Local Conformal Predictions for Calibrated Surrogates

Marginal coverage is what we have, conditional coverage is what we need.

Suprio Dubey[1,2], Henning Bahl[1], Anja Butter[1,3], Jürgen Hesser[2,4,5,6], Tilman Plehn[1,4,6]

[1]Institut für Theoretische Physik, Universität Heidelberg, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Universität Heidelberg, Germany
[3]LPNHE, Sorbonne Université, Université Paris Cité, CNRS/IN2P3, Paris, France
[4]Interdisciplinary Center for Scientific Computing (IWR), Universität Heidelberg, Germany
[5]Central Institute for Computer Engineering (ZITI), Universität Heidelberg, Germany

[6]CZS Heidelberg Initiative for Model-Based AI (MBAI), Universität Heidelberg, Germany

AIPHY Network — funded by the European Union

ML4Jets, September 14, 2026

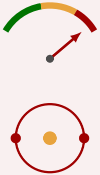
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

MIISM
MANNHEIM INSTITUTE FOR INTELLIGENT SYSTEMS IN MEDICINE

AIPHY

Funded by the European Union

Precision needs trustworthy surrogates

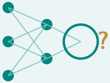


Demand

LHC/HL-LHC demands precision predictions, matching unprecedented experimental accuracy.

Computation

Higher-order amplitude calculations: accurate but computationally very expensive.



Trust Gap

ML surrogates: Fast but... can we trust them everywhere?

Three questions

Q1

Can we achieve calibrated uncertainties **without assumptions on the likelihood distribution**?

Q2

Can this calibration be applied as **post-processing after surrogate training**?

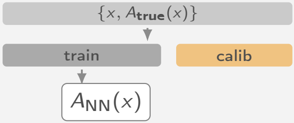
Q3

Can we get reliable uncertainties in challenging kinematic regions, not just on average?

Conformal prediction, the recipe

1. Training

what you would expect



train, then freeze; calib. set held out

Conformal prediction, the recipe

1. Training

what you would expect

$\{x, A_{\mathbf{true}}(x)\}$

train | calib

$A_{\mathrm{NN}}(x)$

train, then freeze; calib. set held out

2. Calibration

where the magic happens

$$R_i = \left|A_{\mathrm{true}}(x_i) - A_{\mathrm{NN}}(x_i)\right|$$

sorted: certain → uncertain

$$k_\alpha = \lceil (1-\alpha)(N{+}1) \rceil$$

α: *target miscoverage (0.32 → 68% CL)*

$$\boxed{q_\alpha = R_{\mathrm{calib},k_\alpha}}$$

$R(x)$: the **nonconformity score**, any function measuring how badly the surrogate missed the truth, here the absolute residual $|A_{\mathrm{true}}(x) - A_{\mathrm{NN}}(x)|$.

Conformal prediction, the recipe

1. Training

what you would expect

$\{x, A_{\mathbf{true}}(x)\}$

train | calib

$A_{\mathrm{NN}}(x)$

train, then freeze; calib. set held out

2. Calibration

where the magic happens

$$R_i = |A_{\mathrm{true}}(x_i) - A_{\mathrm{NN}}(x_i)|$$

sorted: certain → uncertain

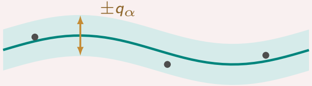
$$k_\alpha = \lceil (1-\alpha)(N{+}1) \rceil$$

α: *target miscoverage (0.32 → 68% CL)*

$$\boxed{q_\alpha = R_{\mathrm{calib},k_\alpha}}$$

3. Prediction

how you use the calibrated scores



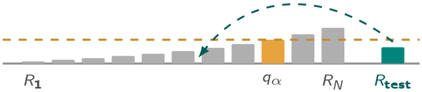
$$C_\alpha(x) = A_{\mathrm{NN}}(x) \pm q_\alpha$$

$$\boxed{P(A_{\mathrm{true}}(x) \in C_\alpha(x)) \geq 1-\alpha}$$

marginal guarantee under exchangeability

$R(x)$: the **nonconformity score**, any function measuring how badly the surrogate missed the truth, here the absolute residual $|A_{\mathrm{true}}(x) - A_{\mathrm{NN}}(x)|$.

Exchangeability



1. **Goal:** $P\left(A_{\text{true}}(x) \in C_\alpha(x)\right) \geq 1-\alpha$
2. **Construction:** $C_\alpha(x) = A_{\text{NN}}(x) \pm q \;\Rightarrow\; A_{\text{true}}(x) \in C_\alpha(x) \Leftrightarrow |A_{\text{true}}(x) - A_{\text{NN}}(x)| \leq q \Leftrightarrow R_{\text{test}} \leq q$
3. **Threshold:** $q = R_{\text{calib},k} \;\Rightarrow\; R_{\text{test}} \leq R_{\text{calib},k}$
4. **Chain:** $P\left(A_{\text{true}}(x) \in C_\alpha(x)\right) = P(R_{\text{test}} \leq q) = P\left(R_{\text{test}} \leq R_{\text{calib},k}\right)$
5. **Exchangeability:** $P\left(R_{\text{test}} \leq R_{\text{calib},k}\right) = \dfrac{k}{N+1}$
6. **Solve for** k**:** $\dfrac{k}{N+1} \geq 1-\alpha \;\Rightarrow\; k_\alpha = \lceil (1-\alpha)(N+1) \rceil$

finite-sample, distribution-free coverage guarantee: Vovk et al. 2005; Angelopoulos & Bates 2021

Part two: adaptive scores get partway

Heteroscedastic (Het)

$$R_\sigma(x) = \frac{|A_{\mathrm{true}}(x) - A_{\mathrm{NN}}(x)|}{\sigma(x)}$$

$$C_\alpha(x) = [A_{\mathrm{NN}}(x) - q_\alpha \sigma(x),\ A_{\mathrm{NN}}(x) + q_\alpha \sigma(x)]$$

width varies with x; intervals stay symmetric

Conformalized QR (CQR) [Romano et al. 2019]

$$R_{\mathrm{CQR}}(x) = \max[q_{\mathrm{lo}}(x) - A_{\mathrm{true}}(x),\ A_{\mathrm{true}}(x) - q_{\mathrm{hi}}(x)]$$

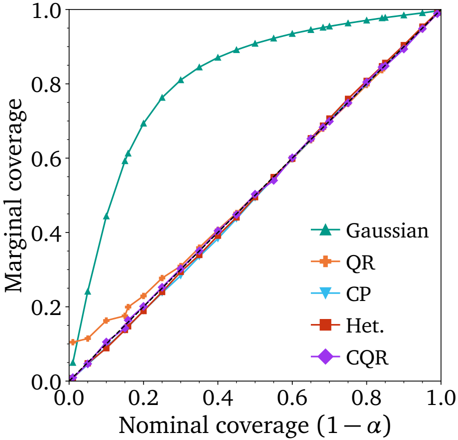
$$C_\alpha(x) = [q_{\mathrm{lo}}(x) - q_\alpha,\ q_{\mathrm{hi}}(x) + q_\alpha]$$

asymmetric; shape inherited from the learned quantile band

- Both depend on model accuracy, which degrades in sparse, challenging regions.

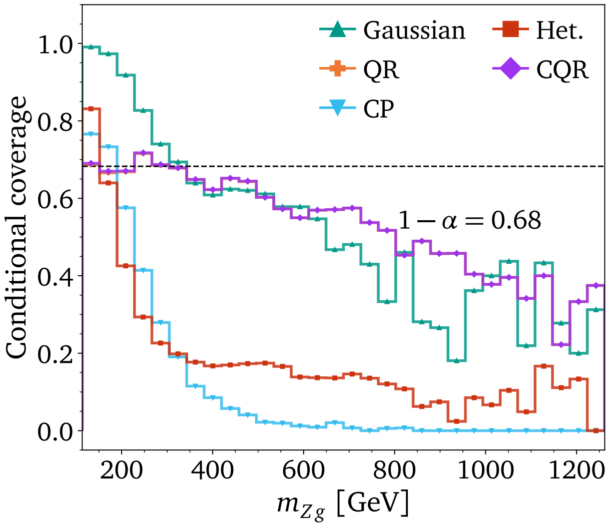
Marginal coverage is what we have, conditional coverage is what we need

- **Marginalization:** marginal coverage averages over all x, so over-coverage here can hide under-coverage there.
  $c_{1-\alpha}^{marg} = \int dx\, p(x)\, c_{1-\alpha}^{cond}(x)$
- **Blind spots:** high-energy tails, soft/collinear regions, and phase-space corners, exactly where event statistics are thin.
- **Non-Gaussian residuals:** no central-limit help at low multiplicity, so Gaussian likelihoods are mis-specified.
  [Bahl et al. 2026]



$q\bar{q} \to Zg$

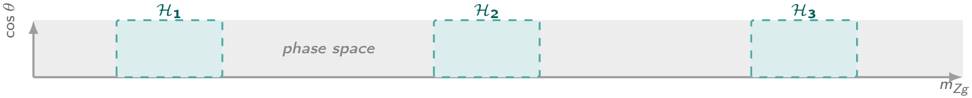
Marginal coverage is what we have, conditional coverage is what we need



- CP and Het collapse toward zero above $\sim$400 GeV
- Gaussian and CQR drift down into the 0.2–0.4 range
- No baseline method holds the 0.68 target conditionally

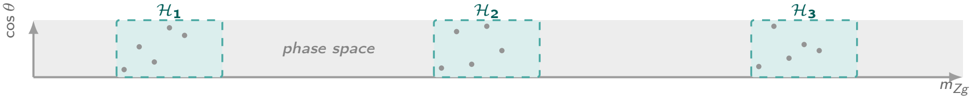
FALCON, field-adapted local conformal prediction

1 · Probe regions



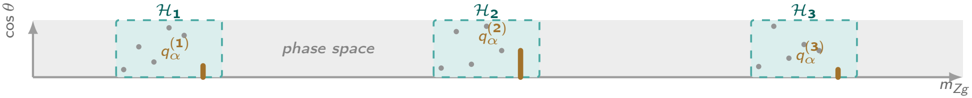
FALCON, field-adapted local conformal prediction

1 · Probe regions → 2 · Probe & evaluate



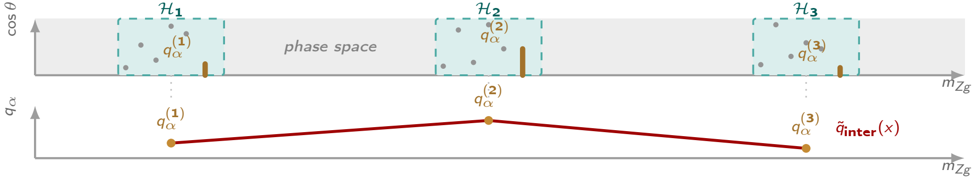
FALCON, field-adapted local conformal prediction

1 · Probe regions → 2 · Probe & evaluate → 3 · Local threshold



FALCON, field-adapted local conformal prediction

1 · Probe regions → 2 · Probe & evaluate → 3 · Local threshold → 4 · Interpolate



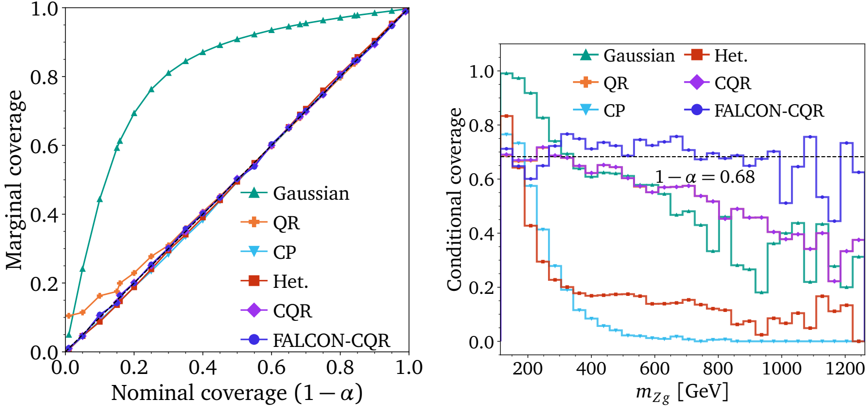
- Naively substituting $q_\alpha \to \tilde{q}_{\text{inter}}(x)$ **breaks the guarantee**: the quantile must be a fixed scalar, not a function of x.
- Instead, $\tilde{q}_{\text{inter}}$ is **absorbed into the nonconformity score**, so the ordinary quantile step runs unchanged.

$$R'_{\text{CQR}}(x) = R_{\text{CQR}}(x) - \tilde{q}_{\text{inter}}(x)$$

$$C(x) = [q_{\text{lo}} - \tilde{q}_{\text{inter}} - q_\alpha,\ q_{\text{hi}} + \tilde{q}_{\text{inter}} + q_\alpha]$$

FALCON-CQR

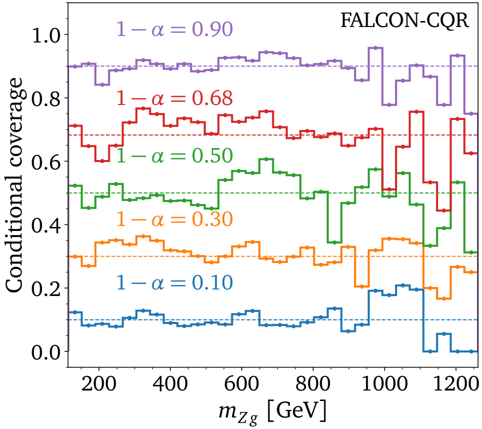
Results, FALCON improves conditional coverage



Left: marginal coverage vs. nominal. Right: conditional 68% CL coverage vs. m_{Zg}, no noise (6 methods).

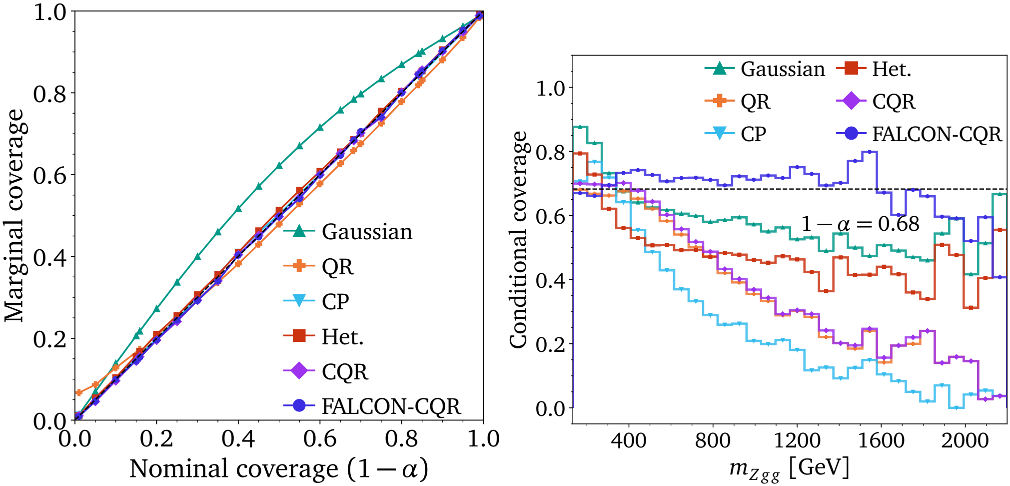
- Marginally, all the conformal prediction methods sit on the diagonal
- Conditionally: CP and Het collapse toward zero above $\sim$400 GeV; Gaussian and CQR drift down into the 0.2–0.4 range
- **FALCON-CQR** is the only method fluctuating around the 0.68 target across the full m_{Zg} range

FALCON at different confidence levels, $q\bar{q} \to Zg$



Conditional coverage from FALCON-CQR for different CLs vs. m_{Zg}, no artificial noise. ***Question answered:*** *does FALCON hold at CLs other than 68%?*

Scaling to $q\bar{q} \to Zgg$



Scaling to $q\bar{q} \to Zgg$: FALCON tracks nominal coverage closely as multiplicity grows. ***Question:*** *does local calibration survive a higher-dimensional phase space?*

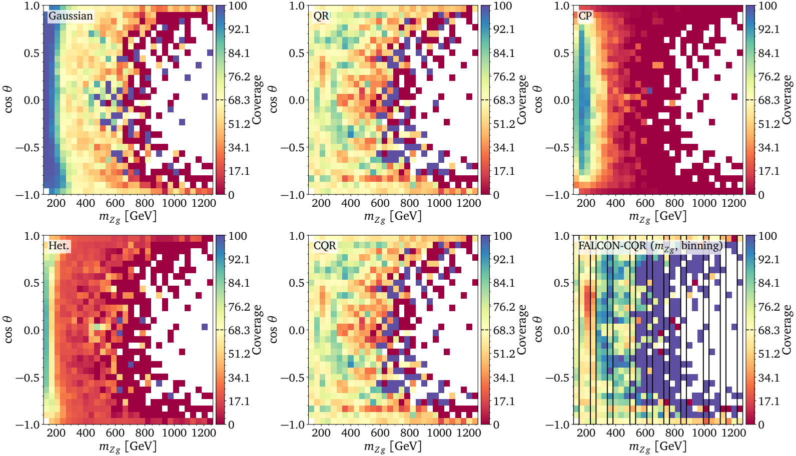
Summary

- ✓ **Q1 Distribution-free**: marginal guarantee, no likelihood assumed
- ✓ **Q2 Post-hoc**: no retraining, no architecture change
- ✓ **Q3 FALCON** substantially improves conditional coverage via probe-field interpolation, empirically, not a new formal guarantee

Thank you!

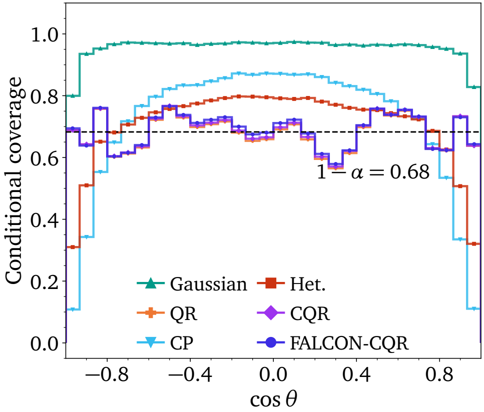
arXiv:2607.01354

Backup: 2D phase-space coverage



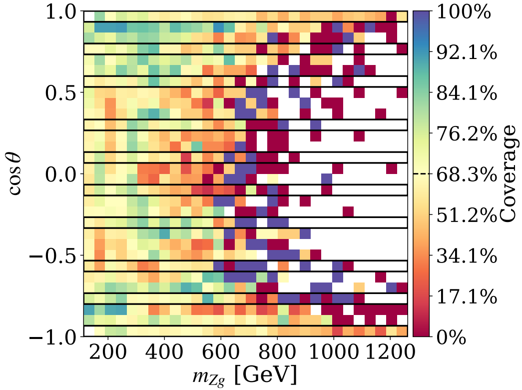
68% CL coverage in the $(m_{Zg}, \cos\theta^*)$ plane, no noise (black lines: FALCON probe windows). ***Question:*** *does coverage hold across the full 2D phase space, not just along m_{Zg}?*

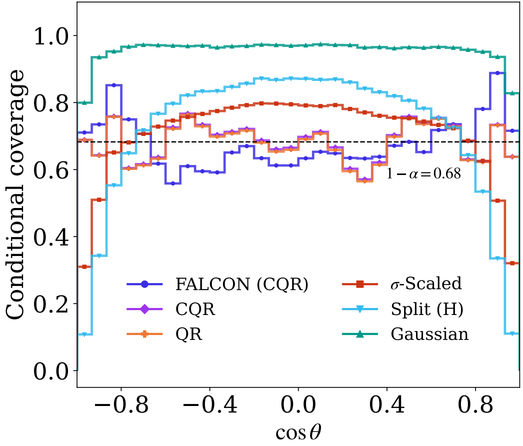
Backup: does FALCON miscalibrate the orthogonal direction?



Conditional 68% CL coverage for $q\bar{q} \to Zg$ vs. $\cos\theta^*$; FALCON is calibrated at 10 evenly spaced probe windows along m_{Zg}. ***Question:*** ***does calibrating along*** ***m_{Zg}*** ***accidentally miscalibrate the orthogonal*** ***$\cos\theta^*$*** ***direction?***

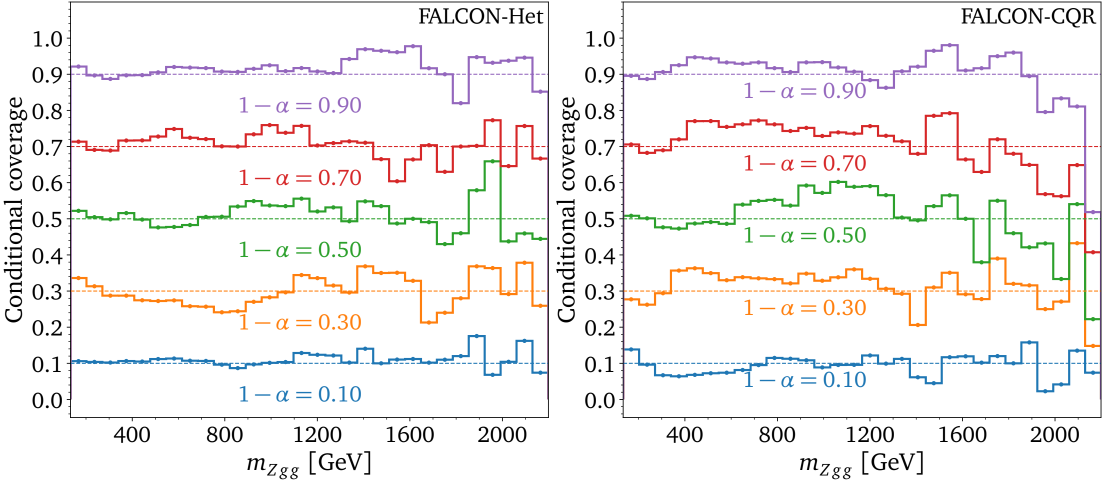
Backup: FALCON with 1D probes along $\cos\theta^*$





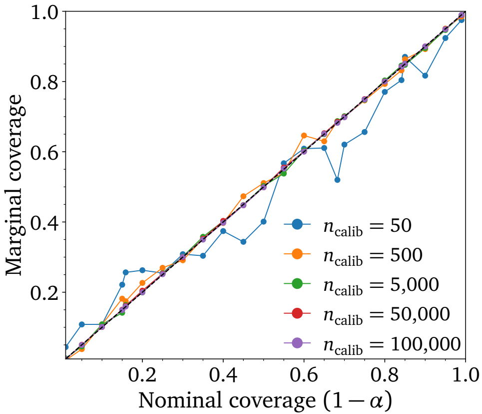
Left: probe windows are bins in $\cos\theta^*$, spanning the full m_{Zg} range (axes swapped relative to the main scheme). Right: resulting conditional 68% CL coverage vs. $\cos\theta^*$.

Backup: FALCON at different confidence levels, $q\bar{q} \to Zgg$



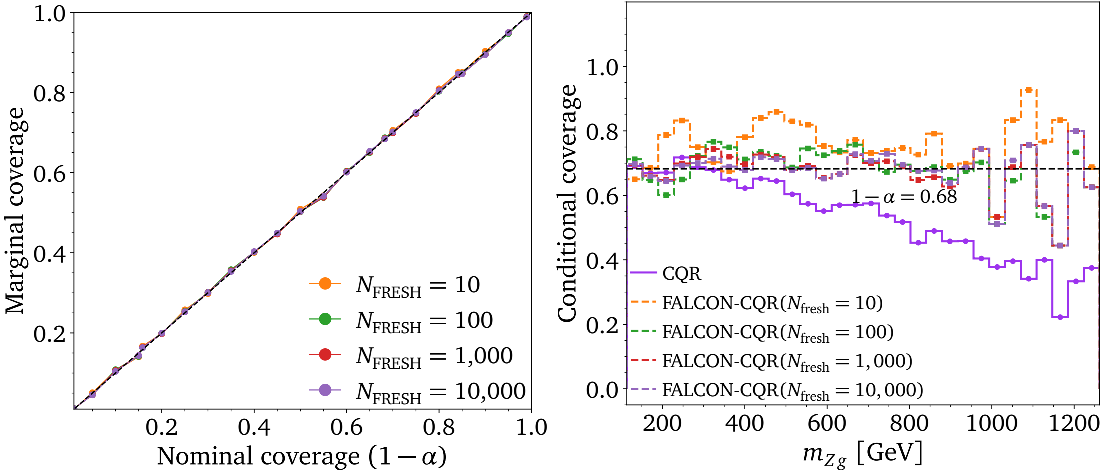
Conditional coverage from FALCON for different CLs vs. m_{Zgg}, no artificial noise. Left: FALCON-Het. Right: FALCON-CQR. ***Question answered:*** *does FALCON hold at CLs other than 68%, for the higher-multiplicity process?*

Backup: sensitivity to calibration-set size



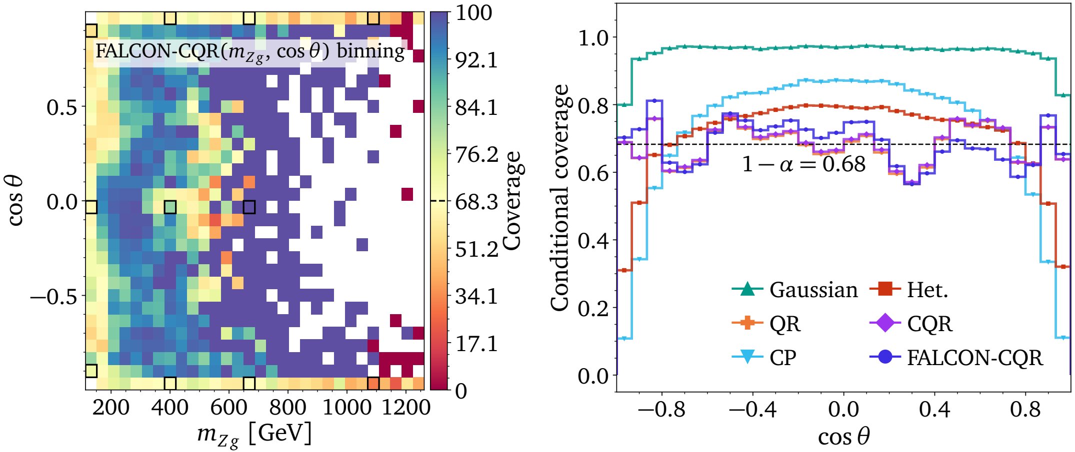
Marginal vs. nominal coverage for FALCON-CQR on $q\bar{q} \to Zg$, varying $n_{\mathbf{calib}}$ with $n_{\mathbf{fresh}} = 100$ fixed; dashed diagonal is ideal. ***Question:*** *how sensitive is FALCON to calibration-set size?*

Backup: how many probe points does FALCON need?



FALCON-CQR on $q\bar{q} \to Zg$ varying $n_{\textit{fresh}}$ with $n_{\textit{calib}} = 5000$ fixed. Left: marginal vs. nominal. Right: conditional 68% CL vs. m_{Zg}. ***Question answered:*** *how many probe points per hotspot ($n_{\textit{fresh}}$) does FALCON actually need?*

Backup: why not a 2D $(m_{Zg}, \cos\theta^*)$ probe scheme?



Left: alternative probe scheme, 11 small boxes in $(m_{Zg}, \cos\theta^*)$ instead of m_{Zg}-only strips. Right: resulting conditional 68% CL coverage vs. $\cos\theta^*$.

***Question:** does probing along $\cos\theta^*$ too improve further?*