

Isolating Unisolated Upsilons with Anomaly Detection in CMS Open Data

Radha Mastandrea

in collaboration with Rikab Gambhir, Benjamin Nachman, Jesse Thaler

QCD@LHC 2026

01/10/2025



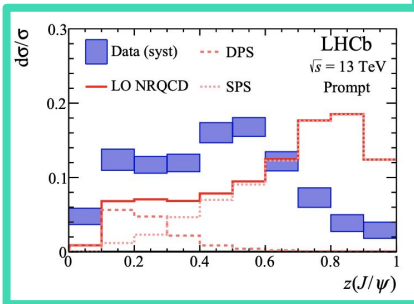
The elephant in the room: is this novel?

Isolating **Unisolated Upsilon**s with **Anomaly Detection** in **CMS Open Data**

The elephant in the room: is this novel?

Isolating **Unisolated Upsilon**s with **Anomaly Detection** in **CMS Open Data**

hep-ph/1701.05116

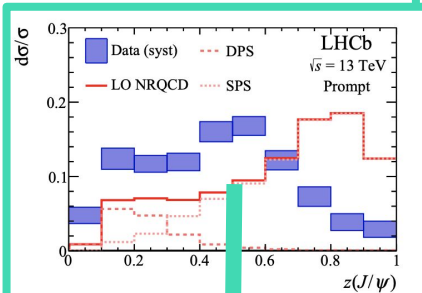


Studying quarkonia
produced within jets
probes the transition
between perturbative and
nonperturbative QCD

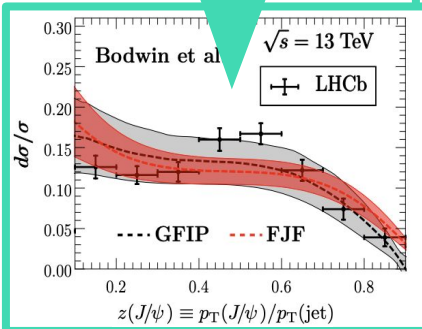
The elephant in the room: is this novel?

Isolating **Unisolated Upsilon**s with **Anomaly Detection** in **CMS Open Data**

hep-ph/1701.05116



Studying quarkonia produced within jets probes the transition between perturbative and nonperturbative QCD



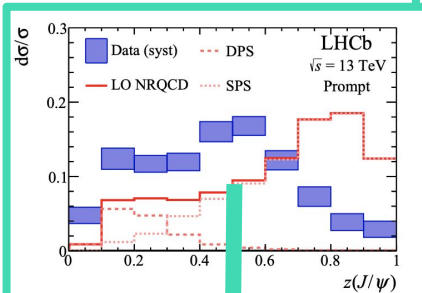
Jet fragmentation functions are improved by comparing to observations J/ψ 's within jets

hep-ph/1702.05525

The elephant in the room: is this novel?

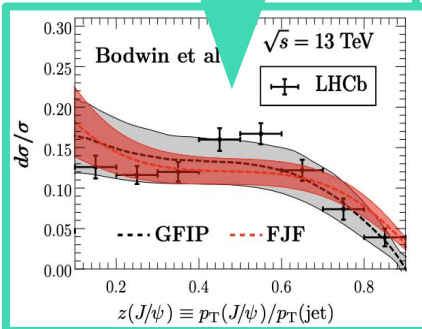
Isolating **Unisolated Upsilon**s with **Anomaly Detection** in **CMS Open Data**

hep-ph/1701.05116



Studying quarkonia produced within jets probes the transition between perturbative and nonperturbative QCD

We develop a new method for elevating significances that uses the full S / B likelihood



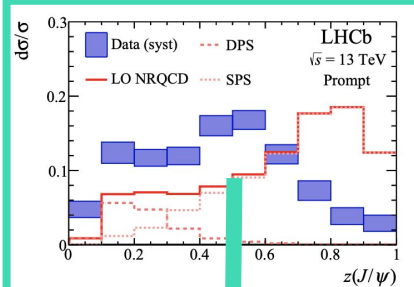
Jet fragmentation functions are improved by comparing to observations J/ψ 's within jets

hep-ph/1702.05525

The elephant in the room: is this novel?

Isolating **Unisolated Upsilon**s with **Anomaly Detection** in **CMS Open Data**

hep-ph/1701.05116

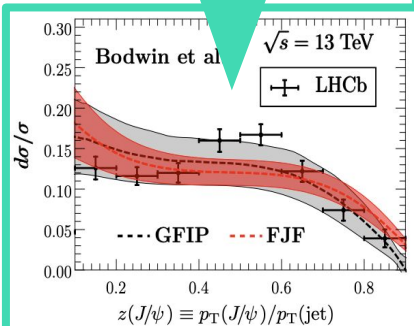


Studying quarkonia produced within jets probes the transition between perturbative and nonperturbative QCD

We develop a new method for elevating significances that uses the full S / B likelihood



Try it yourself!



Jet fragmentation functions are improved by comparing to observations J/ψ 's within jets

This is the first application of anomaly detection to open collider data

hep-ph/1702.05525



Dataset and event selection

We pull our data from the [DoubleMuon primary](#) dataset, made available on the [CERN Open Data Portal](#).

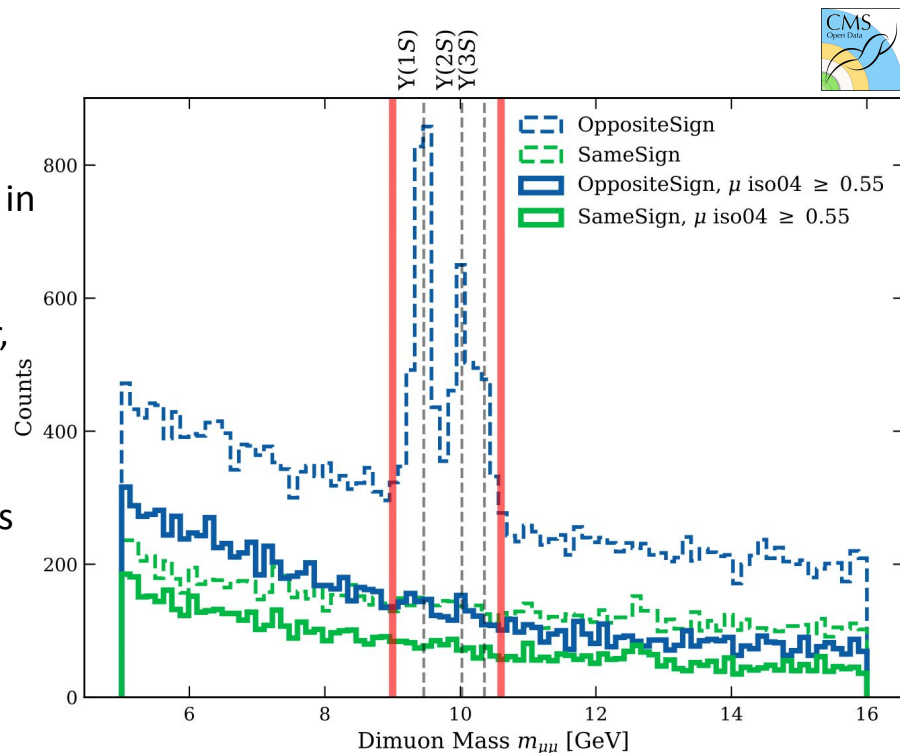
The dataset consists of $\mu\mu$ pairs produced during Run 2 in 2016 of the LHC at the CMS experiment, that

- ❖ pass the HLT_TrkMu15_DoubleTrkMu5NoFiltersNoVtx trigger,
- ❖ pass the p_T cuts 17 GeV, 8 GeV,
- ❖ pass the TightID criteria.

We further impose an **anti-isolation cut** to select for Y's produced in jets:

$$\mu_{\text{iso04}} = \sum_{\text{hadronic particles within } \Delta R=0.4} \frac{p_{T\text{hadronic}}}{p_{T\mu}} \geq 0.55$$

We are left with $\sim 12,000$ dimuon events.



Dataset and event selection

We pull our data from the [DoubleMuon primary](#) dataset, made available on the [CERN Open Data Portal](#).

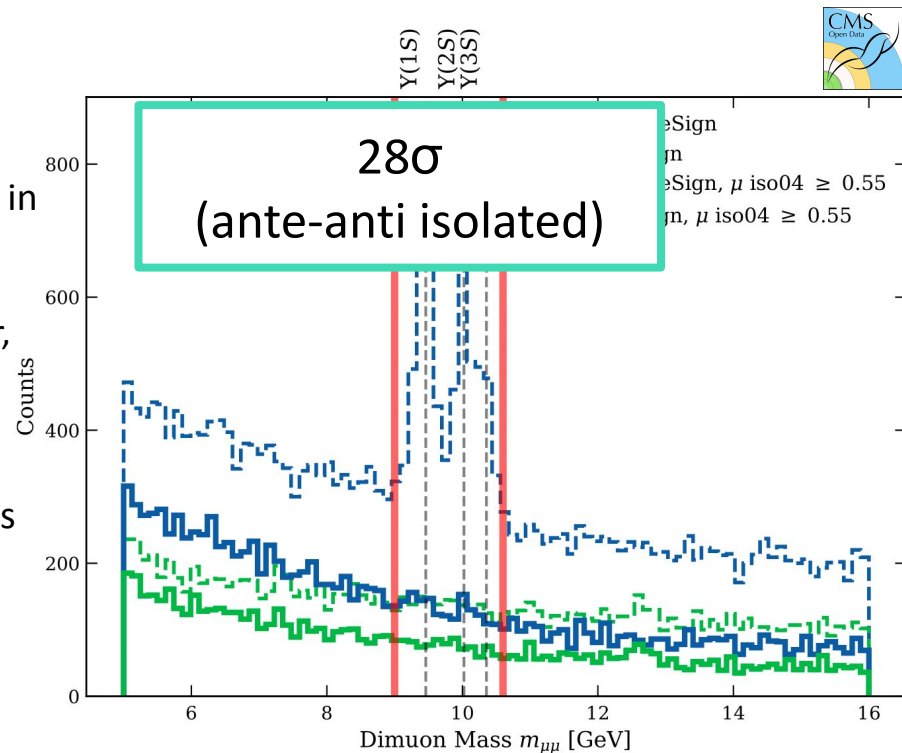
The dataset consists of $\mu\mu$ pairs produced during Run 2 in 2016 of the LHC at the CMS experiment, that

- ❖ pass the HLT_TrkMu15_DoubleTrkMu5NoFiltersNoVtx trigger,
- ❖ pass the p_T cuts 17 GeV, 8 GeV,
- ❖ pass the TightID criteria.

We further impose an **anti-isolation cut** to select for Y 's produced in jets:

$$\mu_{\text{iso04}} = \sum_{\text{hadronic particles within } \Delta R=0.4} \frac{p_{T\text{hadronic}}}{p_{T\mu}} \geq 0.55$$

We are left with $\sim 12,000$ dimuon events.



Dataset and event selection

We pull our data from the [DoubleMuon primary](#) dataset, made available on the [CERN Open Data Portal](#).

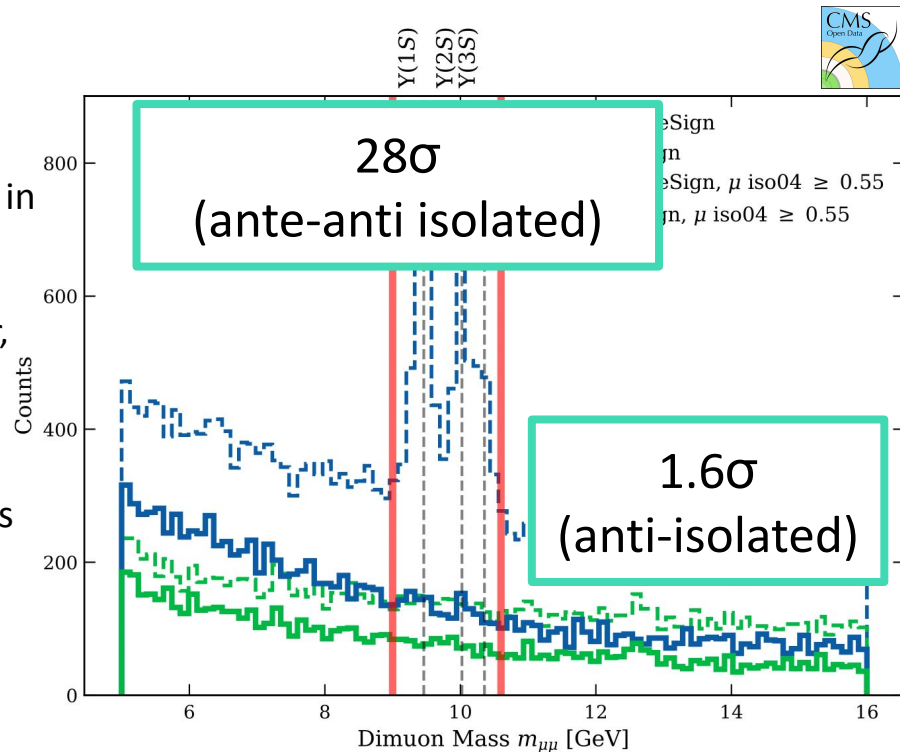
The dataset consists of $\mu\mu$ pairs produced during Run 2 in 2016 of the LHC at the CMS experiment, that

- ❖ pass the HLT_TrkMu15_DoubleTrkMu5NoFiltersNoVtx trigger,
- ❖ pass the p_T cuts 17 GeV, 8 GeV,
- ❖ pass the TightID criteria.

We further impose an **anti-isolation cut** to select for Y's produced in jets:

$$\mu_{\text{iso04}} = \sum_{\text{hadronic particles within } \Delta R=0.4} \frac{p_{T\text{hadronic}}}{p_{T\mu}} \geq 0.55$$

We are left with $\sim 12,000$ dimuon events.

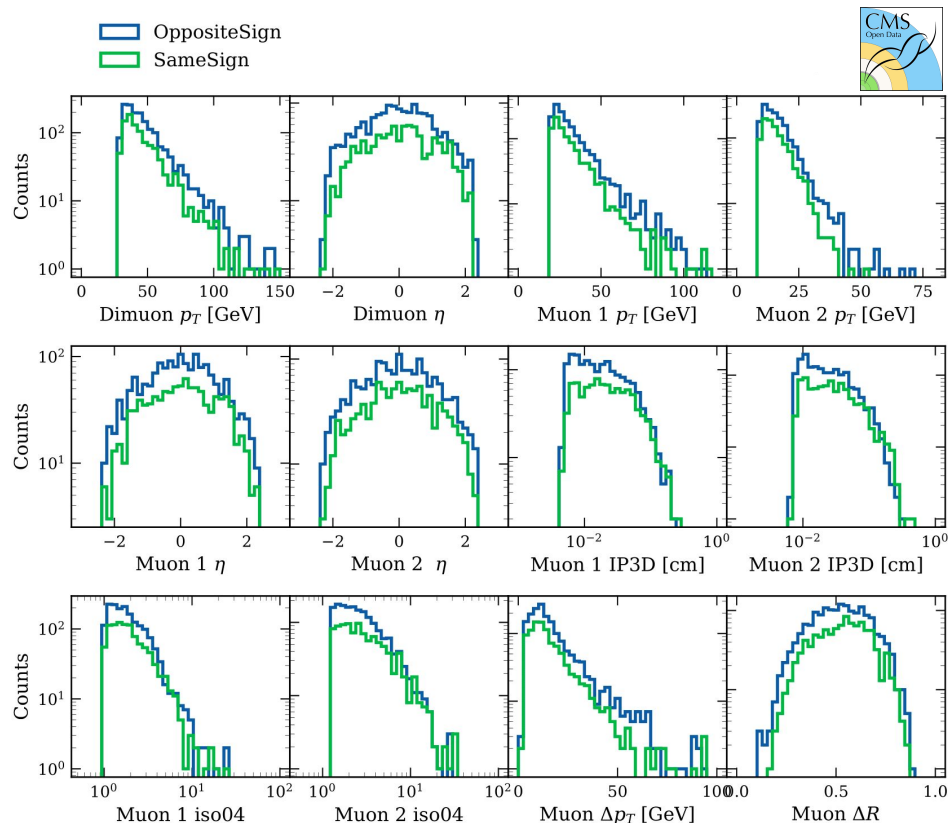


Auxiliary feature selection

We considered a large number of auxiliary observable sets, but one set worked particularly well*:

- ❖ Dimuon p_T
- ❖ Muon 1 3D impact parameter
- ❖ Muon 2 3D impact parameter

* e.g. no background sculpting,
good signal elevation



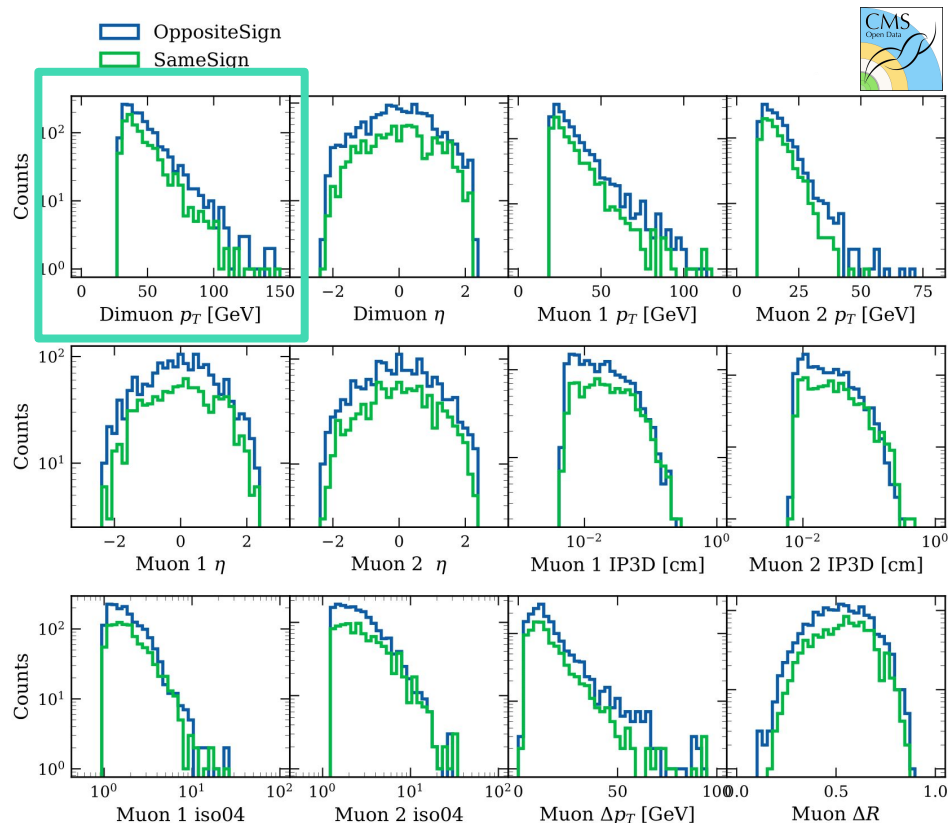
[Here](#) I show the results of 8 other feature sets, as well as their performance on a rediscovery of the η resonance.

Auxiliary feature selection

We considered a large number of auxiliary observable sets, but one set worked particularly well*:

- ❖ Dimuon p_T
- ❖ Muon 1 3D impact parameter
- ❖ Muon 2 3D impact parameter

* e.g. no background sculpting,
good signal elevation



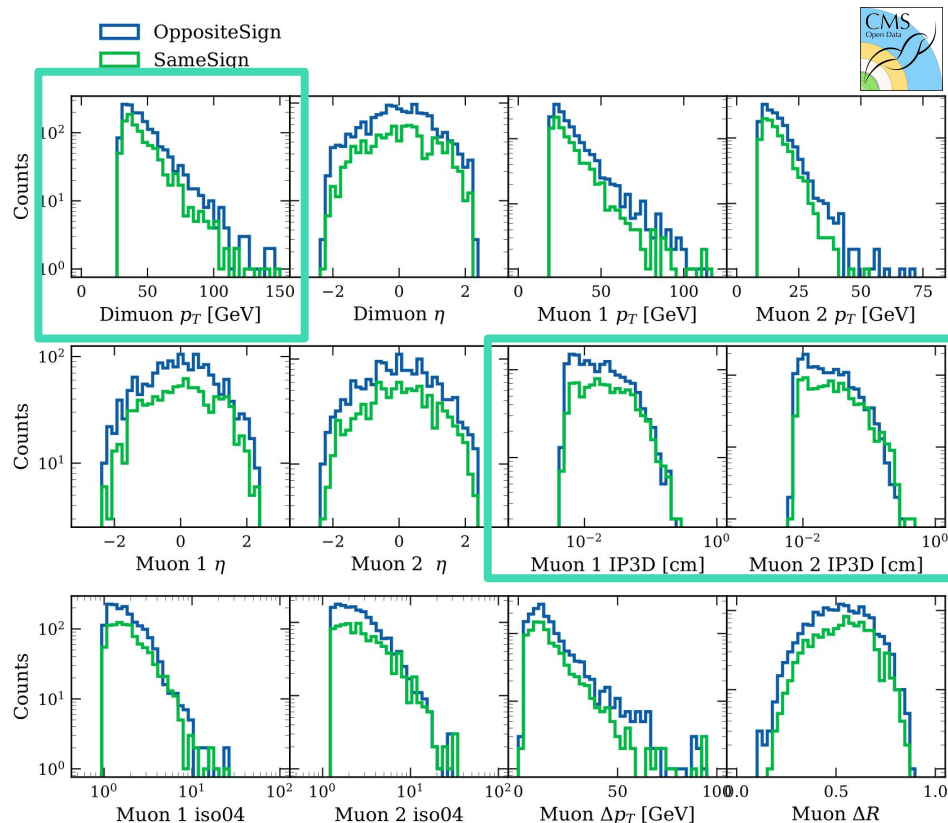
[Here](#) I show the results of 8 other feature sets, as well as their performance on a rediscovery of the η resonance.

Auxiliary feature selection

We considered a large number of auxiliary observable sets, but one set worked particularly well*:

- ❖ Dimuon p_T
- ❖ Muon 1 3D impact parameter
- ❖ Muon 2 3D impact parameter

* e.g. no background sculpting,
good signal elevation



[Here](#) I show the results of 8 other feature sets, as well as their performance on a rediscovery of the η resonance.

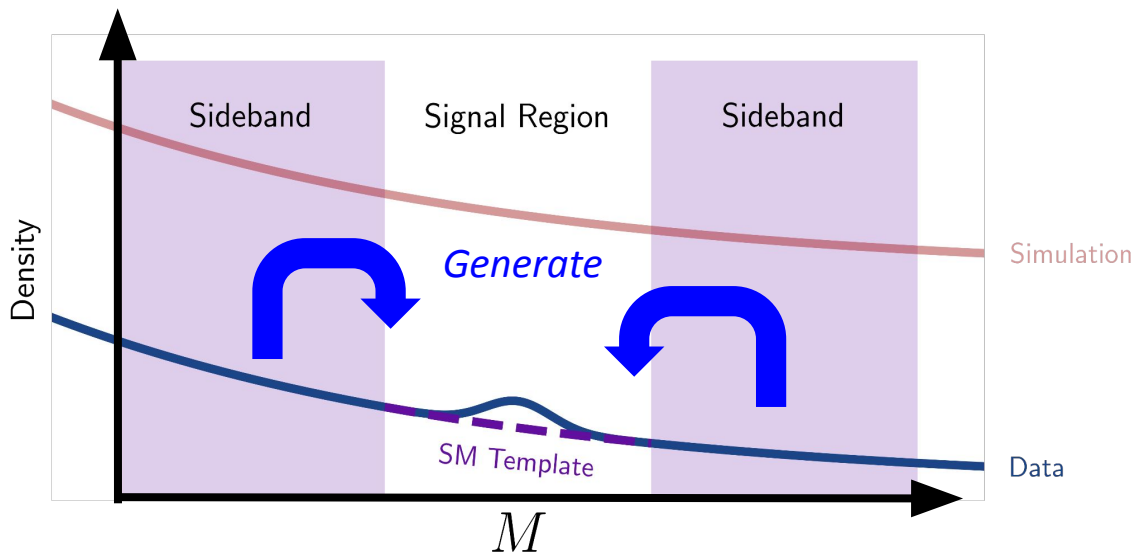
Analysis procedure: the extended bump hunt

- ❖ Use **CATHODE** to generate a set of background-like (i.e. non-Y-like) samples

See backup slides for architecture details.

Generating background samples with CATHODE

Classifying Anomalies THrough Outer Density Estimation
([hep-ph/2109.00546](#))



- ❖ A normalizing flow learns the **distribution** of **data** (i.e. the auxiliary observables) in sidebands, conditioned on the dimuon mass
- ❖ The flow is interpolated into the signal region to **generate** realistic background samples

Analysis procedure: the extended bump hunt

- ❖ Use **CATHODE** to generate a set of background-like (i.e. non-Y-like) samples
- ❖ Train Boosted Decision Trees to discriminate **CATHODE** samples from data. (i.e. Classification Without Labels, [hep-ph/1708.02949](https://arxiv.org/abs/hep-ph/1708.02949)). The classifiers learn:

$$C(x) = \frac{p_{\text{Data}}(x)}{p_{\text{Samples}}(x)} = \frac{\mu p_{\text{Signal}}(x) + (1 - \mu) p_{\text{Background}}(x)}{p_{\text{Background}}(x)} = \mu \frac{p_{\text{Signal}}(x)}{p_{\text{Background}}(x)} + (1 - \mu)$$

- ❖ Cut on the BDT score, which preferentially keeps high S / B events
- ❖ Compare signal significances from cutting on the BDT score with cutting on kinematic observables (i.e. classical tests)

See backup slides for architecture details.

Analysis procedure: full likelihood reweighting

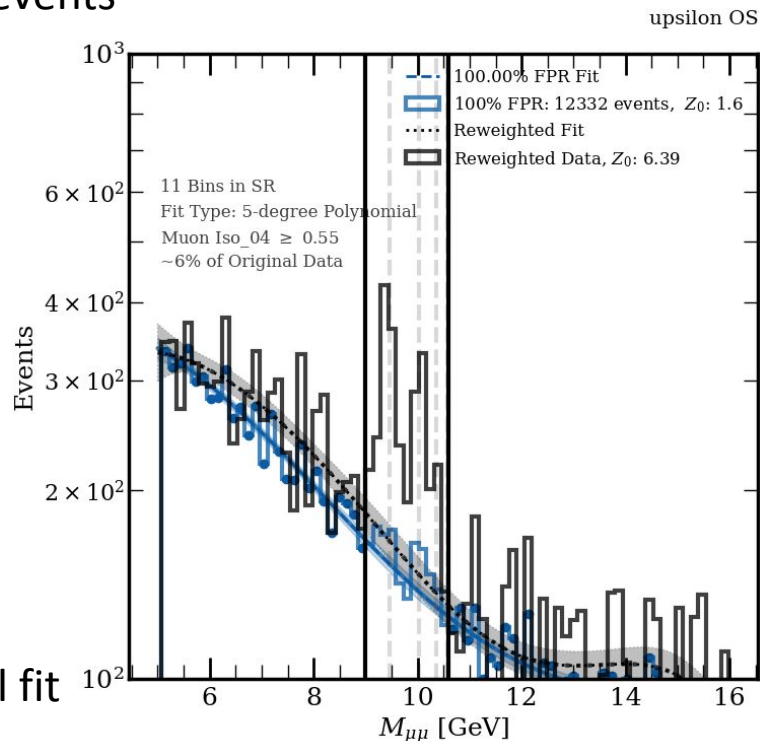
- ❖ For a more powerful test, we can reweight the events with the classifier-derived S / B:

$$w(x) = \frac{p_{\text{Signal}}(x)}{p_{\text{Background}}(x)} = \frac{C(x) - (1 - \mu)}{\mu}$$

where

$$\mu = \frac{N_{\text{Signal}}}{N_{\text{Signal}} + N_{\text{Background}}}$$

and we estimate $N_{\text{background}}$ from a quintic polynomial fit



Full-likelihood reweighting: a heuristic proof*

- ❖ We can consider a single-bin experiment where:
- ❖ To elevate the signal significance, we are free to introduce weights:
- ❖ Maximizing Z with respect to $w(x)$, we find that:

$$Z_{\text{unweighted}} = \frac{S}{\delta B}$$

$$Z_{\text{weighted}} = \frac{\int dx S(x) w(x)}{\sqrt{\int dx B(x) w^2(x)}}$$

$$w(x) = \frac{S(x)}{B(x)} \propto \frac{p_{\text{Signal}}(x)}{p_{\text{Background}}(x)}$$

*Full proof sketched out in [hep-ph/0909.2862](https://arxiv.org/abs/hep-ph/0909.2862)

Full-likelihood reweighting: a heuristic proof*

Actually, this isn't so different from the cut-and-count test! From [Rikab's previous talk](#):

~~Weight~~-and-Count Anomaly Detection (Details)

The effect of weighting is to replace *Poissons* with *Weighted (Compound¹) Poissons*

$$N_{\text{cut}} = \sum_i^N \Theta(f(x_i) - f_{\text{cut}}) \longrightarrow N_w = \sum_i^N w(x_i)$$

Observable: Bin Count N

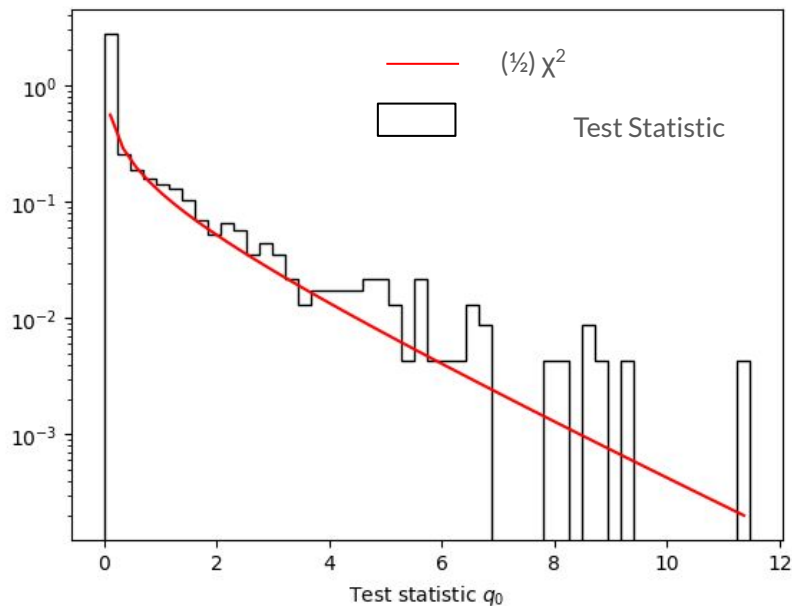
Observable: Weighted Bin Count, depends on N and $\langle w \rangle$

*Full proof sketched out in [hep-ph/0909.2862](https://arxiv.org/abs/hep-ph/0909.2862)

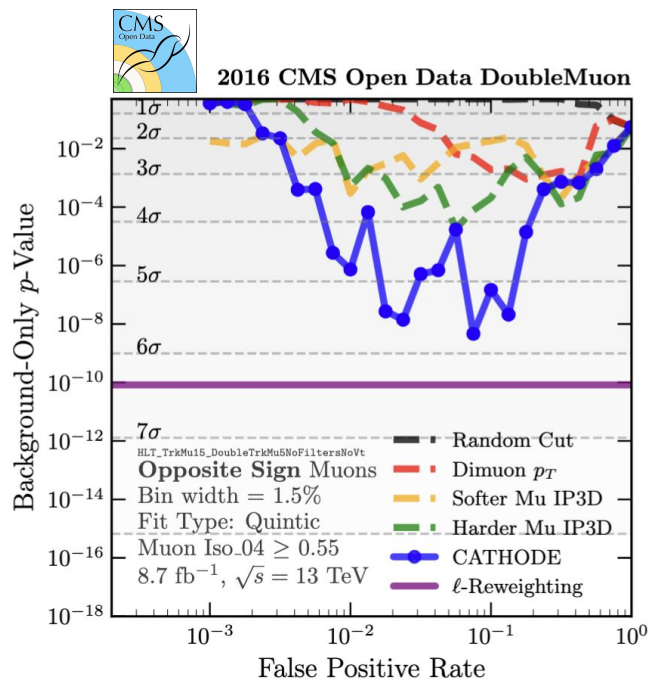
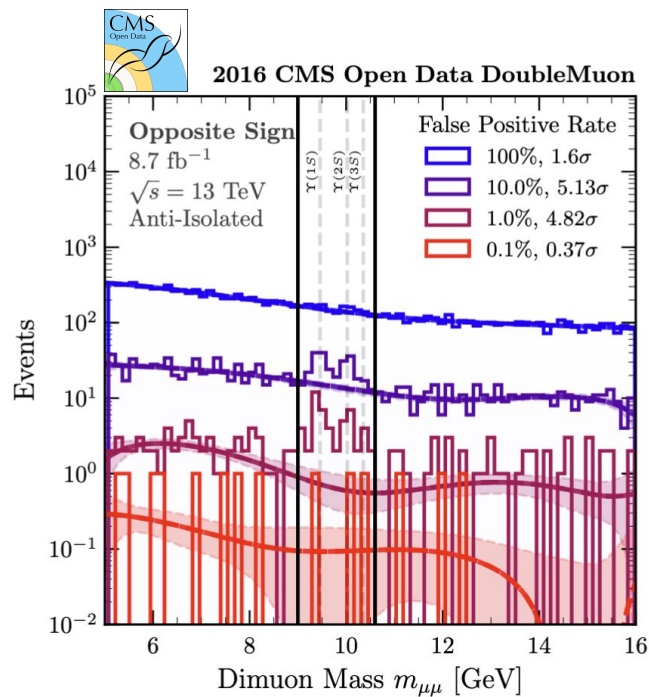
Calibration study: test statistic distribution

We can calculate the test statistic for the full likelihood-reweighted test statistic on 1000 same-sign pseudoexperiments.

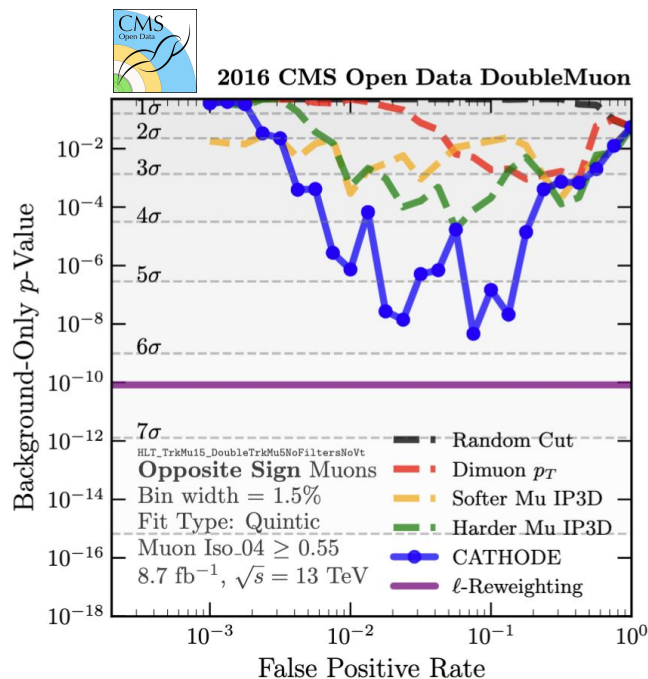
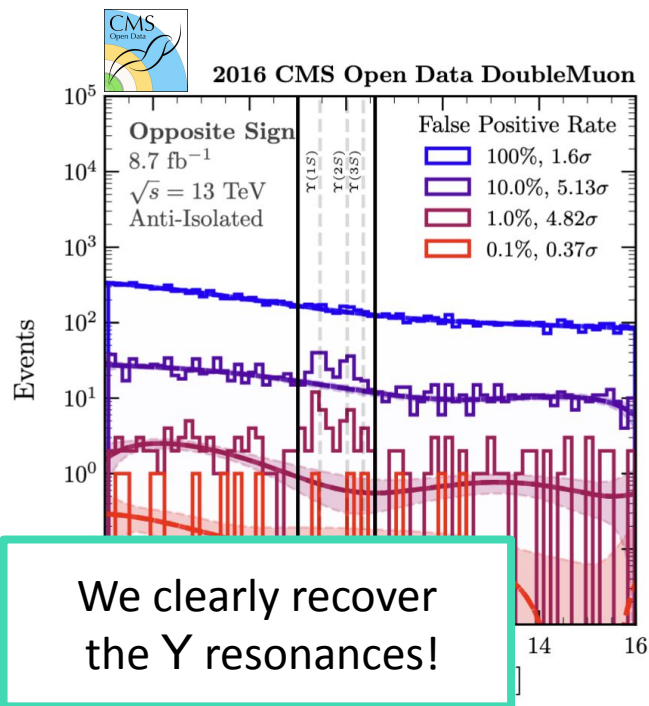
The distribution follows a half-chi2 distribution, so the significance is calibrated!



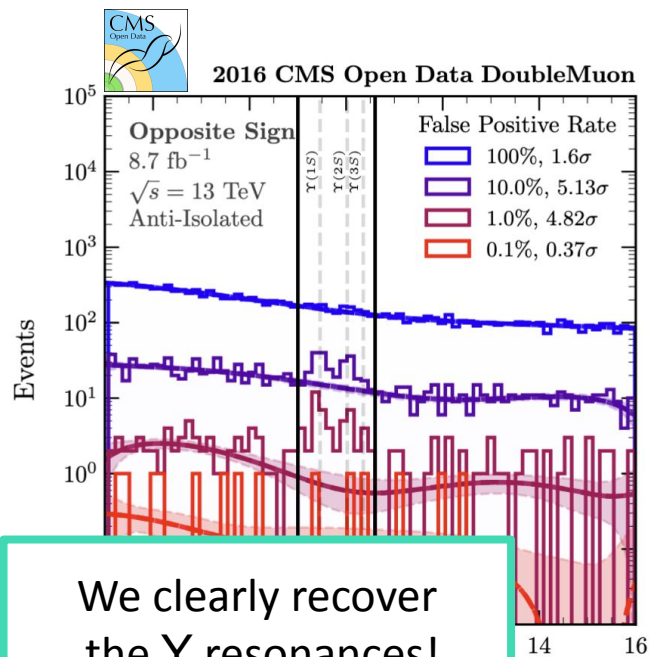
Results: finding Y 's within jets



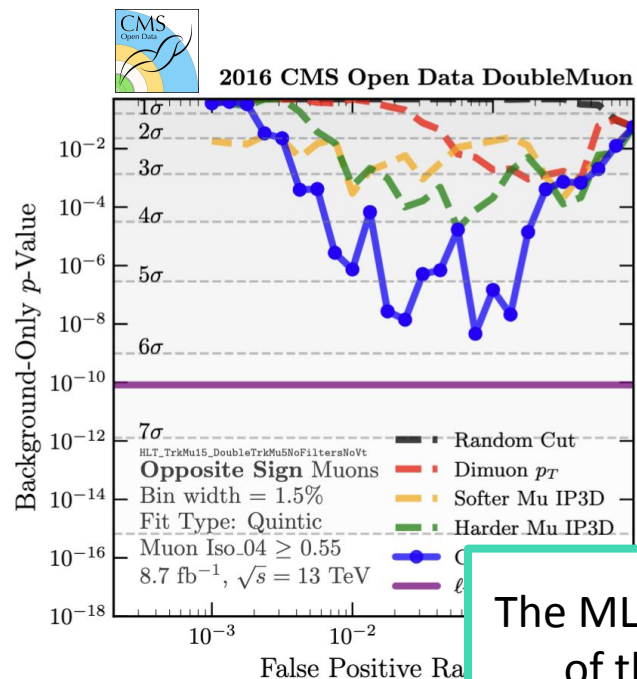
Results: finding Y 's within jets



Results: finding Y 's within jets



We clearly recover
the Y resonances!



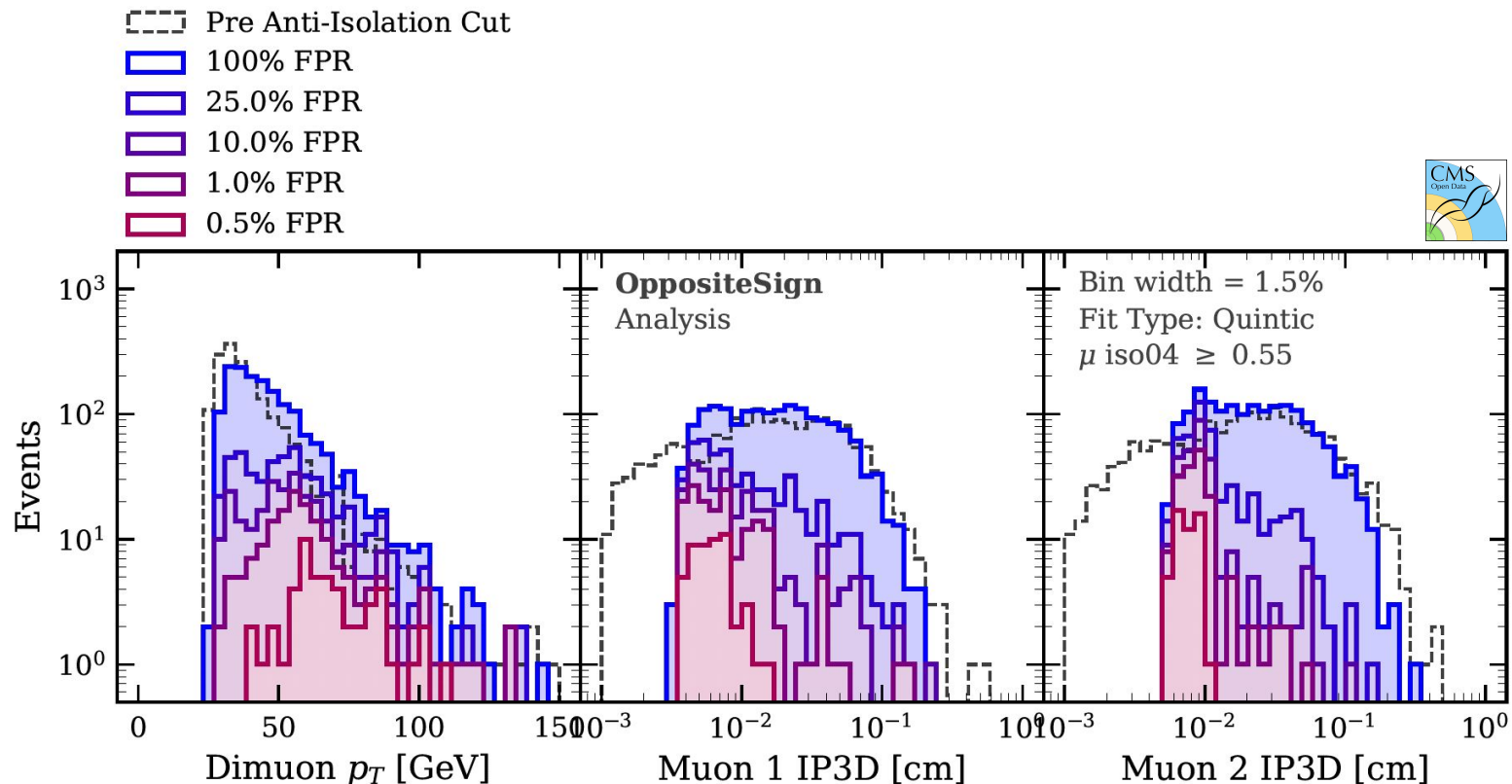
best classical test: 4.1σ

Cut-and-count: 5.7σ

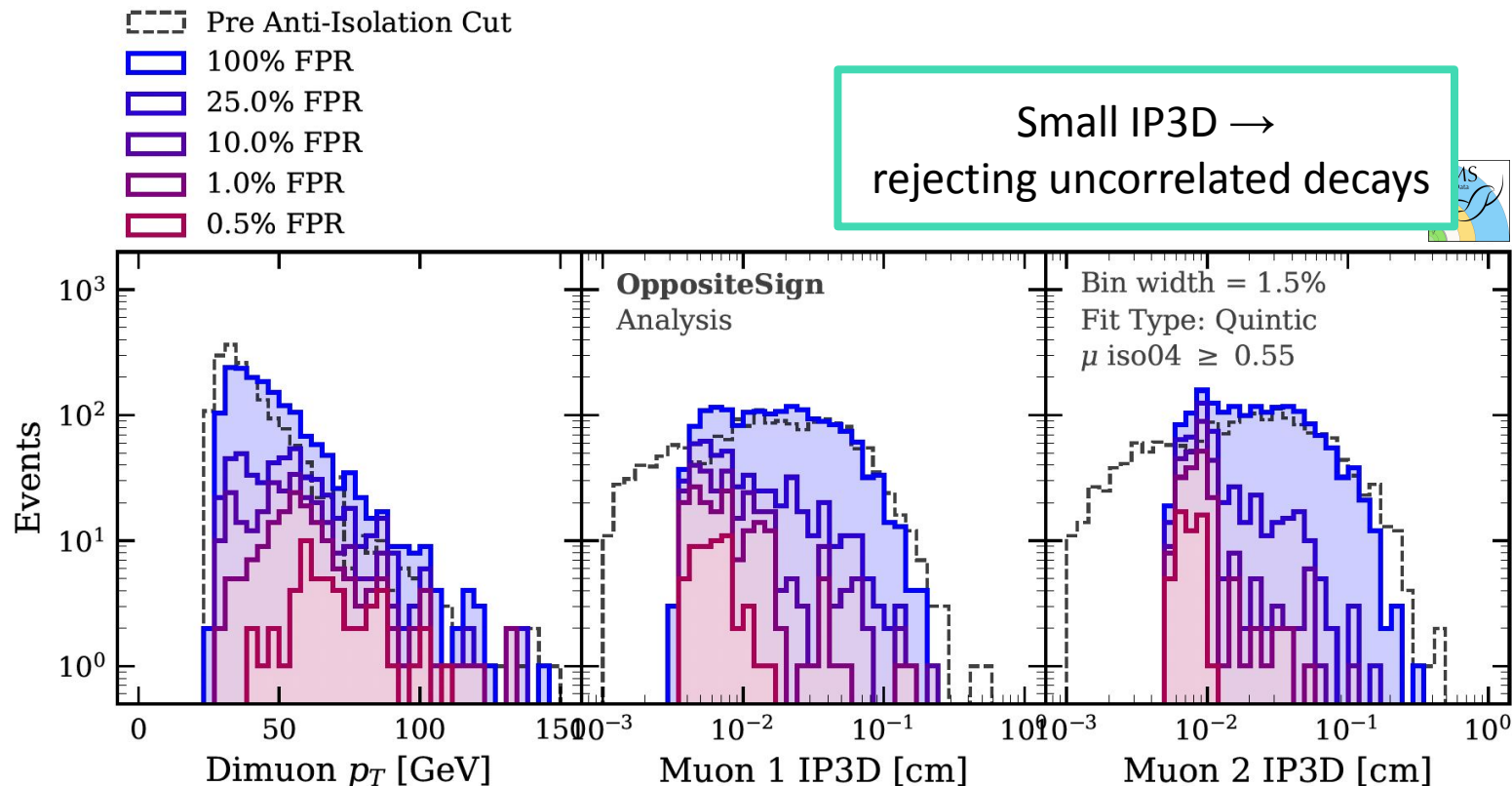
Full-likelihood
reweighting: 6.4σ

The ML tests outperform all
of the classical tests!

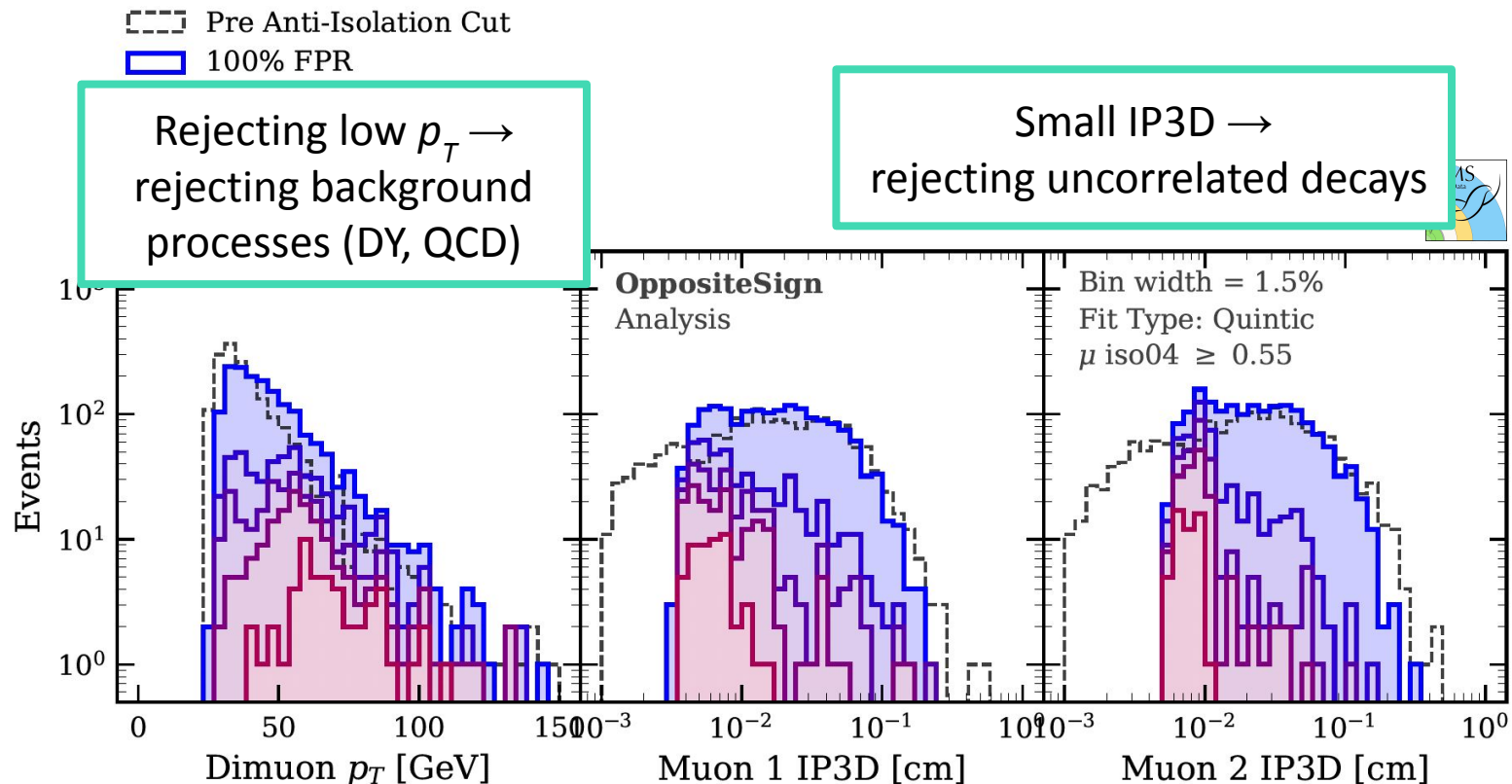
The BDT learns genuine physical properties of Y 's



The BDT learns genuine physical properties of Y 's

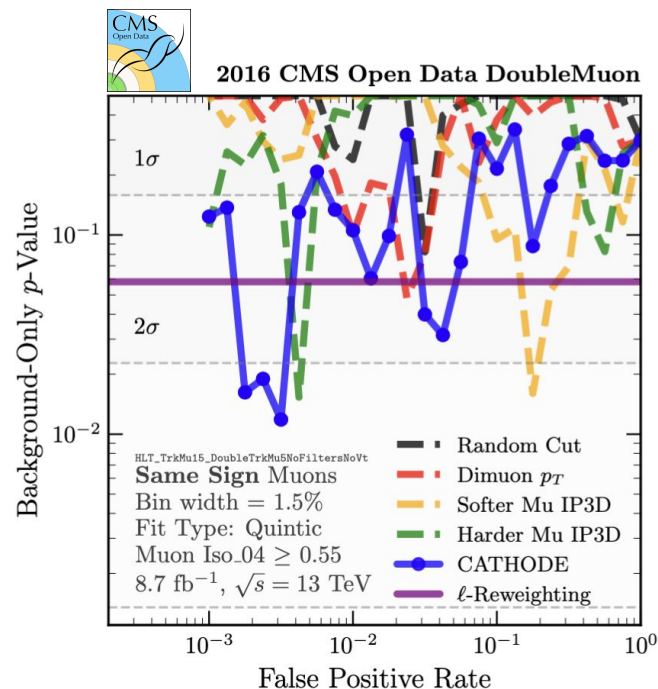
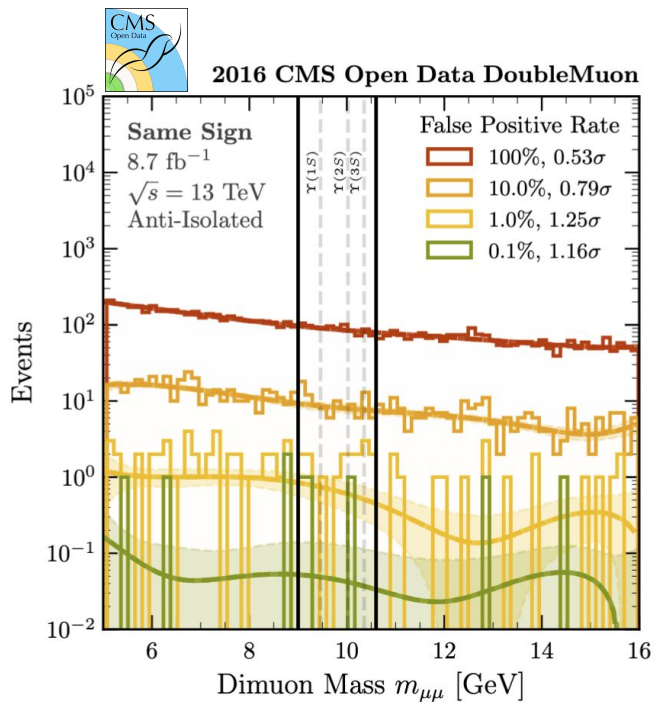


The BDT learns genuine physical properties of Y 's



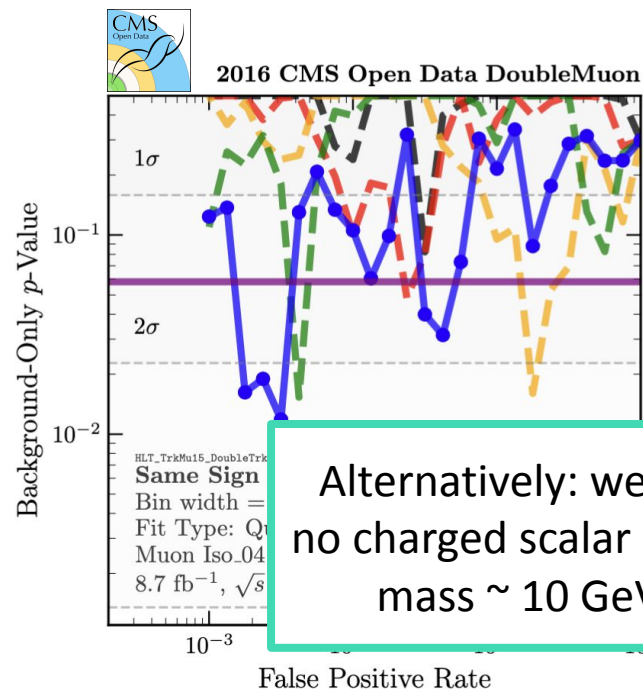
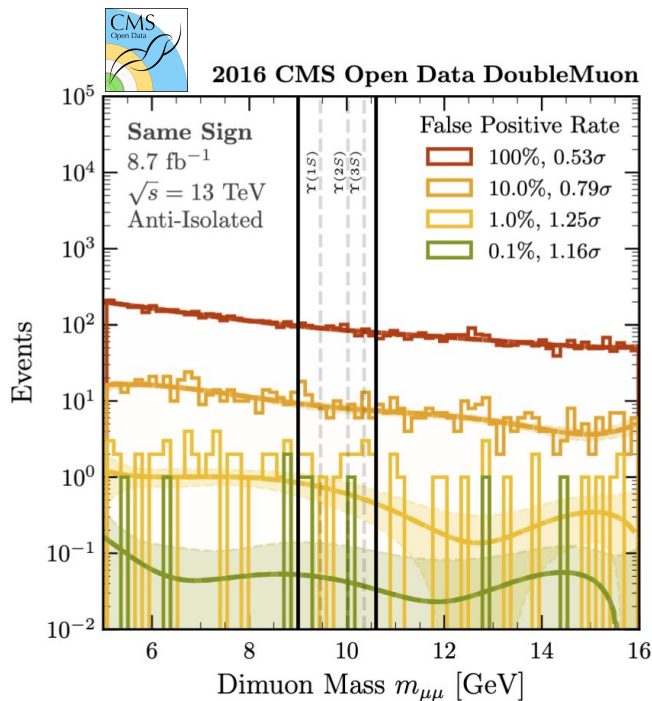
Validating the analysis: same sign resonance “search”

We rerun the entire analysis using same-sign dimuon pairs and find no evidence of sculpting

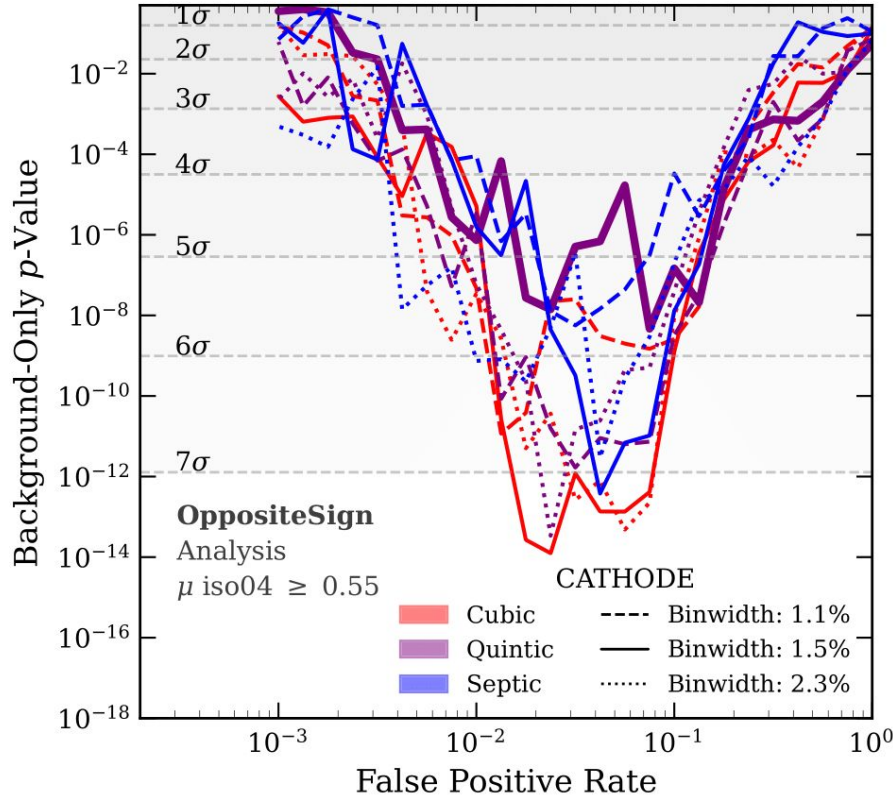


Validating the analysis: same sign resonance “search”

We rerun the entire analysis using same-sign dimuon pairs and find no evidence of sculpting



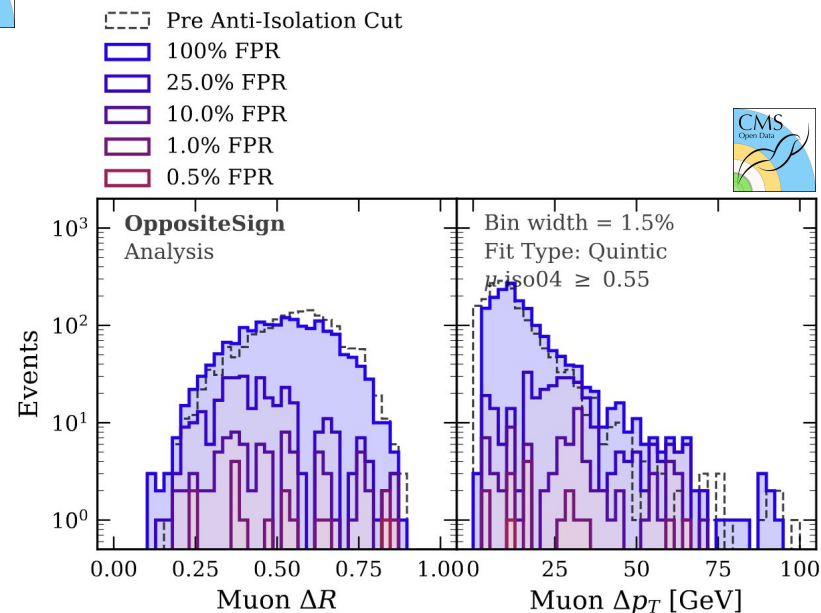
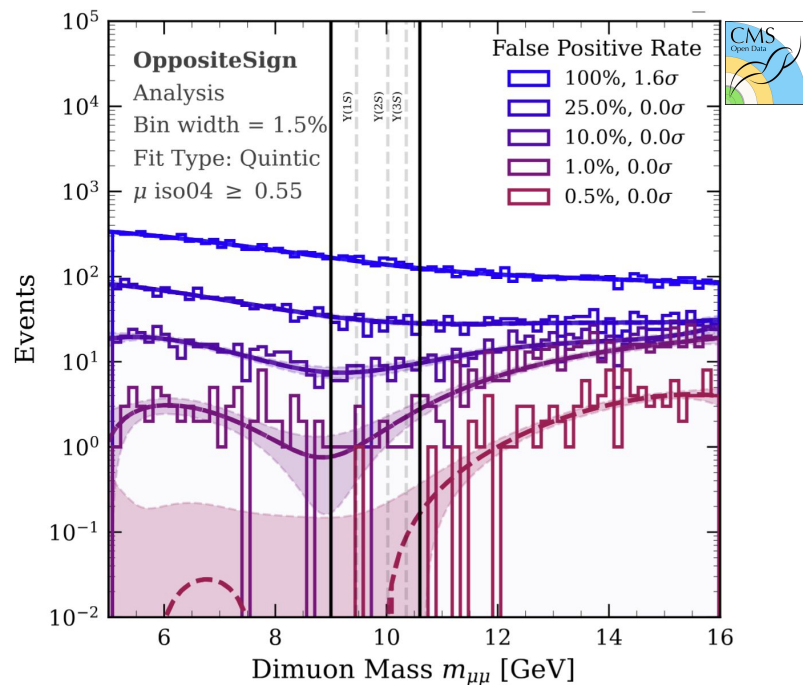
Validating the analysis: systematic variations



The ML cut-and-count (and full likelihood reweighting, though not shown here) analyses are robust with respect to the polynomial fit parameters

Not all observables led to Υ discovery!

Many feature sets had significant background sculpting and / or did not elevate the Υ signal



[Here](#) I show the results of 8 other feature sets, as well as their performance on a rediscovery of the η resonance.

Conclusions

1. We carry out the first study of anti-isolated upsilons (Y 's) produced within jets, which can probe the perturbative-nonperturbative QCD transition
2. We develop a new method for elevating signal significances within an extended bump hunt framework that uses the full S / B likelihood
3. We show that it is possible to use ML to find nontrivial signals in real collider data

Most importantly, this study highlights the work left to be done!

- Real data is messy and has irrelevant features.
- Background sculpting is a huge problem!

We encourage the community to test out new methods on our distilled Y dataset →

GitHub



Data



**Please fill out
this survey!**



Backup slides

Muon TightID

Plain-text description	Technical description	Comments
The candidate is reconstructed as a Global Muon	<code>recoMu.isGlobalMuon()</code>	
Particle-Flow muon id	<code>recoMu.isPFMuon()</code>	the exclusive effect of this requirement is very small, i.e. PFMuon is keeping almost all Tight Muons without this cut
χ^2/ndof of the global-muon track fit < 10	<code>recoMu.globalTrack()->normalizedChi2() < 10.</code>	To suppress hadronic punch-through and muons from decays in flight (see CMS AN 2008/098). This cut might need to be re-tuned due to the change to fully segment based global fit in 50X releases and later. It will need a retuning when muon APEs will be activated in the global track fit.
At least one muon chamber hit included in the global-muon track fit	<code>recoMu.globalTrack()->hitPattern().numberOfValidMuonHits() > 0</code>	To suppress hadronic punch-through and muons from decays in flight.
Muon segments in at least two muon stations This implies that the muon is also an arbitrated tracker muon, see SWGGuideTrackerMuons	<code>recoMu.numberOfMatchedStations() > 1</code>	To suppress punch-through and accidental track-to-segment matches. Also makes selection consistent with the logic of the muon trigger, which requires segments in at least two muon stations to obtain a meaningful estimate of the muon p_T .
Its tracker track has transverse impact parameter $d_{xy} < 2$ mm w.r.t. the primary vertex	<code>fabs(recoMu.muonBestTrack()->dxy(vertex->position())) < 0.2</code> Or <code>dB() < 0.2</code> on <code>pat::Muon [1]</code>	To suppress cosmic muons and further suppress muons from decays in flight (see CMS AN 2008/098). The 2 mm cut preserves efficiency for muons from decays of b and c hadrons. It is a loose cut and can be tightened further with minimal loss of efficiency for prompt muons if background from cosmic muons is an issue. Another way to obtain a better cosmic-ray suppression is to complement the d_{xy} cut with a cut on the opening angle α or use a dedicated cosmic-id algorithm (see Section 7.1 of MUO-10-004). <code>innerTrack()</code> is also supported for dxy cut, as the performance of the two is very close.
The longitudinal distance of the tracker track wrt. the primary vertex is $d_z < 5$ mm	<code>fabs(recoMu.muonBestTrack()->dz(vertex->position())) < 0.5</code>	Loose cut to further suppress cosmic muons, muons from decays in flight and tracks from PU. <code>innerTrack()</code> is also supported for dz cut, as the performance of the two is very close.
Number of pixel hits > 0	<code>recoMu.innerTrack()->hitPattern().numberOfValidPixelHits() > 0</code>	To further suppress muons from decays in flight.
Cut on number of tracker layers with hits > 5	<code>recoMu.innerTrack()->hitPattern().trackerLayersWithMeasurement() > 5</code>	To guarantee a good p_T measurement, for which some minimal number of measurement points in the tracker is needed. Also suppresses muons from decays in flight.

Network architectures and training specifications

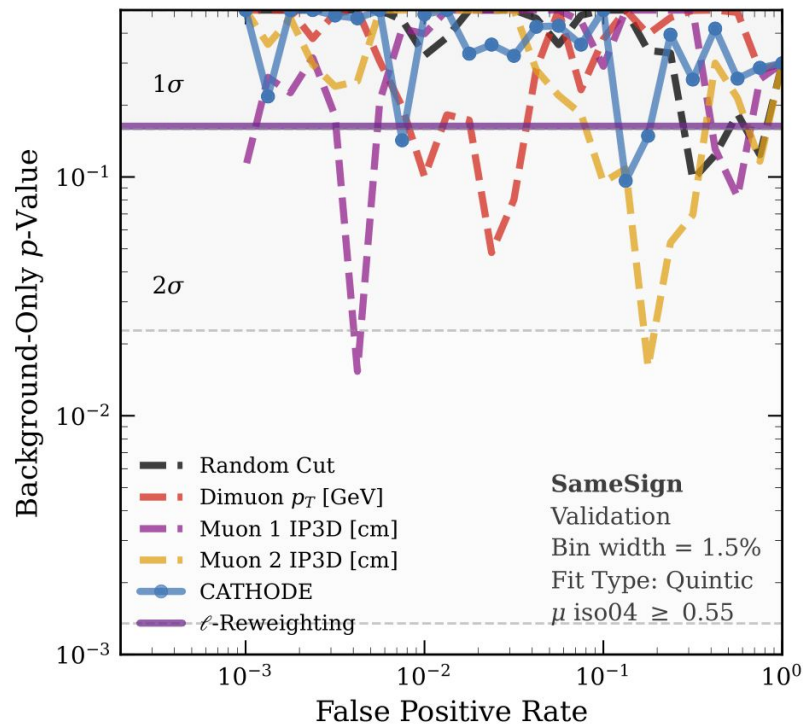
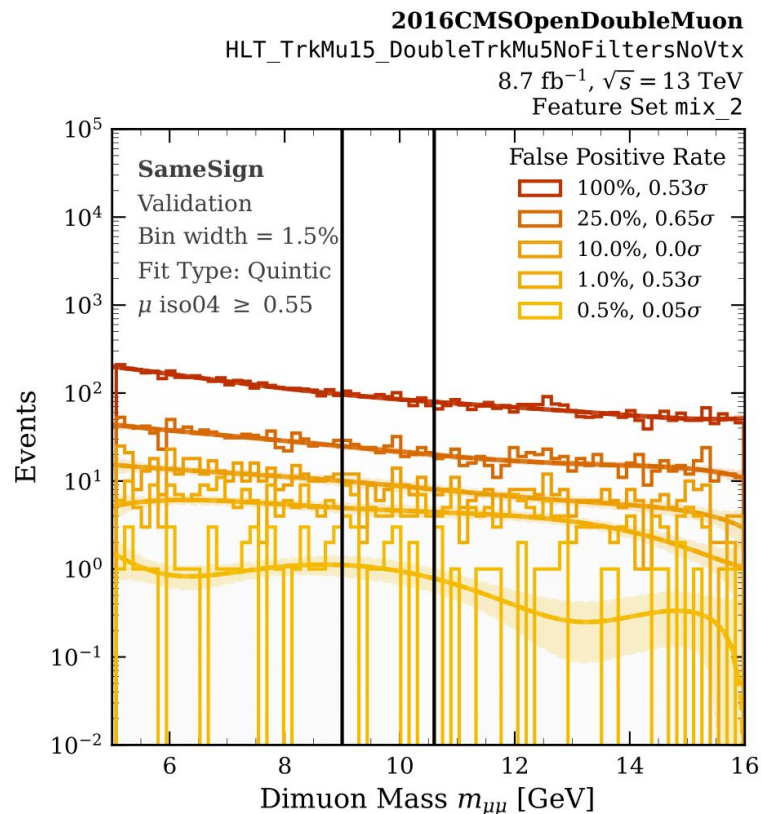
Normalizing flow architecture

- Masked Autoregressive Flow with Affine transformations
 - 1 layer of 6 MADE blocks each with 128 hidden features and each followed by a Reverse Permutation layer.
 - We use tanh activation functions and apply batch normalization after each epoch with momentum 1.
- The flow is trained for up to 1000 epochs with a learning rate of $7.5e-4$ and weight decay of $1e-6$. The network is evaluated at (i.e. sampled from) the epoch of lowest validation loss.
- To mitigate random model initialization and stochastic gradient descent fluctuations, we train an ensemble of five CATHODE models and draw SM proxy samples from each one (oversampling factor of 5).

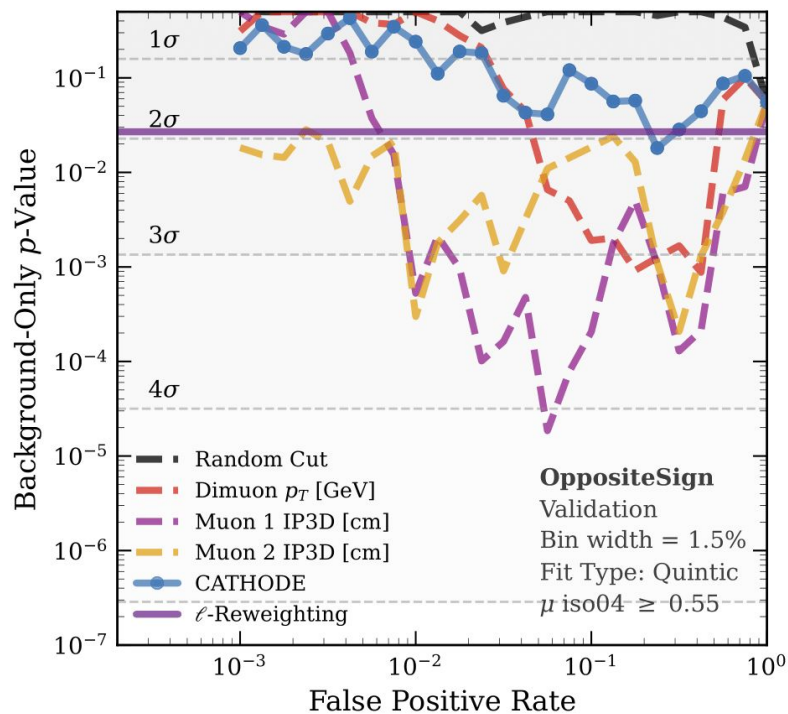
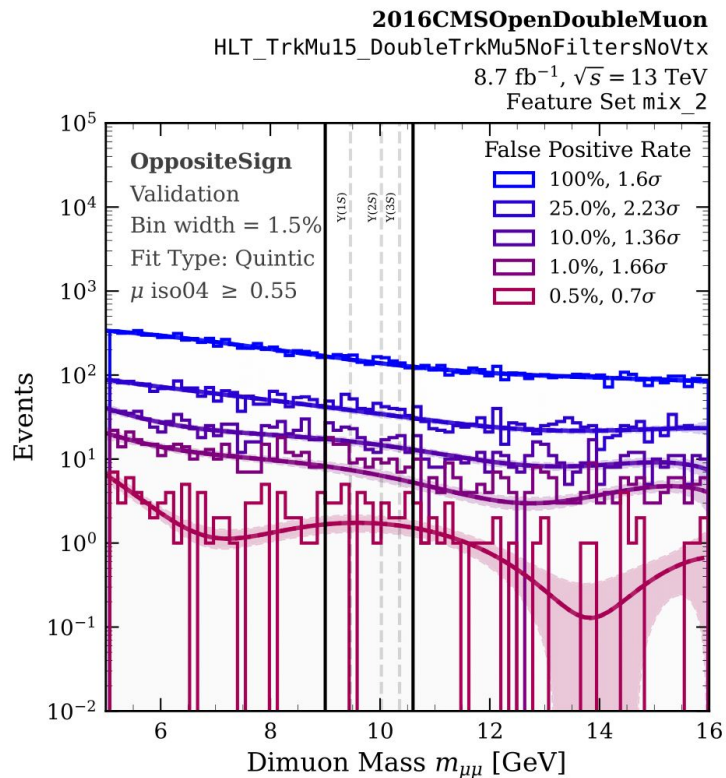
BDT architecture

- Individual trees have a max depth of 3 and are trained on a random 70% subsample of the data.
- We train for up to 300 epochs with a learning rate of $1e-1$, stopping when the validation loss does not decrease for 10 epochs. Trees are evaluated at the epoch of lowest validation loss.
- We ensemble over 100 BDT networks (which may contain an arbitrary number of trees depending on when the early stopping kicked in).

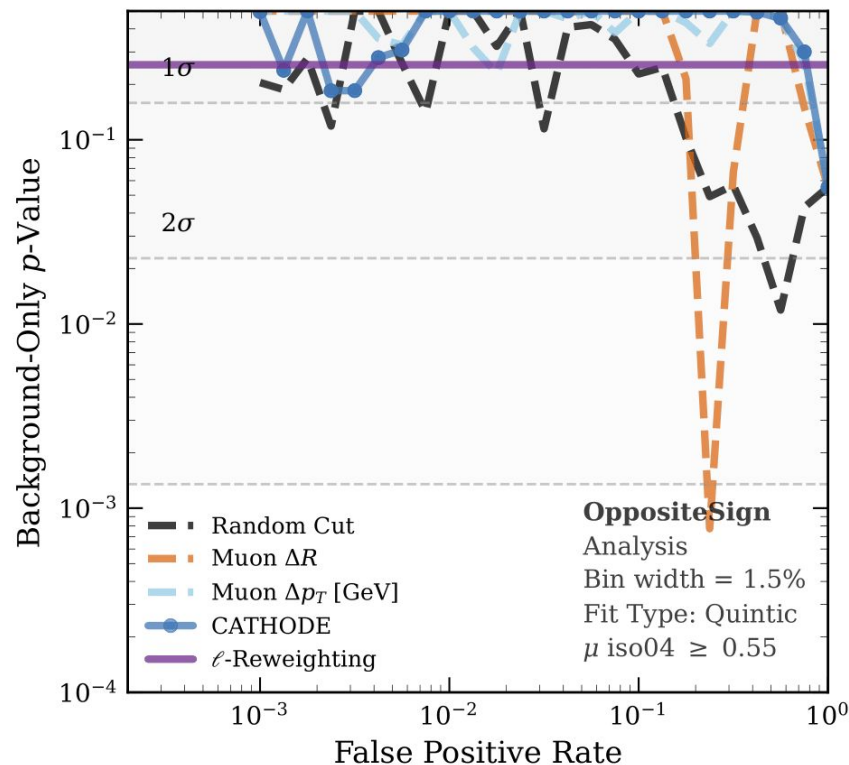
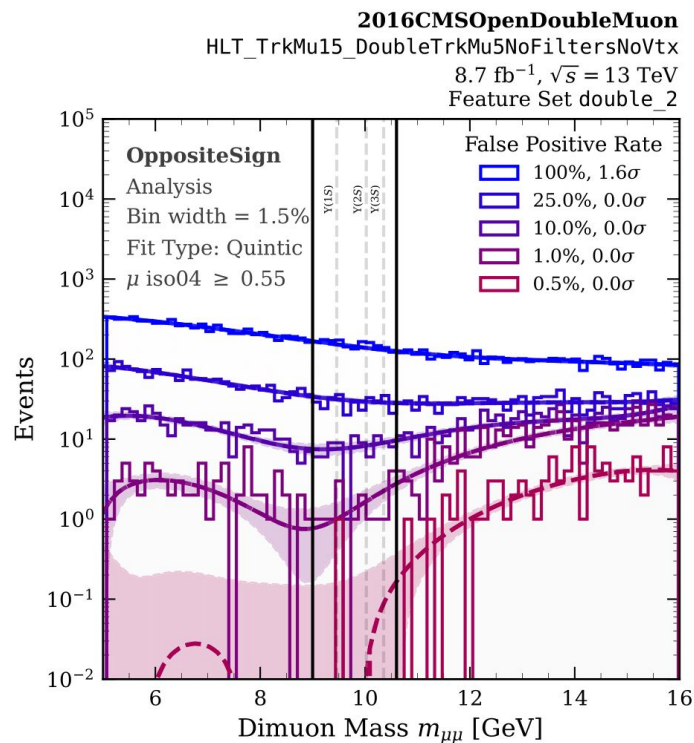
Validation: applying the OS BDT's to SS data



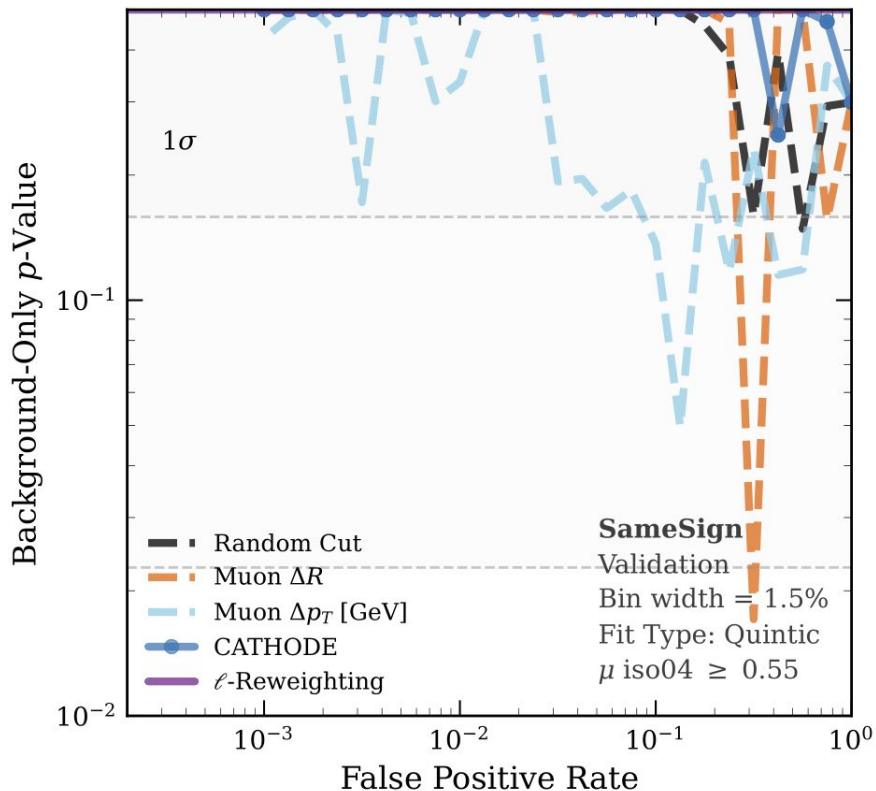
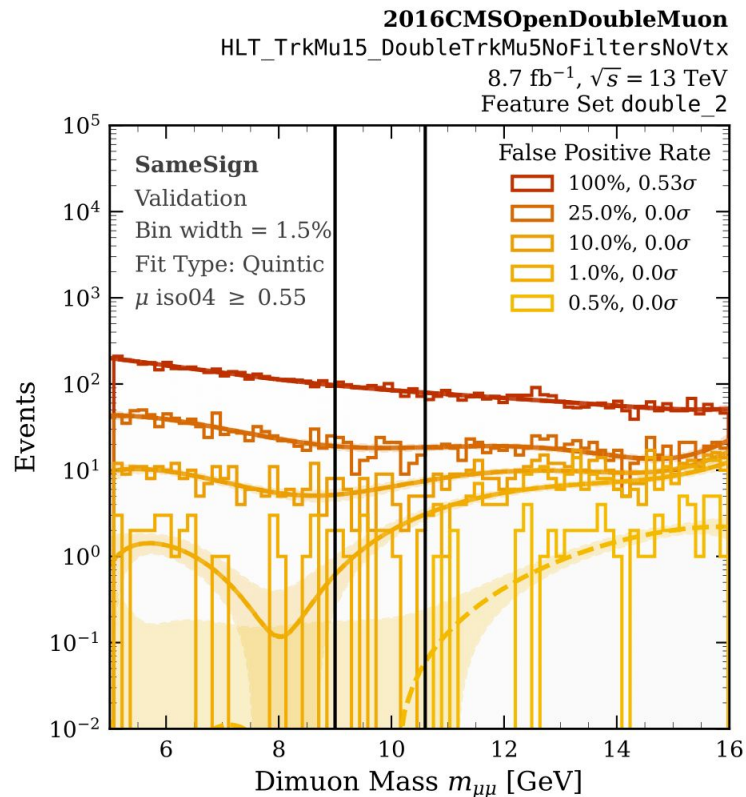
Validation: applying the SS BDT's to OS data



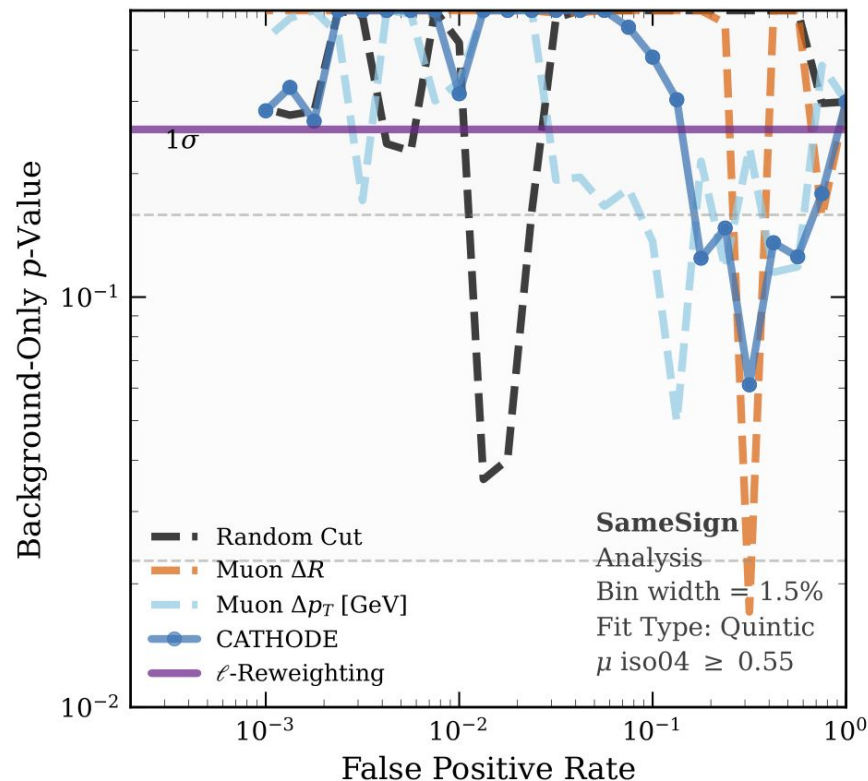
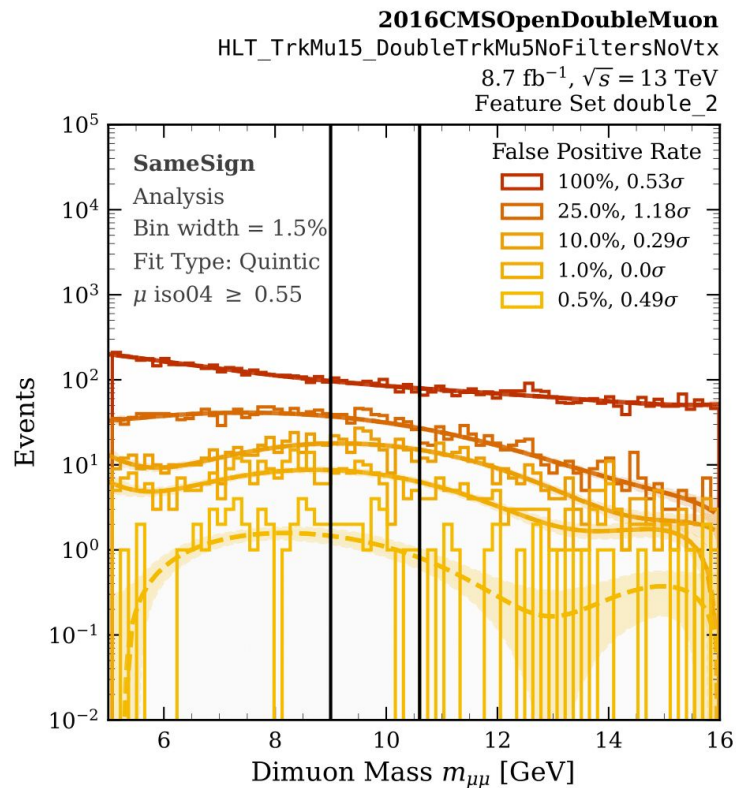
Alternative feature set: OS analysis



Alternative feature set: SS validation



Alternative feature set: SS analysis



Alternative feature set: OS validation

