Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s}$ = 13 TeV

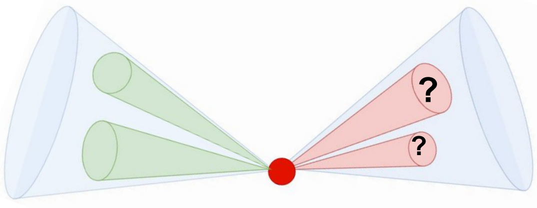
ML4Jets 2024, Paris

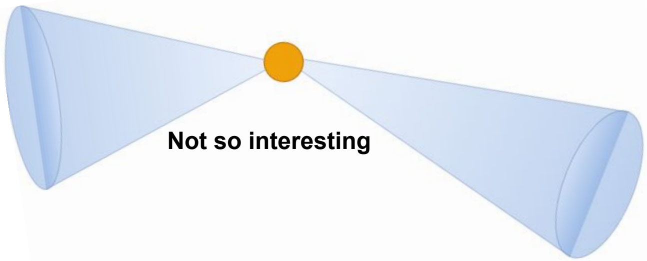
Aritra Bal (Karlsruhe Institute of Technology)

For the CMS Collaboration

Overview

- Target dijet topologies with both jets clustered using anti-k_T (R=0.8)
- Search for a narrow resonance of the form A →BC
- Exploit jet substructure - anomalous jets may have (multiple) prongs
- Total of 5 complementary machine learning techniques - each with its own strengths





Designing an Anomaly Tagger

- Supervised learning - Train on MC with labelled examples
- Unsupervised approach - Train directly on data to avoid specific signal model bias
- All but one of the five methods use only data for training

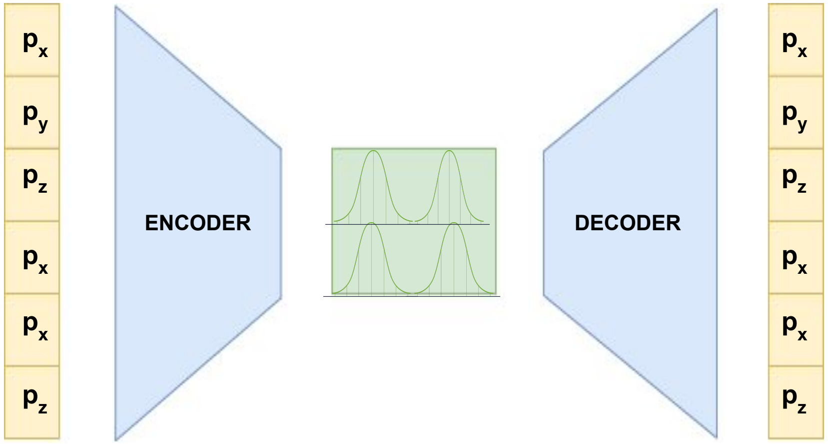
Weak Supervision:	Unsupervised (Autoencoder based)
• *CWoLa* Hunting • Tag N' Train (*TNT*) • Classifying Anomalies THrough Outer Density Estimation (*CATHODE*)	• *VAE-QR* Semi-supervised • *QU*asi *A*nomalous *K*nowledge (*QUAK*) Trained on sideband, learns QCD distribution

Unsupervised Learning with Autoencoders (VAE-QR)

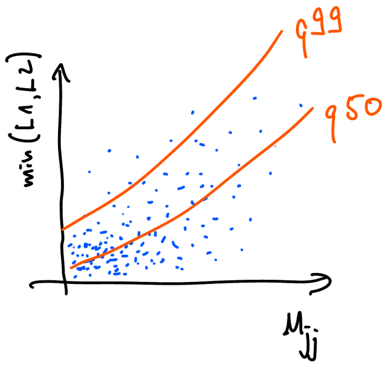
- Autoencoder-based anomaly search - train a network to “reconstruct” jets from a QCD-dominated control region and apply to data from signal region

- Anomaly metric = network loss

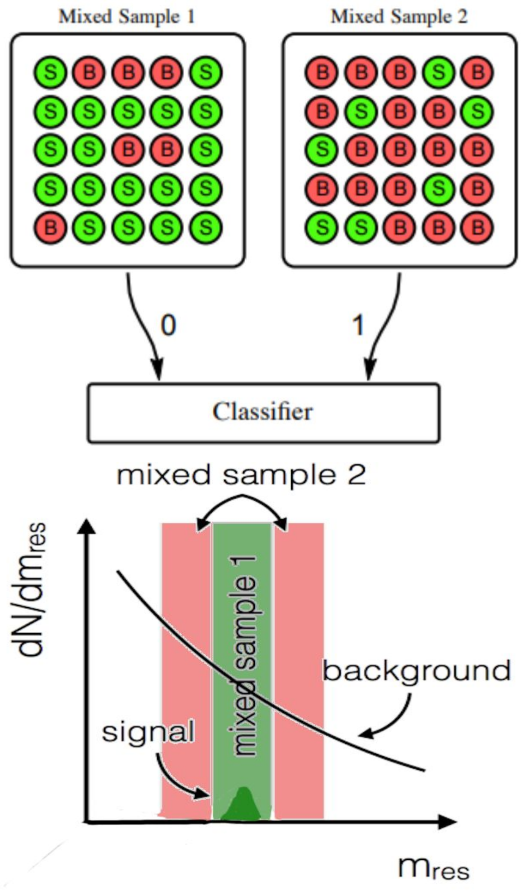
- Decorrelate loss from m_{JJ} using a DNN based Quantile Regression (QR) - reduces background “sculpting”



QR



Weak Supervision Paradigm



- Train classifier to distinguish **data** from a **background**-like sample → different proportions of signal
  - In practice: two sidebands defined on either side of a narrow signal region
- No signal → Classifier learns random noise
- Three methods in total:
  - **CWoLa**: background events selected from sideband defined on either side of narrow signal region
  - **TNT**: Additional autoencoder preselection, designed for events with 2 anomalous jets
  - **CATHODE**: Uses normalizing flows to interpolate background from sideband into signal region

Semi-supervised learning: QUAK

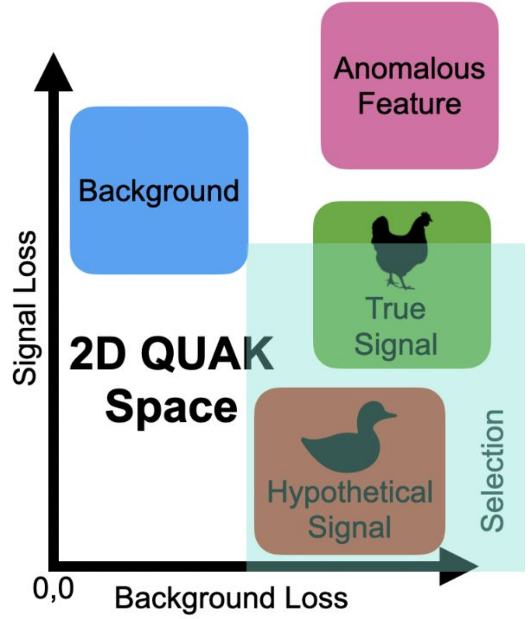
Idea: train separate normalizing flows on background and signal MC

Use losses to construct a 2D QUAK space

- Every event mapped into a unique point in a 2D QUAK space
  - Use different normalizing flows trained on QCD background MC and (mixture of) signal MC

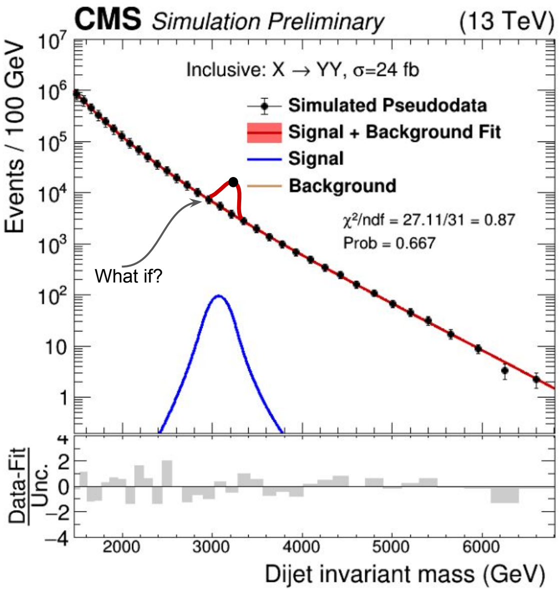
- The signal lies somewhere in that space and the background lies somewhere else

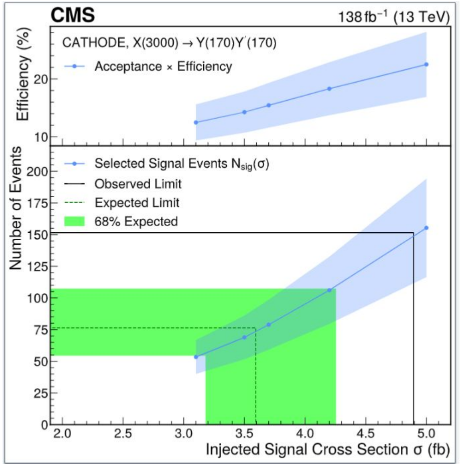
- Select events by creating a unique 2D contour for each signal mass hypothesis designed to exclude background events



What can we do with this?

- Choose a working point and select events to look at
- Perform a bump hunt on the m_{JJ} spectrum and look for interesting deviations
  - QCD background is smoothly falling
  - Signal is a narrow resonance - can be modelled using a Double Crystal Ball function



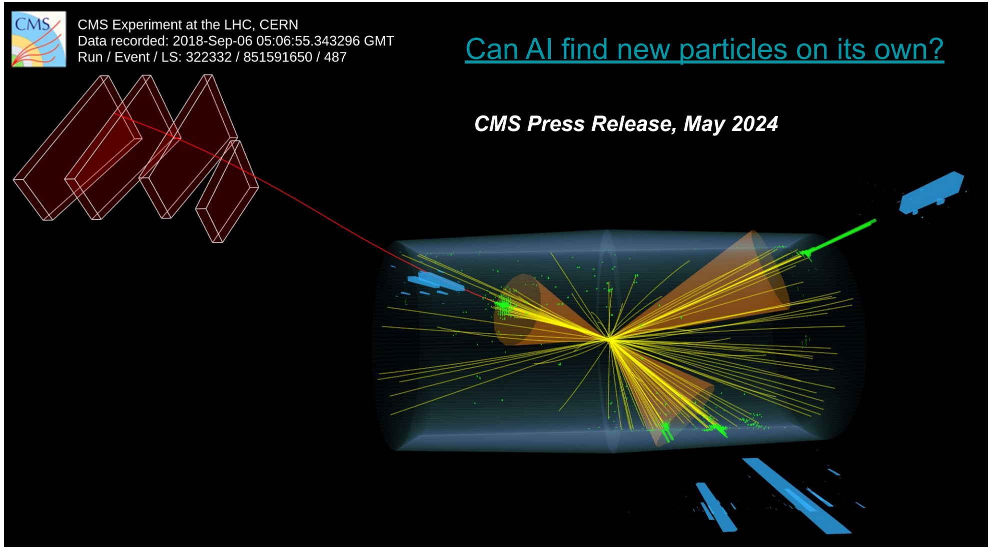


We use (almost) no MC for training

- But can use it to set limits on various signal models
- Never been done before for most models we look at
- For weakly supervised methods - do this by injecting various cross sections of signal into data and training a classifier each time

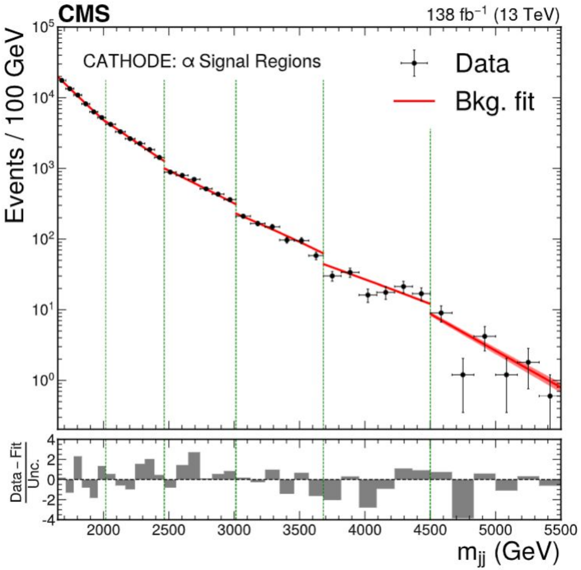
What did we see?

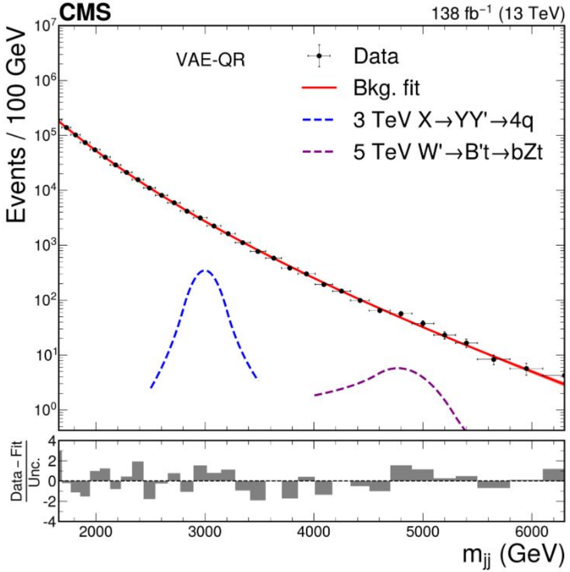
- All methods report no significant deviation from the Standard Model in CMS Run II data (recorded during the period 2016-18) at a total integrated luminosity of 137 fb^{-1}
- Remember that these searches are model agnostic - goal is to show broad sensitivity by setting limits on a range of signals



Bump hunting

- Use generic signal shape to scan for potential anomalies across entire dijet mass spectrum
- No significant deviation observed by any method



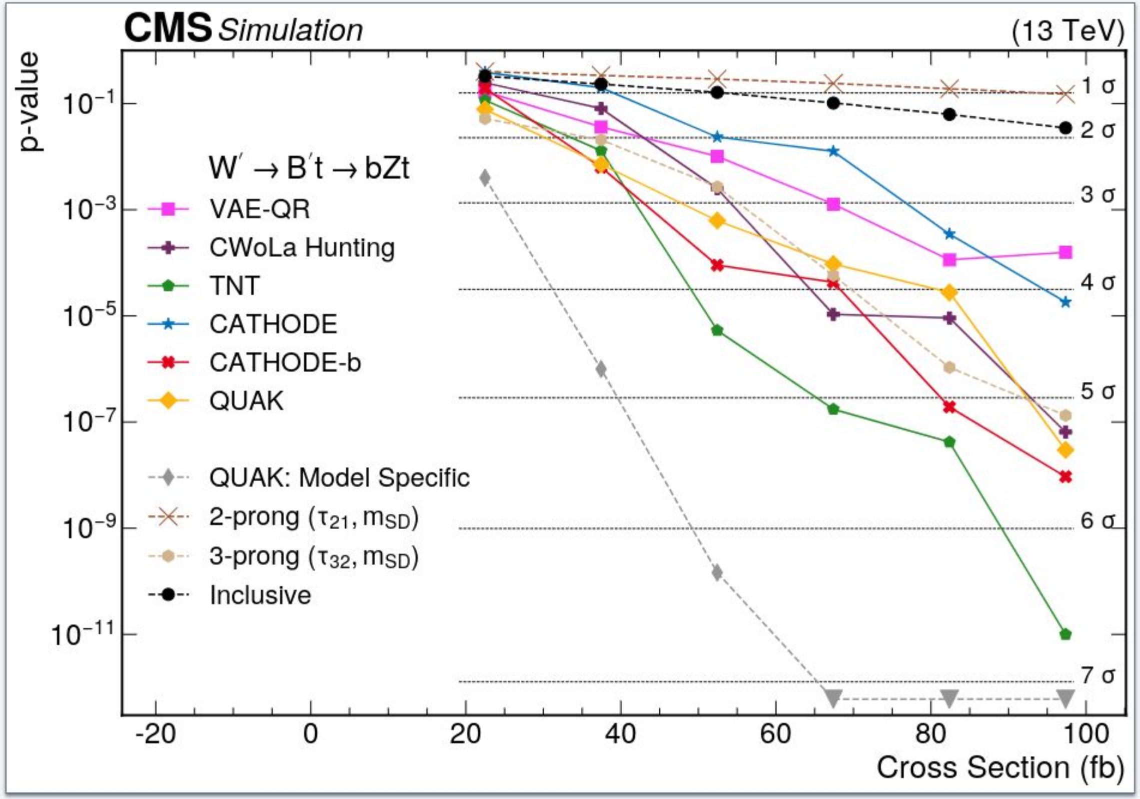


9

Expected Significances

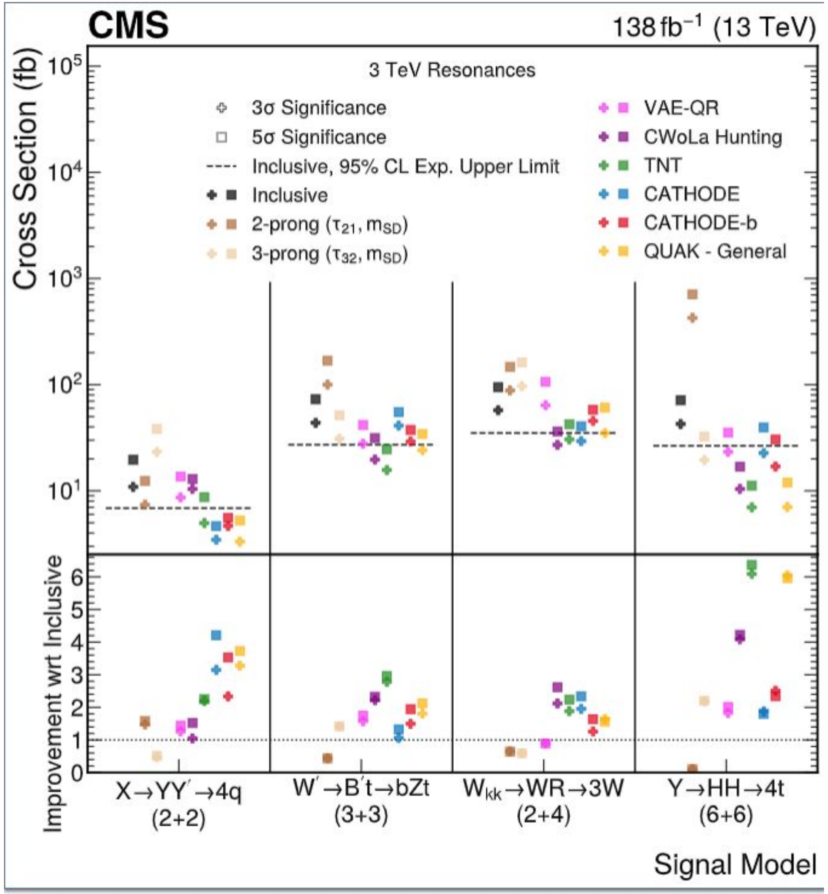
Karlsruhe Institute of Technology

- Inject signal into a toy background MC dataset and calculate expected significances
- Improved performance with higher daughter particle masses in general
- Test with a 3 pronged signal:
  - $W \to B' \, t \to qqq \; qqq$



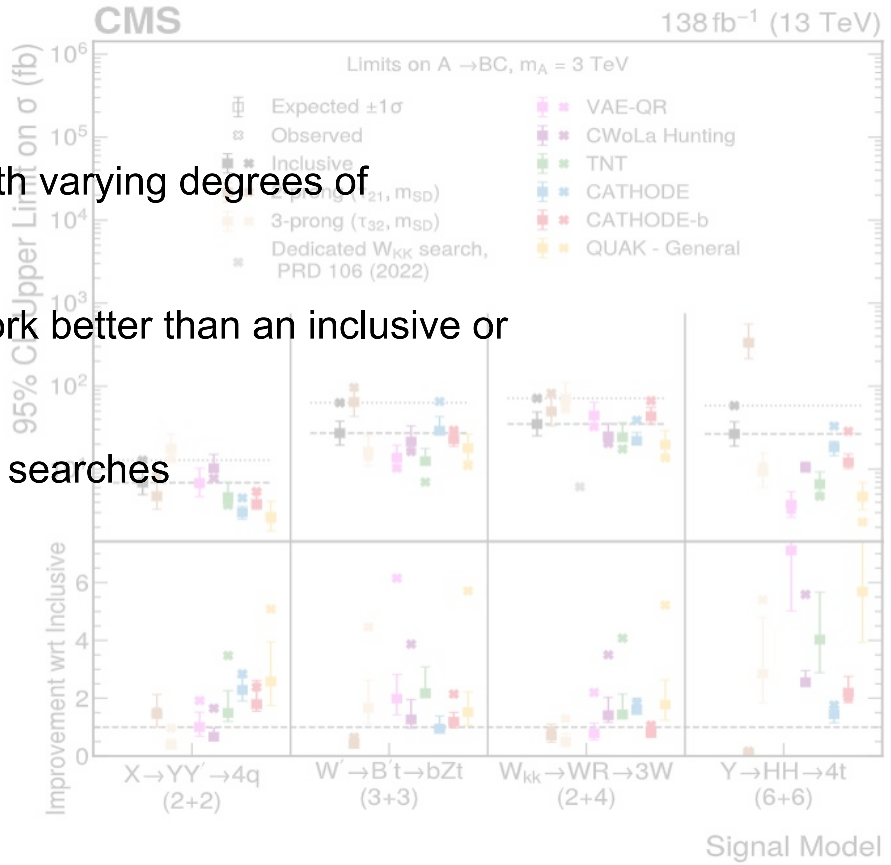
Discovery Potential at 3 TeV

- Compare methods by benchmarking on several signal models
- Find what injected cross section of signal would lead to a 3σ/5σ significance
- Better than inclusive, or simple cuts

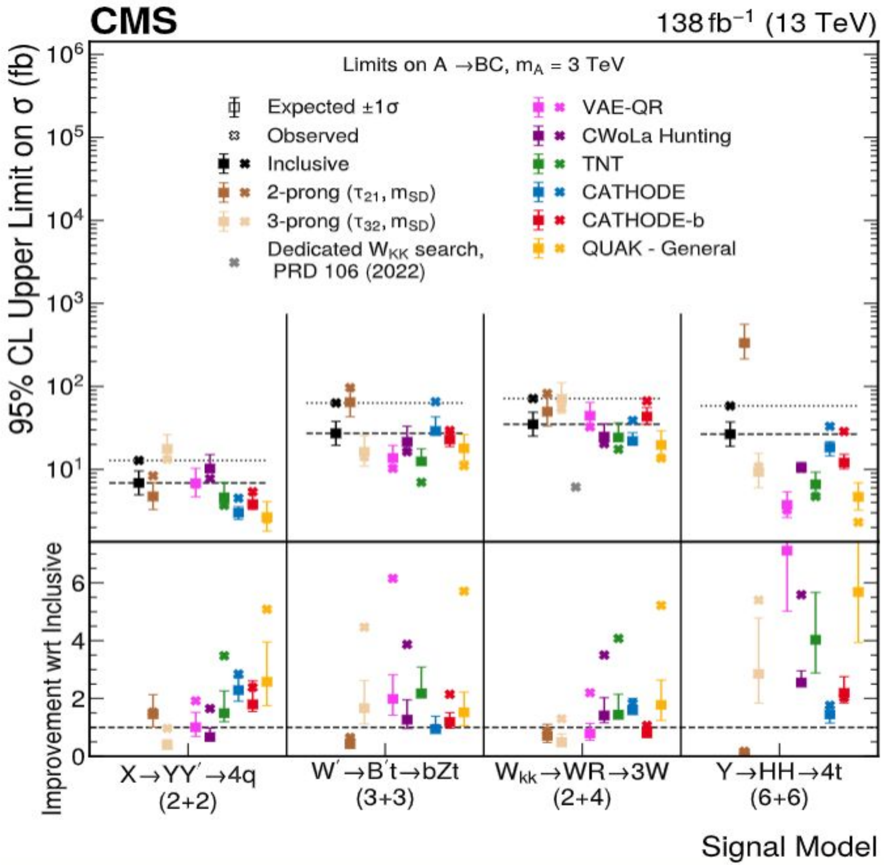


Limits - 3 TeV

- ❖ Look at all sorts of signals with varying degrees of substructure and pronginess
- ❖ All unsupervised methods work better than an inclusive or cut-based approach
- ❖ Not comparable to dedicated searches



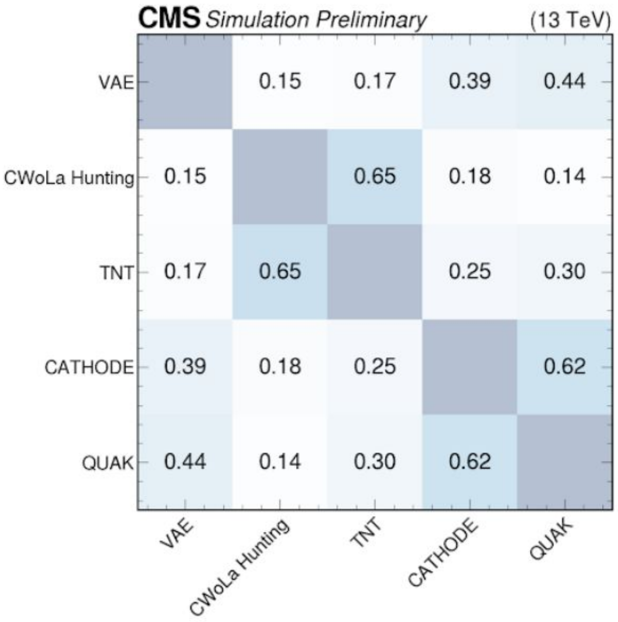
Limits - 3 TeV

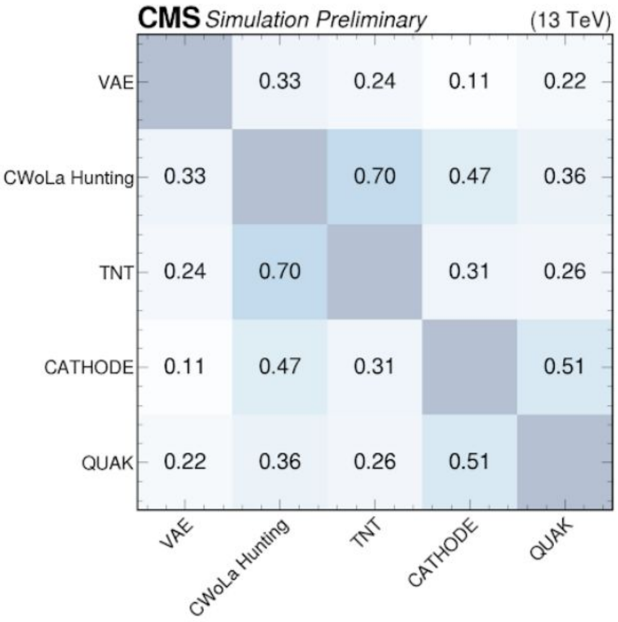


CMS
138 fb^{-1} (13 TeV)
Limits on A →BC, m_A = 3 TeV
Expected ±1σ
Observed
Inclusive
2-prong (τ_{21}, m_{SD})
3-prong (τ_{32}, m_{SD})
Dedicated W_{KK} search, PRD 106 (2022)
VAE-QR
CWoLa Hunting
TNT
CATHODE
CATHODE-b
QUAK - General
95% CL Upper Limit on σ (fb)
Improvement wrt Inclusive
X→YY′→4q (2+2)
W′→B′t→bZt (3+3)
W_{kk}→WR→3W (2+4)
Y→HH→4t (6+6)
Signal Model

Comparing Methods

- In general - no strong correlations between methods
- TNT and CWoLa are the most correlated → expected since the difference lies in the autoencoder preselection





Summary and Conclusions

- First results on data from the CMS Detector, using Unsupervised Anomaly Detection techniques
- Methods are sensitive to a broad range of signals - could flag any interesting deviations to direct dedicated searches
- Lots of scope for future work in anomaly detection with CMS - this was just the beginning
- Results already available on CDS [**CMS-EXO-22-026**] and will soon appear in a journal - stay tuned!

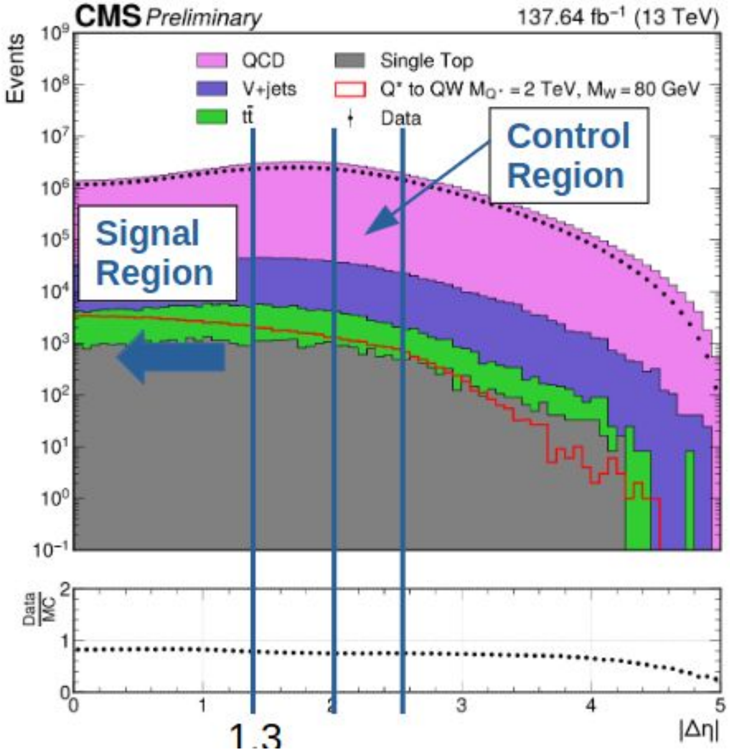
CMS-EXO-22-026

BACKUP

CASE: Input Features

- VAE: pT, η, ϕ of leading 100 particle flow constituents (per jet)
- CWoLa, TNT: mSD, τ_{21}, τ_{32}, τ_{43}, nPF, LSF3, b-tagging score (per jet)
- CATHODE: mSD1, mSD1 – mSD2, $\tau_{41,1}$, $\tau_{41,2}$ (per event)
- QUAK: mSD, τ_{21}, τ_{32}, τ_{43}, $\sqrt{\tau_{21}}/\tau_1$, M/pT (for each jet, per event)

Control Region Definition



- Signal region $|\Delta\eta| < 1.3$
- Control region $2.0 < |\Delta\eta| < 2.5$ + additional cuts
  - Extra cuts further suppress signal contam
  - Ensure signal reduction is at least 10x

Full Control Region Selection

$$\text{AND}\begin{cases} 2.0 < \Delta\eta < 2.5 \\ \text{No jet extra with } p_{\mathrm{T}} > 300\,\text{GeV} \\ \text{OR}\begin{cases} \left|\frac{p_{\mathrm{T},1}-p_{\mathrm{T},2}}{p_{\mathrm{T},1}+p_{\mathrm{T},2}}\right| > 0.1 \\ A = p_{\mathrm{T},1}p_{\mathrm{T},2}(2\cosh\Delta\eta + 2)/m_{jj}^2 \notin [0.95, 1] \end{cases} \end{cases}$$

Weak Supervision

Bin Name	Range (GeV)	Eff. Cut	Signal Masses (GeV)	Num. data events
A0	1350-1650	-	-	13.8M
A1	1650-2017	1%	1800, 1900	4.5M
A2	2017-2465	1%	2200, 2300	1.4M
A3	2465-3013	1%	2600, 2700, 2800	400k
A4	3013-3682	3%	3200, 3300, 3400, 3500	100k
A5	3682-4500	3%	3900, 4100, 4200, 4300	22k
A6	4500-5500	5%	4800, 4900, 5000, 5100, 5200	3.9k
A7	5500-8000	-	-	479
B0	1492-1824	-	-	6.6M
B1	1824-2230	1%	2000, 2100	2.1M
B2	2230-2725	1%	2400, 2500	630k
B3	2725-3331	1%	2900, 3000, 3100	170k
B4	3331-4071	3%	3600, 3700, 3800	42k
B5	4071-4975	3%	4400, 4500, 4600, 4700	8.5k
B6	4975-6081	5%	5300, 5400, 5500, 5600, 5700, 5800	1.3k
B7	6081-8000	-	-	144

m_{sd}, τ_{21}, τ_{32}, τ_{43}, n_{PF}, LSF_3, DeepB

CWoLa

Karlsruhe Institute of Technology

- Reweight events in SR and SB:
  - Upper and low mass sidebands reweighted to have same weight
  - Signal region also re-weighted to have weight equal to both SBs
  - Finally, reweight SR jets to have same p_T distribution as SB jet

- Two different network architectures used in different signal regions to prevent overfitting
  - Smaller network with O(3.6k) parameters used when SR events < 10k
  - Larger network with O(30k) parameters used otherwise

- Combining CWoLa scores (since there are 2 per-jet classifiers):
  - Convert each score to %ile using their distributions
  - Event anomaly score = max(S_1,S_2)
  - Finally define threshold as anomaly score that selects events with given efficiency (see table) in weighted average of sidebands, and use across whole mass spectrum for that SR

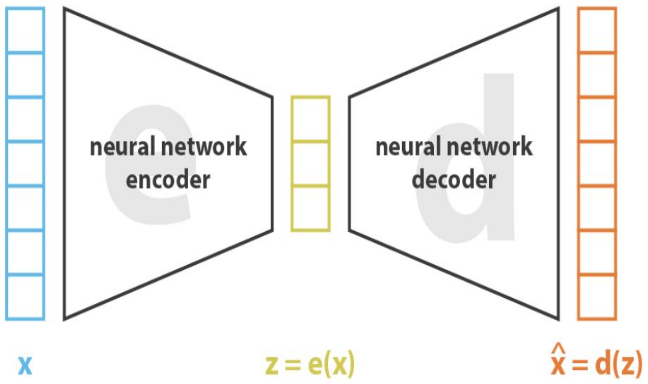
CWoLa + TNT inputs:

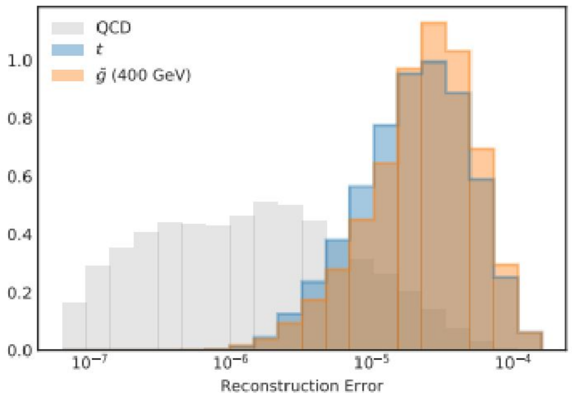
m_{sd}, τ_{21}, τ_{32}, τ_{43}, n_{PF}, LSF_3, DeepB

CATHODE

- Conditional normalizing flow - uses m_{JJ} as conditional input
- Train separate density estimator for m_{JJ} using a Gaussian Kernel Density Estimator
- $f^{-1}(z,m)$ with $z \sim N^n(0,1)$ and $m \sim KDE(m_{JJ})$ is used to generate synthetic samples

Autoencoders: Basics





- Goal: Pass through information bottleneck to reconstruct input
- Hidden (Latent) space: learns most important features
- Train on QCD sideband so network learns background but not signal - use reconstruction loss as anomaly metric
- Signal - high reconstruction loss
- Background - low reconstruction loss
- Variational Autoencoder: Gaussian latent space

VAE

- Latent space size of 12
- Training uses Chamfer loss + Kullback-Leibler divergence of between latent space & Gaussian
- Cross validation with 4 folds used for Quantile Regression
  - Average QR fit of other 3 folds used when selecting events on 4^{th}
- QR fits use dense NN with 5 layers and 30 nodes per layer
- Three categories used in limit setting
  - Cat1: Most anomalous 1% (>99%)
  - Cat2: Next most anomalous 4% (95-99%)
  - Cat3: Next most anomalous 5% (90-95%)
- In model-indep search, use single category, >90%

3-category fit: Use above defined three categories, fully correlate the backgrounds and fit with a single function

QUAK

- Uses Masked Autoregressive Rational Quadratic Spline (RQS) flows
- Chain of analysis:

1. Calculate the spline parameters:

$$\theta^j_{w,i}, \theta^j_{h,i}, \theta^j_{d,i} = \mathrm{NN}(z_i^{1:j-1})$$

where θ_w, θ_h, and θ_d specify the bin widths along the input (w), output (h) dimensions, and the internal derivatives (d).

2. Use the parameters to evaluate the spline and update the input:

$$z_i^j = \mathrm{RQS}_{\theta^j_{w,i},\theta^j_{h,i},\theta^j_{d,i}}(z^j_{i-1})$$

3. Repeat for all $j = 1, .., D$ (D = dimensionality of input $\mathbf{z}$).

QUAK selection

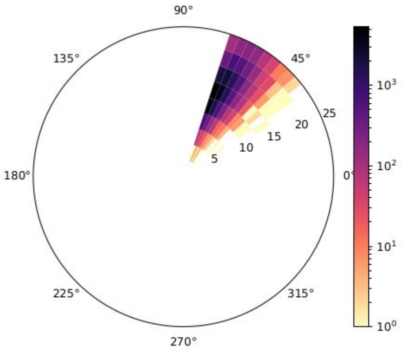
- Evaluate NLL Loss of each different model (1 bkg + 6 signal) on inputs
- Perform loss reduction on signal losses to get 2D loss vector

1. $(M_B, M_C) = (80, 80)$: the only signal sample used here was XYY2000_Y80_Yp80.
2. $(M_B, M_C) = (80, 170)$: combination of Wkk2000_R170, Wkk3000_R170, Wp2000_B80_T170, Wp3000_B80_T170, and XYY2000_Y80_Yp170 events.
3. $(M_B, M_C) = (80, 400)$: combination of Wkk2000_R400, Wkk3000_R400, XYY2000_Y400_Yp80, XYY2000_Y80_Yp400, and XYY3000_Y80_Yp400 events.
4. $(M_B, M_C) = (170, 170)$: combination of Wp2000_B170_T170, Wp3000_B170_T170, and XYY3000_Y170_Yp170 events.
5. $(M_B, M_C) = (170, 400)$: combination of Wp2000_B400_T170, Wp3000_B400_T170, XYY2000_Y170_Yp400, XYY2000_Y400_Yp170, and XYY3000_Y400_Yp170 events.
6. $(M_B, M_C) = (400, 400)$: combination of YHH2000_H400, YHH3000_H400, ZTT2000_Tp400, and ZTT3000_Tp400 events.

QUAK Selection

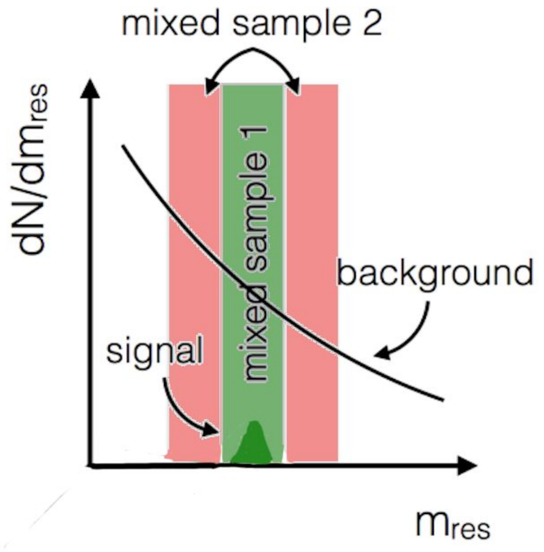
- Construct 2D QUAK Space with bkg and sig losses as described
- Select top X% of events with highest bkg. Loss and bin surviving in 2D QUAK space
- For given m_H define
  - SR: $[m_H - 400, m_H + 200]$ GeV
  - SBs: $[m_H - 900, m_H - 400]$ GeV and $[m_H + 200, m_H + 700]$ GeV

- Background template: Bin sideband in polar coordinates with $r < 10$ and $\theta \in [-0.1\pi, 0.4\pi]$
- Consider bins that are least populated in background in this template
  - Loop over these bins and select events from SR in these bins until at least 200 events selected



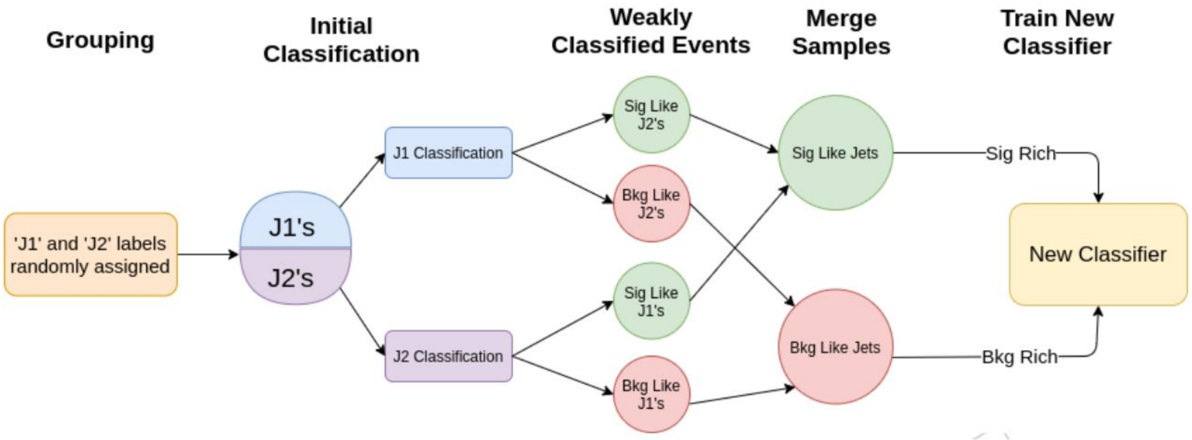
AD1 - CWoLa Hunting

- Assume signal is a narrow resonance and choose a mass window that is defined as SR (signal region) - signal enriched
- Define sidebands (SB) on either side of SR - background dominated
- Train classifier to distinguish SR from SB
- Use separate per-jet classifiers for heavier and lighter jet in each event
- Select events as per defined anomaly metric - function of classifier scores
- Jet features must be uncorrelated with m_{JJ}
  - Reweight SR events accordingly to match jet p_T in SB



AD2 - Tag N' Train (TNT)

- Similar to CWoLa - but uses a CNN-based Autoencoder for creating purer samples
- Tag first (second) jet in event as signal/background like using autoencoder score
  - Create mixed samples of second (first) jet in the event
- Samples can be combined since J1 and J2 labels are random
- Train new NN classifier using weak supervision

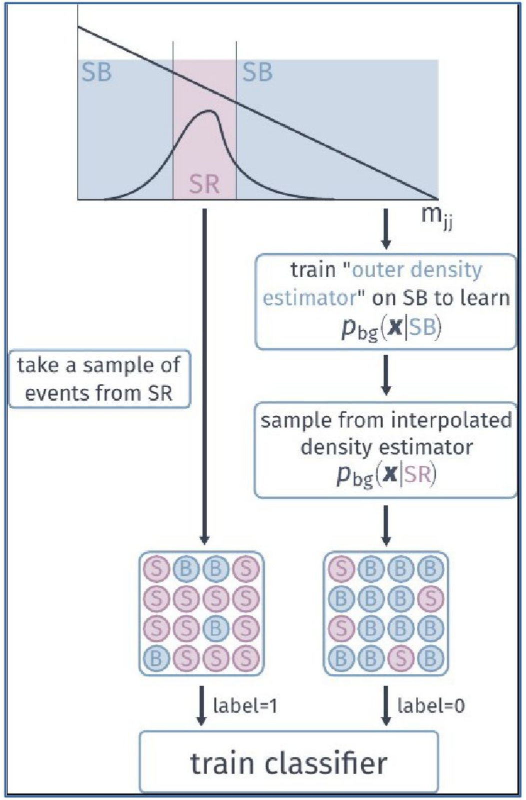


Works only if both jets in event are anomalous

Same p_T reweighting procedure as CWoLa

AD3 - CATHODE

- Train conditional normalizing flow to learn $p_{Bkg}(x|SB)$
- Interpolate into SR: $p_{bkg}(x|SR)$ using flow
- Train classifier to distinguish data in SR: $p_{Sig+Bkg}(x|SR)$ from interpolated events
- Noticeable improvement in classification performance



CATHODE-b: Uses DeepB scores as additional feature for training normalizing flow

AD5 - QUAK

Idea: train separate normalizing flows on background and signal MC

Use losses to construct a 2D QUAK space

- Every event mapped into a unique point in a 2D QUAK space
    - X-axis value comes from log-likelihood of event in normalizing flow trained on simulated QCD background events
    - Y-axis value comes from combining log-likelihood of event passed through 6 normalizing flows trained on different signal priors
    - Values normalized so background centered at (0,0)
- Select events by creating a unique 2D contour for each signal mass hypothesis designed to exclude background events
    - Contour created by using sidebands around hypothesis mass window (should be dominated by background)