

Autoencoder-based anomaly detection and the normalized autoencoder (NAE)

Lorenz Vogel

20 October 2026

Institute for Theoretical Physics
Heidelberg University, Germany

IRN Terascale Fall Meeting 2026, Heidelberg University

NAE, *Statistically*, arXiv:2608.19317 [hep-ph]

R. Das, J. Ostertag-Henning, T. Plehn, and L. Vogel



**UNIVERSITÄT
HEIDELBERG**
ZUKUNFT
SEIT 1386

Searching for BSM physics



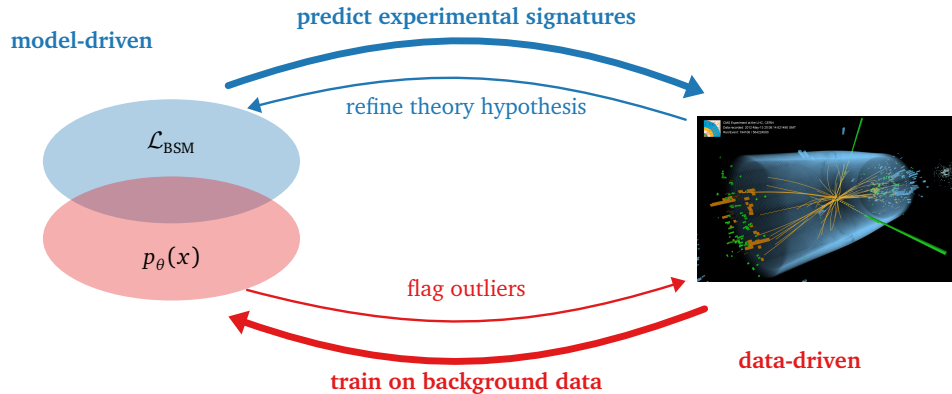
[arXiv:2211.01421]

three generations of matter (fermions)			interactions / forces (bosons)	
	I	II	III	
mass charge spin	$\approx 2.16 \text{ MeV}$ $+\frac{2}{3}$ $\frac{1}{2}$ u up	$\approx 1.27 \text{ GeV}$ $+\frac{2}{3}$ $\frac{1}{2}$ c charm	$\approx 173 \text{ GeV}$ $+\frac{2}{3}$ $\frac{1}{2}$ t top	0 0 1 g gluon
QUARKS	$\approx 4.7 \text{ MeV}$ $-\frac{1}{3}$ $\frac{1}{2}$ d down	$\approx 94 \text{ MeV}$ $-\frac{1}{3}$ $\frac{1}{2}$ s strange	$\approx 4.18 \text{ GeV}$ $-\frac{1}{3}$ $\frac{1}{2}$ b bottom	0 0 1 γ photon
	$\approx 0.511 \text{ MeV}$ -1 $\frac{1}{2}$ e electron	$\approx 106 \text{ MeV}$ -1 $\frac{1}{2}$ μ muon	$\approx 1.78 \text{ GeV}$ -1 $\frac{1}{2}$ τ tau	$\approx 80.4 \text{ GeV}$ ± 1 1 W W boson
LEPTONS	$\approx 0.8 \text{ eV}$ 0 $\frac{1}{2}$ ν_e electron neutrino	$< 0.17 \text{ MeV}$ 0 $\frac{1}{2}$ ν_μ muon neutrino	$< 18.2 \text{ MeV}$ 0 $\frac{1}{2}$ ν_τ tau neutrino	$\approx 91.2 \text{ GeV}$ 0 1 Z Z boson
				SCALAR BOSONS VECTOR BOSONS

- **Standard Model (SM)** is remarkably successful — but incomplete
- experiments have confirmed SM predictions to great precision, yet many fundamental questions remain
- **observational problems, conceptual shortcomings, and theoretical puzzles**
 - cosmic enigmas: dark matter, dark energy, matter–antimatter asymmetry
 - conceptual questions: origin of EWSB, hierarchy problem, flavor puzzle, strong CP problem, gauge coupling unification, gravity, ...
- goal of LHC: **search for signatures of new physics**

BSM searches — model-driven versus data-driven

[arXiv:2211.01421]

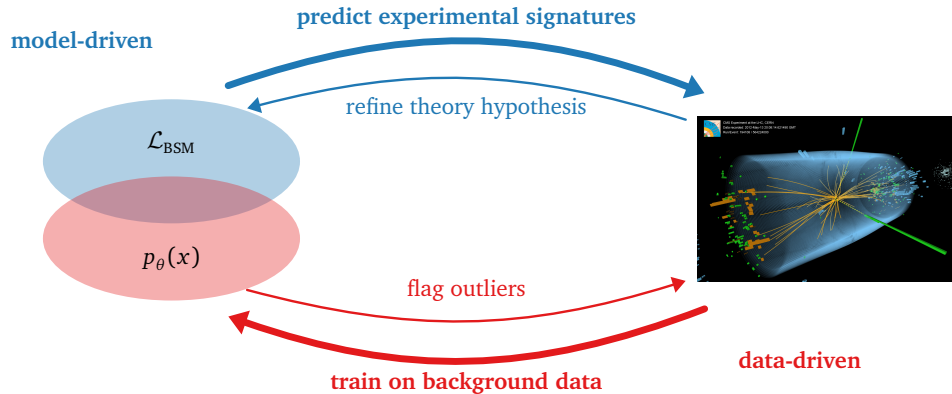


traditional approaches start from a **pre-defined theory hypothesis** and compare its predicted experimental signatures to data → targeted searches are **limited and inefficient**

BSM searches — model-driven versus data-driven

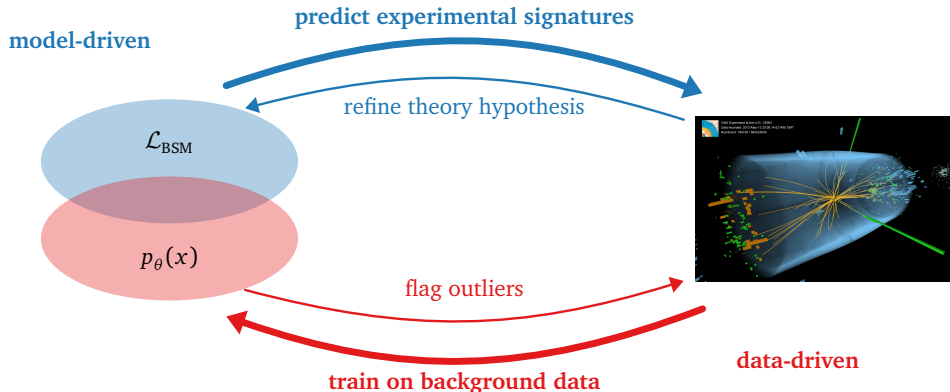


[arXiv:1808.08979, arXiv:1808.08992, arXiv:2010.14554, arXiv:2104.09051, arXiv:2202.00686]

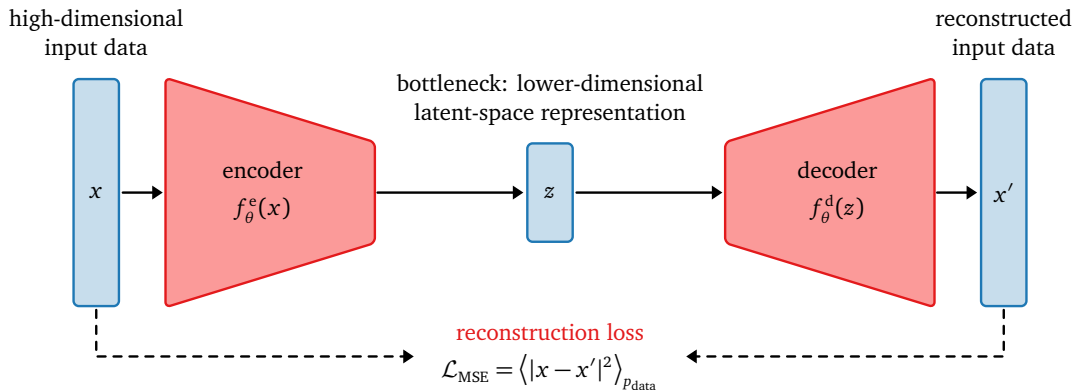


unsupervised anomaly detection looks for rare events that differ from bulk of data, without committing to a signal hypothesis (**outlier detection**)

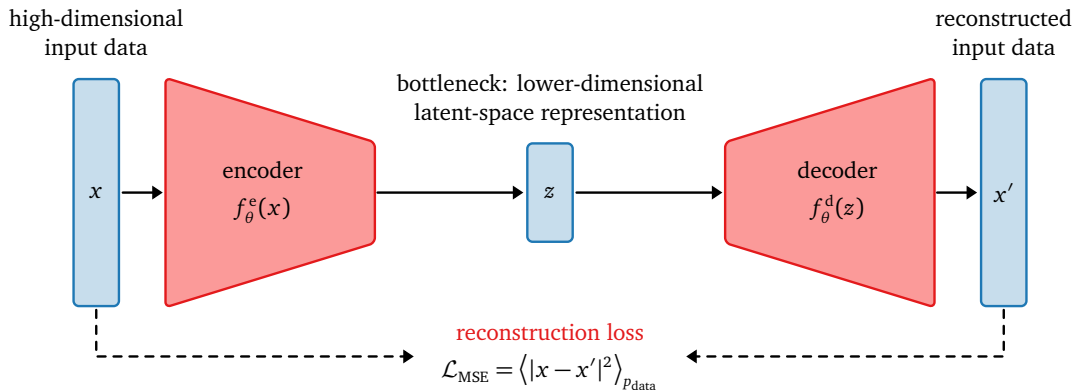
NAE — two main questions



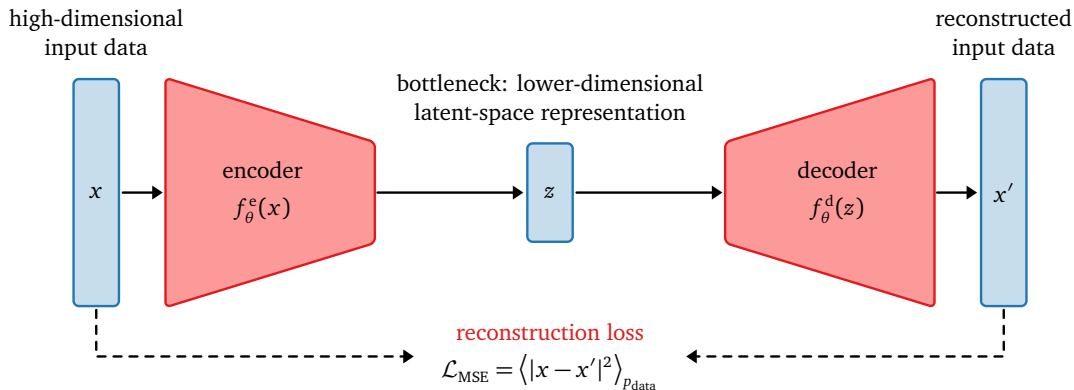
What is **statistical interpretation of anomaly score** of a normalized autoencoder?
Can we extract anomaly score with an **uncertainty estimate**?



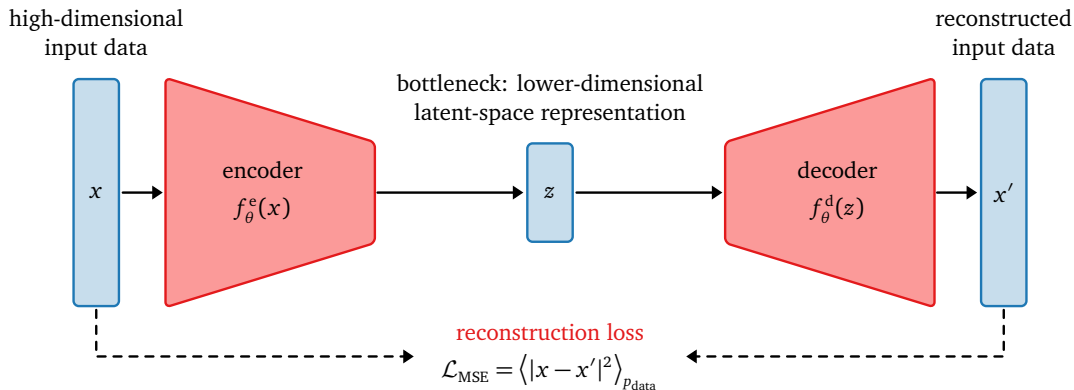
ML-based unsupervised anomaly detection often relies on **autoencoders**:
compress input x into latent space z and decode into reconstruction x'



information bottleneck prevents AE from learning identity mapping,
instead AE learns intrinsic features of training data



expose AE only to **non-anomalous (background)** events during **network training**,
AE learns to compress and reconstruct them well (**small reconstruction error**)

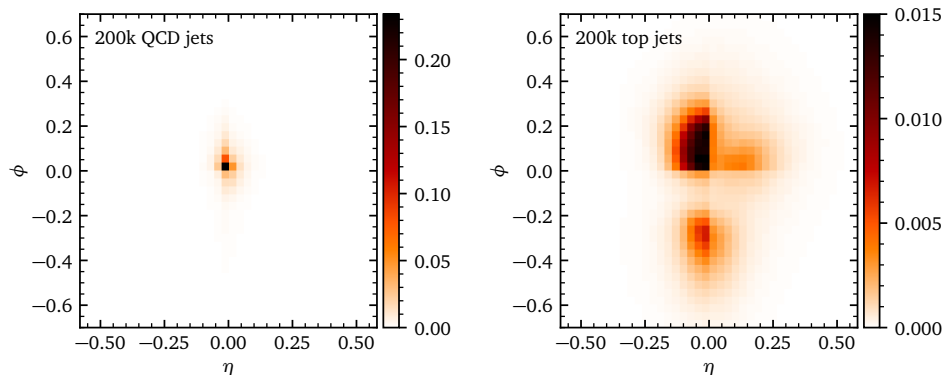


anomalous (signal) events are not part of training data, AE is expected to reconstruct them poorly: **reconstruction error serves as anomaly score**
→ but without a proper **statistical interpretation...**

Top-tagging reference dataset and jet images



[arXiv:1707.08966, arXiv:1808.08979, arXiv:1902.09914, arXiv:2107.00656]

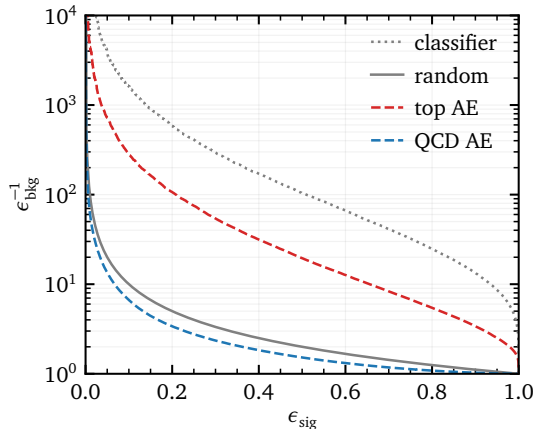


widely used and well-established benchmark for jet tagging and anomaly detection:
QCD jets versus jets from boosted hadronic top-quark decays ($t \rightarrow bW^+ \rightarrow bq\bar{q}'$)

AE — inverse training and complexity bias



[arXiv:1902.09914, arXiv:2104.09051, arXiv:2110.06948]

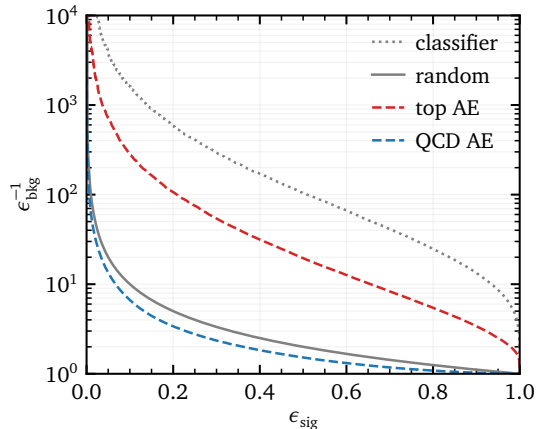


- single-pronged QCD jets versus three-pronged top jets
- standard task is **top-tagging**: train on QCD jets, top jets are anomalous
- **inverse training** is simple test for model-independence
- **QCD-tagging** as robustness test: train on top jets, QCD jets are anomalous
- **ROC curves** show performance: inverse background efficiency (1/FPR) versus signal efficiency (TPR)

AE — inverse training and complexity bias



[arXiv:1902.09914, arXiv:2104.09051, arXiv:2110.06948]



- top-trained network is able to interpolate simpler QCD features
- **MSE is not a fail-proof anomaly score:** reconstruction error tracks complexity rather than abnormality (**complexity bias**)
- MSE can separate an intrinsically high-dimensional (more complex) signal from a low-dimensional (less complex) background, but not vice versa
- training objective does not penalize good reconstruction outside of training data support



[arXiv:2101.03288, arXiv:2105.05735, arXiv:2206.14225, arXiv:2510.02168, arXiv:2608.19317]

- energy-based network assigns **scalar energy** to each phase-space point

$$E_{\theta}(x) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$$

- learned energy defines a learned probability density through **Boltzmann distribution**

$$p_{\theta}(x) = \frac{1}{Z_{\theta}} \exp\left(-\frac{E_{\theta}(x)}{T}\right)$$

- **partition function** Z_{θ} ensures that $p_{\theta}(x)$ is properly normalized over x

$$Z_{\theta} = \int dx \exp\left(-\frac{E_{\theta}(x)}{T}\right)$$

Energy-based network — loss function



[arXiv:2101.03288, arXiv:2105.05735, arXiv:2206.14225, arXiv:2510.02168, arXiv:2608.19317]

- train by minimizing **negative log-likelihood**

$$\mathcal{L}(x|\theta) = -\log p_{\theta}(x) = \frac{E_{\theta}(x)}{T} + \underbrace{\log Z_{\theta}}_{\text{intractable}}$$

- gradient splits into two terms, training becomes **min-max problem**,
at minimum $p_{\theta}(x) = p_{\text{data}}(x)$

$$\langle \nabla_{\theta} \mathcal{L}(x|\theta) \rangle_{p_{\text{data}}} = \frac{1}{T} \underbrace{\langle \nabla_{\theta} E_{\theta}(x) \rangle_{p_{\text{data}}}}_{\text{positive energy}} - \frac{1}{T} \underbrace{\langle \nabla_{\theta} E_{\theta}(x) \rangle_{p_{\theta}}}_{\text{negative energy}}$$

- positive energy (**pushed down**): evaluated with automatic differentiation on training data
- negative energy (**pushed up**): requires samples from $p_{\theta}(x)$, generated with **Langevin Monte Carlo**
- loss reduces energy on $p_{\text{data}}(x)$ while accounting for normalization of $p_{\theta}(x)$
→ NAE explicitly penalizes well-reconstructed regions not covered by training dataset



[arXiv:2105.05735, arXiv:2206.14225, arXiv:2510.02168, arXiv:2608.19317]

- combine bottleneck-based AE with energy-based description:

MSE reconstruction error is energy

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

- architecture is same as standard autoencoder, but with energy-based training
- NAE anomaly score $E_{\theta}(x)$ is related to likelihood

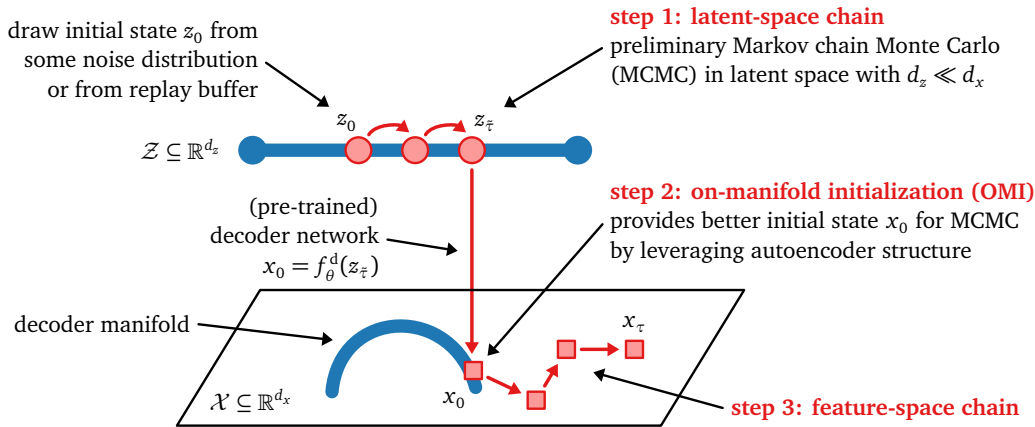
$$-\log p_{\text{data}}(x)^{\text{train}} \approx -\log p_{\theta}(x) = \frac{E_{\theta}(x)}{T} + \log Z_{\theta}$$

- for sparse data, NAE score can be unreliable → Bayesian NAE (BNAE) captures statistical uncertainty, replacing deterministic network parameters with parameter distributions
- but first: in high-dimensional feature space, chains are slow to reach high-density regions...

OMI — on-manifold initialization



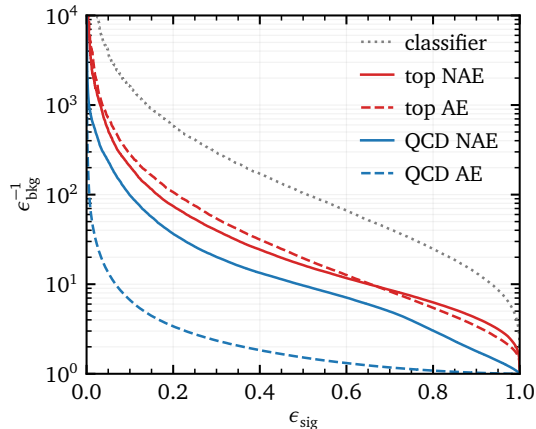
[arXiv:2105.05735, arXiv:2206.14225, arXiv:2510.02168, arXiv:2608.19317]



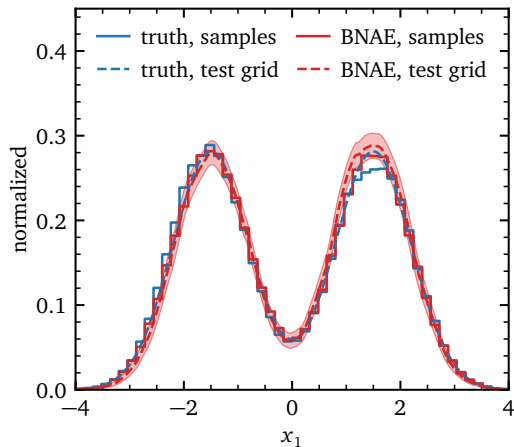
NAE — first AE that works for symmetric training



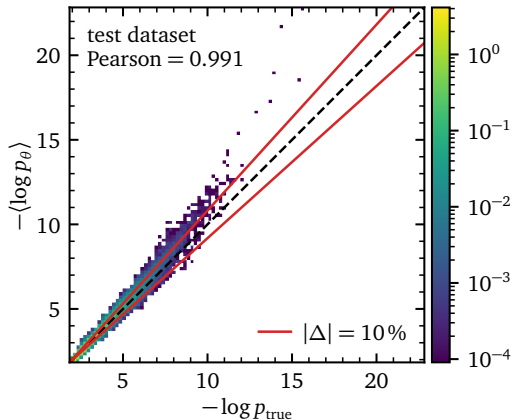
[arXiv:2206.14225, arXiv:2608.19317]



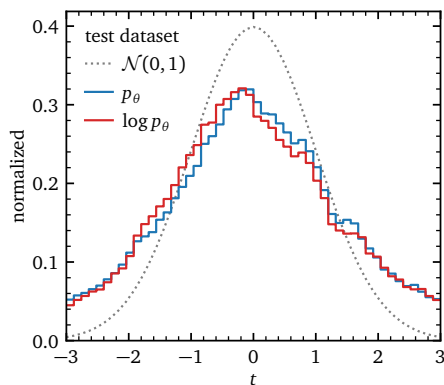
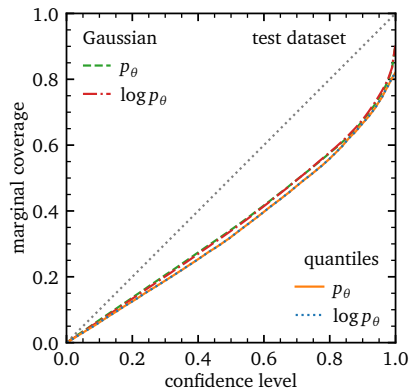
- AE performs worse for QCD-tagging than for top-tagging, while NAE largely **removes this asymmetry**
- **robust (symmetric) anomaly detection:** first AE that identifies jets with higher and with lower complexity
- **but does NAE actually learn a likelihood?**



- **two-dimensional toy model**
with known true likelihood $p_{\text{true}}(x)$
- check relation of BNAE anomaly score
(learned energy) with underlying likelihood
- marginalized true and BNAE-learned
densities show **good agreement**, but with
slight underestimate on Gaussian peaks
- **classifier test** between samples from
 $p_{\text{true}}(x)$ and $p_{\theta}(x)$: AUC = 0.522



- BNAE-learned versus true log-likelihood on test dataset
- **diagonal correlation**: BNAE density, learned via energy, tracks true density over full range of likelihood values



pulls are approximately Gaussian and **marginal coverage** shows slight under-coverage, i.e., learned uncertainties are slightly too small



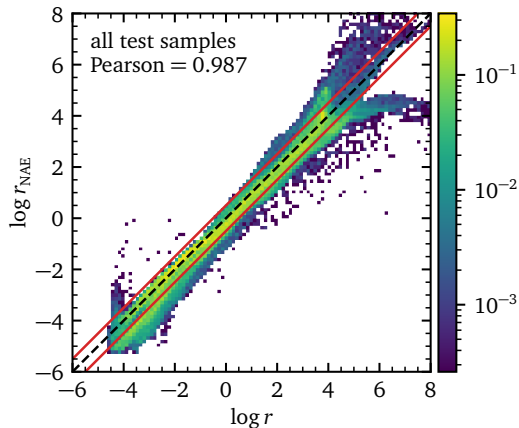
- for realistic LHC applications, **true likelihood is unknown**
- **train two NAEs on two different densities**, e.g., top jets and QCD jets: their energies give log-likelihood ratio up to additive constant

$$\log r_{\text{NAE}}(x) = \log \frac{p_{\theta, \text{top}}(x)}{p_{\theta, \text{QCD}}(x)} = \frac{E_{\theta, \text{QCD}}(x)}{T_{\text{QCD}}} - \frac{E_{\theta, \text{top}}(x)}{T_{\text{top}}} + \text{const}$$

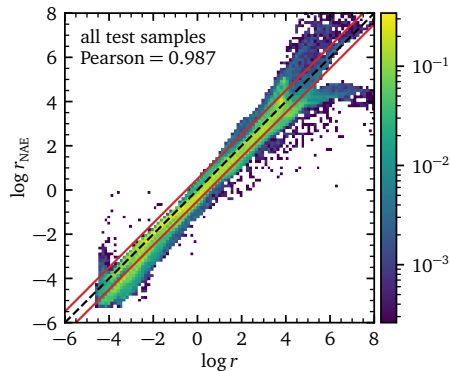
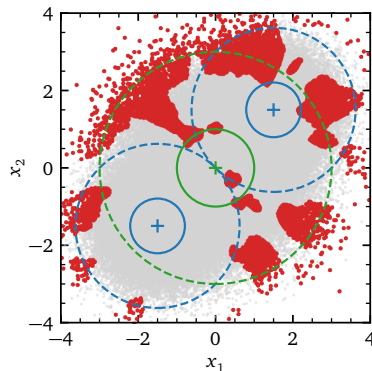
- **supervised classifier** $C(x)$ approximates likelihood ratio

$$r_{\text{class}}(x) = \frac{C(x)}{1 - C(x)} \approx \frac{p_{\text{top}}(x)}{p_{\text{QCD}}(x)}$$

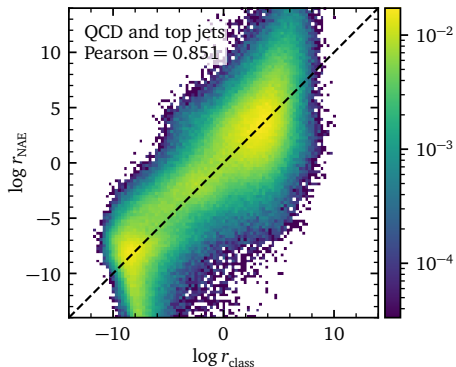
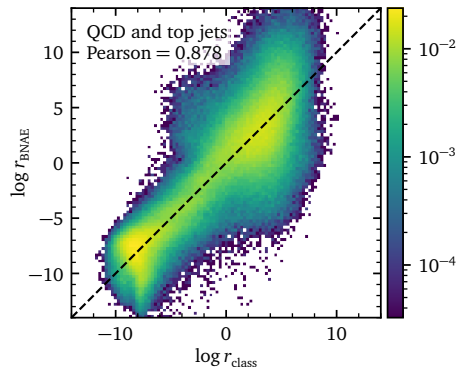
- **dual-NAE test**: correlate (B)NAE-extracted log-likelihood ratio with reference (toy model: true likelihood ratio, jets: classifier)



- toy model: one NAE trained on a single Gaussian, one on two Gaussians
- NAE-extracted versus true log-likelihood ratio on test samples from both densities
- **diagonal correlation**: each NAE learns correct likelihood of its training sample, so their combination predicts correct likelihood ratio



NAEs reproduce true likelihood ratio in regions with sufficient support:
large deviations where at least one density has little support, in tails and between Gaussians



test whether (B)NAEs learn underlying top-jet and QCD-jet densities:
compare classifier-based likelihood ratios with (B)NAE-extracted likelihood ratios

Thank you!

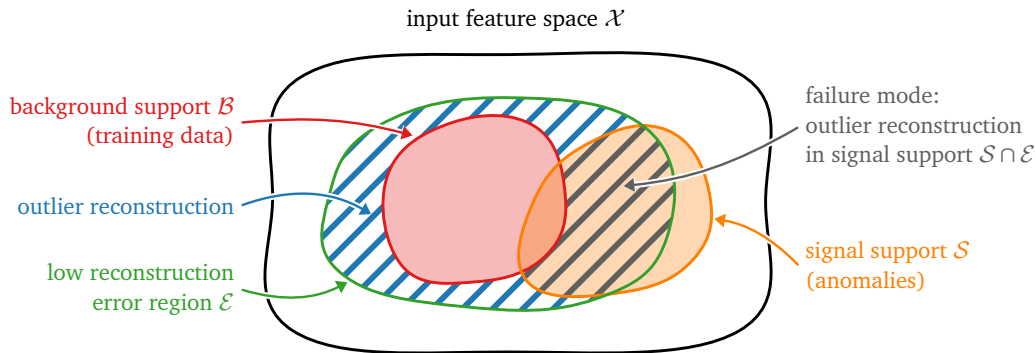
Summary

- AE reconstruction error has no link to likelihood
- NAE is energy-based description of AE
- NAE trains reconstruction error as energy of Boltzmann distribution
- reconstruction error and log-likelihood are directly related
- no complexity bias: symmetric anomaly detection
- (B)NAE learns correct likelihood (ratio)
- BNAE provides statistical uncertainties

NAE, Statistically

arXiv:2608.19317 [hep-ph]





outlier reconstruction: AE unexpectedly reconstructs inputs well although they lie outside of training distribution — such anomalies get a small anomaly score



[arXiv:2101.03288, arXiv:2105.05735, arXiv:2206.14225, arXiv:2510.02168, arXiv:2608.19317]

- samples from $p_\theta(x)$ through **Langevin Monte Carlo (LMC)**, with step size λ and noise σ

$$x_{t+1} = x_t + \lambda \nabla_x \log p_\theta(x_t) + \sigma \epsilon_t = x_t - \frac{\lambda}{T} \nabla_x E_\theta(x_t) + \sigma \epsilon_t \quad \text{with } \epsilon_t \sim \mathcal{N}(0, \mathbb{1})$$

- chain drifts toward low-energy regions, while noise term allows exploration
- choose λ and σ to place more weight on gradient term than on noise term: sampling at low effective temperature converges more quickly

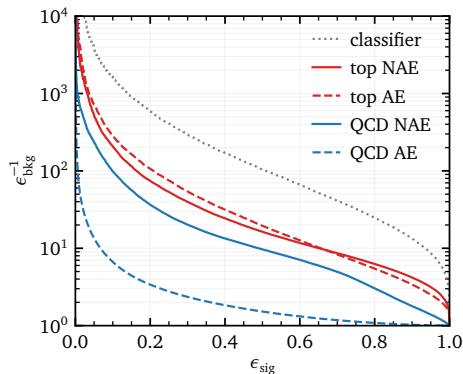
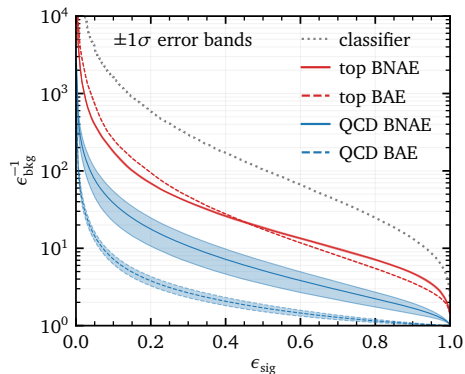
$$T_{\text{eff}} = \frac{\sigma^2}{2\lambda} T$$

- but in high-dimensional feature space, chains are slow to reach high-density regions
→ **on-manifold initialization (OMI)**

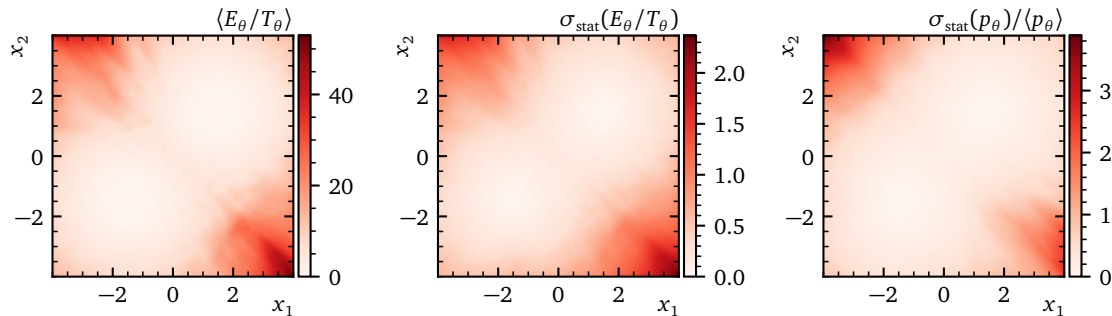
NAE — first AE that works for symmetric training



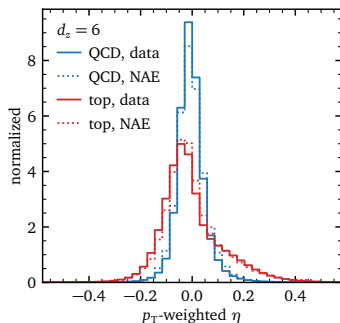
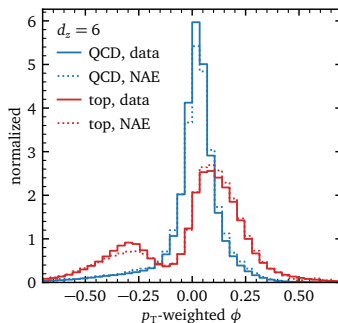
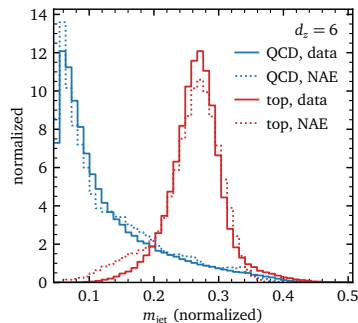
[arXiv:2206.14225, arXiv:2608.19317]



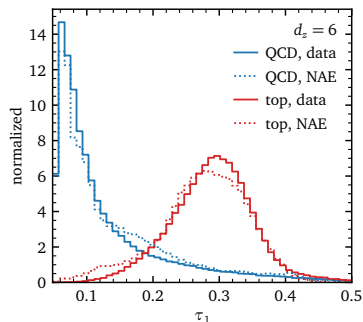
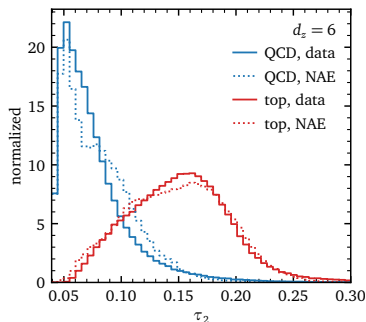
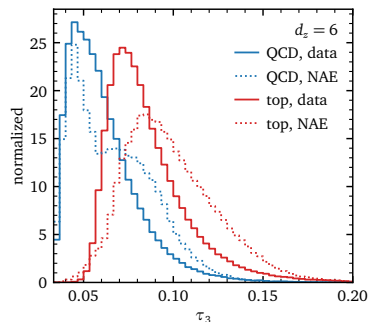
robust (symmetric) method to detect anomalies:
first AE that identifies jets with higher and with lower complexity



learned energy, energy uncertainty, and relative density uncertainty:
both uncertainties are small near data support and increase in low-density regions



compare NAE-generated and true jets in high-level jet-substructure observables:
agreement is strongest for angular distributions and remains good for jet mass



N -subjettiness: agreement remains good for τ_1 and τ_2 , but larger discrepancies in τ_3 (NAE has difficulty capturing three-prong structure)