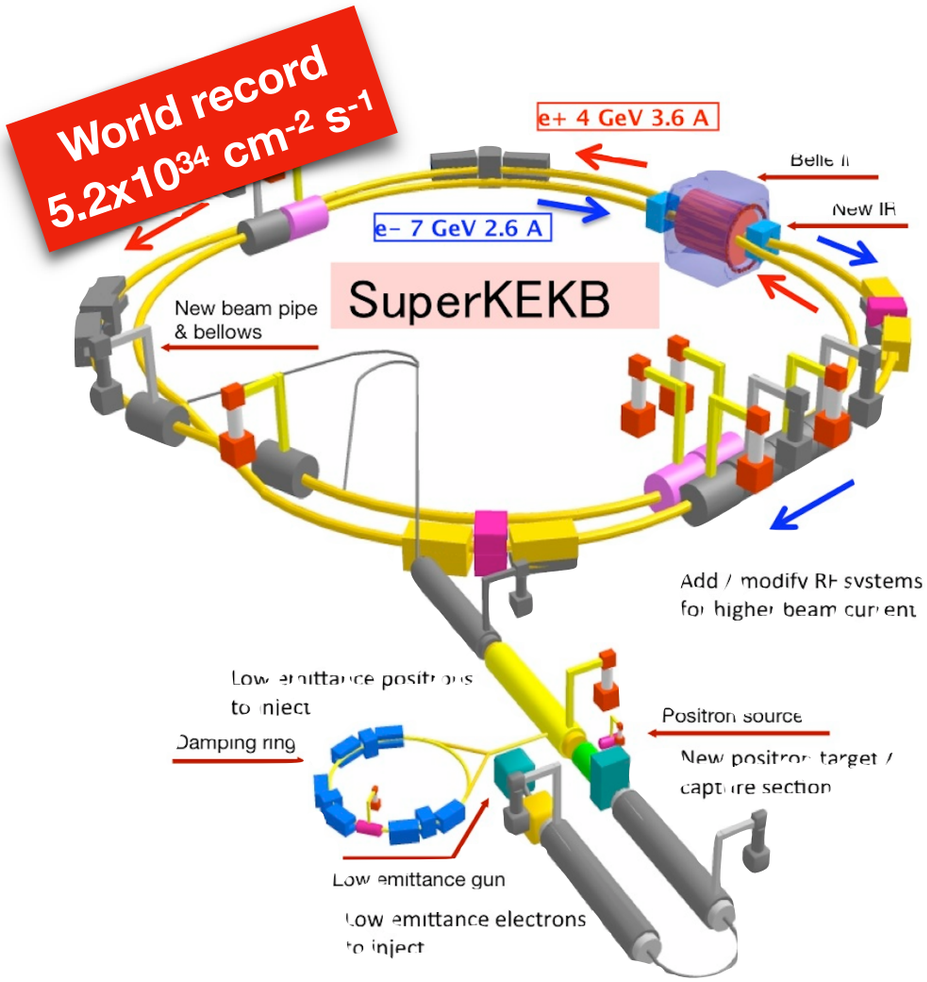
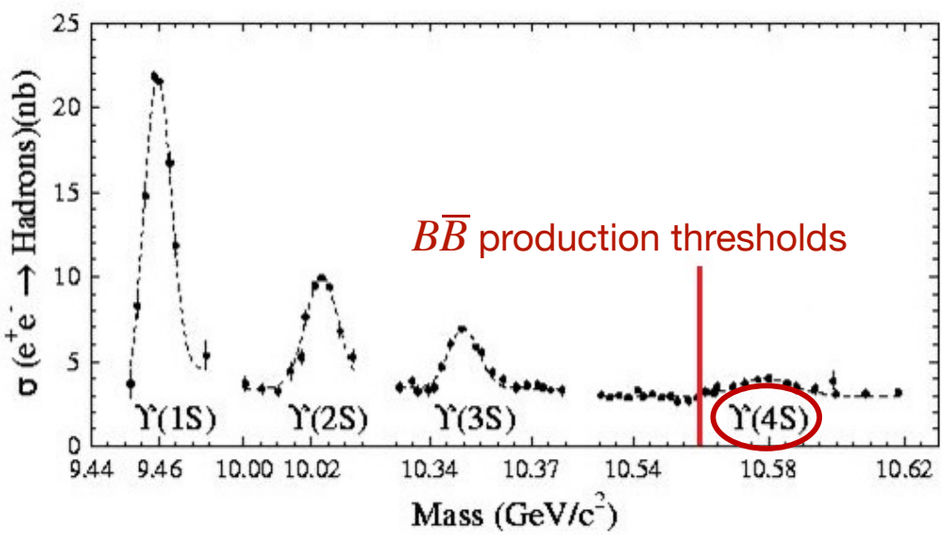
Belle II for TREASURE

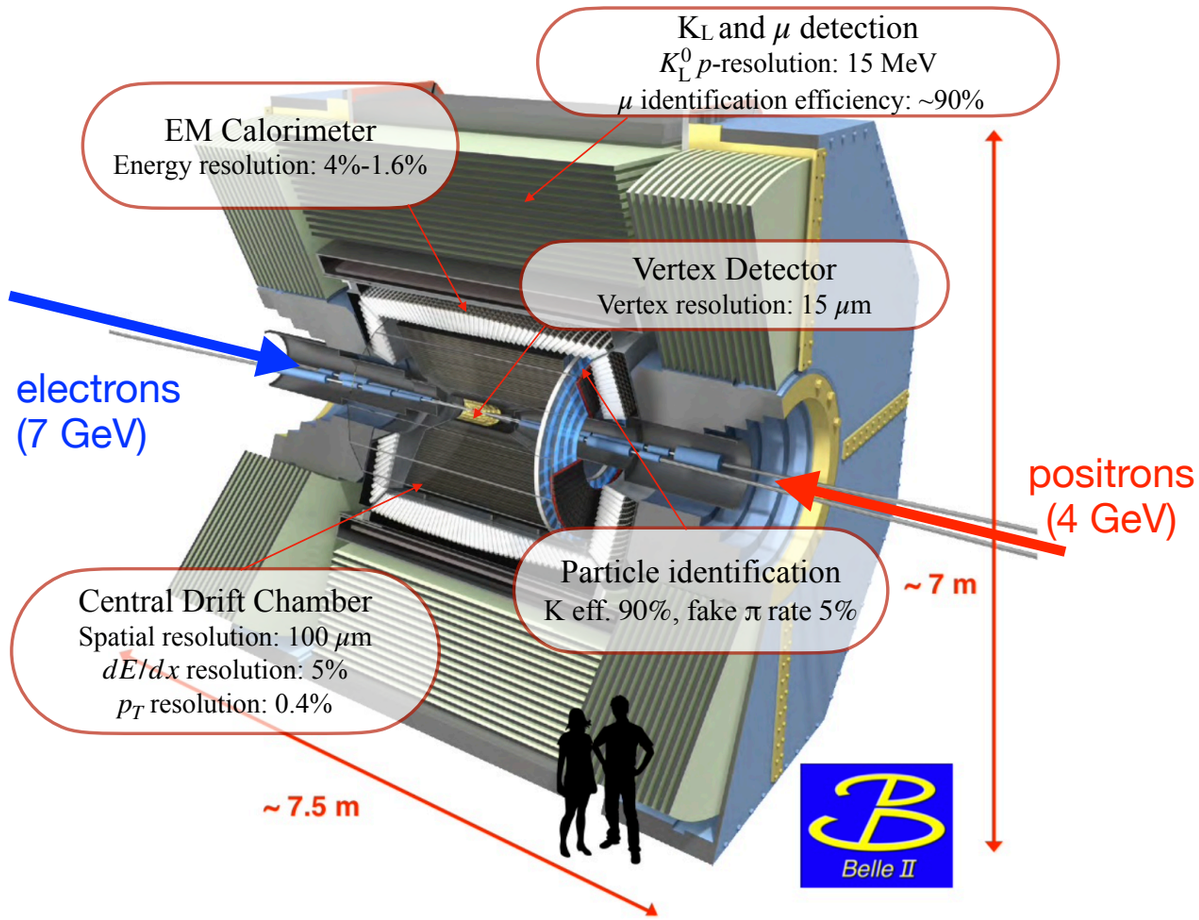
Riccardo Manfredi

April 29, 2026

Belle II @SuperKEKB in a nutshell







~100% of $\Upsilon(4S)$ decay to $B\overline{B}$ pairs

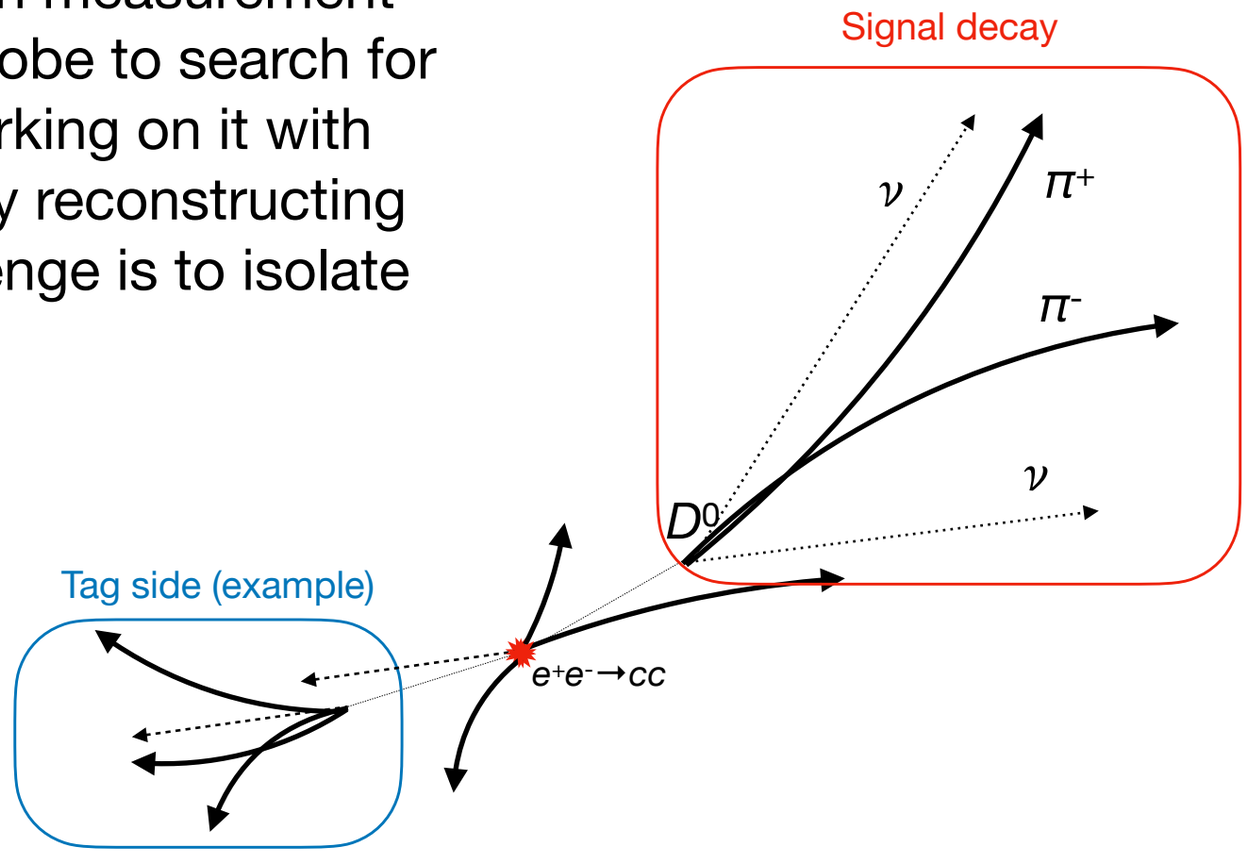
- Low-background production
- Precisely known collision energy
- Coherent evolution of B and $\overline{B}$
- Large production of charm and τ too

Plan

Tokenize Belle II simulated data with simple binning (“discrete” tokenization) and with VQ-VAE, then train classifier to compare signal-background separation performance. Adapt existing code developed for ATLAS samples

Benchmark analysis: branching fraction measurement of $D^0 \to \pi^+\pi^- \nu\nu$ processes, promising probe to search for BSM physics. BNL group currently working on it with an inclusive approach, i.e. not explicitly reconstructing the tag-side charm hadron. First challenge is to isolate $e^+e^- \to cc$ using tag-side info

Today showing: first tokenization of variables to classify events on small private simulation sample
+ discussion on data format and availability



Variables used

Currently tokenizing directly high-level objects, the same one used later in the classifier.
Apply some specific selections to each type of object

Tracks: p^*, θ^*, ϕ^*, dr*, dz*, cosθ_{thr}, πID, KID, μID, recoil mass
Track objects

Photons: E*, p^*, θ^*, ϕ^*, Δt, cosθ_{thr}, beamBkgMVA, fakePhotonMVA
ECL clusters not matched to a track

V^0 objects: M, L_{flight}, L_{flight}/σ
Long-lived secondary vertices found combining tracks

K_L^0 candidates: one-side E^{miss}, one-side p^{tot}, one side cos$\theta(p^{tot})$, cos$\theta_{thr}(p^{tot})$, numbers of: tracks, K, μ, p, e, ɣ, Λ, K_S, K_L, 2-body vertices
KLM clusters not matched to a track

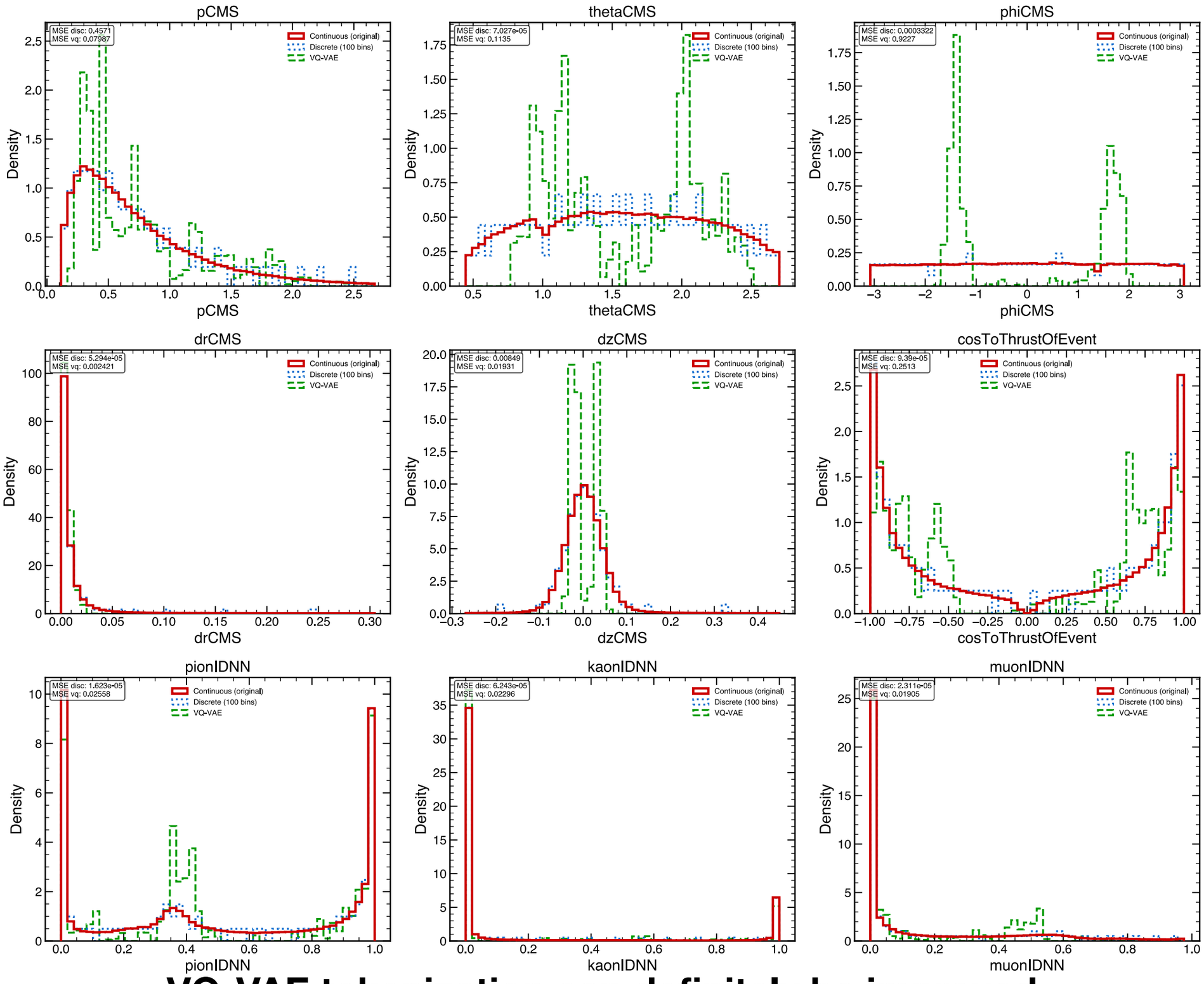
Jets: p_T, p_x, p_y, p_z, η, θ, ϕ, E, number of particles in the jet
Clustered adapting fastJet algorithm to Belle II objects

Event info: √s, counts for different type of particles

* = in the collision c.m.s.

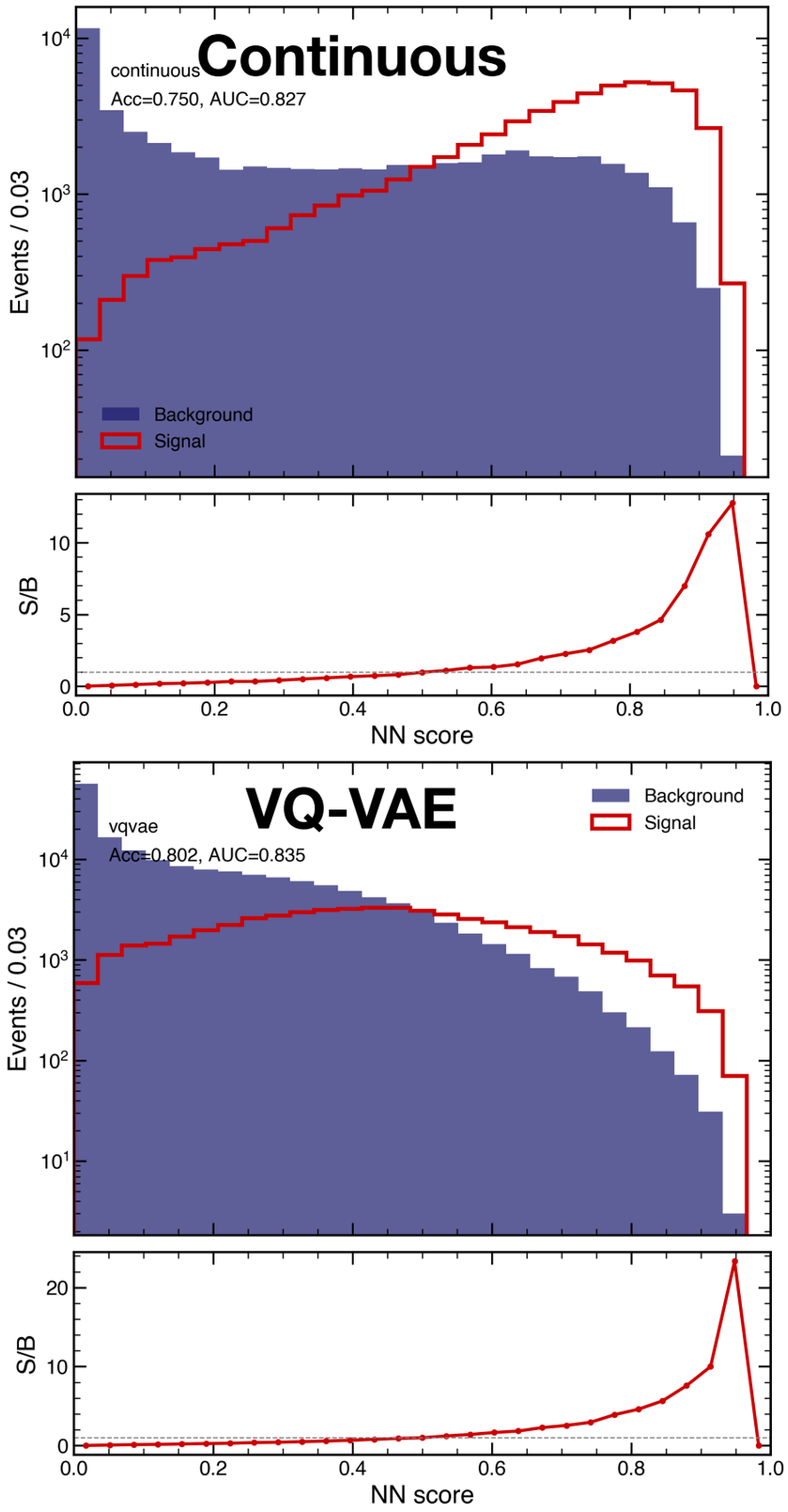
Track variables

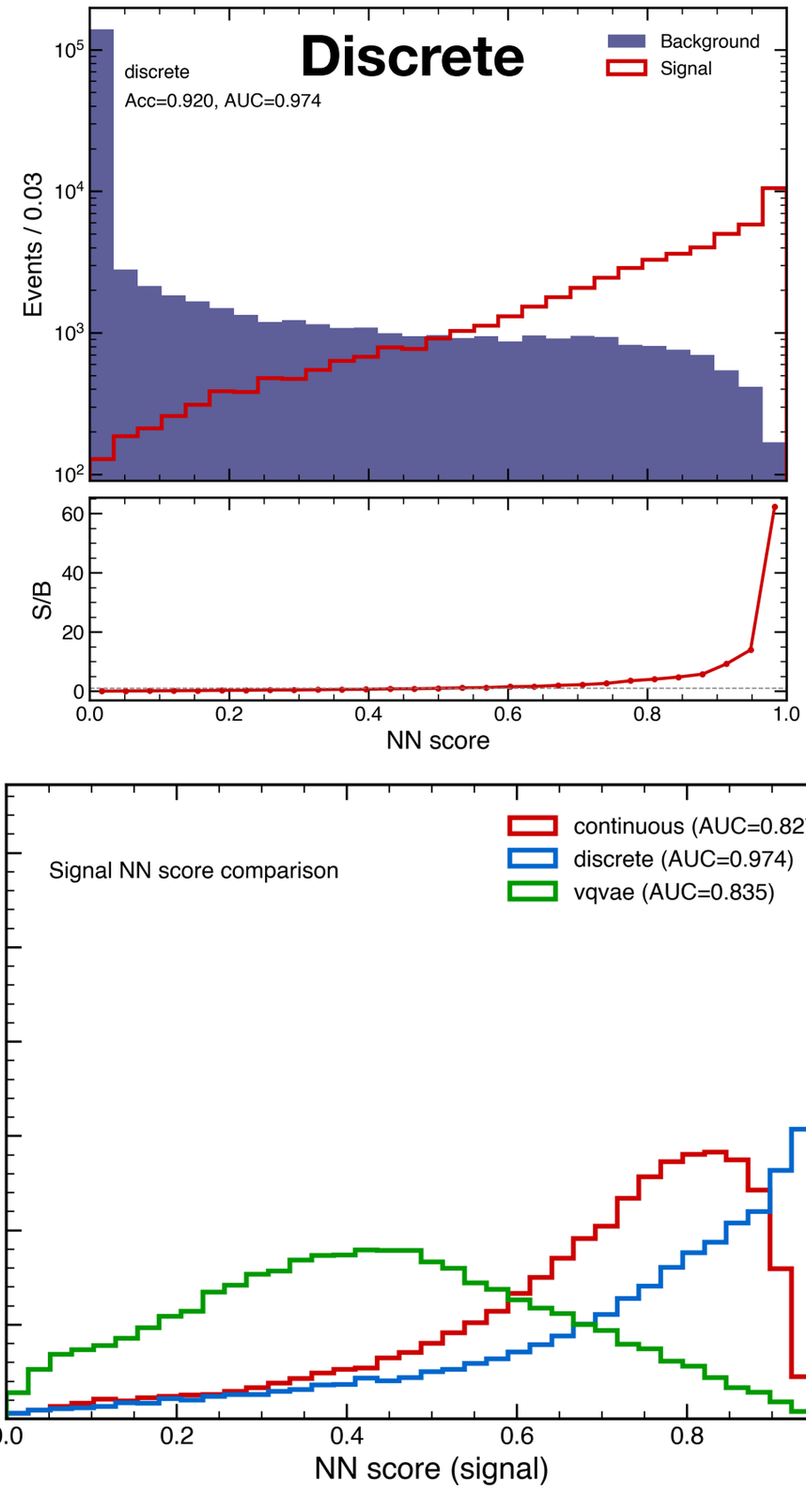
Track features: tokenization comparison (set 1/2)



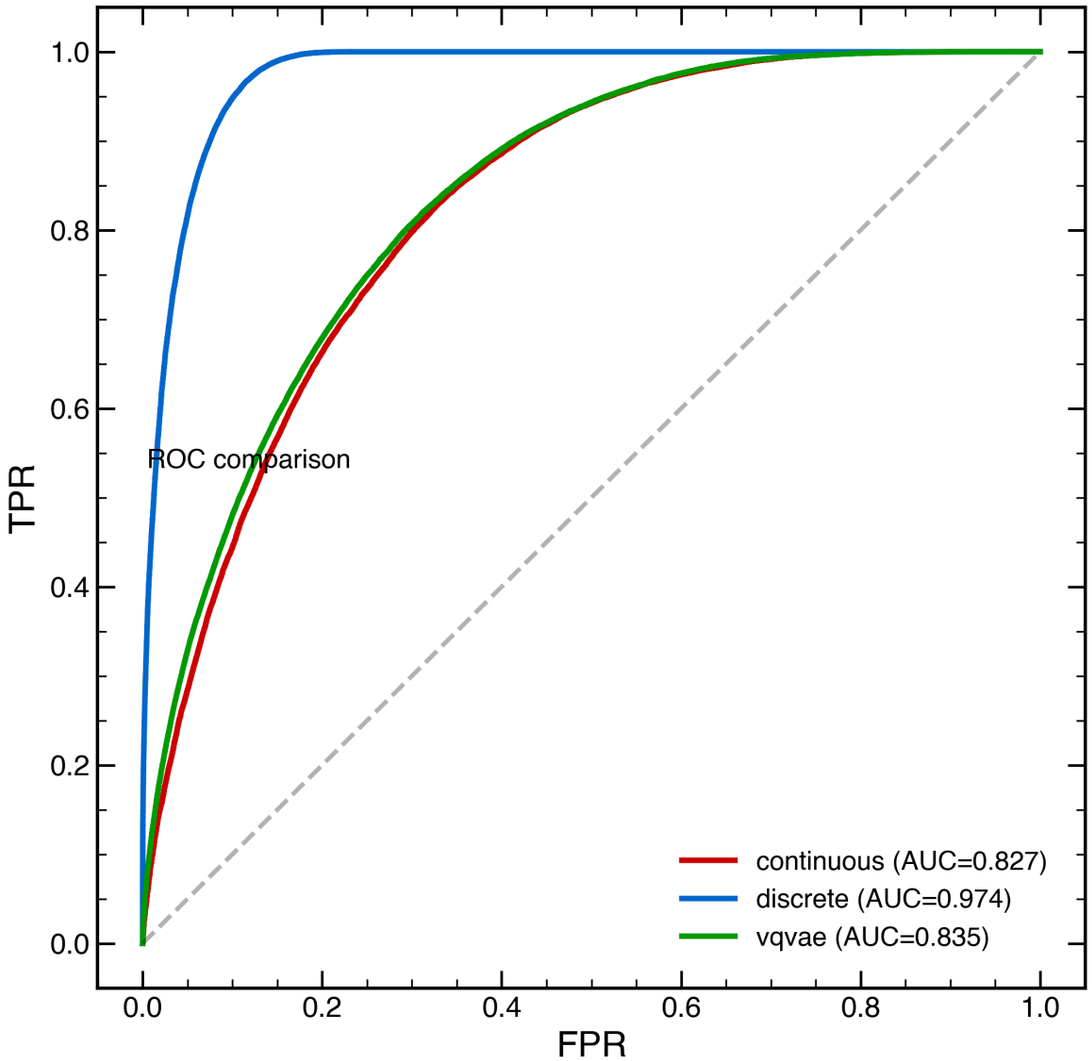
VQ-VAE tokenization can definitely be improved

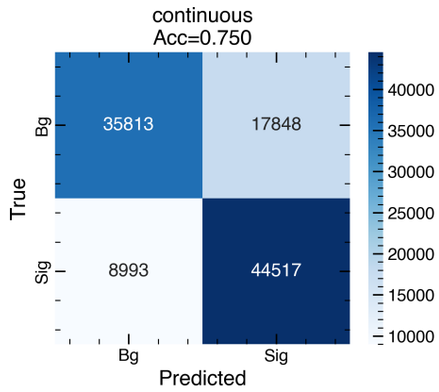
Classifier output

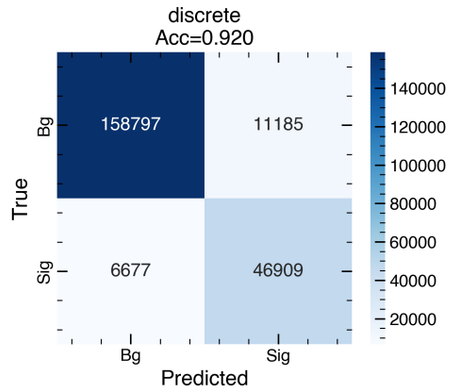


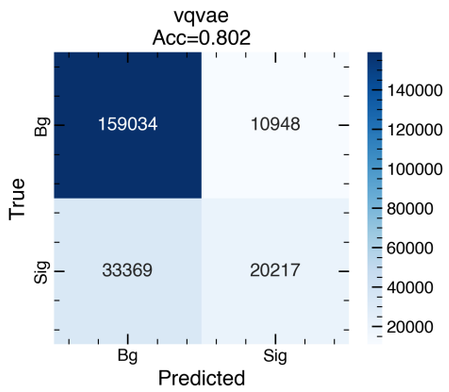


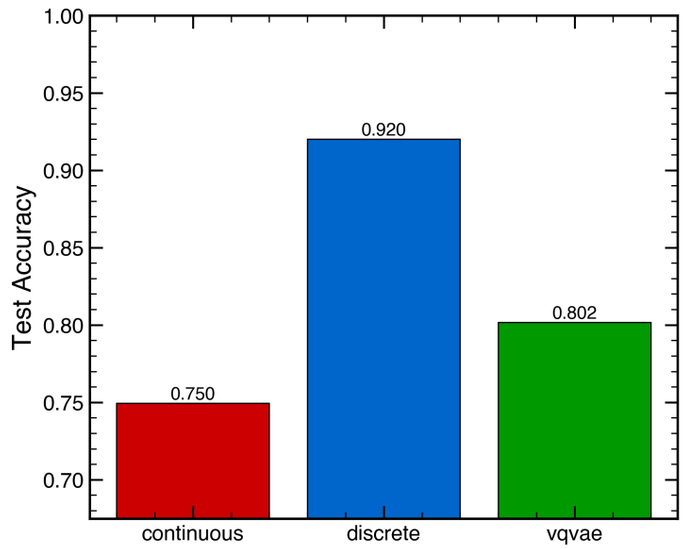
Classifier performances

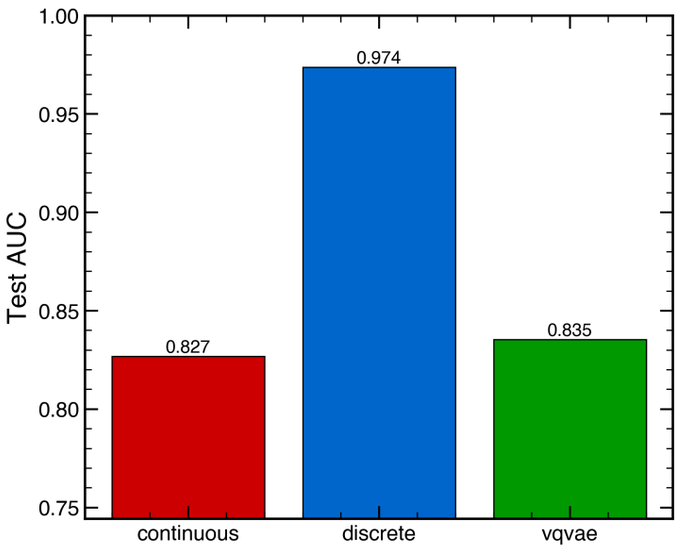












Performance confirm current VQ-VAE underperformance

Availability of Belle II samples

Simulation: Collisions generated with Pythia/KKMC and EvtGen (public softwares), then simulated in the detector and reconstructed with Belle II Analysis Software (basf2), which is also public.
I will investigate if we can make publicly available simulation samples that are already produced

Collision data: there is an official policy to request data to be made public. So far, only a few HEPData entries exist. Case for public data could be made, maybe proposing a multiple-step release

Summary

Belle II, based at the SuperKEKB e^+e^- collider, aims at performing indirect searches of BSM physics in the flavor sector.
The collider environment and data structure can be a proxy for FCCee

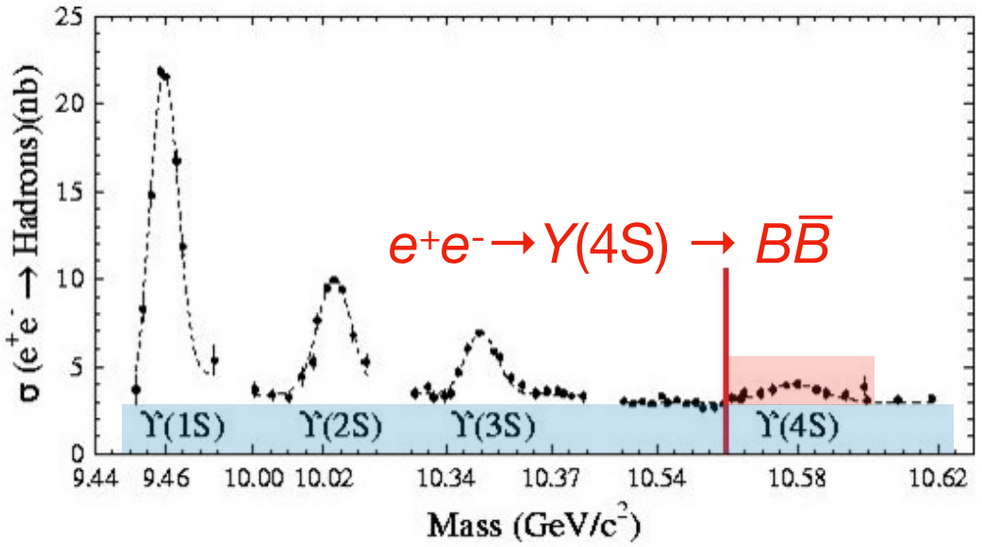
Adapted existing TREASURE tokenization pipeline to Belle II simulated data, to classify charm vs other quark production using tag-side information.
Physics can be improved, but technical machinery works

Started discussion yesterday on which variables to save: for now idea is to save only info of tracks and clusters (both matched to a track and not).
I will work on a way to convert Belle II data formats to h5 or parquet

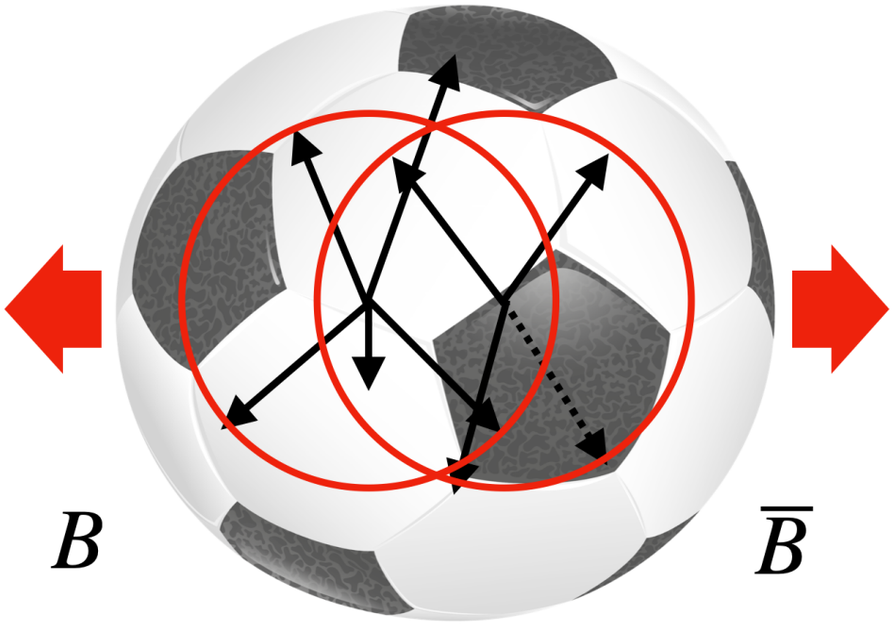
Will talk with Belle II people to see how to make simulated samples public.
Discussion about data will require more inputs, most important how much data would be needed

backup

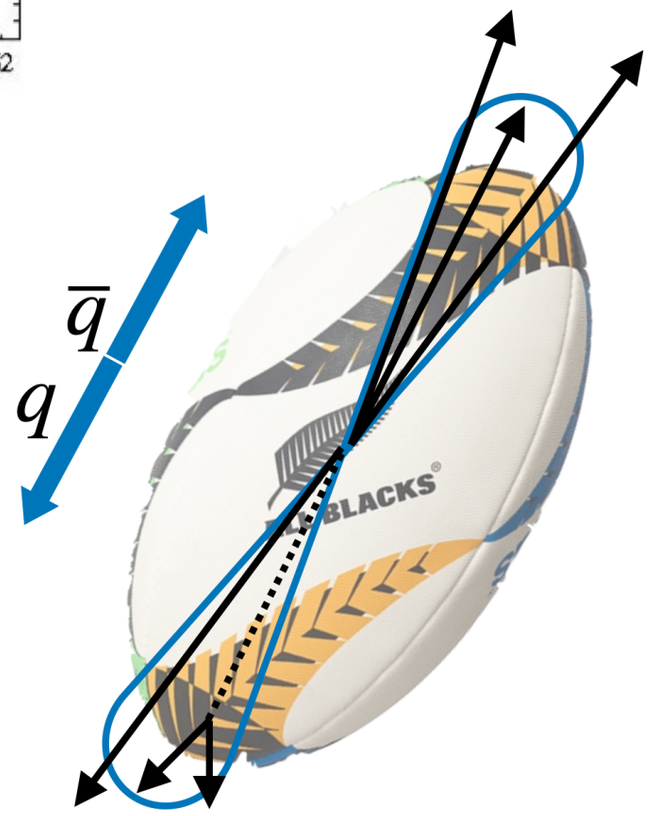
“Jets” at Belle II



$e^+e^- \to q\bar{q}$



$p(B) \approx 0.3$ GeV/c

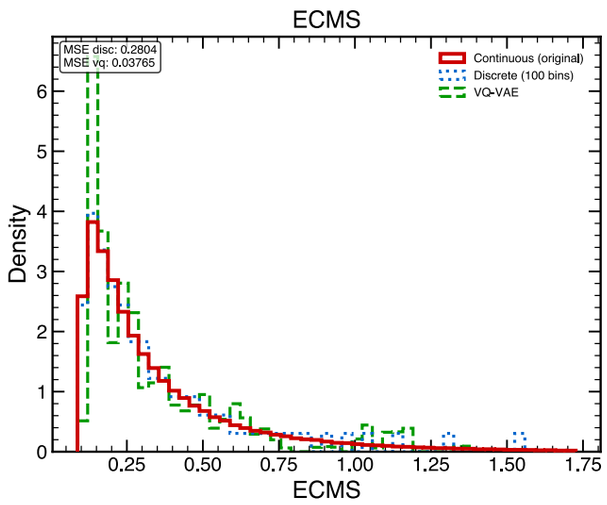


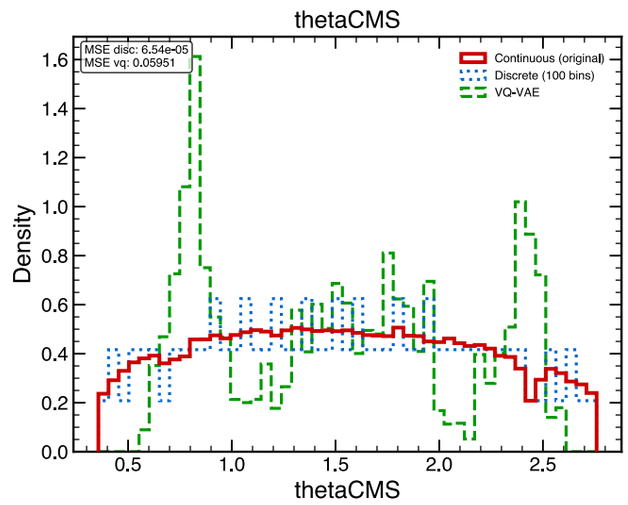
$p(q) \approx 5$ GeV/c

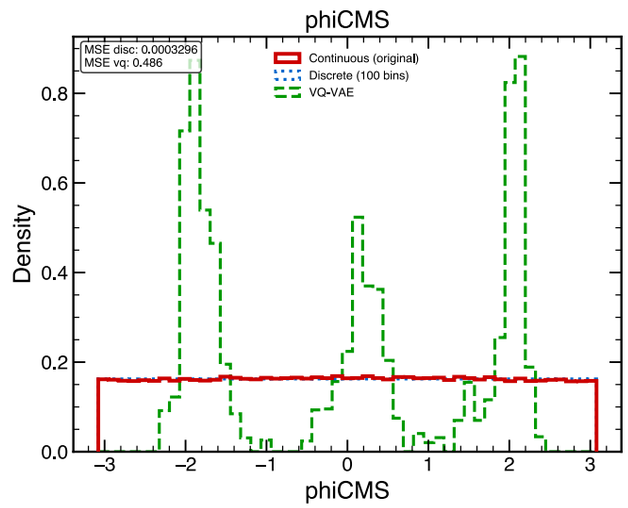
Makes sense to try to reconstruct jets in “*continuum*” events

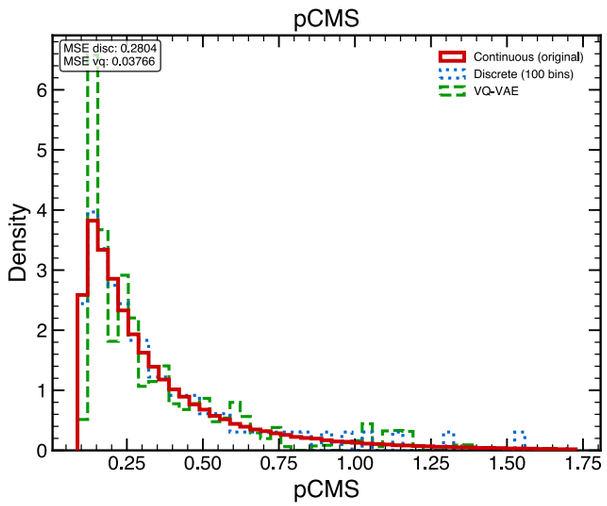
Photon variables

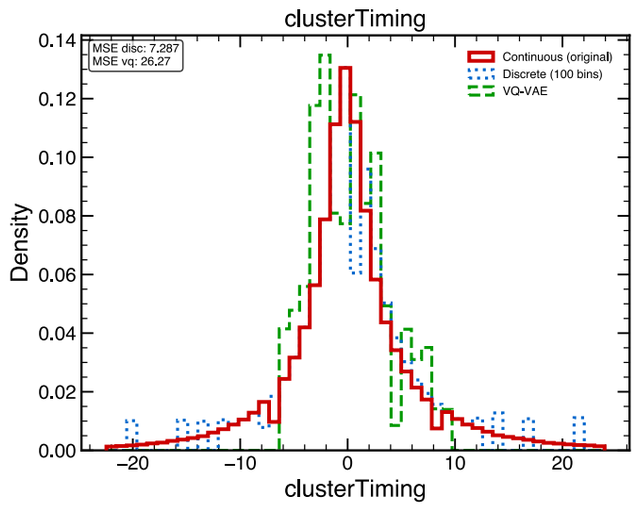
Photon features: tokenization comparison (set 1/1)

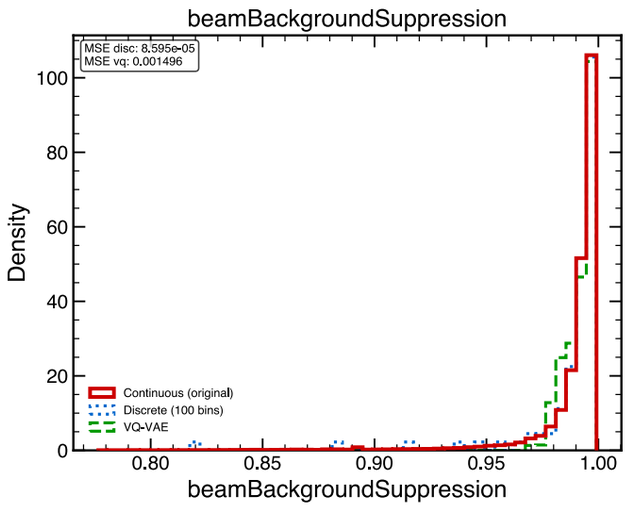


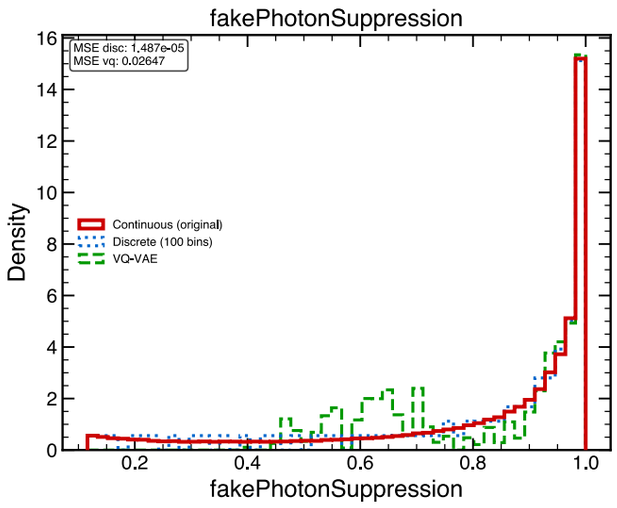


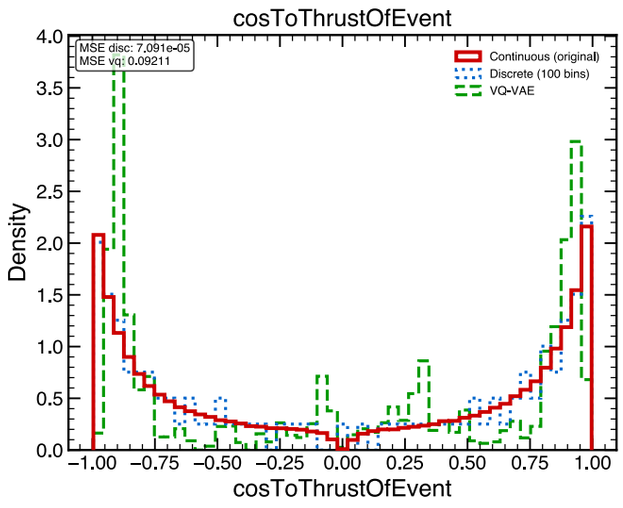






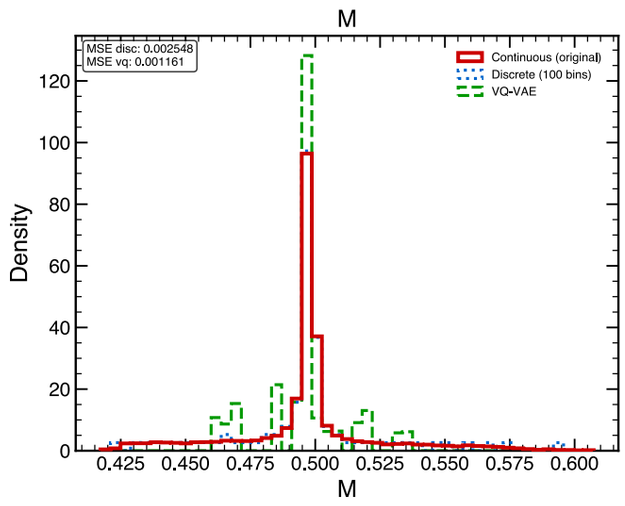


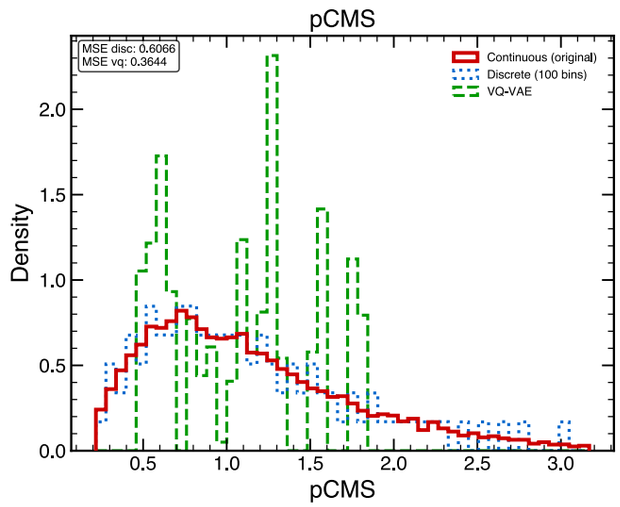


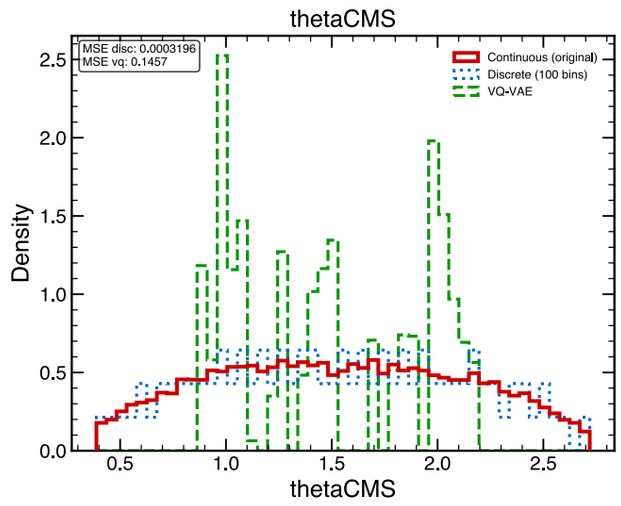


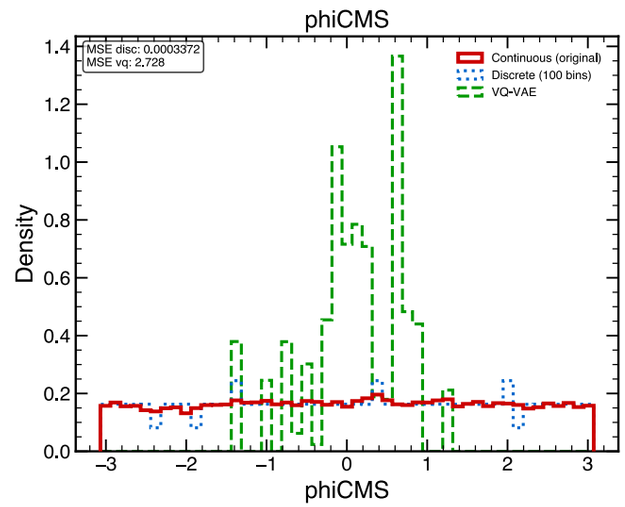
V^0 variables

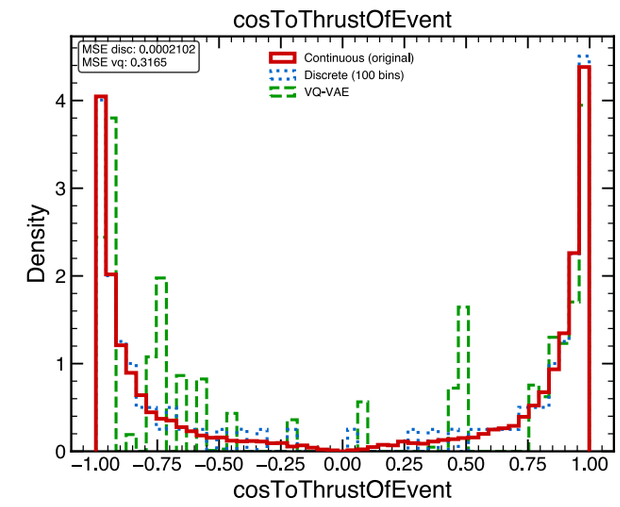
V0 features: tokenization comparison (set 1/2)

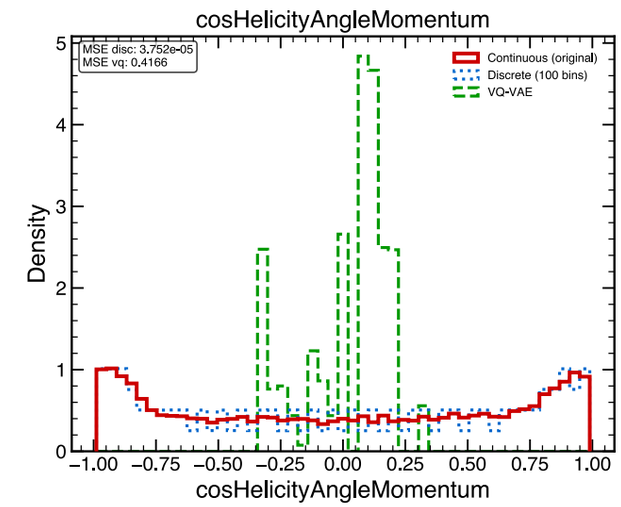


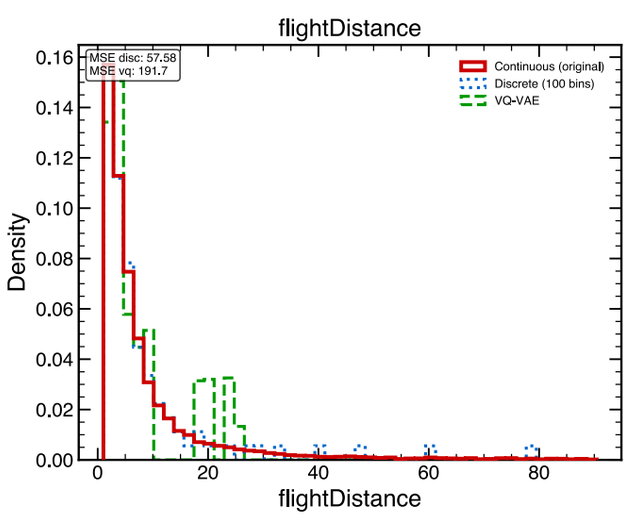


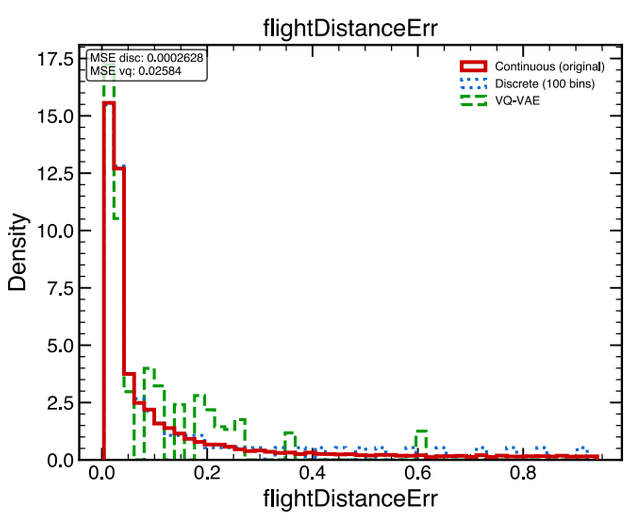


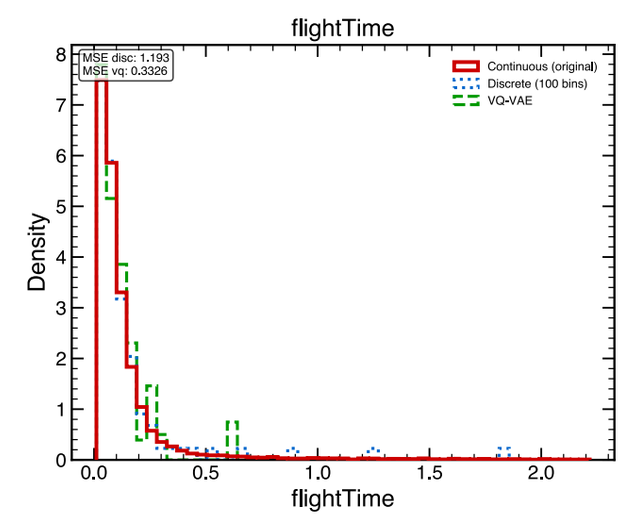






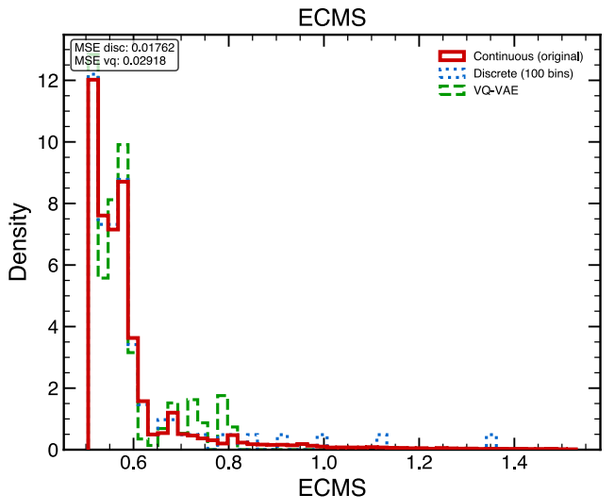


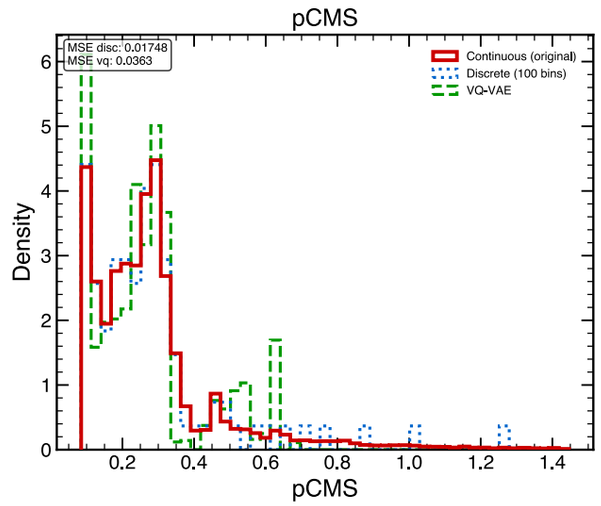


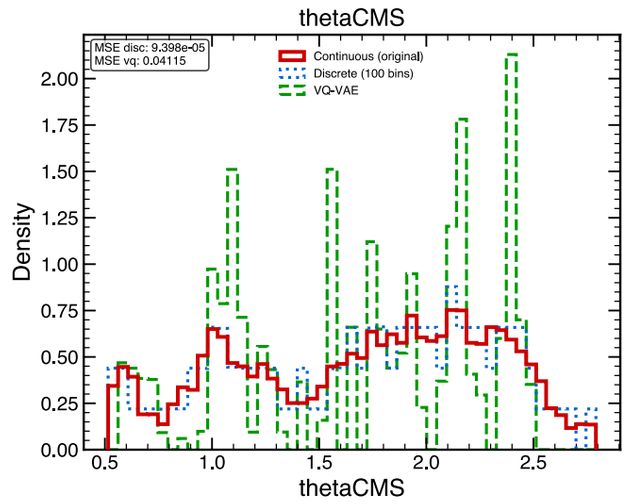


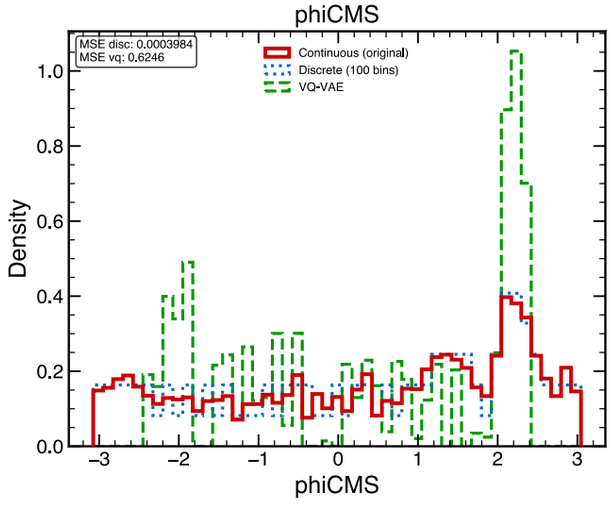
K_L^0 variables

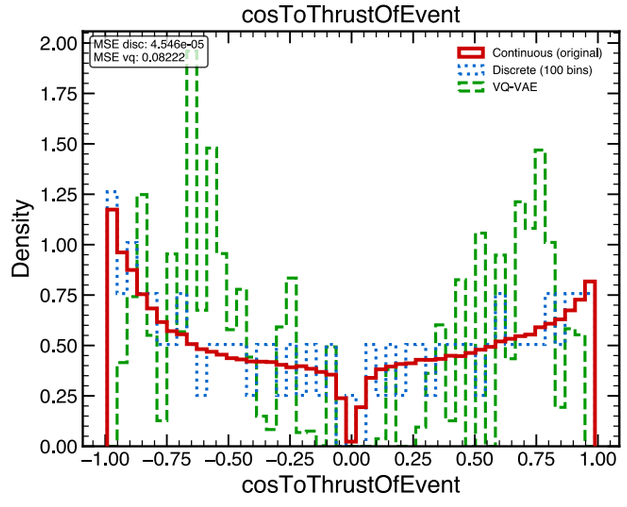
Kl0 features: tokenization comparison (set 1/1)

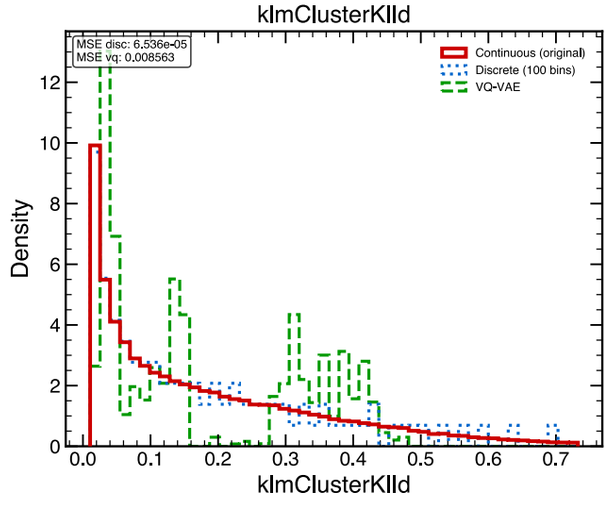


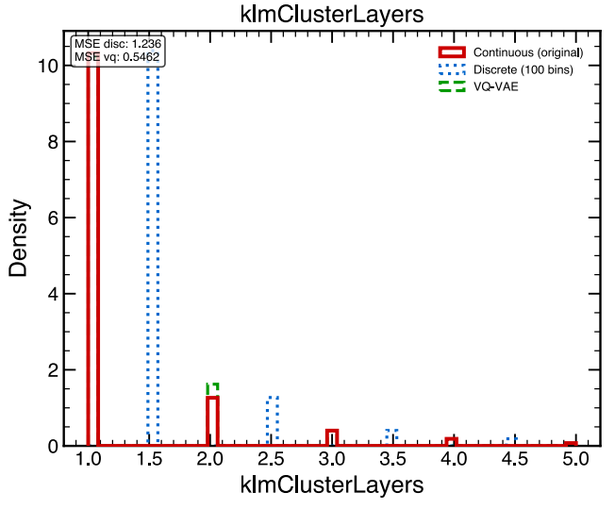






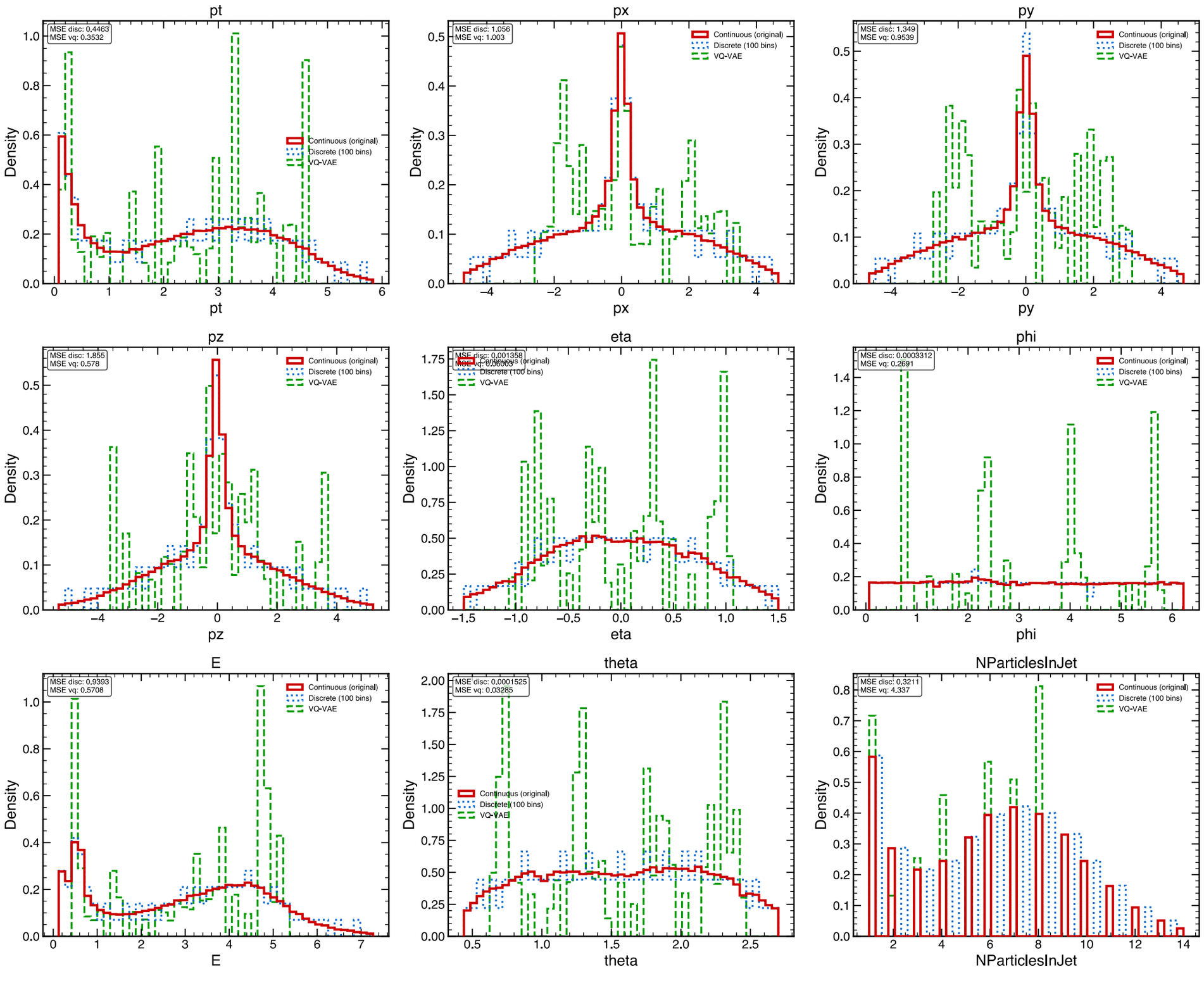






Jet variables

Jet features: tokenization comparison (set 1/1)



Event variables