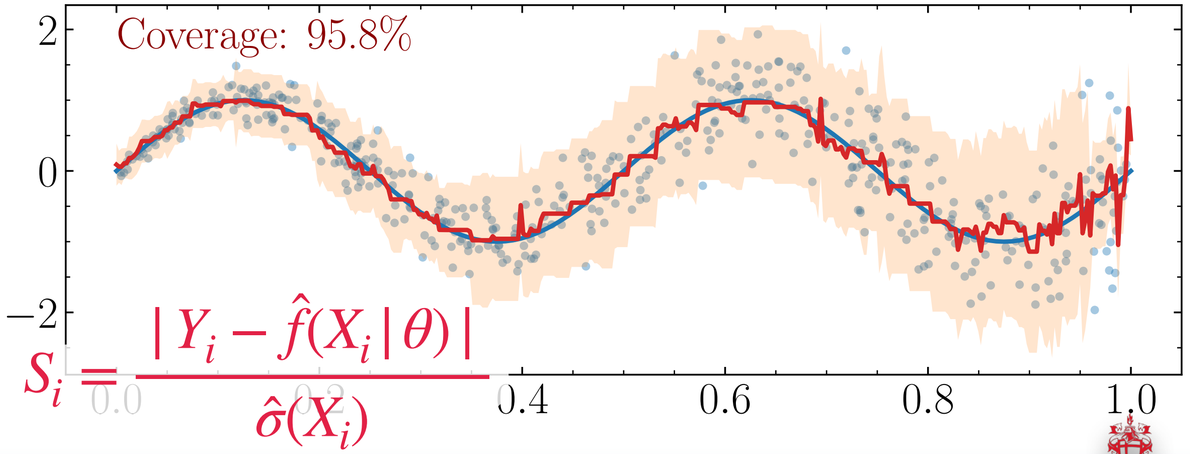
Gaussian Process



Conformal Prediction with an overfitted model



Adaptive Conformal Prediction



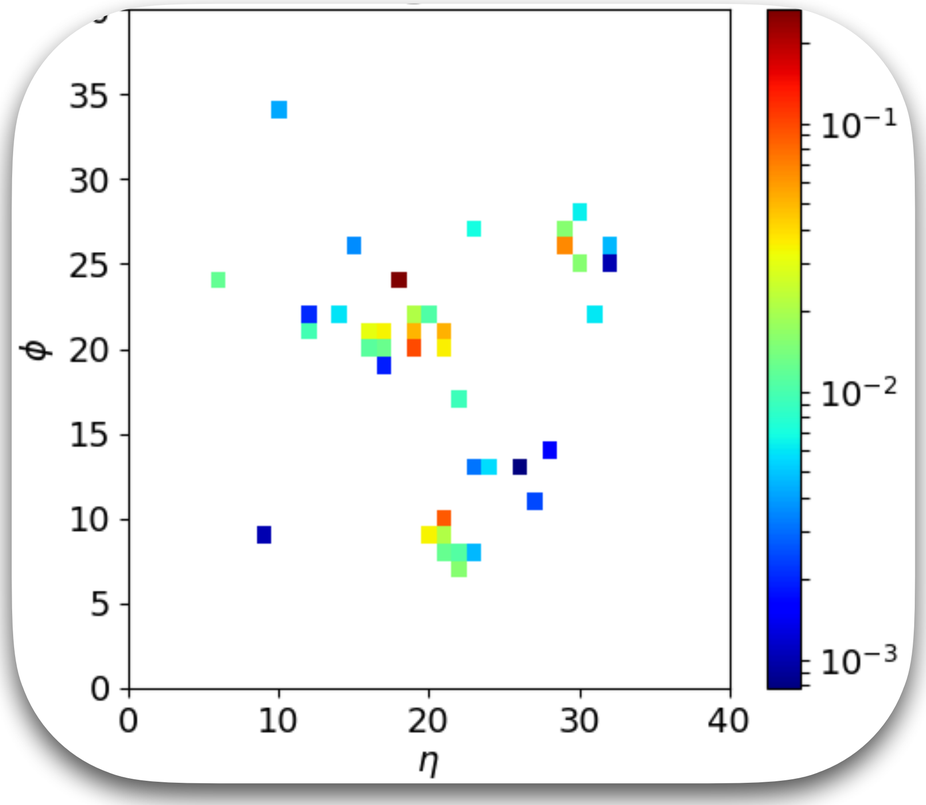
Conformal Prediction for Classification

Jack Y. Araz

CP for Classification

See Yuanda's and Joel's talk from this morning!

Kasieczka *et. al.* SciPost'19

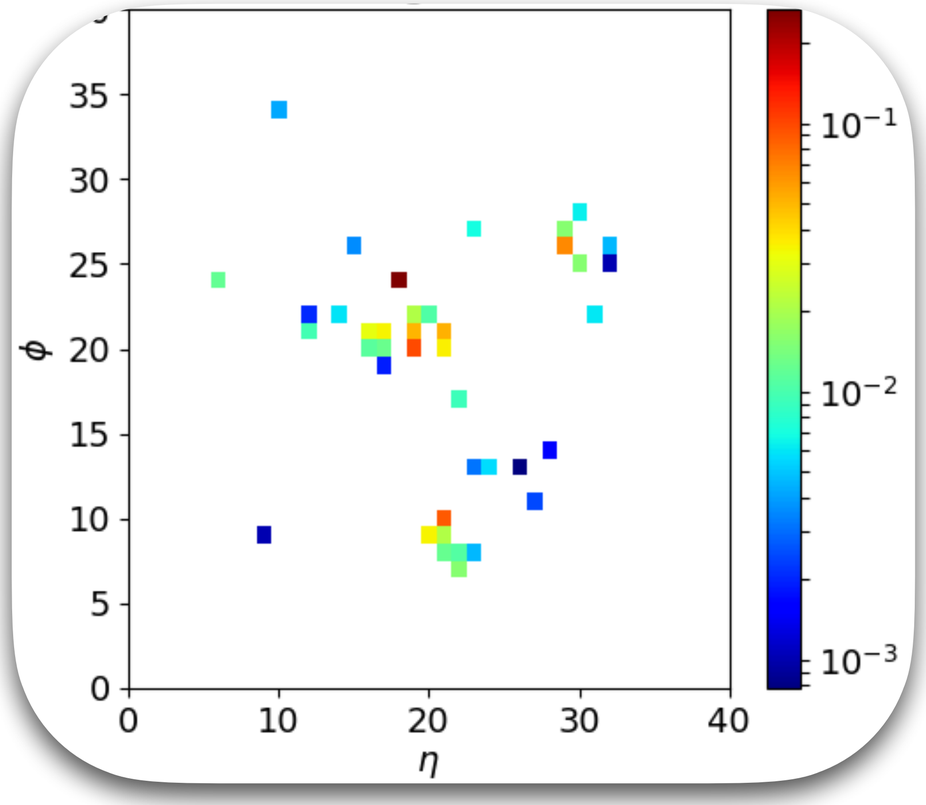


$$\Gamma_\alpha \rightarrow \{\text{Top},\ H \rightarrow bb, \text{QCD}\}$$

CP for Classification

See Yuanda's and Joel's talk from this morning!

Kasieczka *et. al.* SciPost'19



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

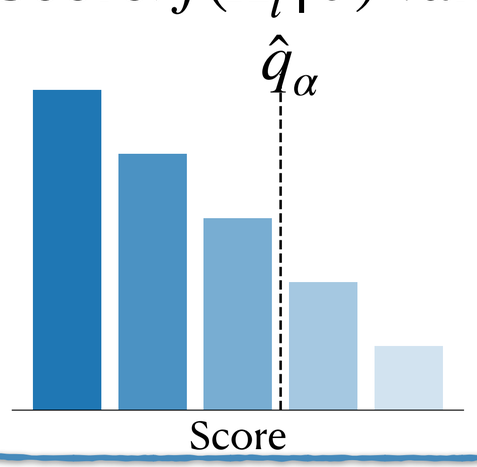
- NN: OmniLearn transformer-based foundation model

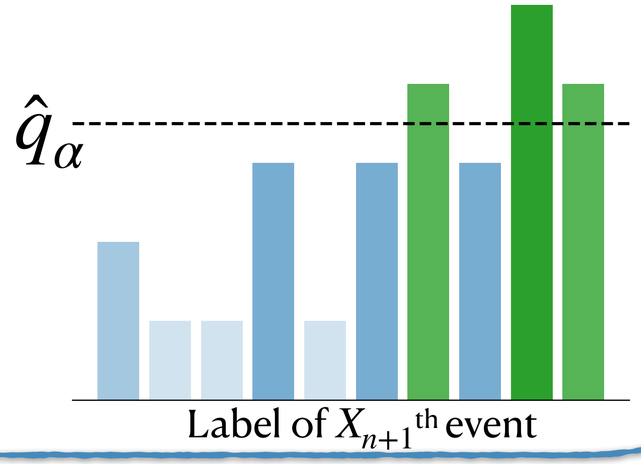
  OmniLearn: [Mikuni, Nachman; PRD '25]

- Dataset: JetClass (10 different classes)

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

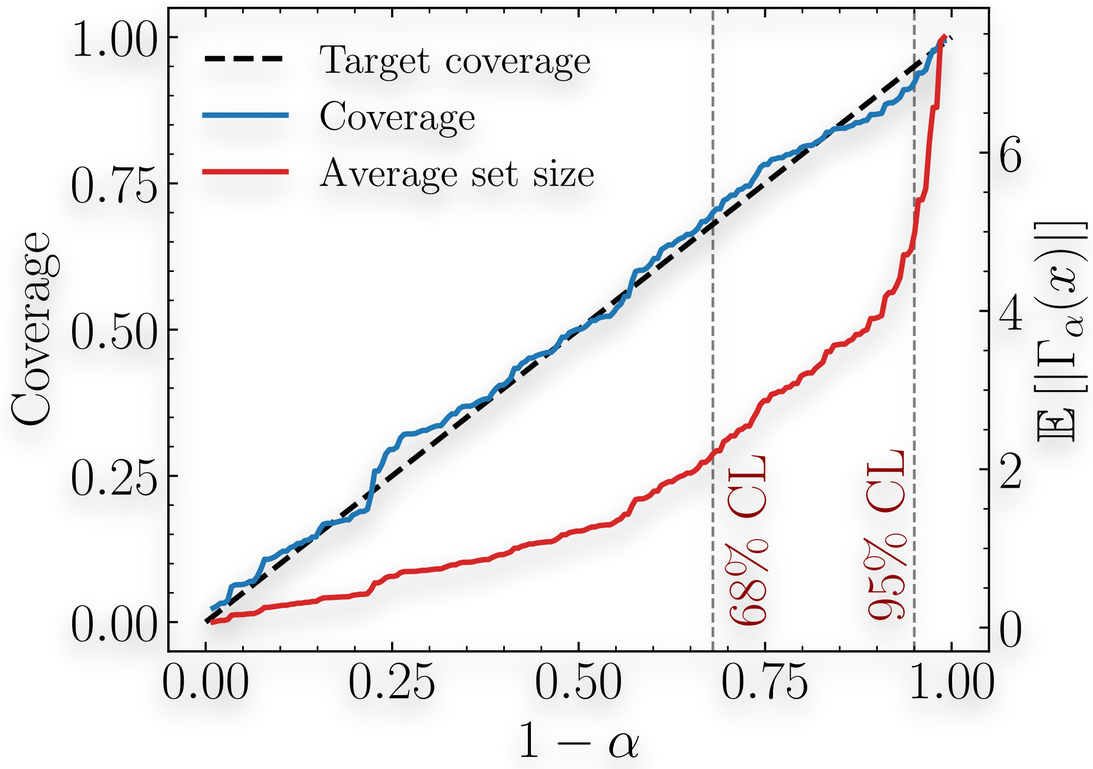
- Score: $\hat{f}(X_i \,|\, \theta)$ value of the true class



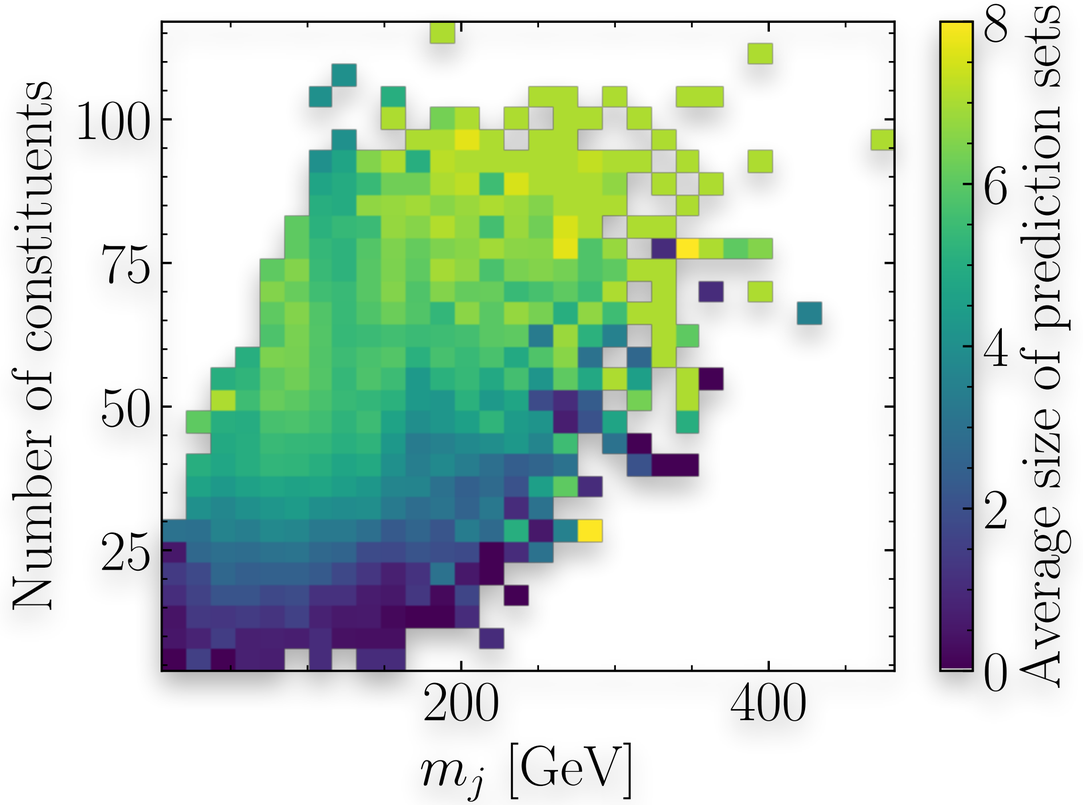


CP for Classification

Reliability



Interpretability



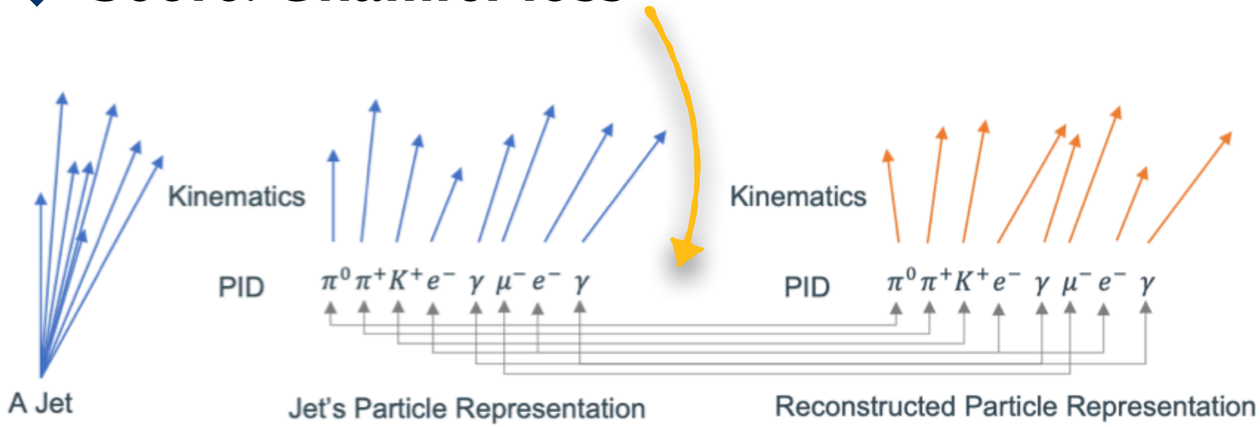
Conformal Prediction for Anomaly Detection

Jack Y. Araz

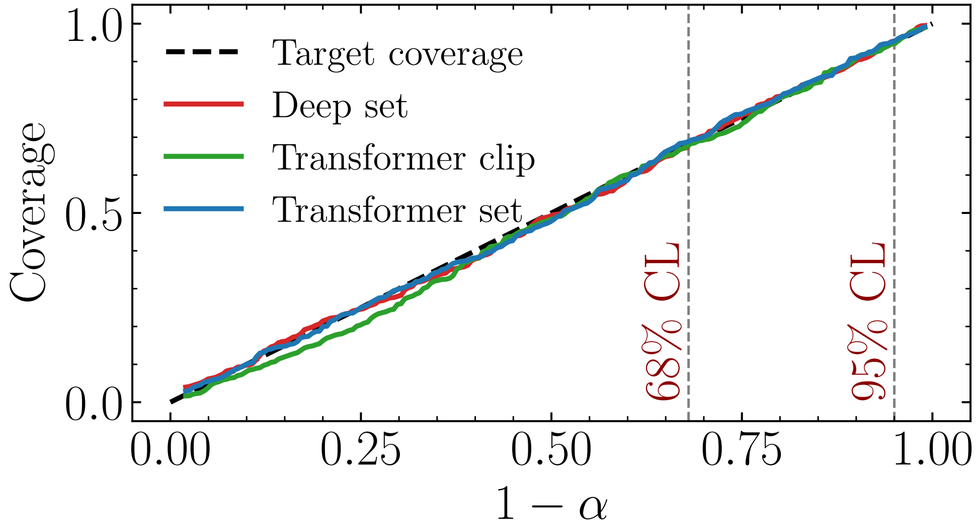
CP for Anomaly Detection

See Mehrnoosh's talk from this morning!

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



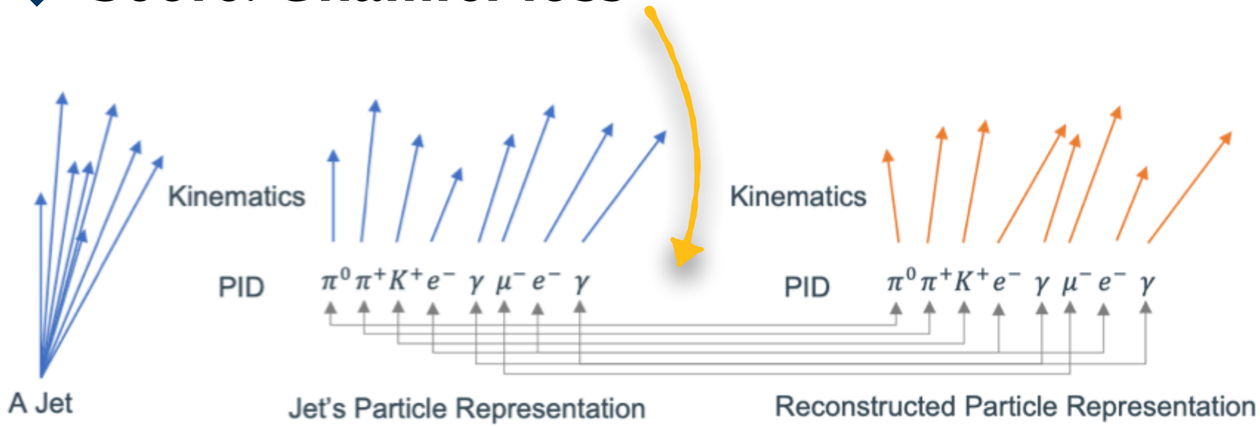
[Liu et. al.; 2311.17162]



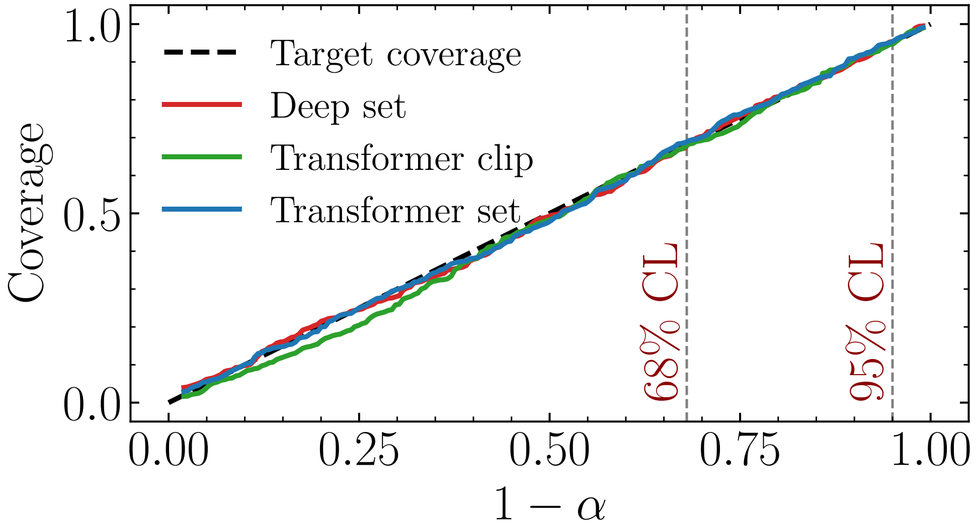
CP for Anomaly Detection

See Mehrnoosh's talk from this morning!

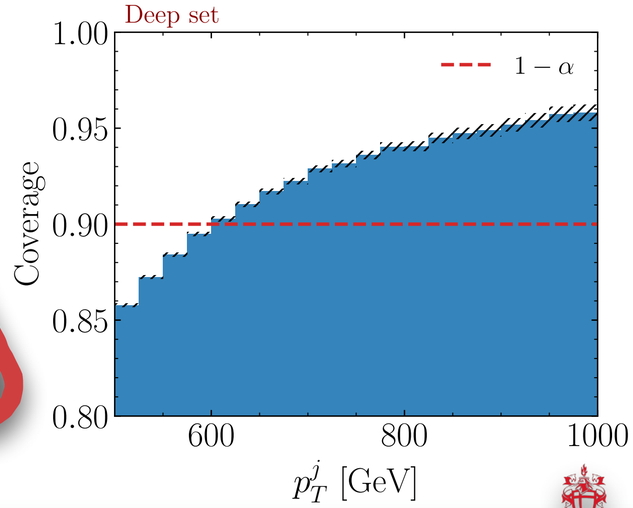
- Particle cloud-based VAE
- Trained only for the QCD dataset from JetClass
- Score: Chamfer loss



[Liu et. al.; 2311.17162]

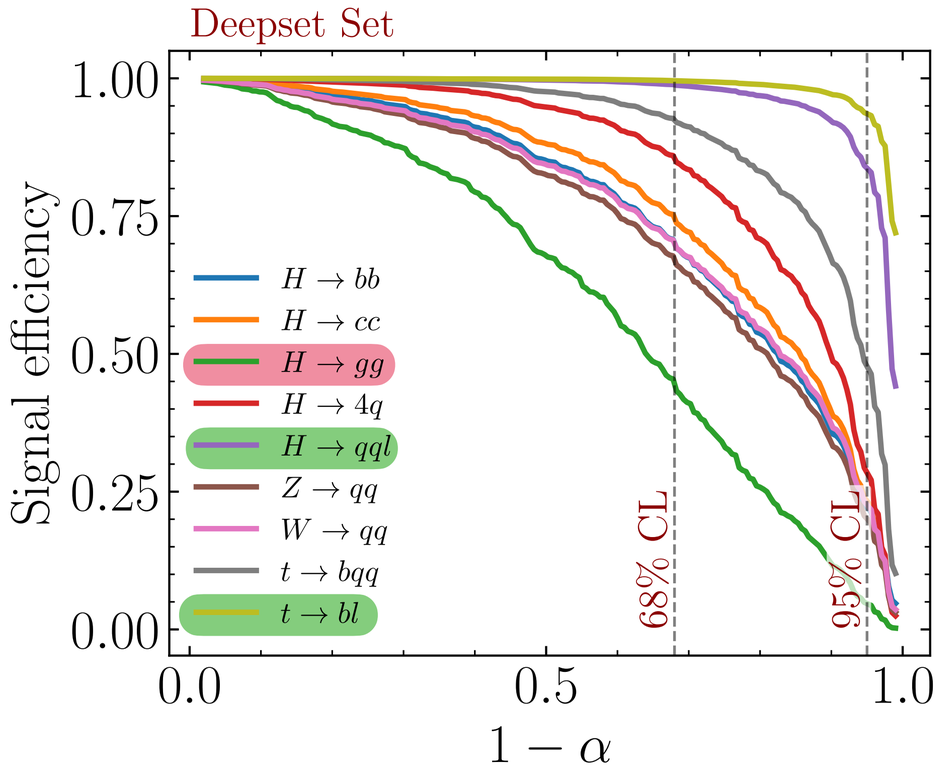


Coverage is only marginal!



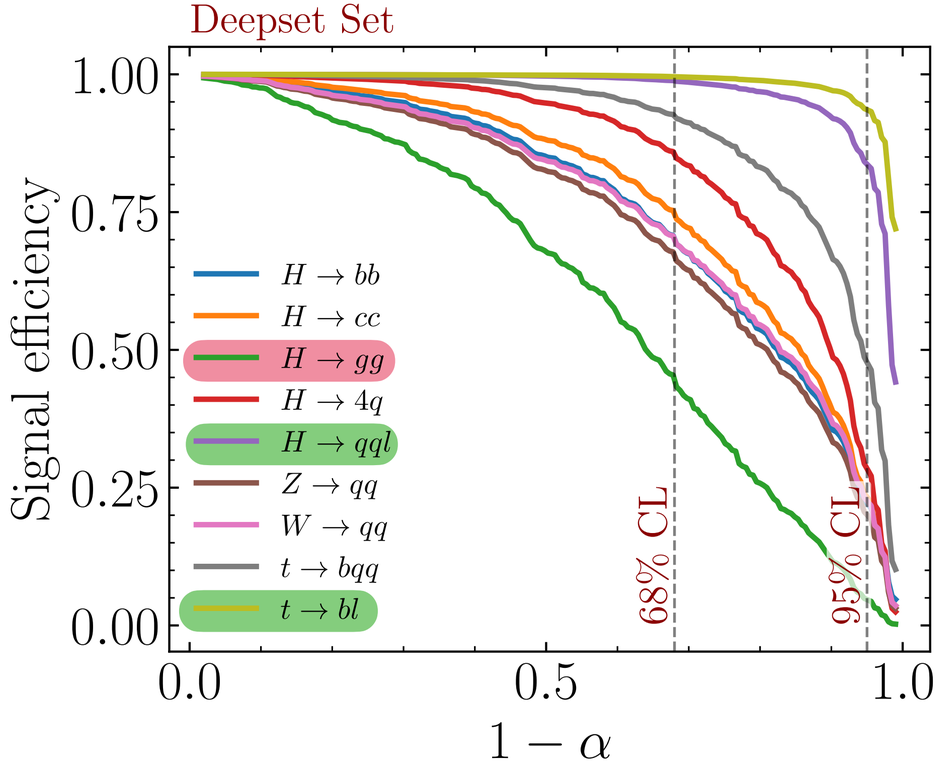
CP for Anomaly Detection

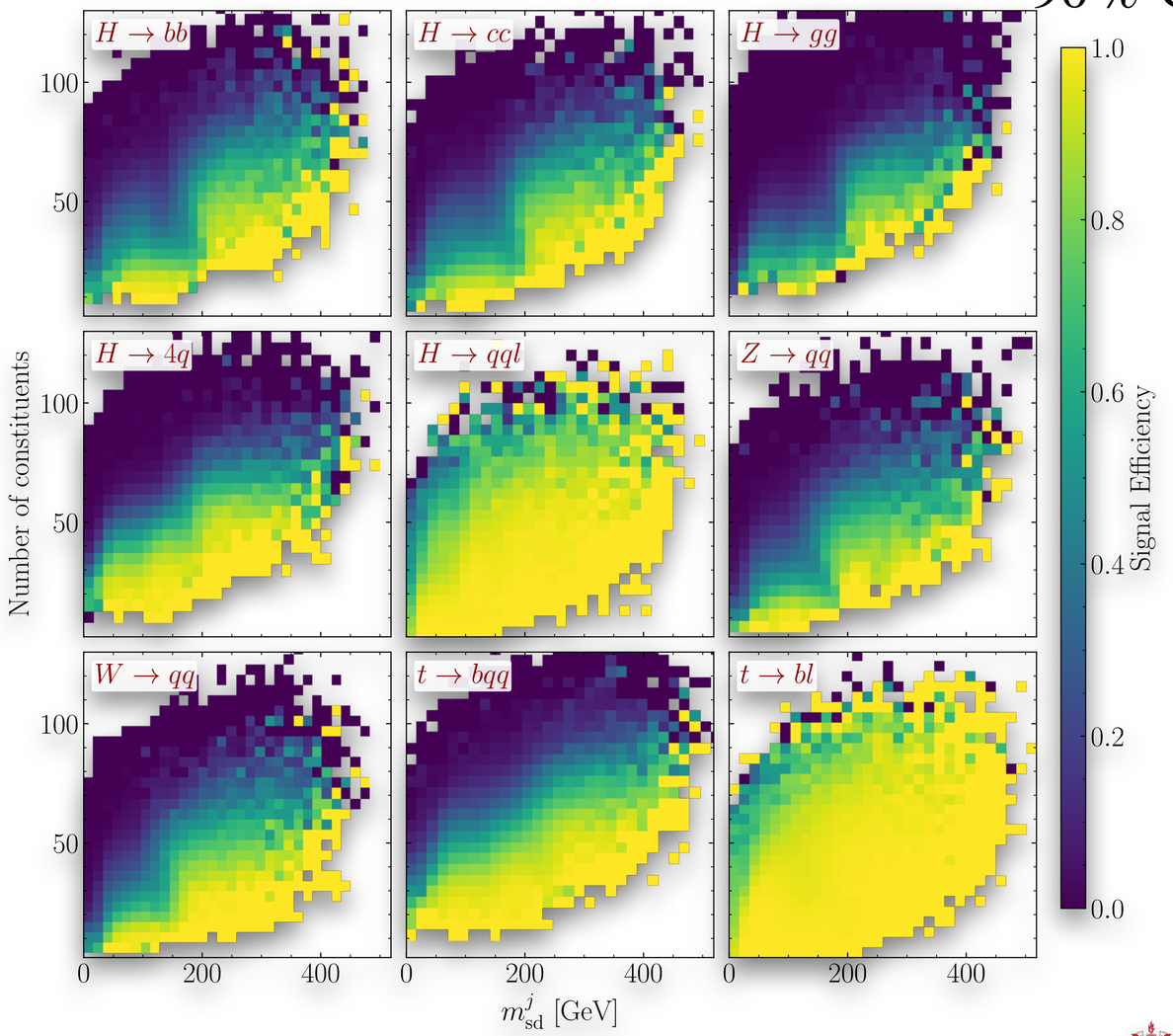
$$\varepsilon_{\text{sig}}(1-\alpha) \;=\; \mathbb{P}_{X\sim P_{\text{sig}}}\left(s(X) > \hat{q}_{1-\alpha}\right)$$



CP for Anomaly Detection

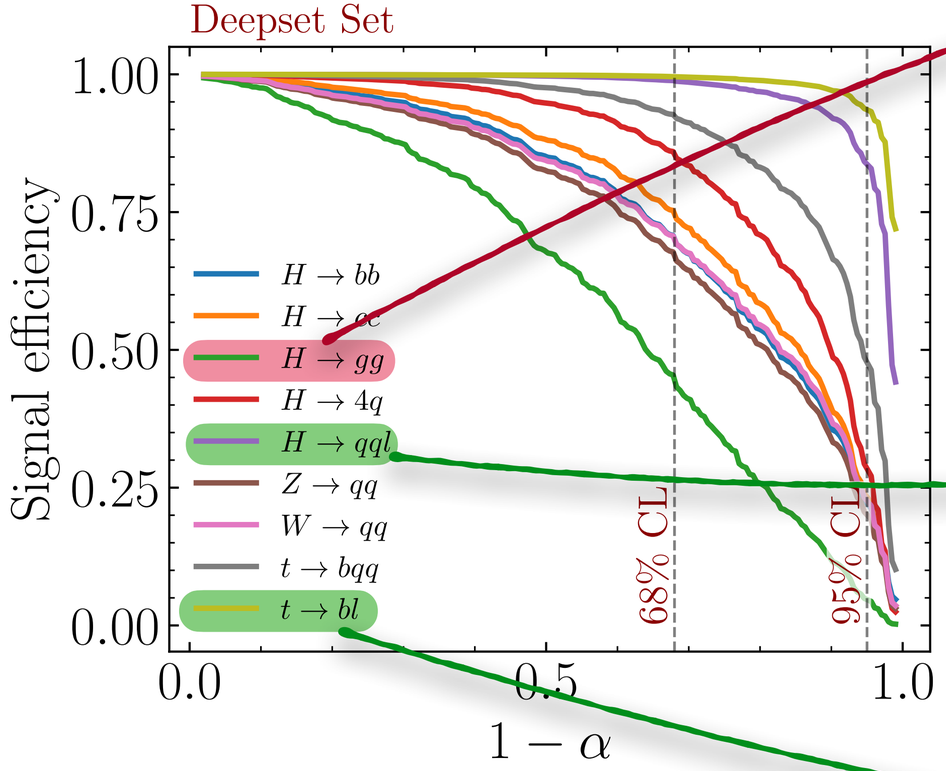
$$\varepsilon_{\text{sig}}(1-\alpha) \;=\; \mathbb{P}_{X\sim P_{\text{sig}}}\left(s(X) > \hat{q}_{1-\alpha}\right)$$

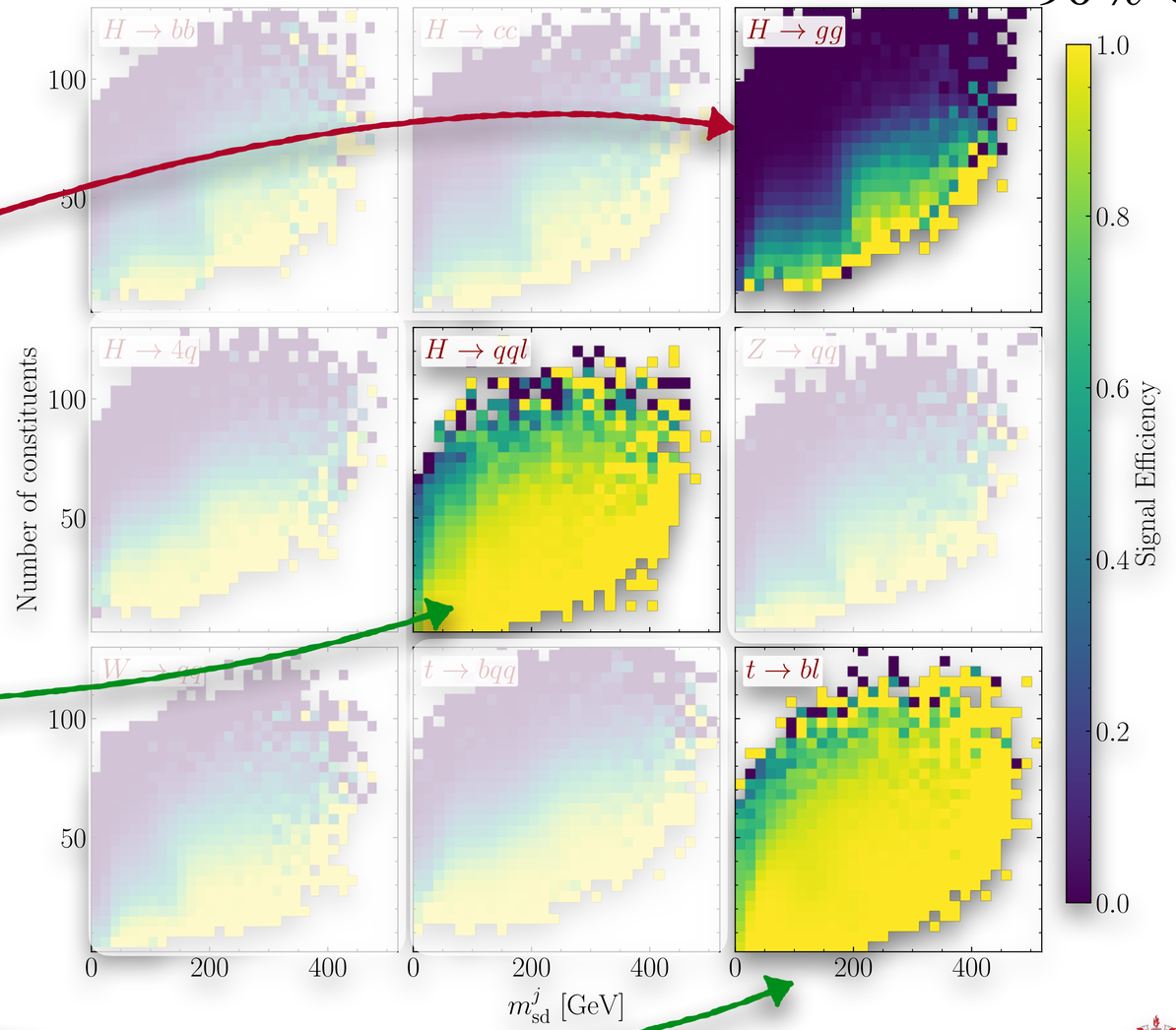




CP for Anomaly Detection

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



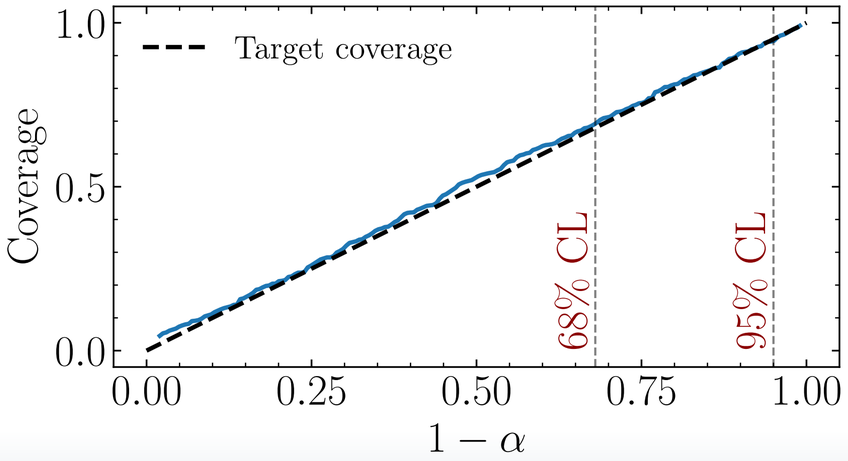


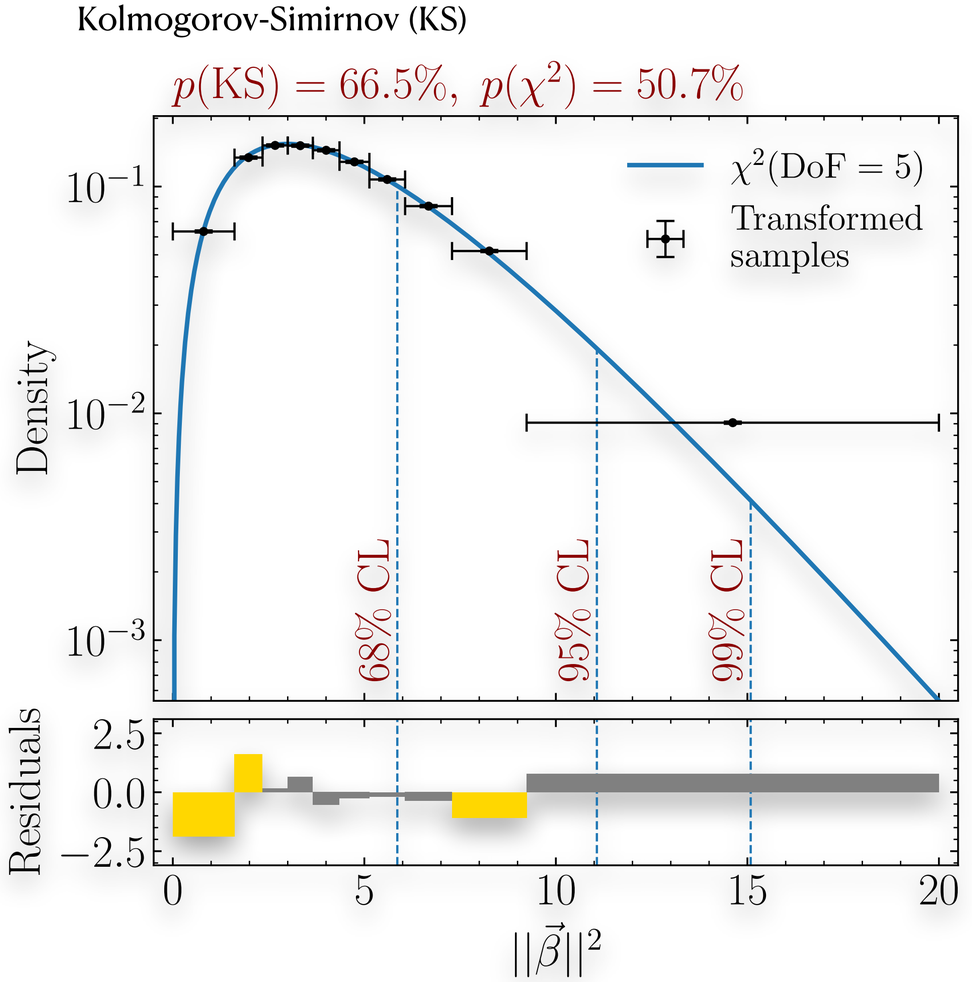
Conformal Prediction for Generative Modelling

Jack Y. Araz

CP for Generative Modelling

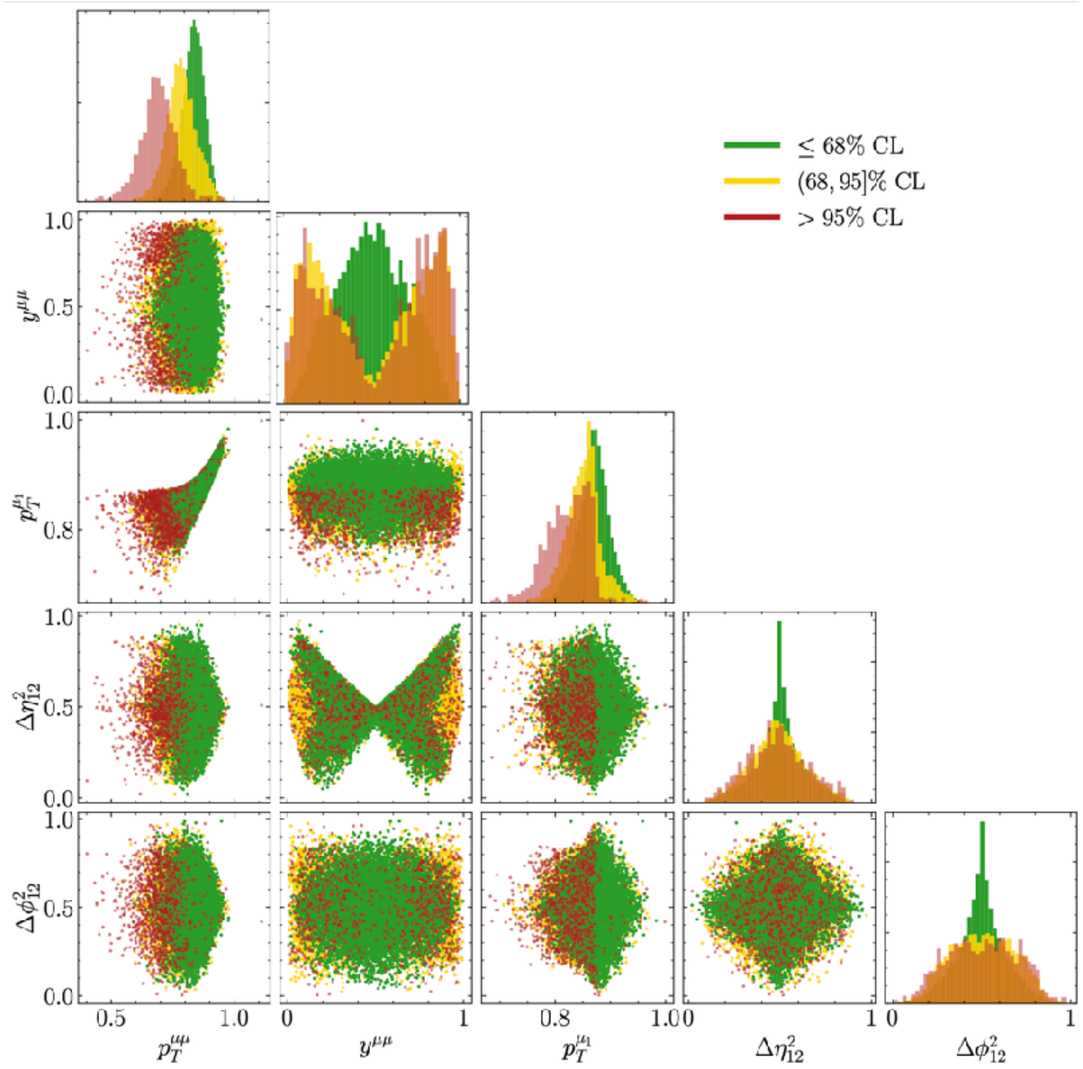
- ATLAS OmniFold 24-Dimensional $Z \to \mu\bar{\mu}$ + jets Open Data
- Dominant kinematic structure of the underlying $2 \to 3$ scattering: $\{p_T^{\mu\mu},\ y^{\mu\mu},\ p_T^{\mu_1},\ \Delta\eta_{\mu\mu},\ \Delta\phi_{\mu\mu}\}$
- Score: negative log-probability



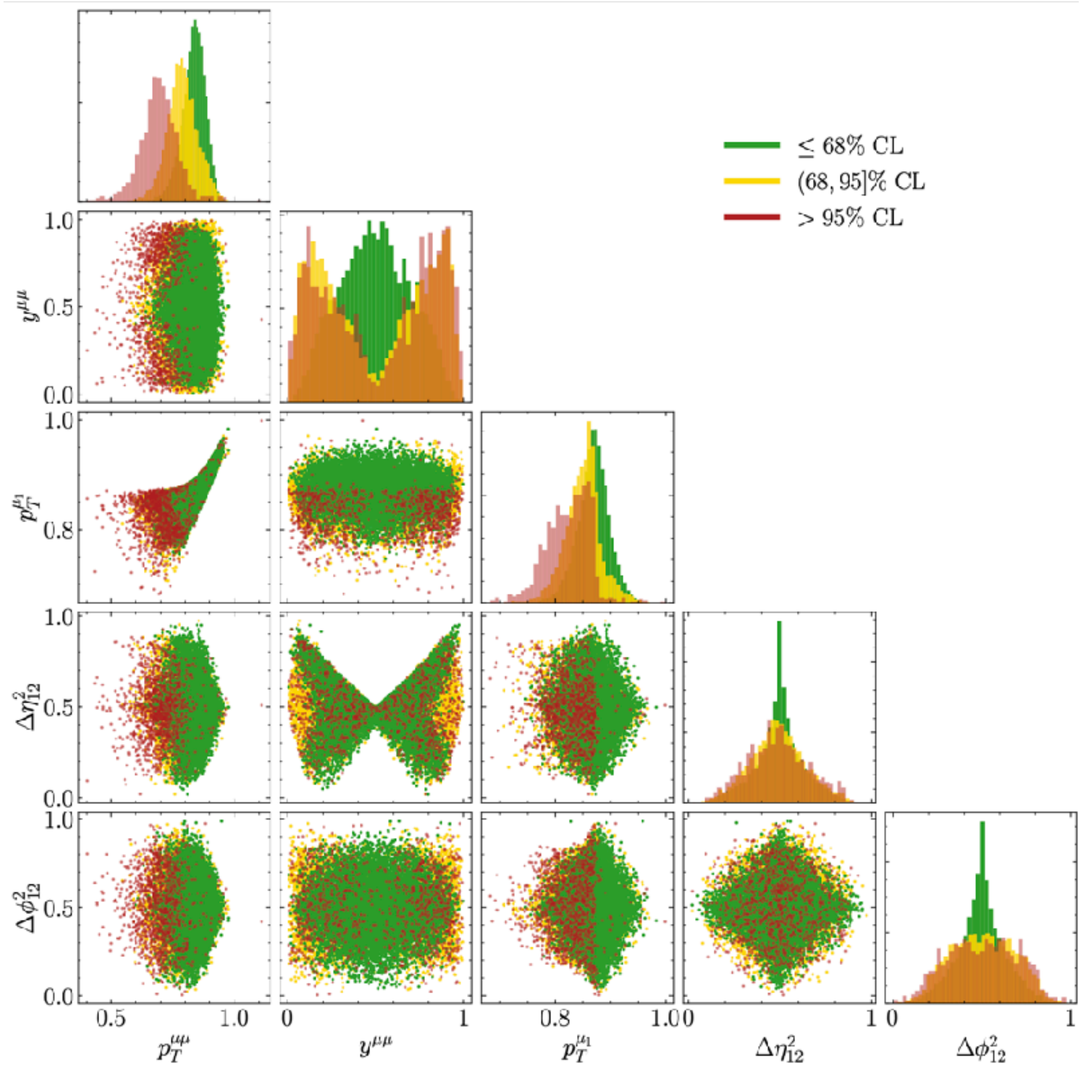


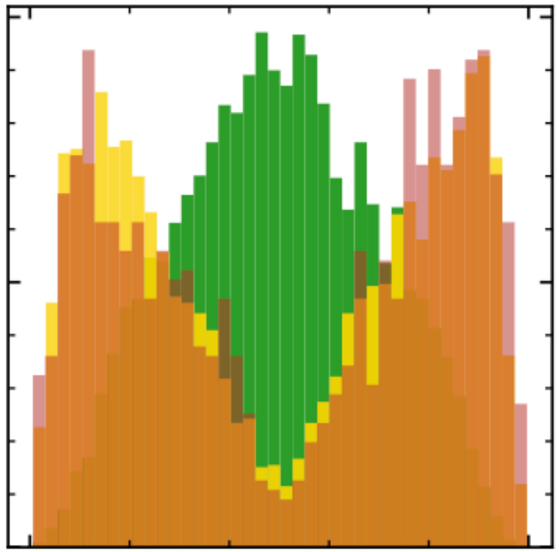
[JYA, Beck, Reboud, Dyk, Spannowsky; 2502.09494]

CP for Generative Modelling

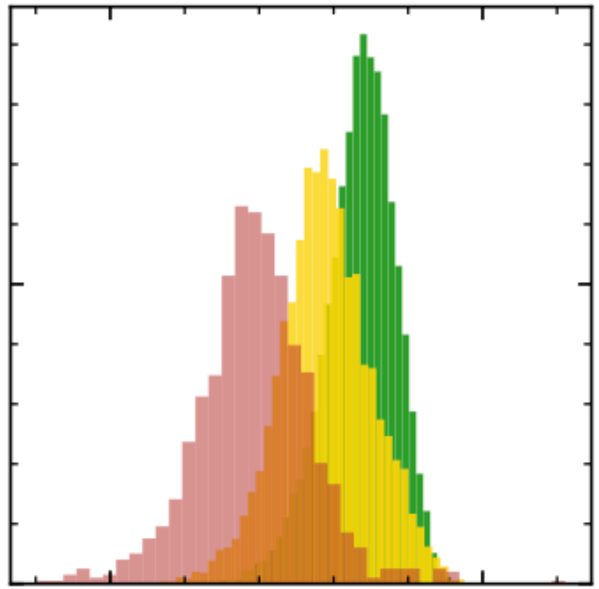


CP for Generative Modelling





$\tilde{y}^{\mu\mu}$

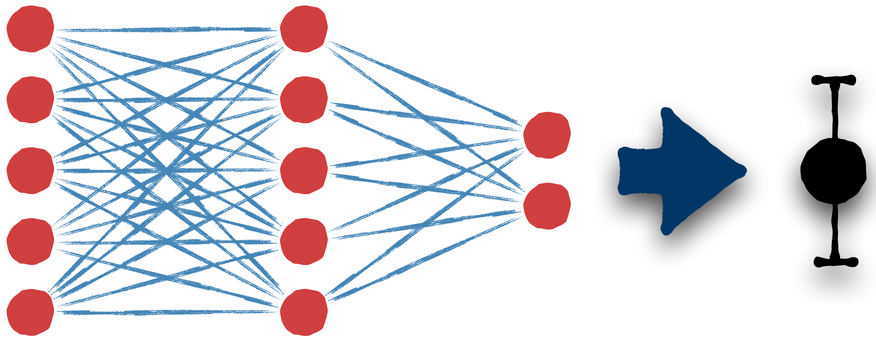


$\tilde{p}_T^{\mu\mu}$

Conclusion & Outlook

Conclusion & Outlook

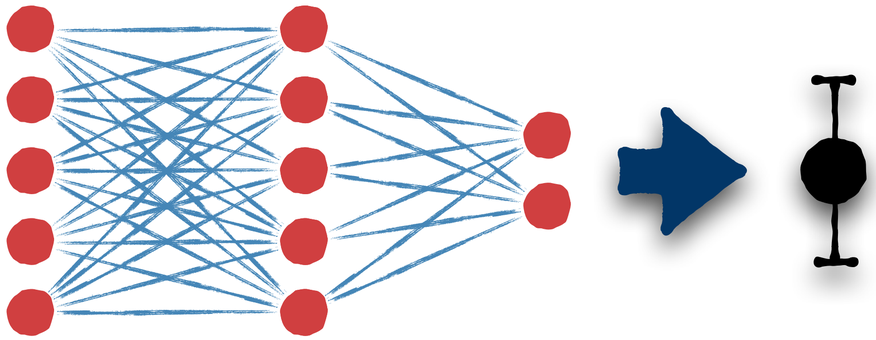
- Conformal prediction
  - Model agnostic
  - Distribution free
  - Coverage guarantee



Sponsors

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