UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

AITP

AI and Theoretical Physics

NAE, Statistically

Ranit Das **Jonathan Ostertag-Henning** Tilman Plehn Lorenz Vogel

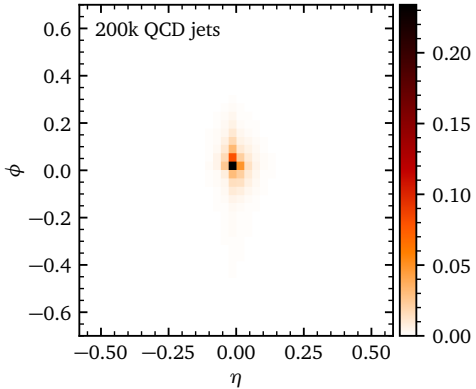
Institut für Theoretische Physik, Universität Heidelberg

ML4Jets 2026 — Vienna

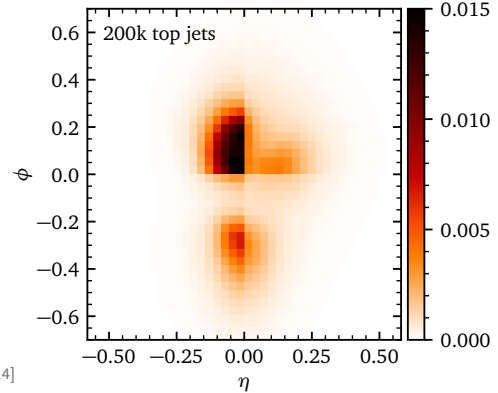
Why Anomaly Scores Need Statistics

Model-agnostic anomaly detection: score events without a signal hypothesis — but the score itself carries no statistical meaning

Lower complexity



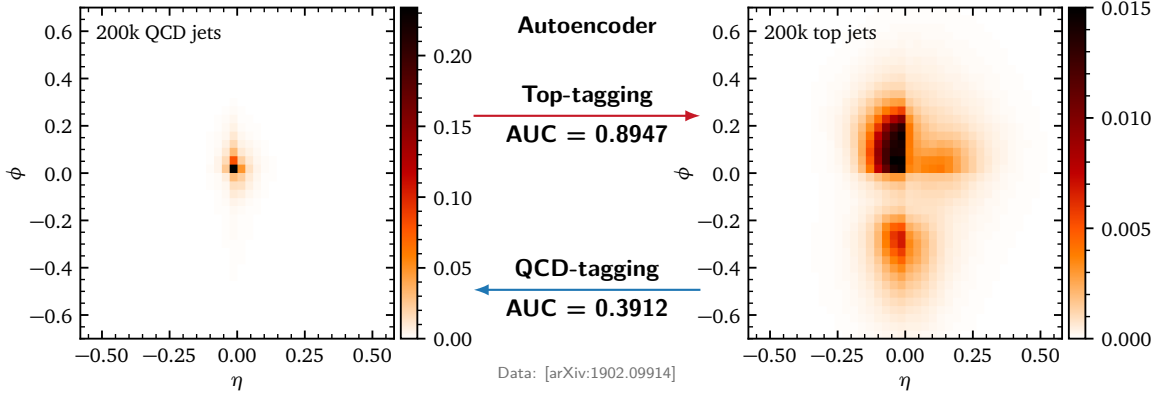
Higher complexity



Data: [arXiv:1902.09914]

Why Anomaly Scores Need Statistics

Model-agnostic anomaly detection: score events without a signal hypothesis — but the score itself carries no statistical meaning



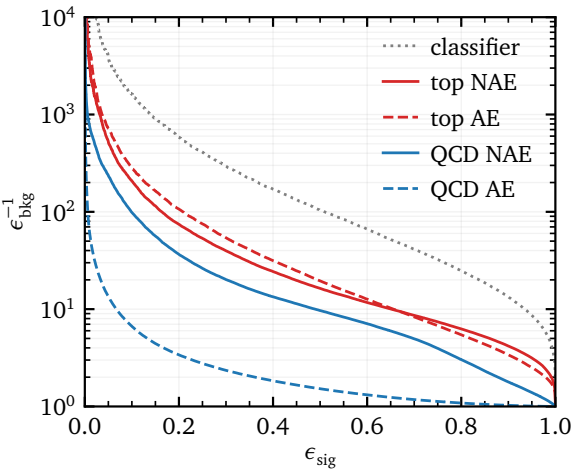
Known failure mode: reconstruction error tracks *compressibility*, not anomalousness

Does It Work? Jet Tagging Performance

Tagging performance:
anomaly-detection AUC

	Top-tagging	QCD-tagging
AE	0.8947	0.3912
NAE	0.9036	0.8071

NAE for LHC triggers: [arXiv:2206.14225]



Learning An Uncertainty

Bayesian Neural Network: weights are distributions, not fixed values

$$w \sim q_\phi(w) \qquad \text{mean-field variational posterior}$$

BNAE: only the final decoder layer is Bayesian for increased training stability

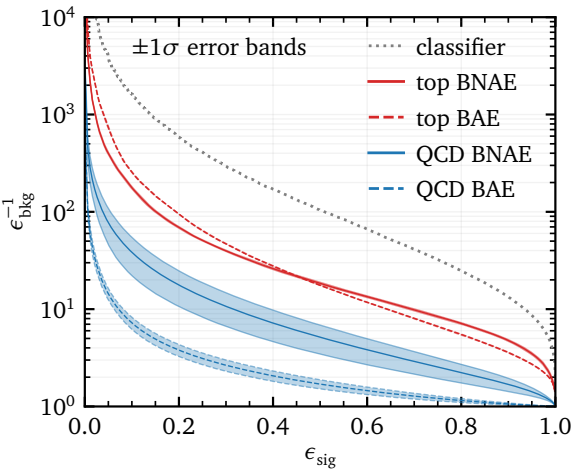
Sampling: draw N weight samples $w^{(i)} \sim q_\phi(w) \Rightarrow$ distribution over score, not a single point estimate

Does It Work? Jet Tagging Performance

Tagging performance: anomaly-detection AUC, $\pm 1\sigma$ from 50 MC samples

	Top-tagging	QCD-tagging
BAE	0.896 ± 0.001	0.432 ± 0.037
BNAE	0.913 ± 0.002	0.729 ± 0.054

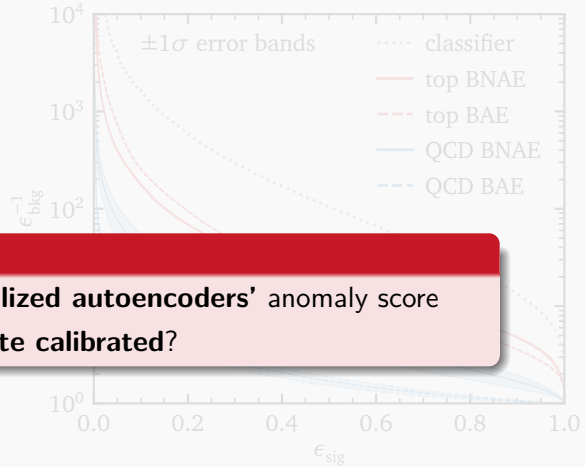
NAE for LHC triggers: [arXiv:2206.14225]



Does It Work? Jet Tagging Performance

Tagging performance: anomaly-detection AUC, $\pm 1\sigma$ from 50 MC samples

	Top-tagging	QCD-tagging
BAE	0.896 ± 0.001	0.432 ± 0.037
BNAE		
NAE for		



Two questions

1. Statistical interpretation of **normalized autoencoders'** anomaly score
2. Is the learned **uncertainty estimate calibrated**?

From Reconstruction Error to Likelihood

Energy-based network: Boltzmann density

$$p_\theta(x) = \frac{e^{-E_\theta(x)/T}}{Z_\theta} \qquad Z_\theta = \int \mathrm{d}x\, e^{-E_\theta(x)/T}$$

NAE: reconstruction error = energy

$$E_\theta(x) = |x - x'|^2$$

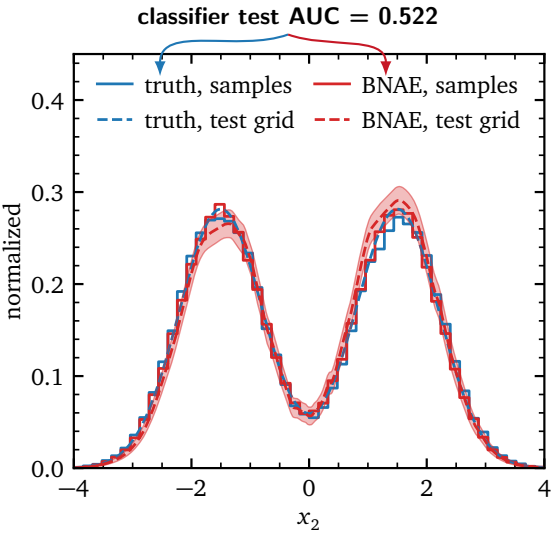
Loss:

$$\mathcal{L} = E_\theta(x_{data}) - E_\theta(x_{NAE})$$

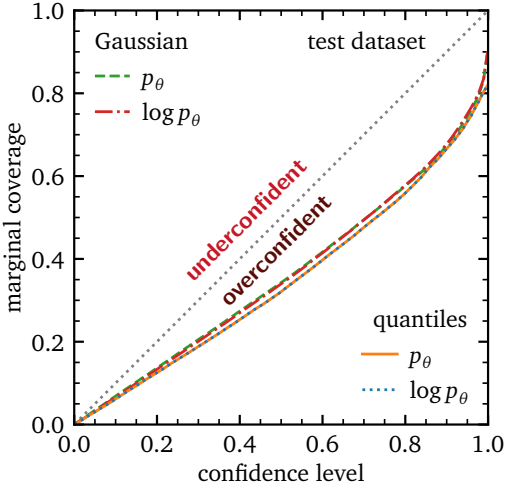
- $p_\theta = p_{\text{data}}$ at minimum

$\Rightarrow$ score = likelihood

BNAE: + uncertainty estimate



Uncertainty Calibration



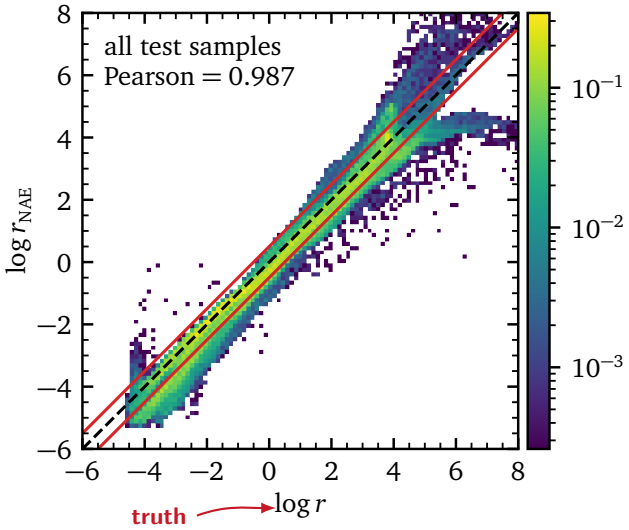
The Dual-NAE Test

Problem: no known p_{true} for real data

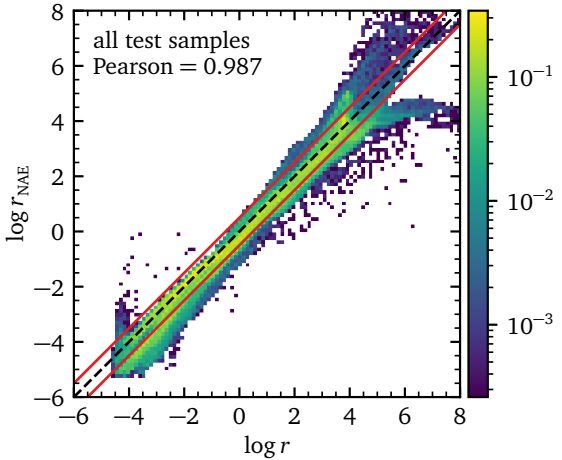
Idea: two NAEs on p_1, p_2 $\Rightarrow$ ratio

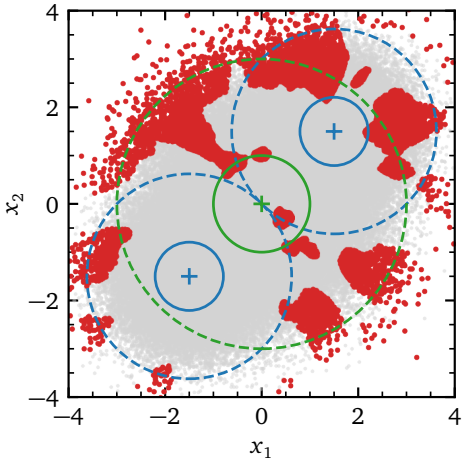
$$\log r_{\text{NAE}} \equiv \log \frac{p_{\theta,1}(x)}{p_{\theta,2}(x)} = \frac{E_{\theta,2}(x)}{T_2} - \frac{E_{\theta,1}(x)}{T_1} + C_\theta$$

C_θ: unknown constant, doesn't affect ranking



Where We Differ





The Dual-NAE Test

Problem: no known p_{true} for real data

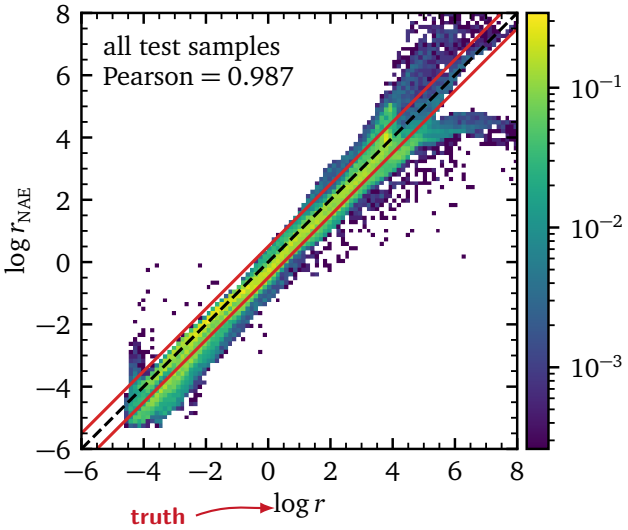
Idea: two NAEs on p_1, p_2 $\Rightarrow$ ratio

$$\log r_{\text{NAE}} \equiv \log \frac{p_{\theta,1}(x)}{p_{\theta,2}(x)} = \frac{E_{\theta,2}(x)}{T_2} - \frac{E_{\theta,1}(x)}{T_1} + C_\theta$$

C_θ: unknown constant, doesn't affect ranking

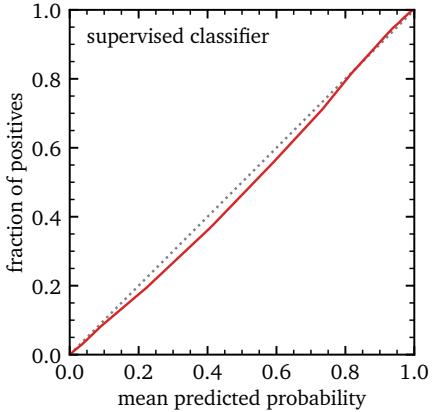
Check: cross-validate against a classifier

Jet tagging: p_1, p_2 = QCD-, top-trained NAEs



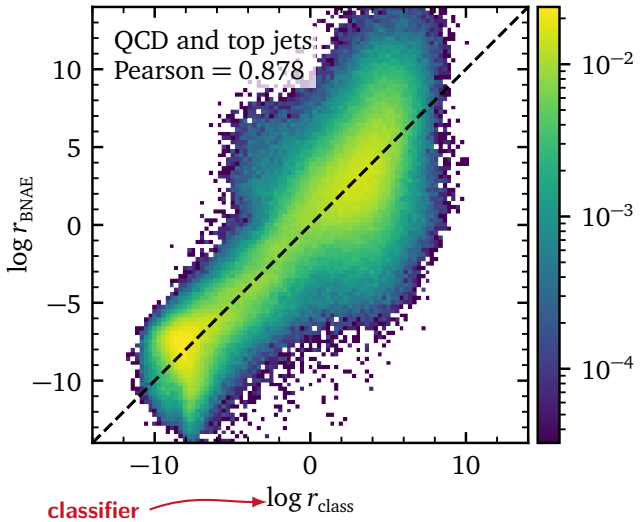
Is the Classifier Trustworthy?

Why: classifier ratio = ground-truth stand-in — needs to be calibrated itself

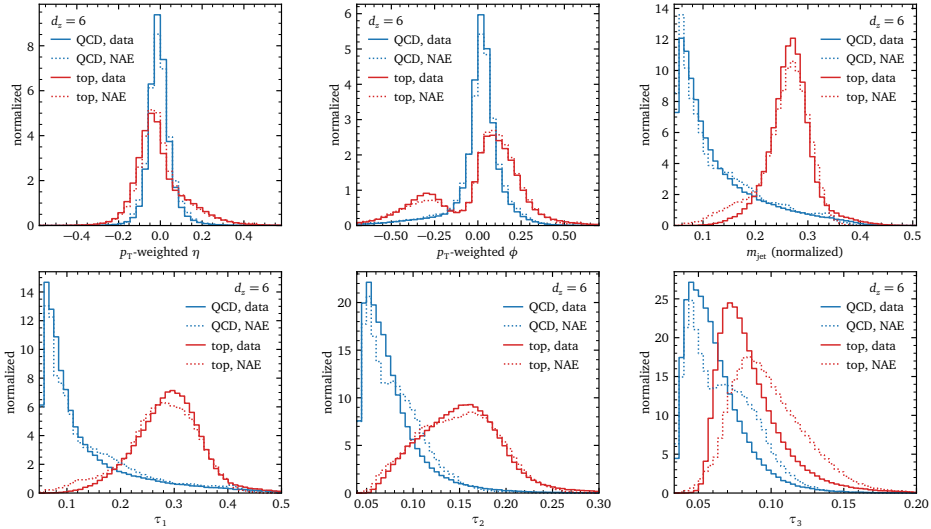


Jet Tagging: Statistical Interpretation

Same test, real jets: BNAE ratio vs. classifier ratio



Do the Generated Samples Look Like Real Jets?



Conclusion

- NAE score $\Rightarrow$ **likelihood**
- BNAE $\Rightarrow$ + **calibrated uncertainty**

arXiv:2608.19317

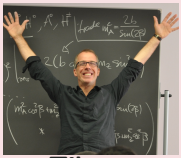
Thank you! Questions welcome.

Find Us

Lorenz

Ranit

Jonathan



Tilman

Backup

Backup: Pull Distributions