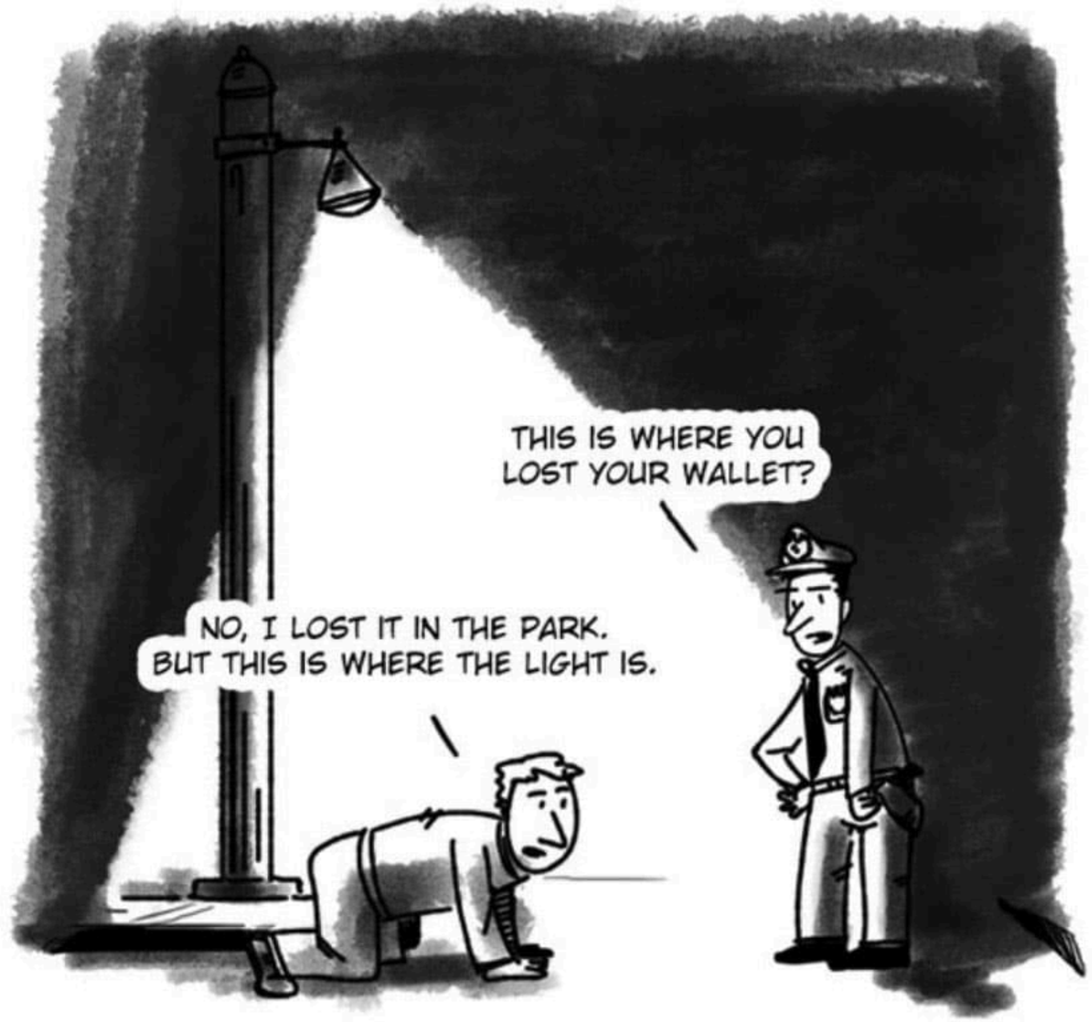
Anomaly Detection Mini-Review

David Shih

May 24, 2021

Pheno 2021

Where is the new physics??



Despite thousands of searches for new physics at the LHC, nothing but limits and null results so far.

What if new physics is hiding in the data but we haven't looked in the right places yet?

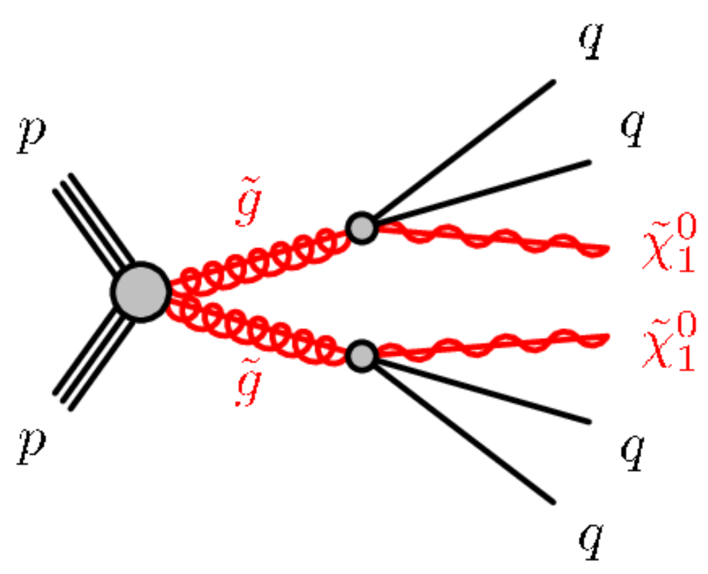
The most common approach

Model specific searches

Most NP searches at the LHC are heavily optimized with specific signals in mind (SUSY, extra dimensions, …)

ATLAS jets+MET 2010.14293

	BDT-GGd1	BDT-GGd2	BDT-GGd3	BDT-GGd4
N_{j}	≥ 4			
$\Delta\phi(j_{1,2,(3)}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})_{\mathrm{min}}$	> 0.4			
$\Delta\phi(j_{i>3}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})_{\mathrm{min}}$	> 0.4			
$E_{\mathrm{T}}^{\mathrm{miss}}/m_{\mathrm{eff}}(N_{\mathrm{j}})$	> 0.2			
m_{eff} [GeV]	> 1400		> 800	
BDT score	> 0.97	> 0.94	> 0.94	> 0.87
$\Delta m(\tilde{g}, \tilde{\chi}_1^0)$ [GeV]	1600–1900	1000–1400	600–1000	200–600



Kinematic cuts (or BDTs) optimized using simulations of signal AND background.

The most common approach

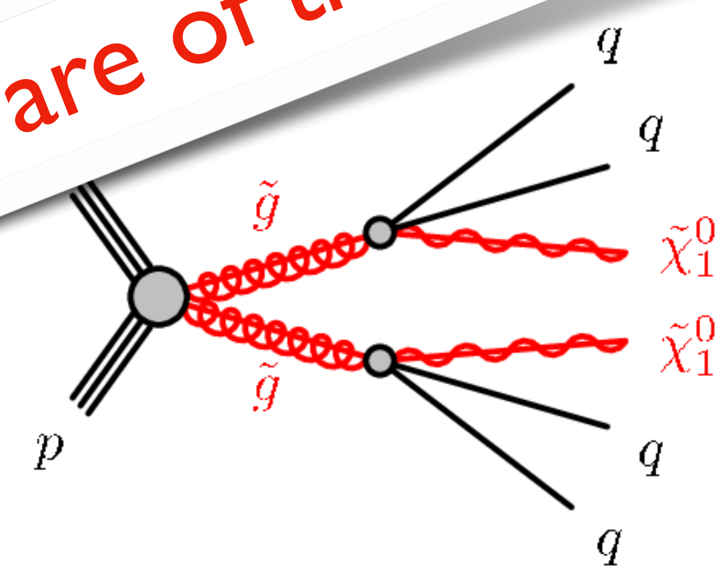
Model specific searches

Most NP searches at the LHC are heavily optimized with spe... signals in mind (SUSY, extra dimensions, …)

> 99% of searches at the LHC are of this type

ATLAS jets+MET 2010.14293

	BDT-GGd1	BDT-GGd2	BDT-GGd3	[illegible]
N_j	≥ 4			
$\Delta\phi(j_{1,2,(3)}, \boldsymbol{p}_T^{miss})_{min}$	[illegible]			
$\Delta\phi(j_{i>3}, \boldsymbol{p}_T^{miss})_{min}$	[illegible]			
$E_T^{miss}/m_{eff}(N_j)$	[illegible]			
m_{eff} [GeV]	[illegible]	[illegible]	> 800	
BDT sc[illegible]	[illegible]	> 0.94	> 0.94	> 0.87
$\Delta m(\tilde{g}, \tilde{\chi}_1^0)$	[illegible]00–1900	1000–1400	600–1000	200–600



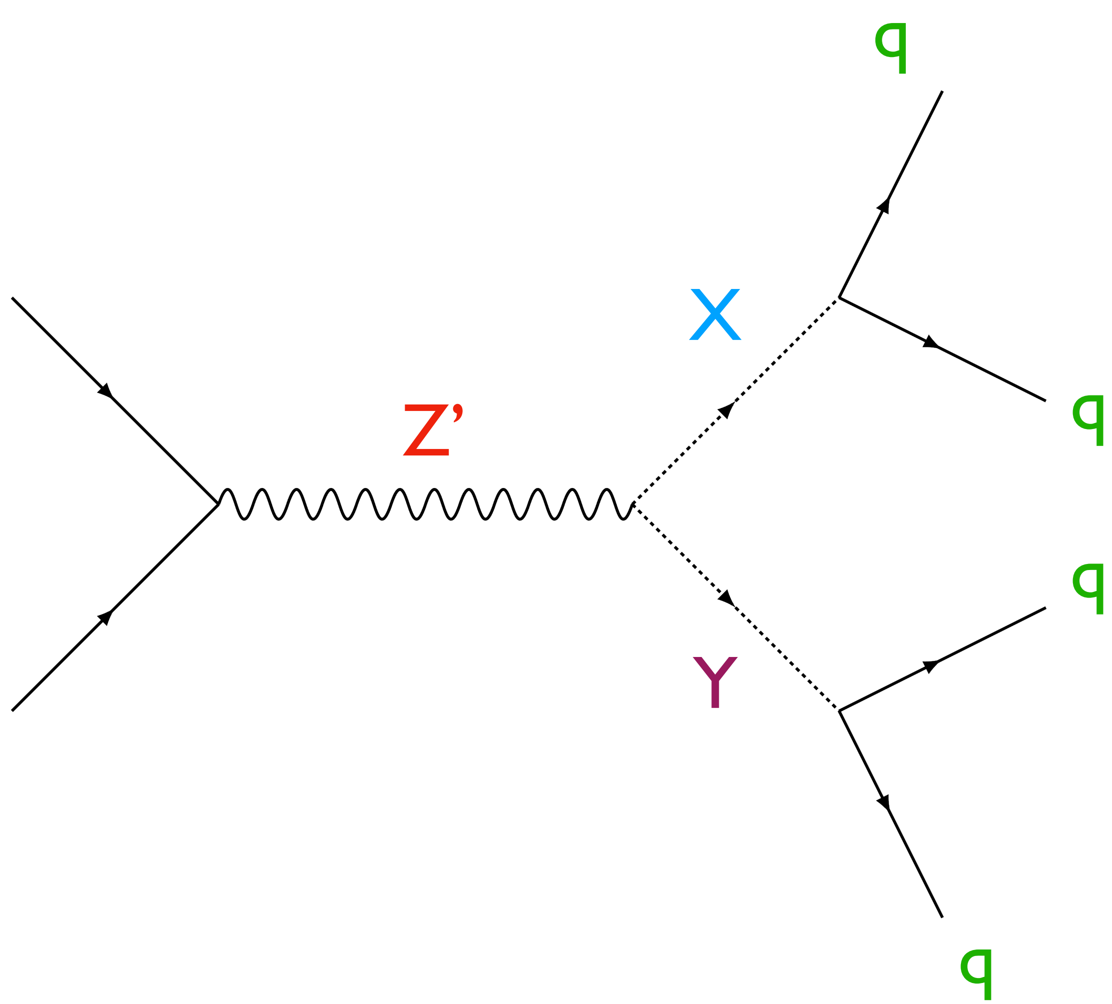
Kinematic cuts (or BDTs) optimized using simulations of signal AND background.

Of course, we should continue to perform these model-specific searches, because NP could always be right around the corner…

But we probably can't cover every possible model this way...

A Benchmark Example

LHC Olympics 2020 R&D Dataset
https://doi.org/10.5281/zenodo.2629072



No explicit search at the LHC for this scenario!

Could be hiding in the dijet resonance search at >5sigma significance!!

General approaches to anomaly detection

Outlier detection

- Look for events where $p_{bg}(x) \lll 1$
- Can find rare signals, can be fully model independent (or at least, may not require very precise background model)
- Uncontrolled, no optimality guarantee — new physics may not be an outlier!

Group anomaly detection

- Look for over-densities in data over background expectation
- **Optimal discriminant:**

$$R(x) = \frac{p_{data}(x)}{p_{bg}(x)}$$

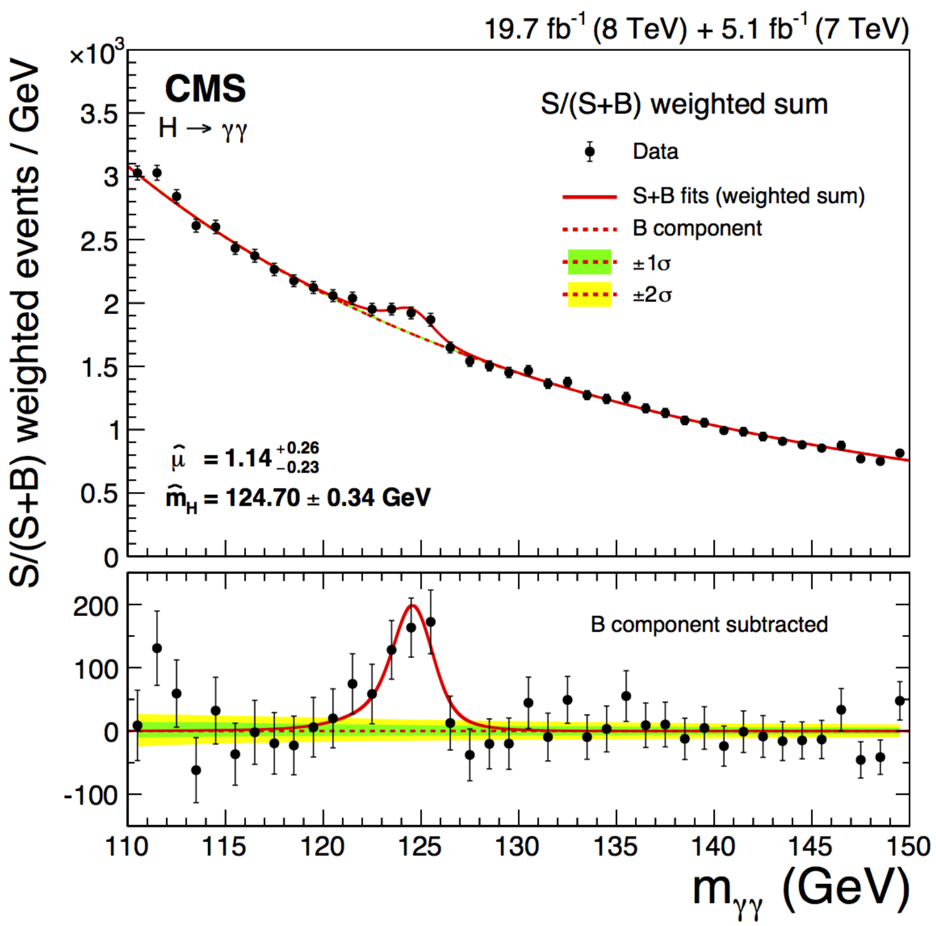
- Generally requires more assumptions on signal and background model — either data driven (eg sideband interpolation, ABCD method) or from simulation

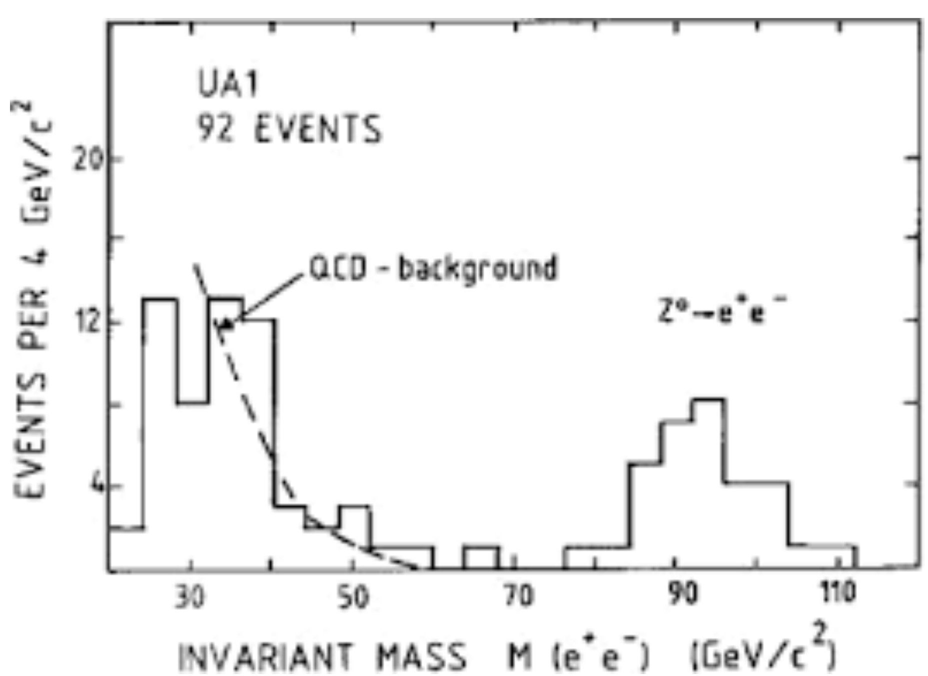
Existing model-independent searches

"the bump hunt"

Idea: assume signal is localized in some feature (usually invariant mass) while background is smooth.

Interpolate from **sidebands** into **signal region**, search for an excess.





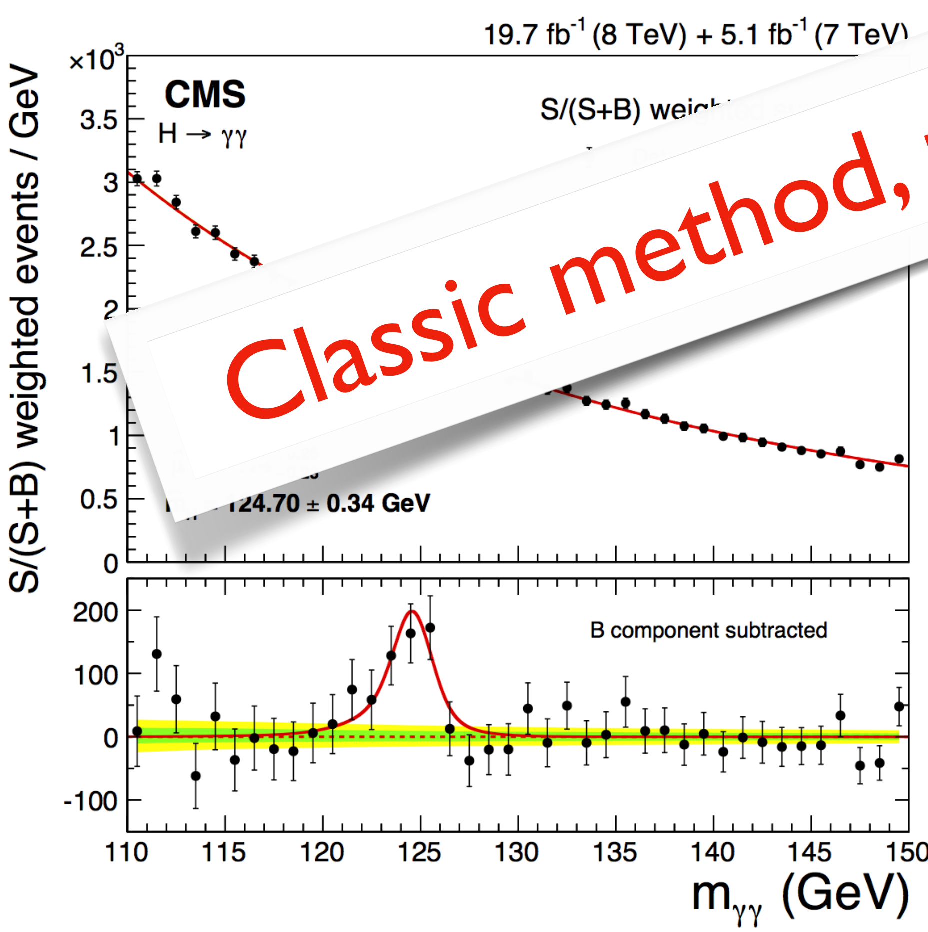
Existing model-independent searches

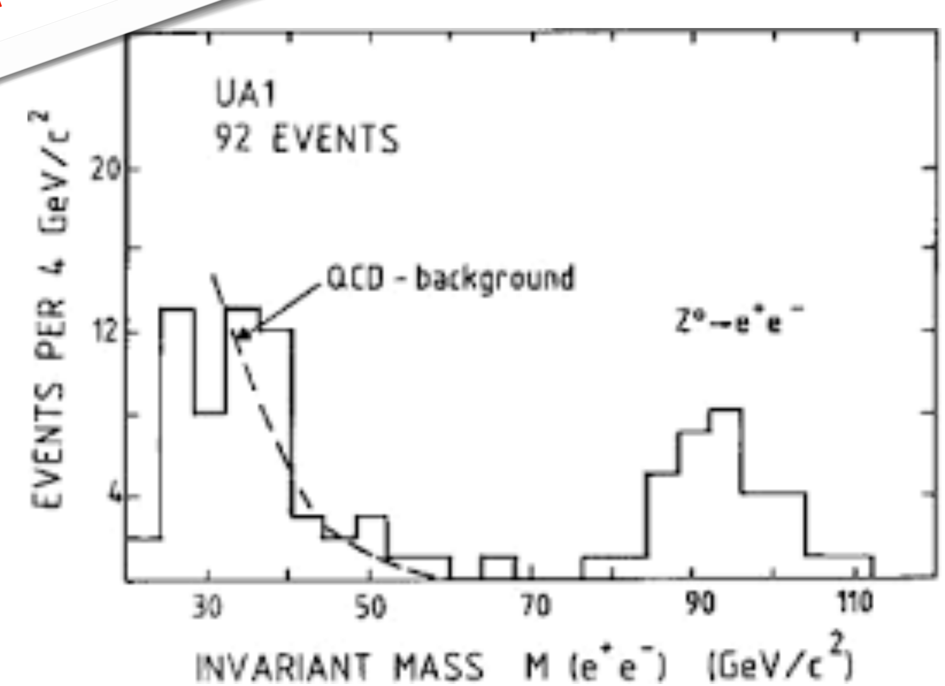
“the bump hunt”

Idea: assume signal is localized in some feature (usually invariant mass) while background is smooth.

Interpolate from **sidebands** into **signal re** … excess.

Classic method, used in many discoveries.

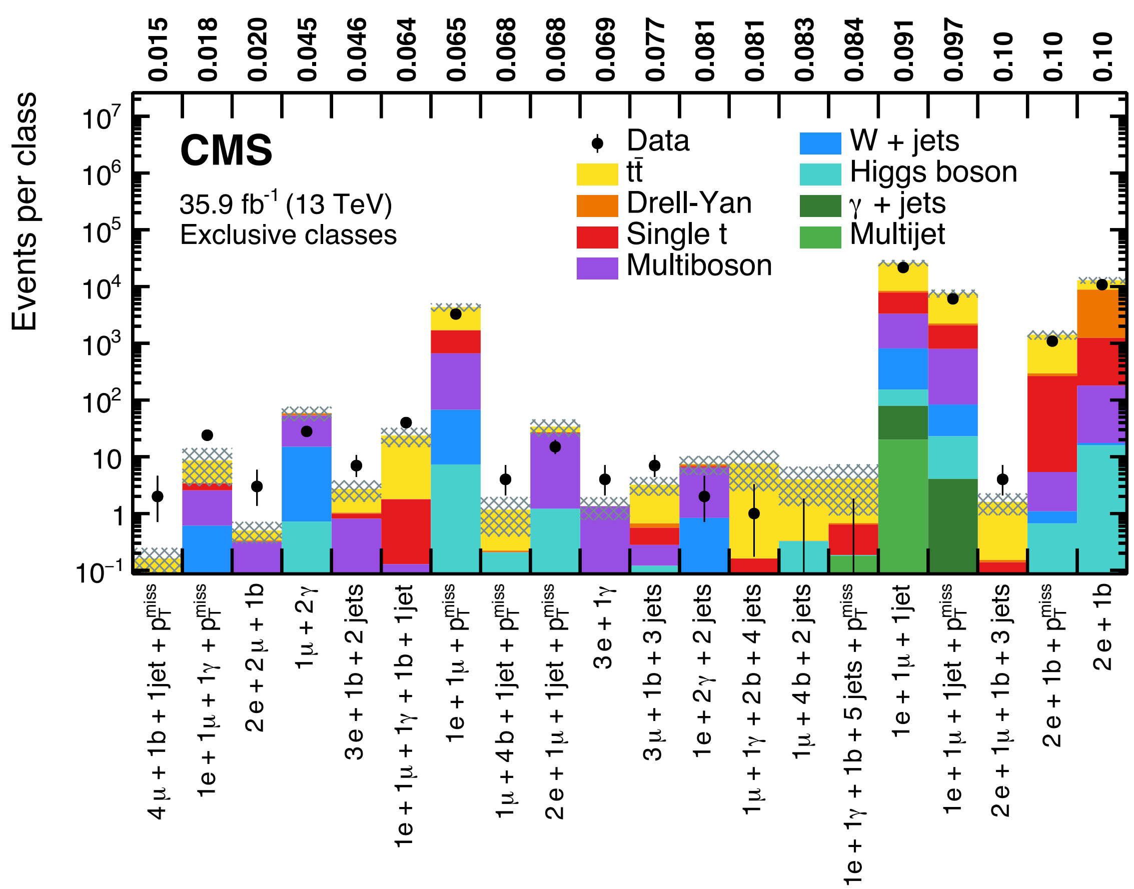




Existing model-independent searches

“the general search”

Idea: divide the phase space up into thousands of bins, compare **data to SM simulation** in each one



CMS

“MUSIC”

CMS-PAS-EXO-14-016

ATLAS

“Model independent general search”

1807.07447
EPJC 79:120 (2019)

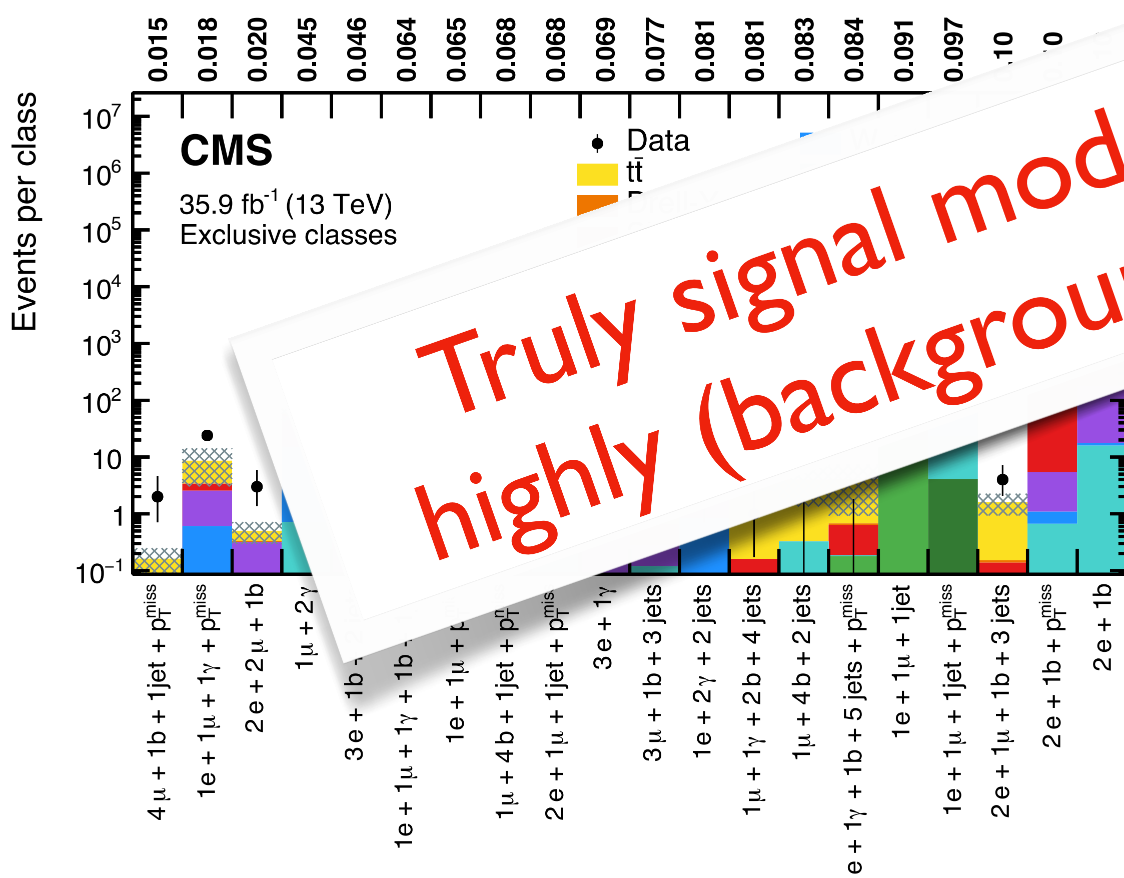
See also proposals by D’Agnolo, Wulzer et al (1806.02350, 1912.12155): train DNN on full phase space to distinguish data from background

Existing model-independent searches

"the general search"

Idea: divide the phase space up into thousands of bins, com[...]
data to SM simulation in each one

Truly signal model independent, but still highly (background) simulation dependent



"MUSIC"

CMS-PAS-EXO-14-016

ATLAS

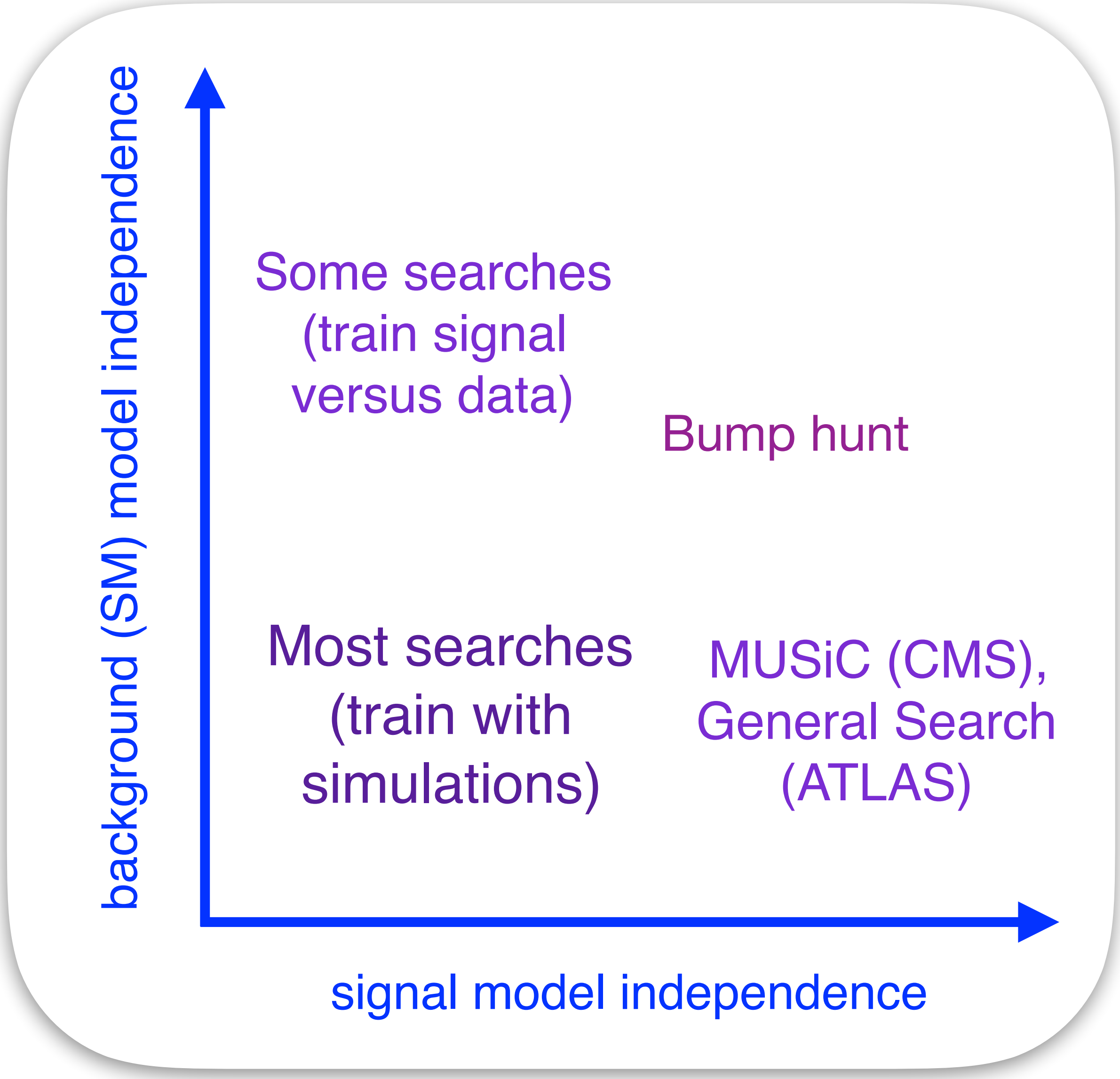
"Model independent general search"

1807.07447
EPJC 79:120 (2019)

See also proposals by D'Agnolo, Wulzer et al (1806.02350, 1912.12155): train DNN on full phase space to distinguish data from background

New paradigms for model-agnostic searches

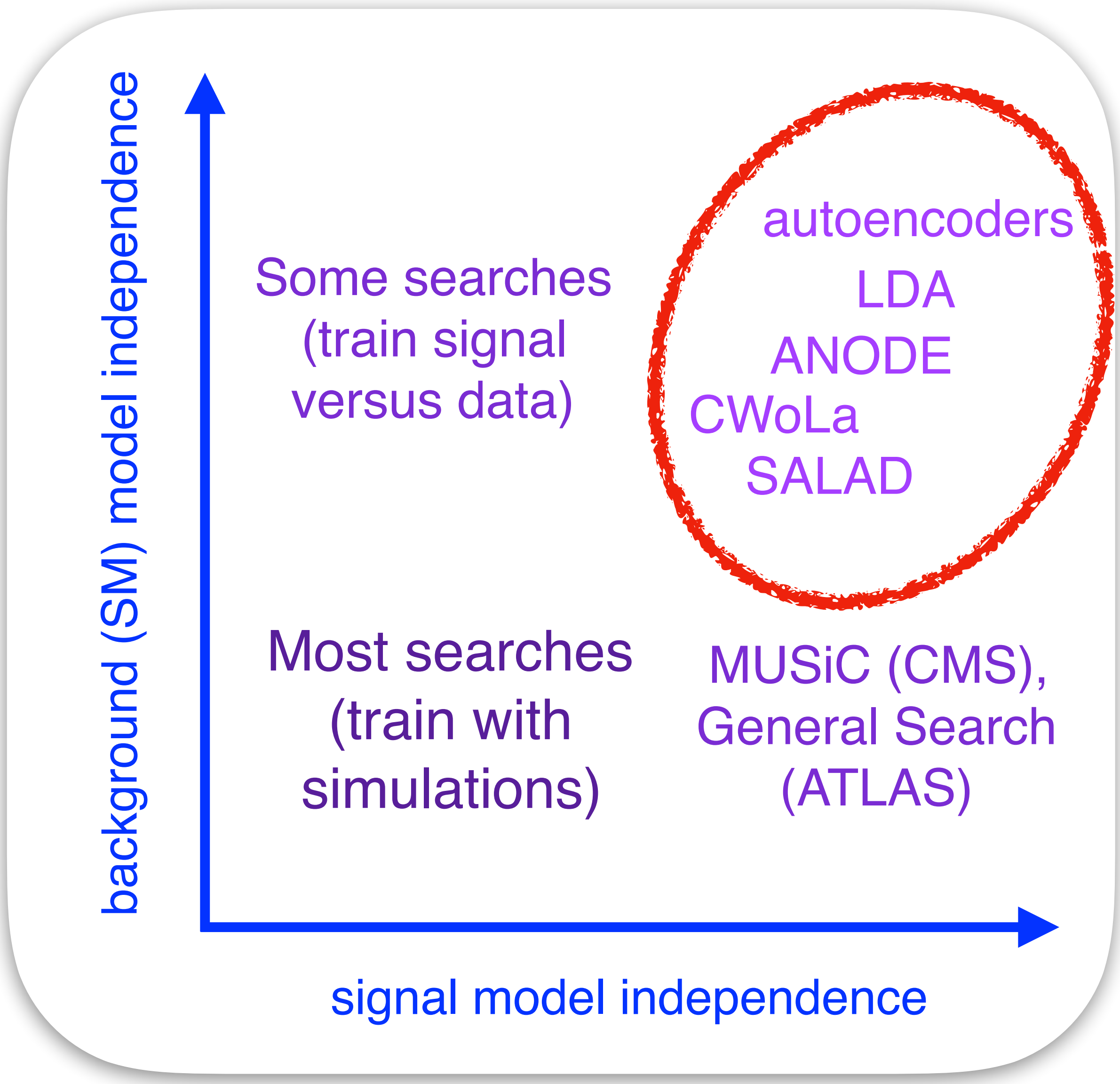
Can advances in machine learning open up new avenues for model-independent searches?



New paradigms for model-agnostic searches

Can advances in machine learning open up new avenues for model-independent searches?

from Nachman & DS 2001.04990



Many new ideas recently!

Many new approaches inspired by the

LHC Olympics 2020 Data Challenge

[G. Kasieczka, B. Nachman & DS, organizers]

It consisted of three "black boxes" of simulated data (bg dominated!):

https://doi.org/10.5281/zenodo.3547721

- 1 million events each
- 4-vectors of every reconstructed particle (all hadronic) in the event
- Particle ID, charge, etc not included
- Single R=1 jet trigger pT>1.2 TeV

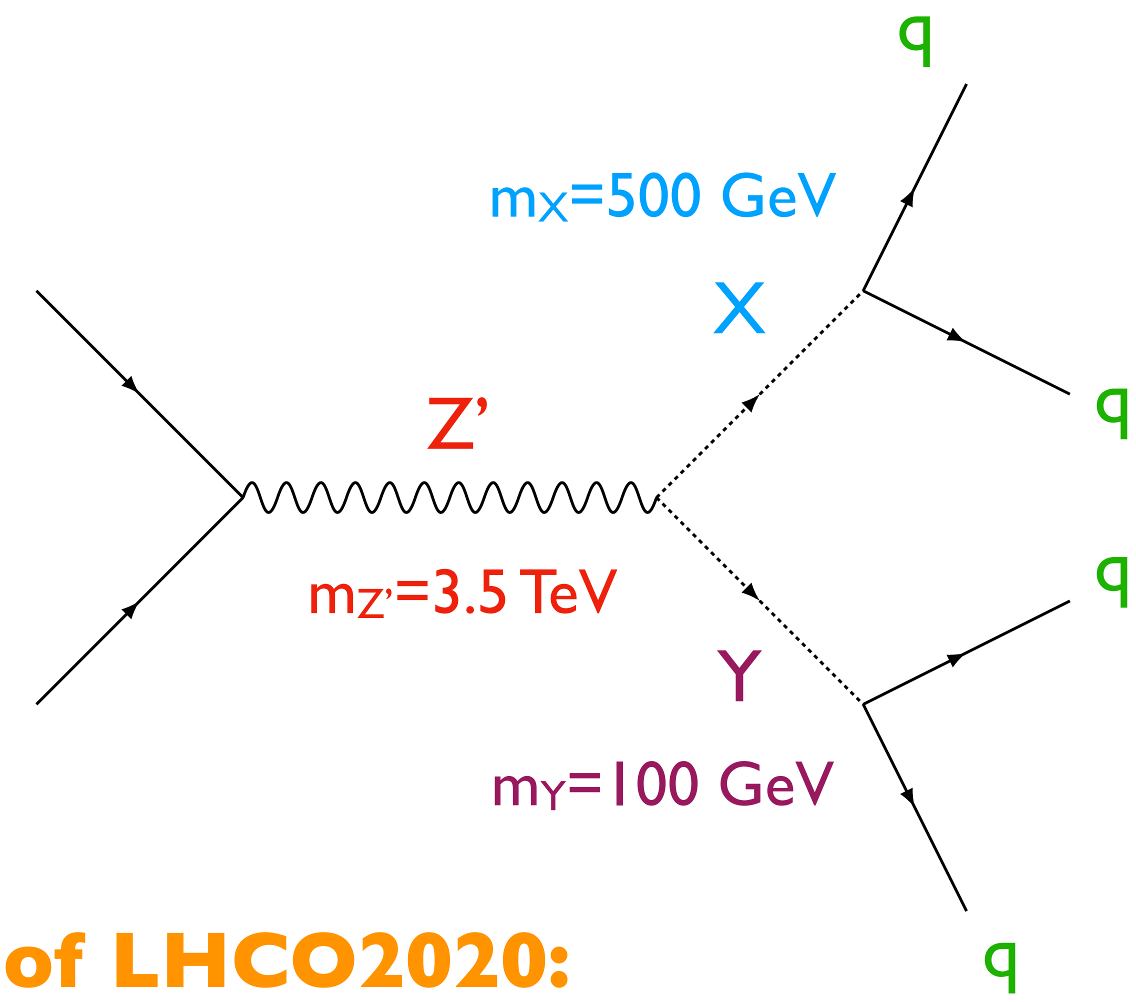
The goal of the challenge was for participants to analyze each box and

1. Decide whether or not it contains new physics
2. Characterize the new physics, if it's there

LHC Olympics 2020: R&D Dataset

https://doi.org/10.5281/zenodo.2629072

Prior to the challenge, we also released a labeled R&D dataset consisting of 1M QCD dijet events and 100k signal events



Unofficial theme of LHCO2020:

"enhancing the bump hunt"

Many new approaches inspired by the **LHC Olympics 2020 Data Challenge**

- 9 groups submitted results on box 1
- 5 groups submitted results on boxes 2 and 3
- (A number of additional groups could not finish the challenge in time but got results on the R&D dataset, or on the black boxes after unblinding)
- Two workshops:
  - "Winter Olympics" — special session of the ML4Jets conference, January 2020, NYU [box 1 opened]
  - "Summer Olympics" — virtual anomaly detection mini-workshop, July 2020, "Hamburg" [boxes 2 & 3 opened]

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

arxiv: 2101.08320

Many new approaches inspired by the

LHC Olympics 2020 Data Challenge

arxiv: 2101.08320

Many new approaches inspired by the **LHC Olympics 2020 Data Challenge**

training with
no labels

arxiv: 2101.08320

Many new approaches inspired by the
LHC Olympics 2020 Data Challenge

training with no labels

training with noisy labels

arxiv: 2101.08320

Many new approaches inspired by the **LHC Olympics 2020 Data Challenge**

training with no labels

training with noisy labels

training with truth labels (from simulation)

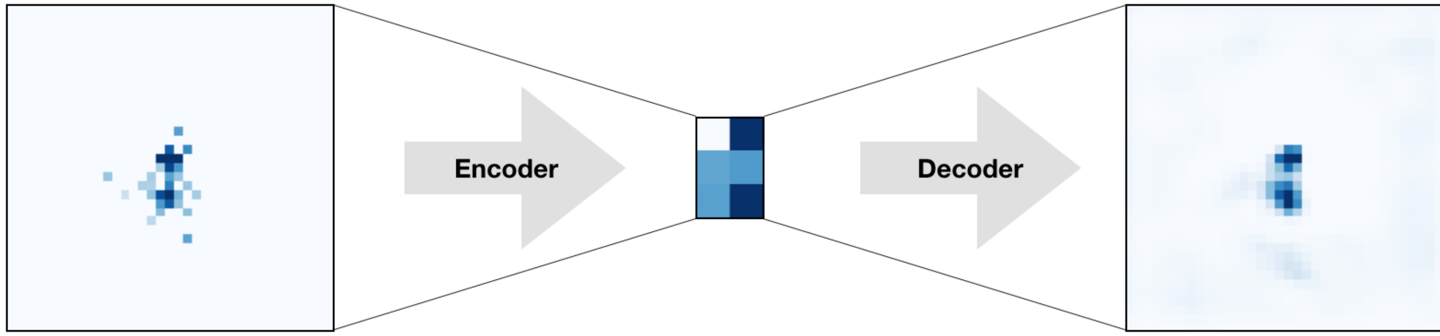
arxiv: 2101.08320

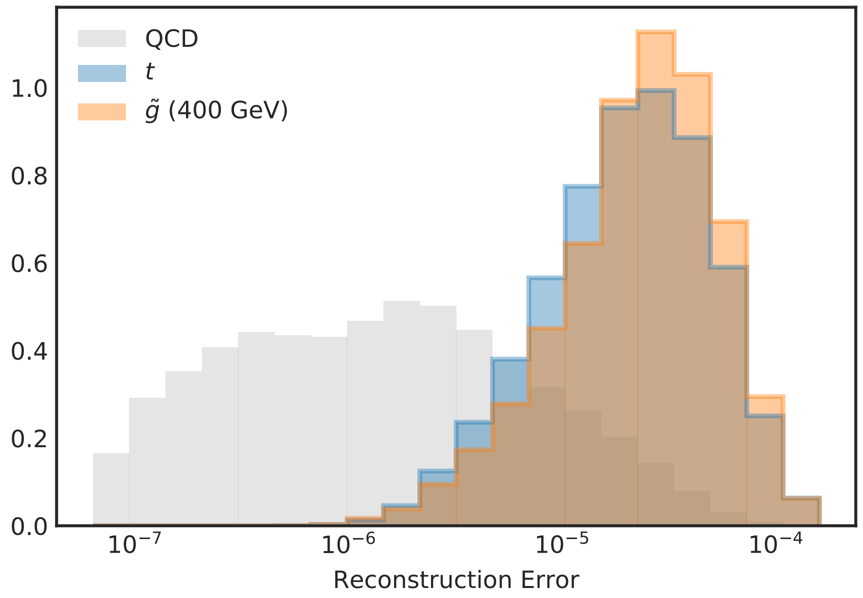
Unsupervised Anomaly Detection

General idea: train ML algorithm directly on (background-dominated) data to identify outliers [events with low $p_{bg}(x)$]

Example: Autoencoders

Train lossy ML algorithm to map data to itself through a compressed latent space.





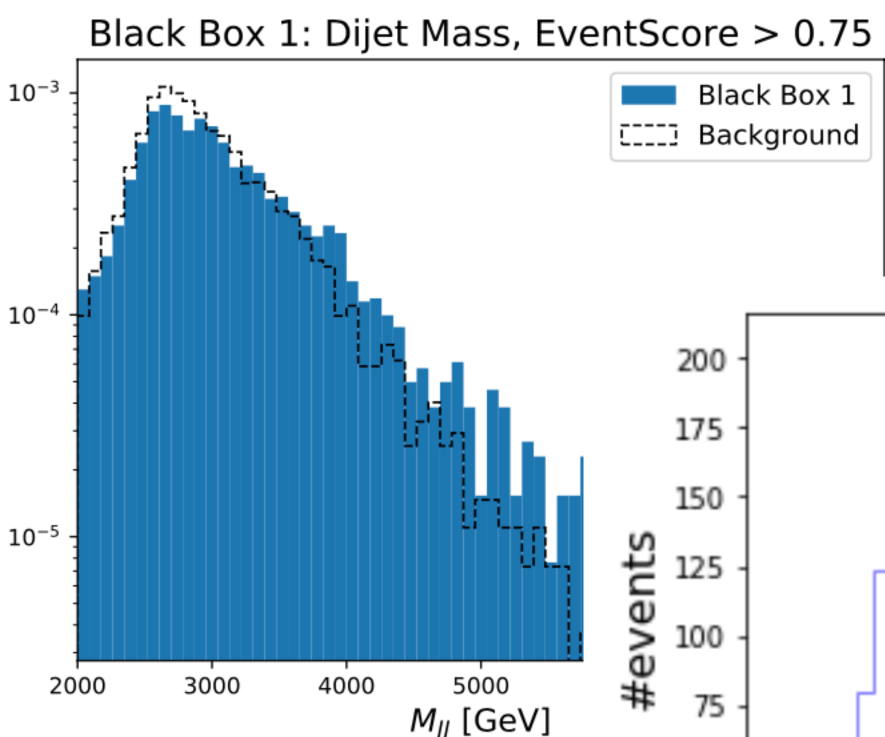
Rare anomalies should be poorly reconstructed

Heimel, Kasieczka, Plehn & Thompson 1808.08979
Farina, Nakai & DS 1808.08992
and many, many more!

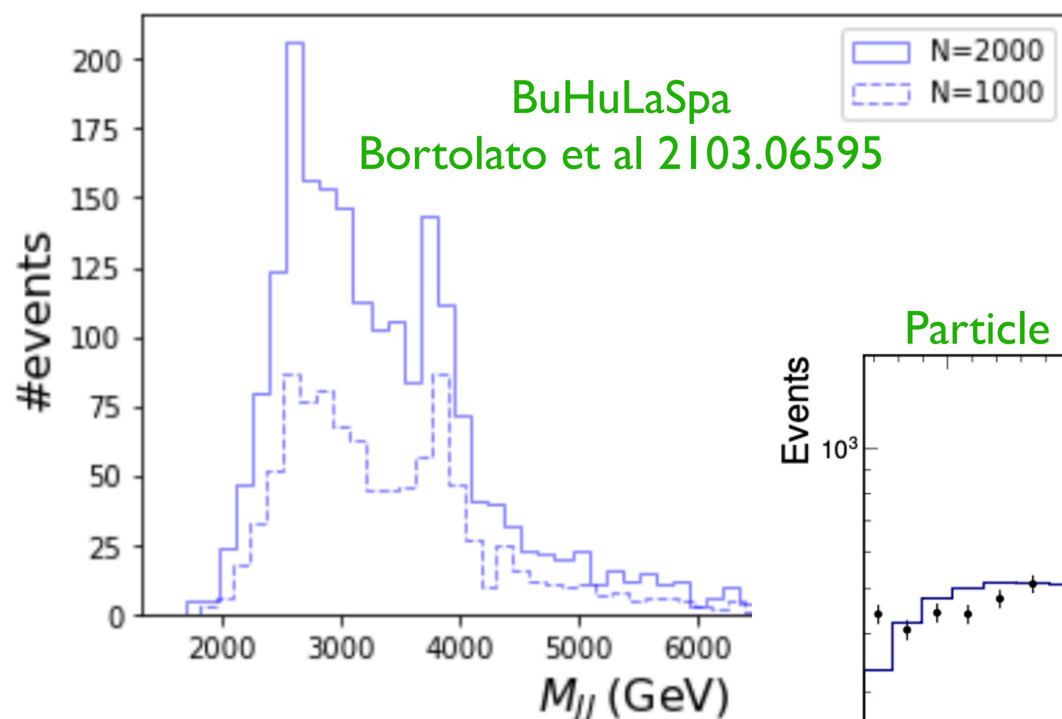
Unsupervised Anomaly Detection

Several successful LHCO2020 approaches were based on AEs.

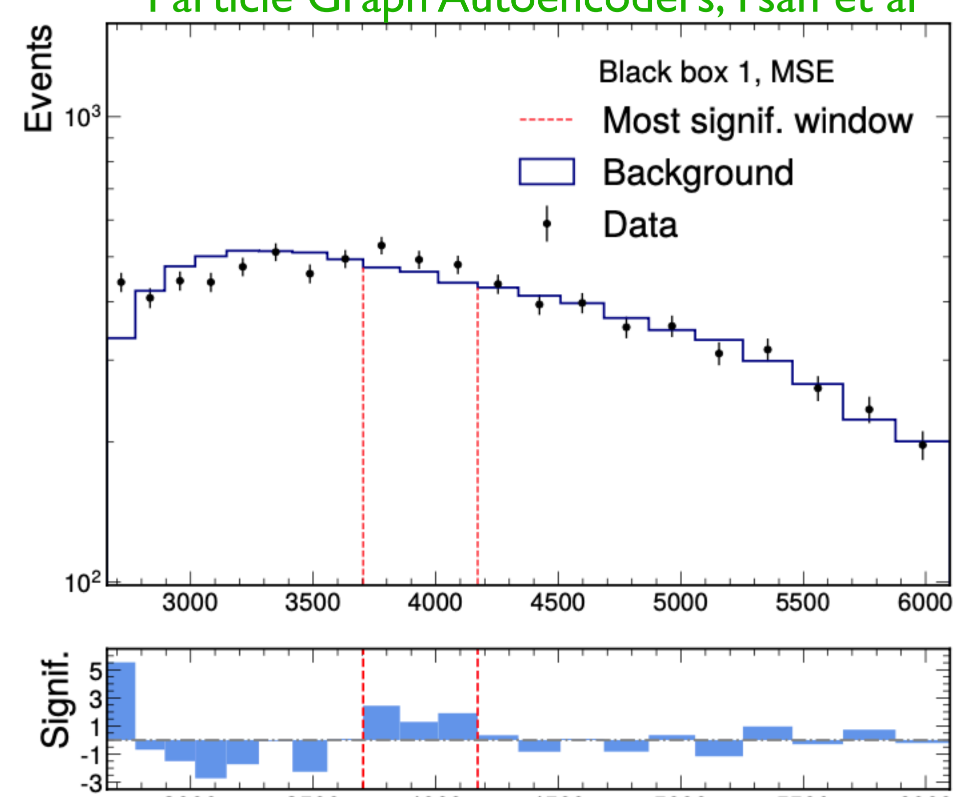
VRNN, Kahn et al 2105.09274



BuHuLaSpa
Bortolato et al 2103.06595



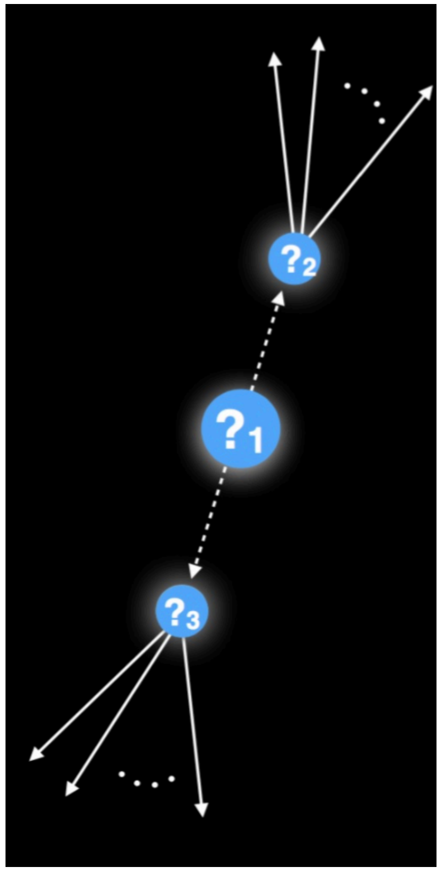
Particle Graph Autoencoders, Tsan et al

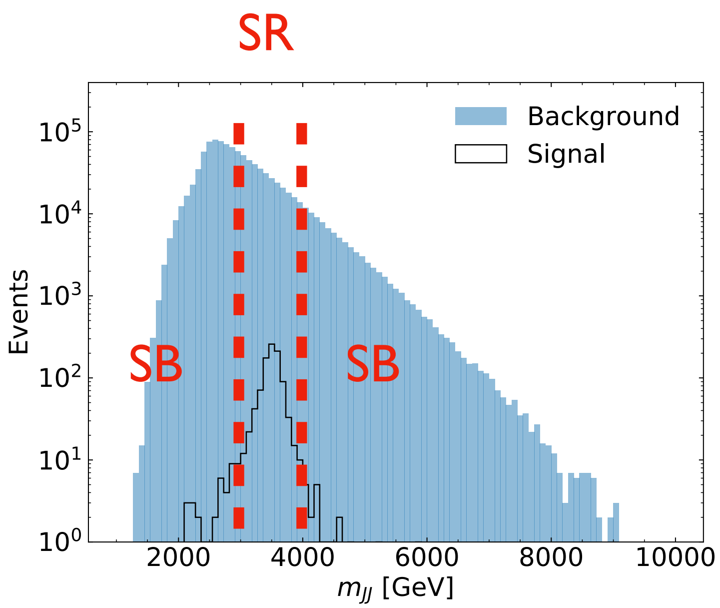


Weakly-supervised Anomaly Detection

General idea: train ML algorithm to compare two datasets with different levels of signal, identify events with high $p_{data}(x)/p_{bg}(x)$

Example: “CwoLa Hunting” [Collins, Howe & Nachman 1805.02664]





Train a binary classifier on additional features $x = m_{j1}, m_{j2}, tau_{21}(j_1), tau_{21}(j_1), \ldots$ to distinguish between signal region and sideband events.

If additional features are uncorrelated with m_{JJ} in the background, should learn $p_{data}(x)/p_{bg}(x)$ [Neyman-Pearson lemma]

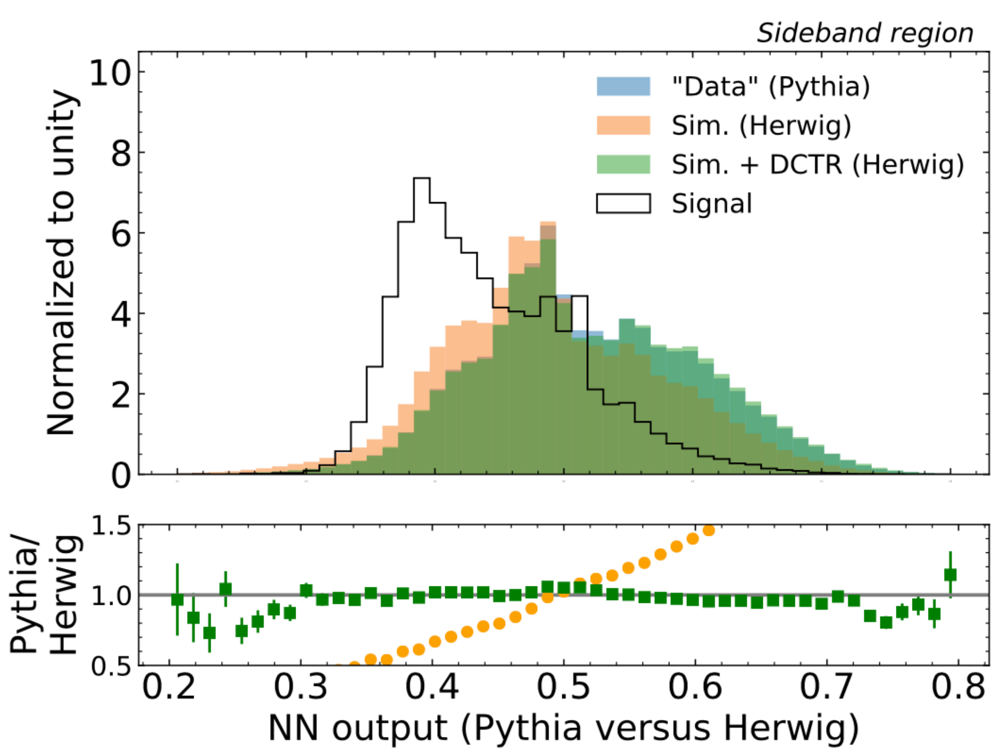
Z’->XY (boosted), X->?, Y->?

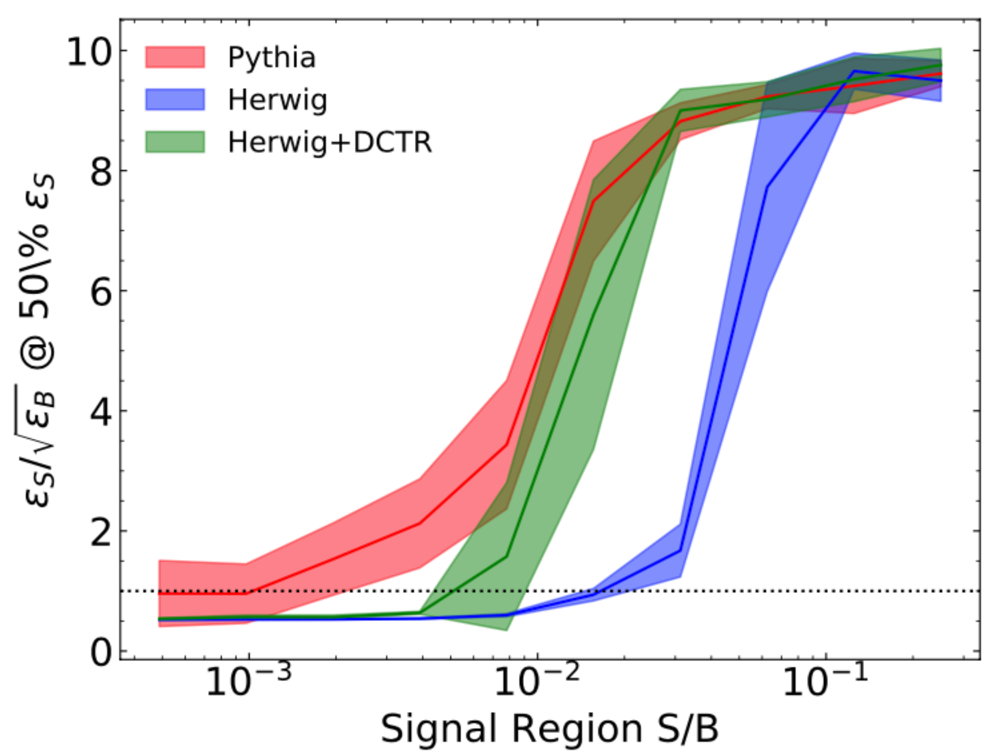
Weakly-supervised Anomaly Detection

Another example: Simulation Assisted Likelihood-free Anomaly Detection (SALAD) [Andreassen, Nachman & DS 2001.05001]

Try to leverage simulated backgrounds for learning $p_{data}(x)/p_{bg}(x)$:

- reweight bg sim to look like data in sideband region using DCTR method [Andreassen & Nachman 1907.08209]
- interpolate into SR
- train classifier on data vs bg.



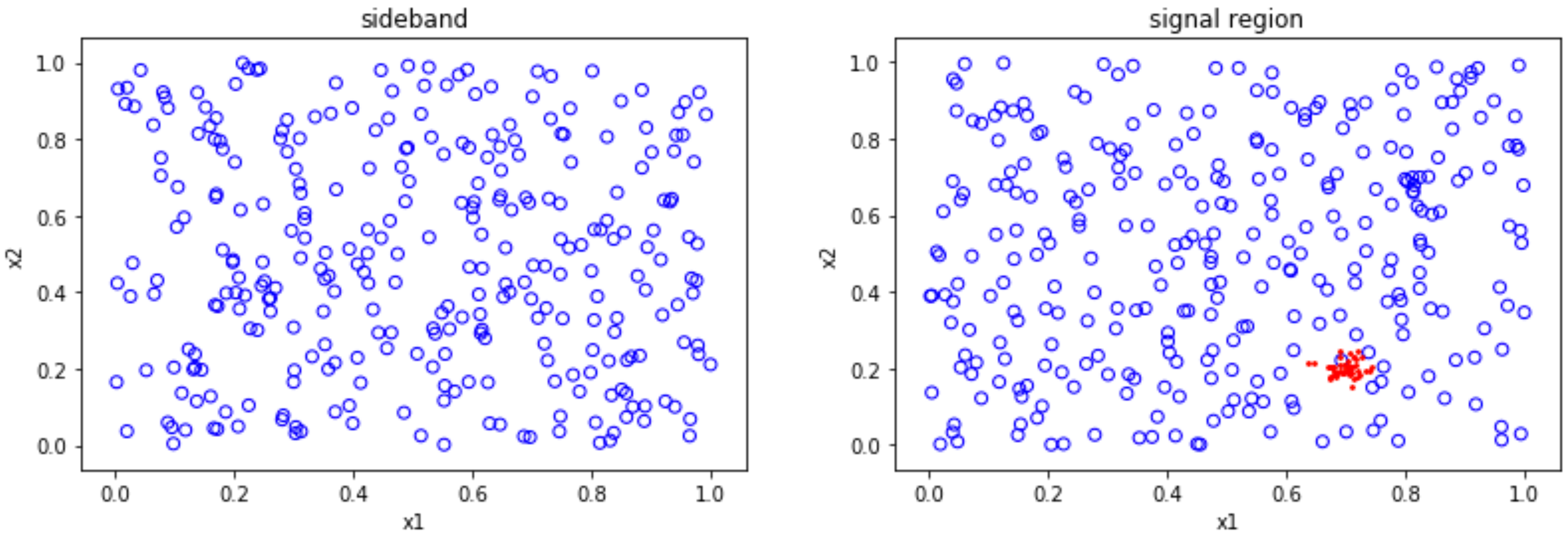


Between weak and un-supervised

ANOmaly detection with Density Estimation (ANODE):
[Nachman & DS 2001.04990]

Use unsupervised approach to learn the likelihood ratio:

- Train density estimator to directly learn $p_{SR}(x)$ and $p_{SB}(x)$
- Interpolate latter in mJJ to obtain $p_{bg}(x)$ in the SR
- Construct likelihood ratio $R(x)=p_{data}(x)/p_{bg}(x)$ explicitly

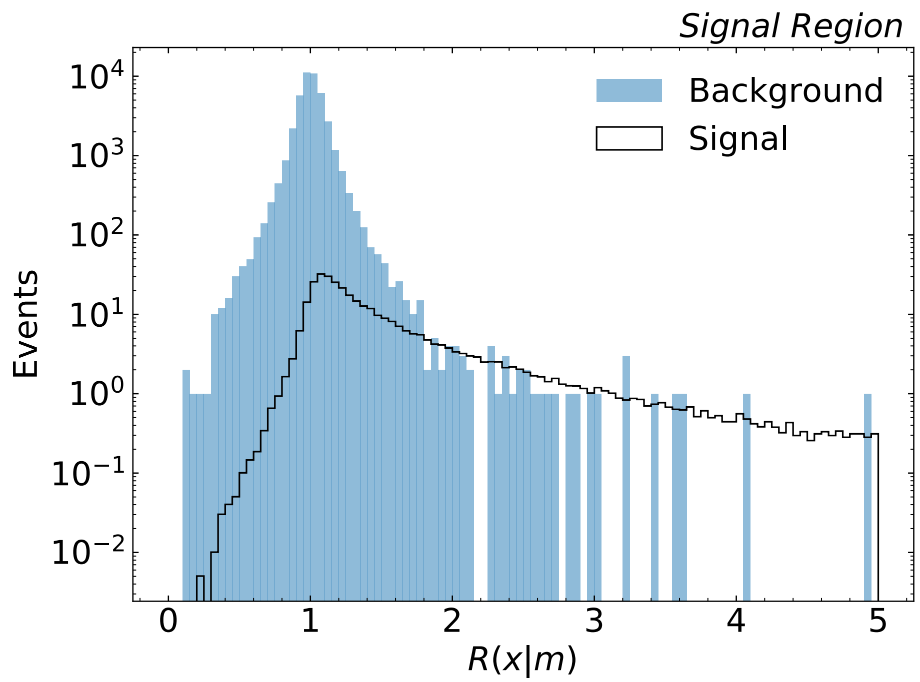


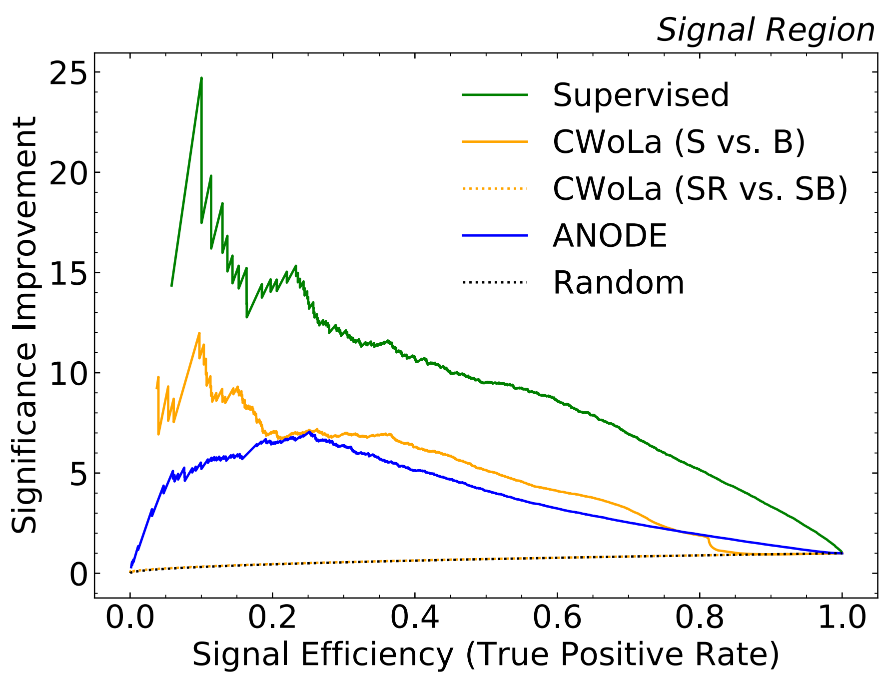
Between weak and un-supervised

ANOmaly detection with Density Estimation (ANODE):
[Nachman & DS 2001.04990]

Use unsupervised approach to learn the likelihood ratio:

- Train density estimator to directly learn $p_{SR}(x)$ and $p_{SB}(x)$
- Interpolate latter in mJJ to obtain $p_{bg}(x)$ in the SR
- Construct likelihood ratio $R(x)=p_{data}(x)/p_{bg}(x)$ explicitly





Between weak and un-supervised

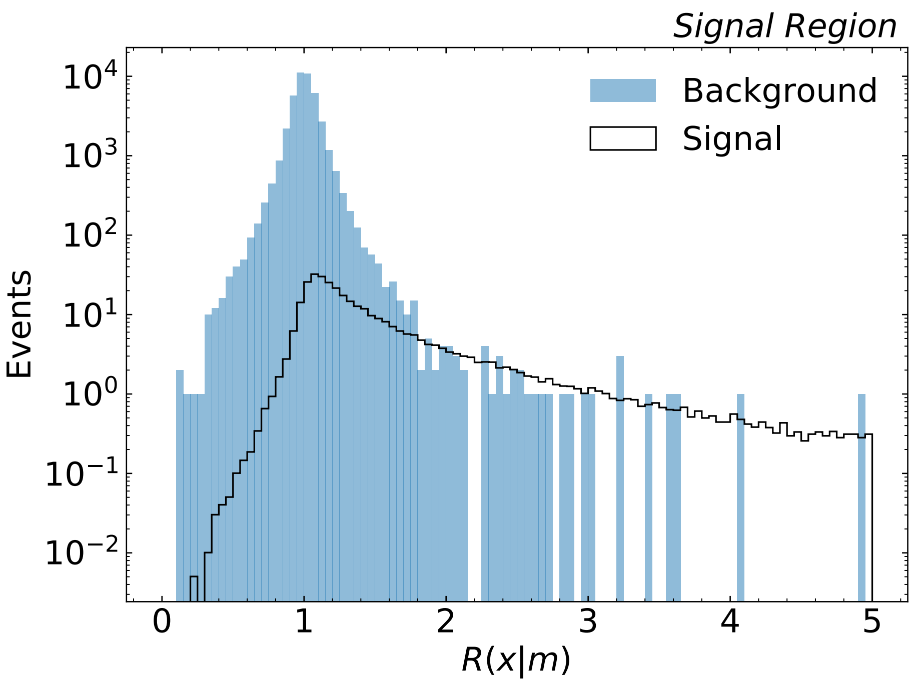
ANOmaly detection with Density Estimation (ANODE):

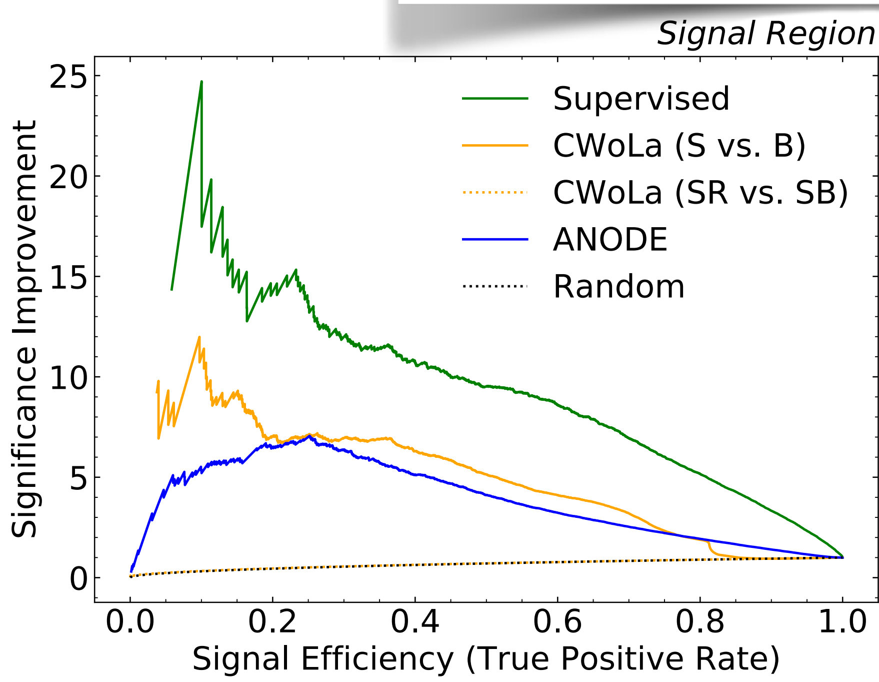
[Nachman & DS 2001.04990]

Use unsupervised approach to learn the likelihood ratio:

- Train density estimator to directly learn $p_{SR}(x)$ and $p_{SB}(x)$
- Interpolate latter in mJJ to obtain $p_{bg}(x)$ in the SR
- Construct likelihood ratio $R(x)=p_{data}(x)/p_{bg}(x)$ explicitly

Can enhance the significance of the bump hunt by a factor of up to 7!
1.5σ (dijet bump hunt)
=> 10σ (ANODE+bump hunt)

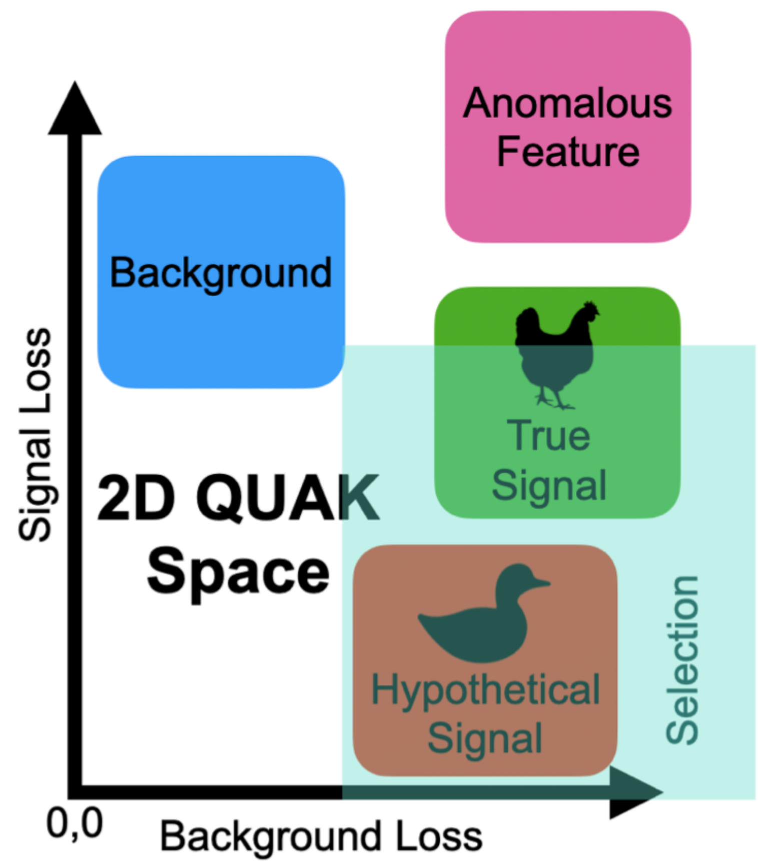


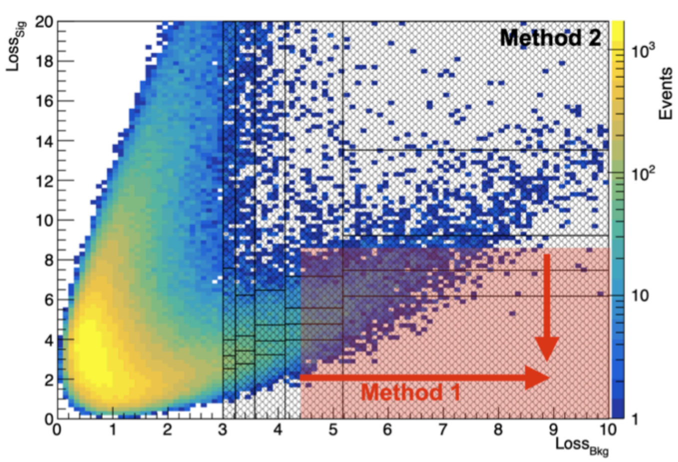


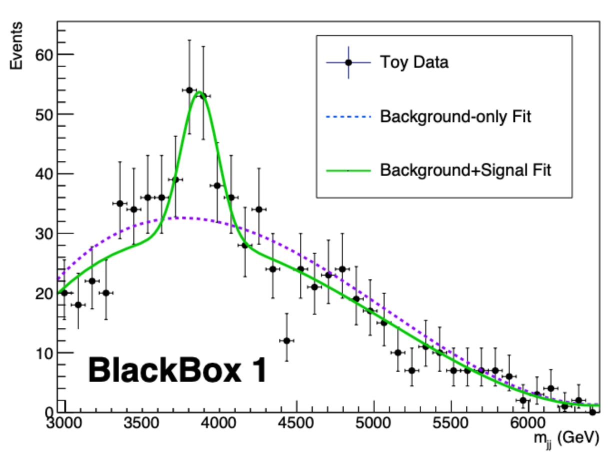
(Semi)Supervised Anomaly Detection

General idea: train ML algorithm on signal and background simulation, apply to data to find "signal-like" events

Example: Quasi Anomalous Knowledge (QUAK)
[Park, Rankin, Udrescu, Yunus, Harris 2011.03550]







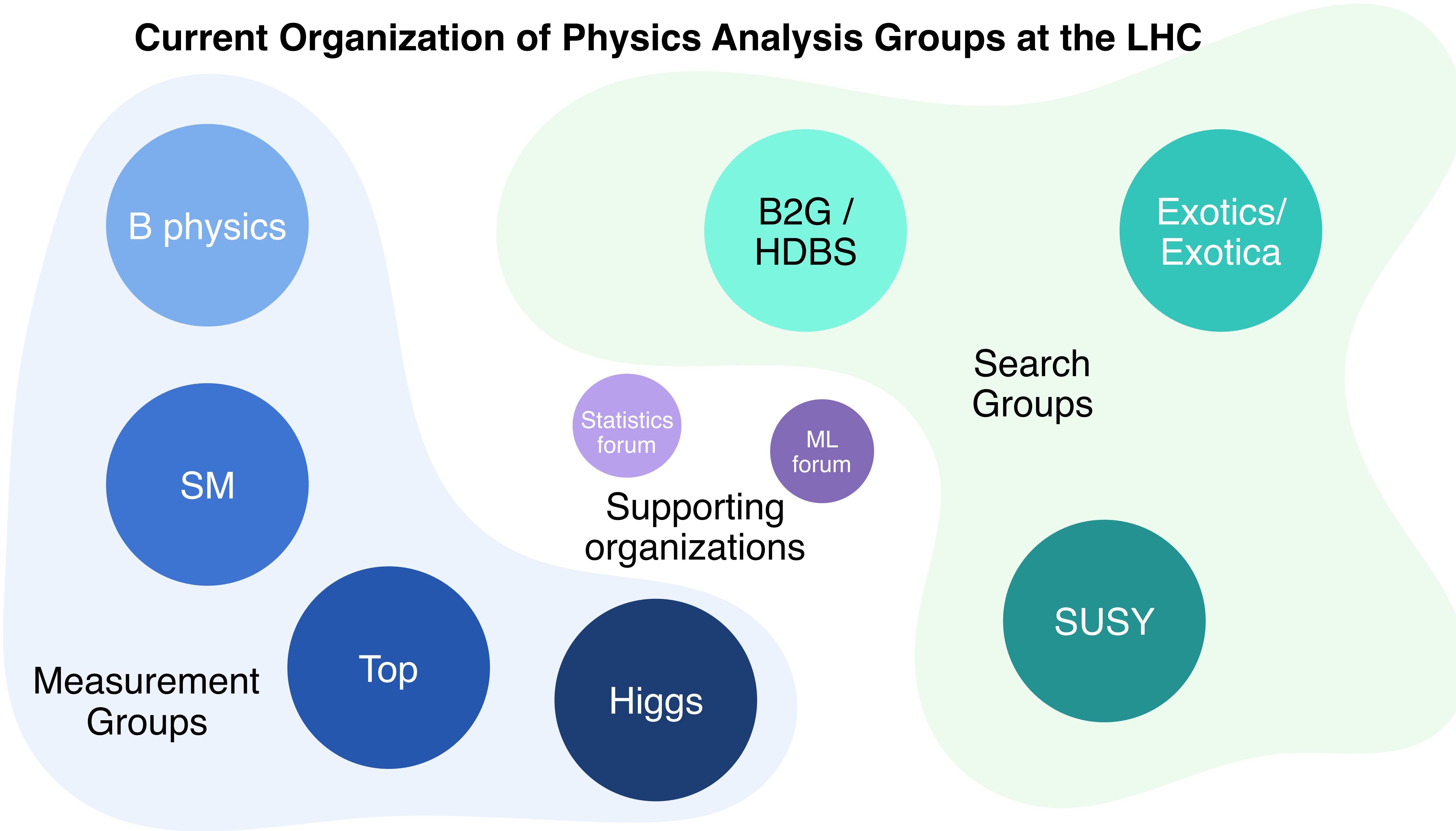
Train separate autoencoders on signal models and background model.

Look for events in data with high background loss and low signal loss

Summary and Outlook

- Advances in machine learning are opening up new and exciting avenues for model independent new physics searches at the LHC.
- The LHC Olympics 2020 provided a very useful testing ground for the development and common benchmarking of new approaches.
- Much work remains to be done in order to port these ideas over to ATLAS and CMS and implement them as actual analyses on real data.
- We need more ideas for model-independent searches at the LHC. This is just the beginning!

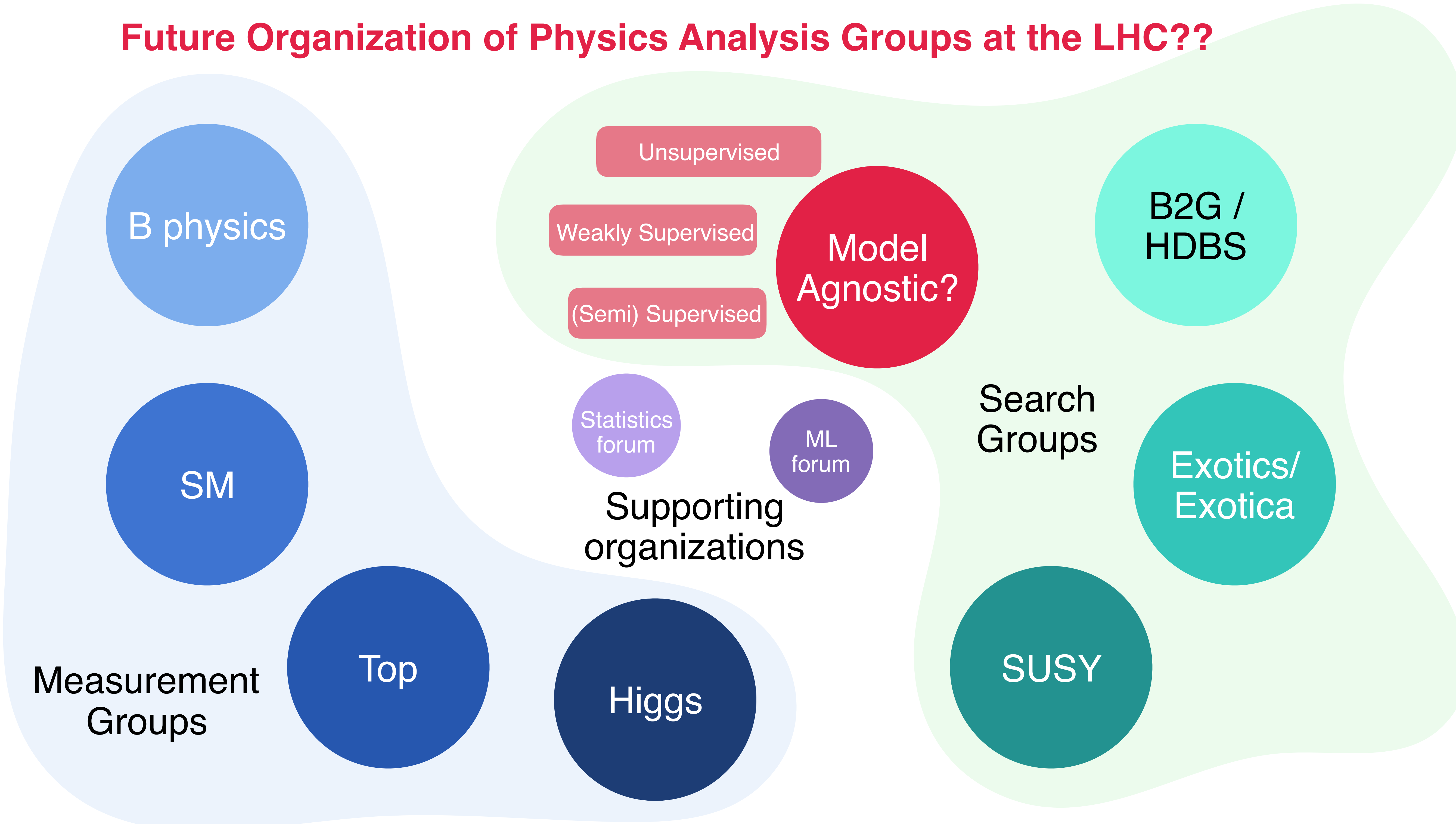
Current Organization of Physics Analysis Groups at the LHC



Q: Why is there no model independent search group???

A vision for the future...

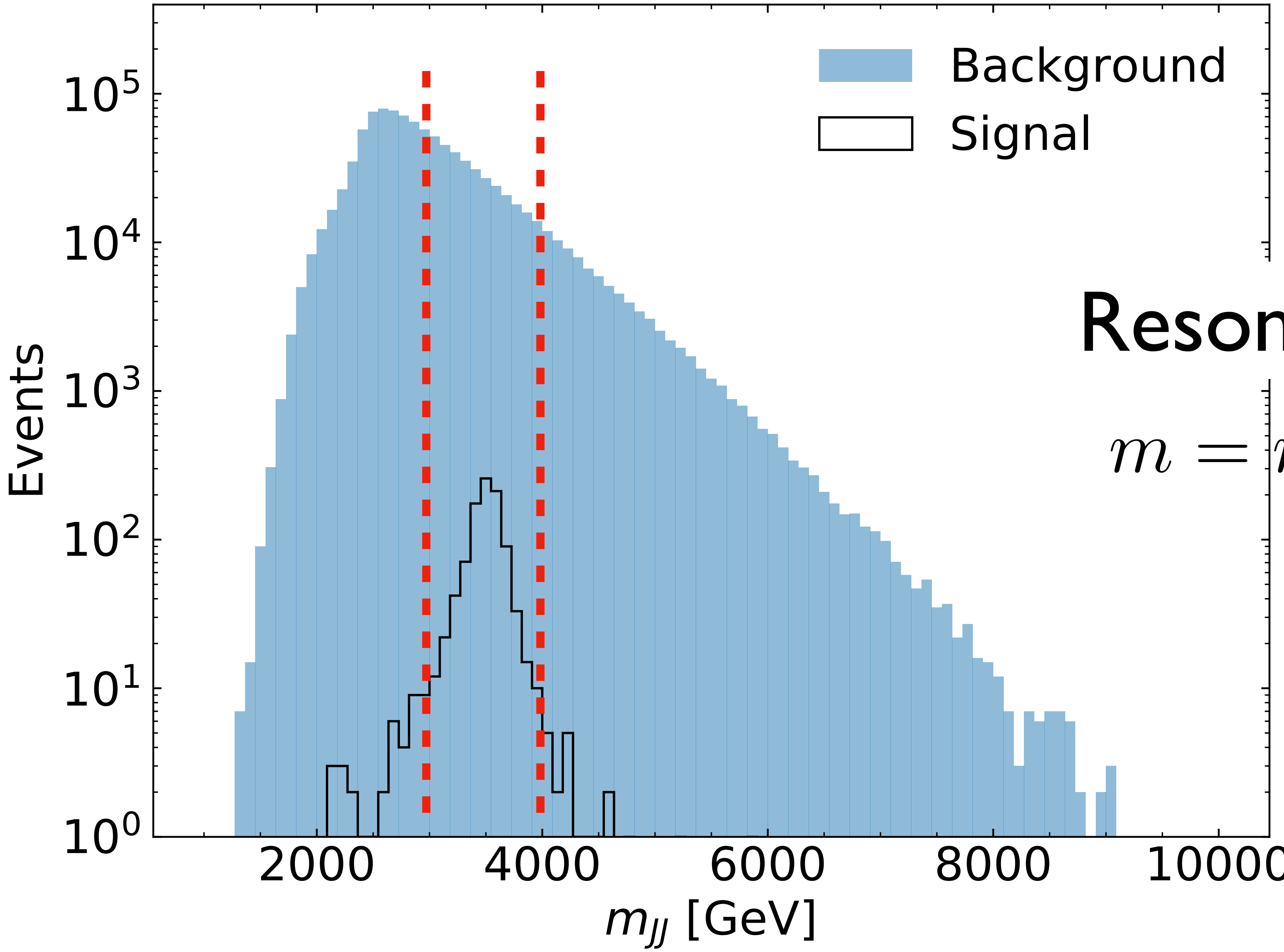
Future Organization of Physics Analysis Groups at the LHC??



from G. Kasieczka, B. Nachman, DS (eds), et al 2101.08320

Thanks for your attention!

LHC Olympics 2020: R&D Dataset



Resonant feature

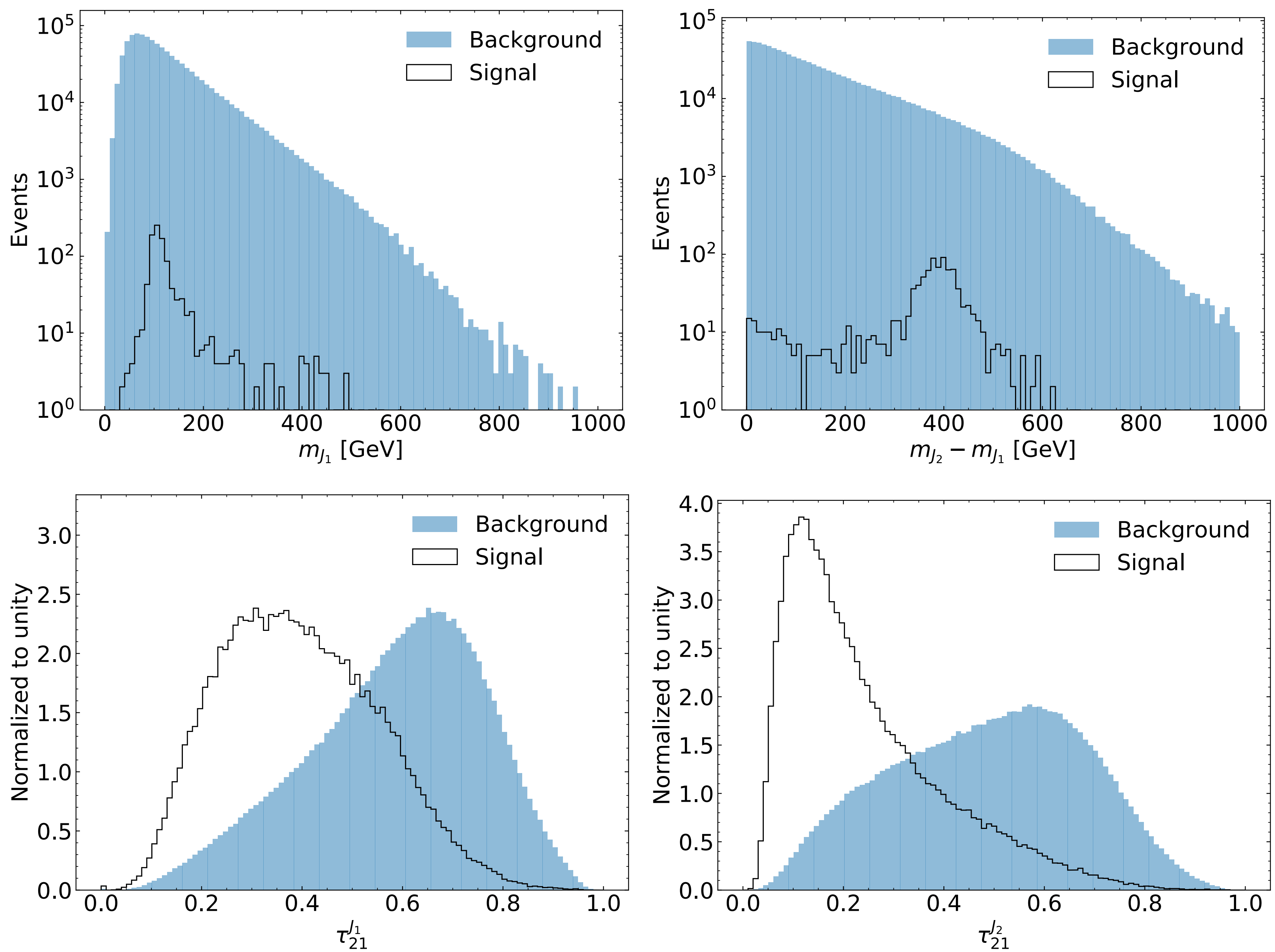
$$m = m_{Z'} = m_{JJ}$$

Benchmark signal strength:

S=500, B=500,000, B_{SR}=61,000

$S/B_{SR} \sim 6 \times 10^{-3}$, $S/\sqrt{B_{SR}} \sim 1.5$

LHC Olympics 2020: R&D Dataset



Additional features: $x = (m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2})$

Box 1

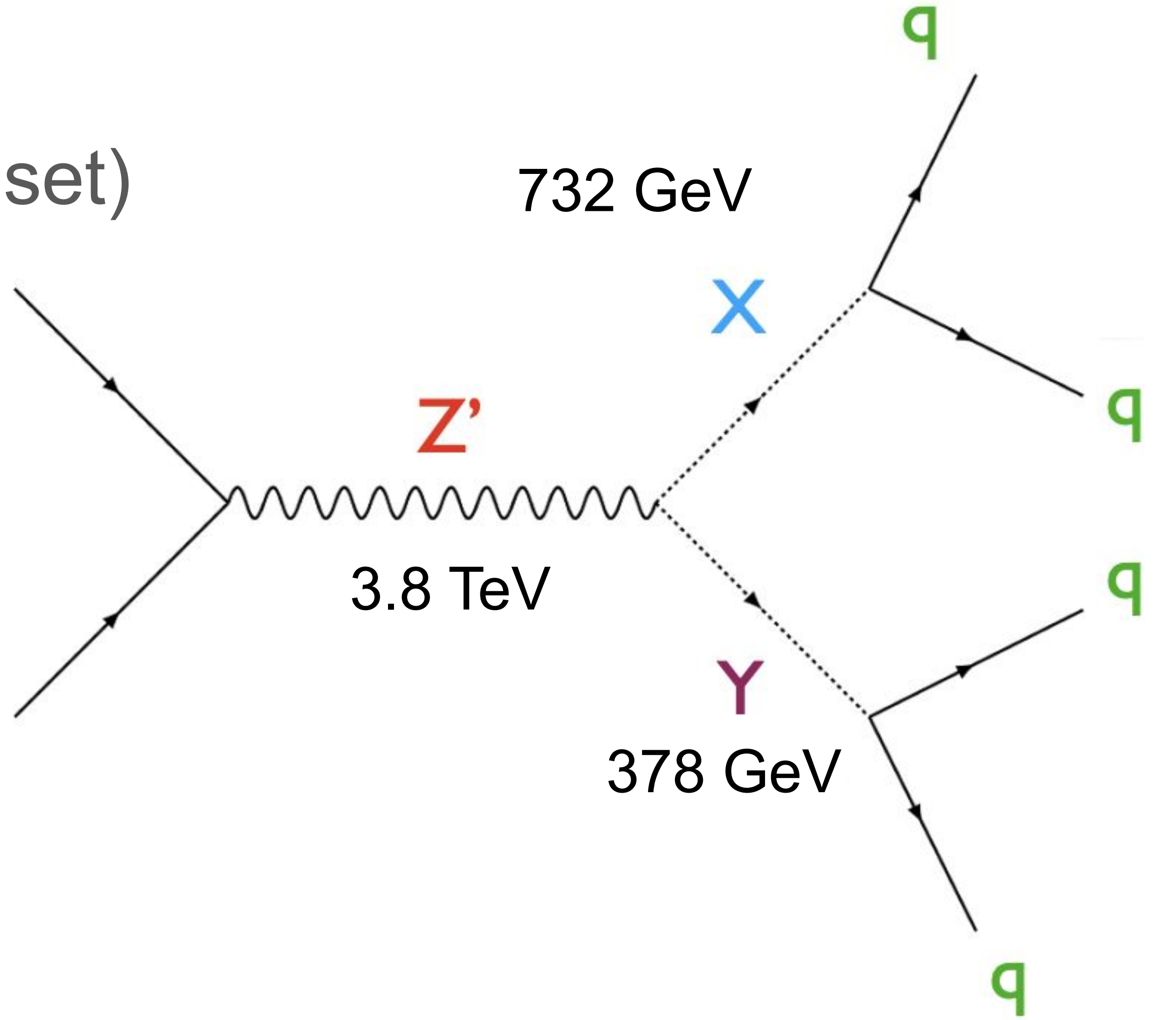
Signal: 834 events

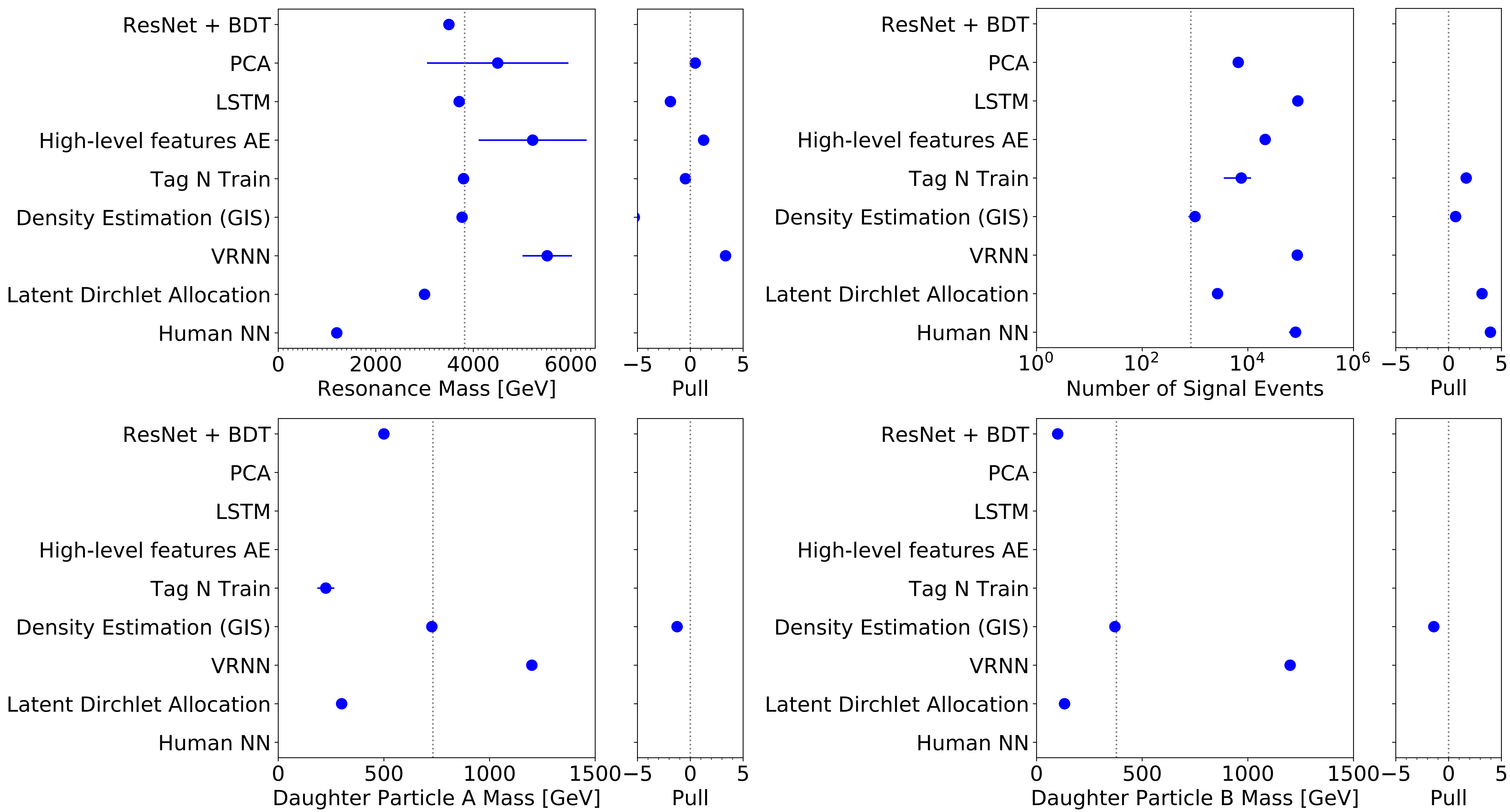
Z’->XY; X,Y->qq
(same topology as R&D dataset)

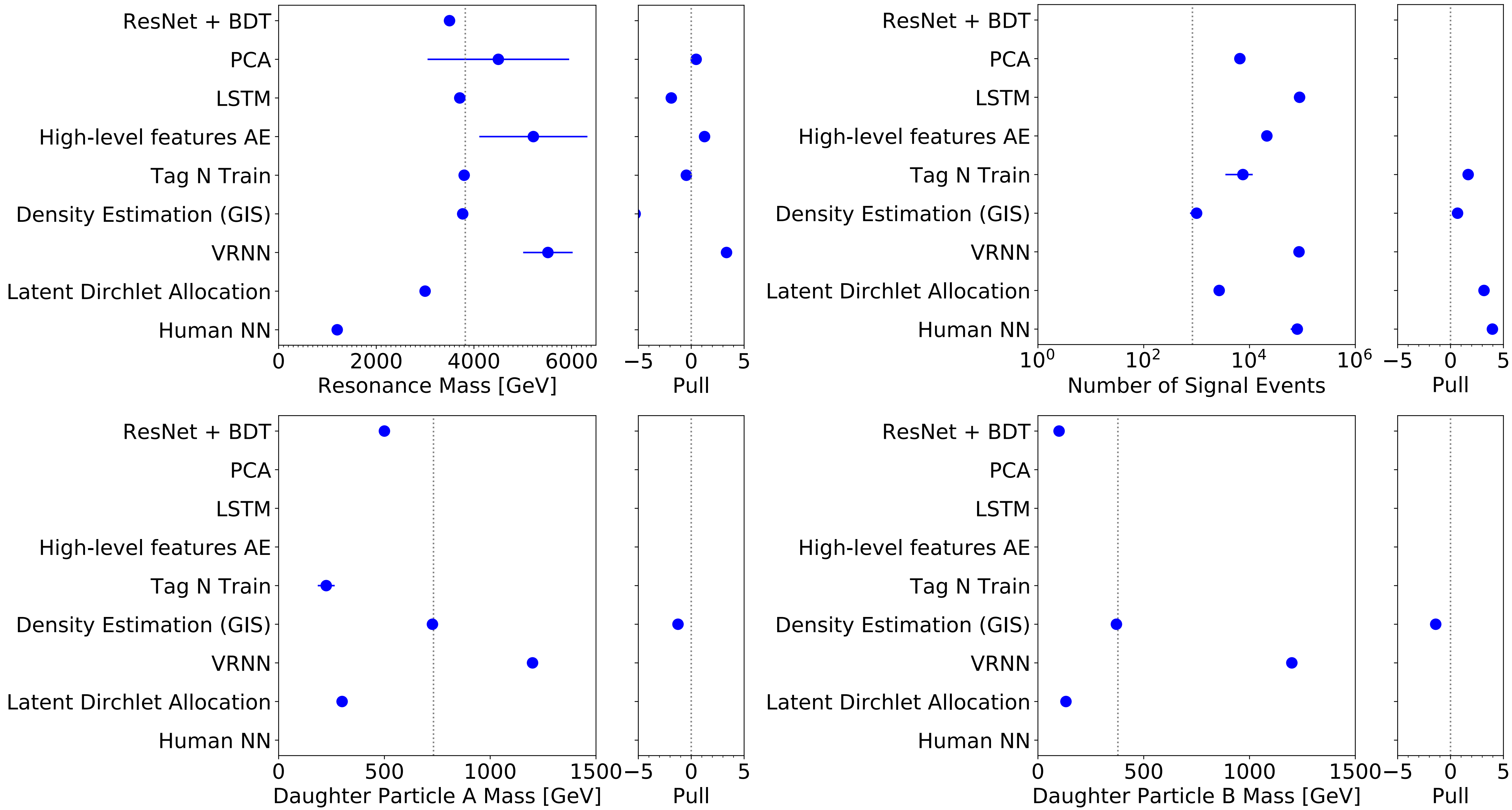
mZ’ = 3823 GeV

mX = 732 GeV

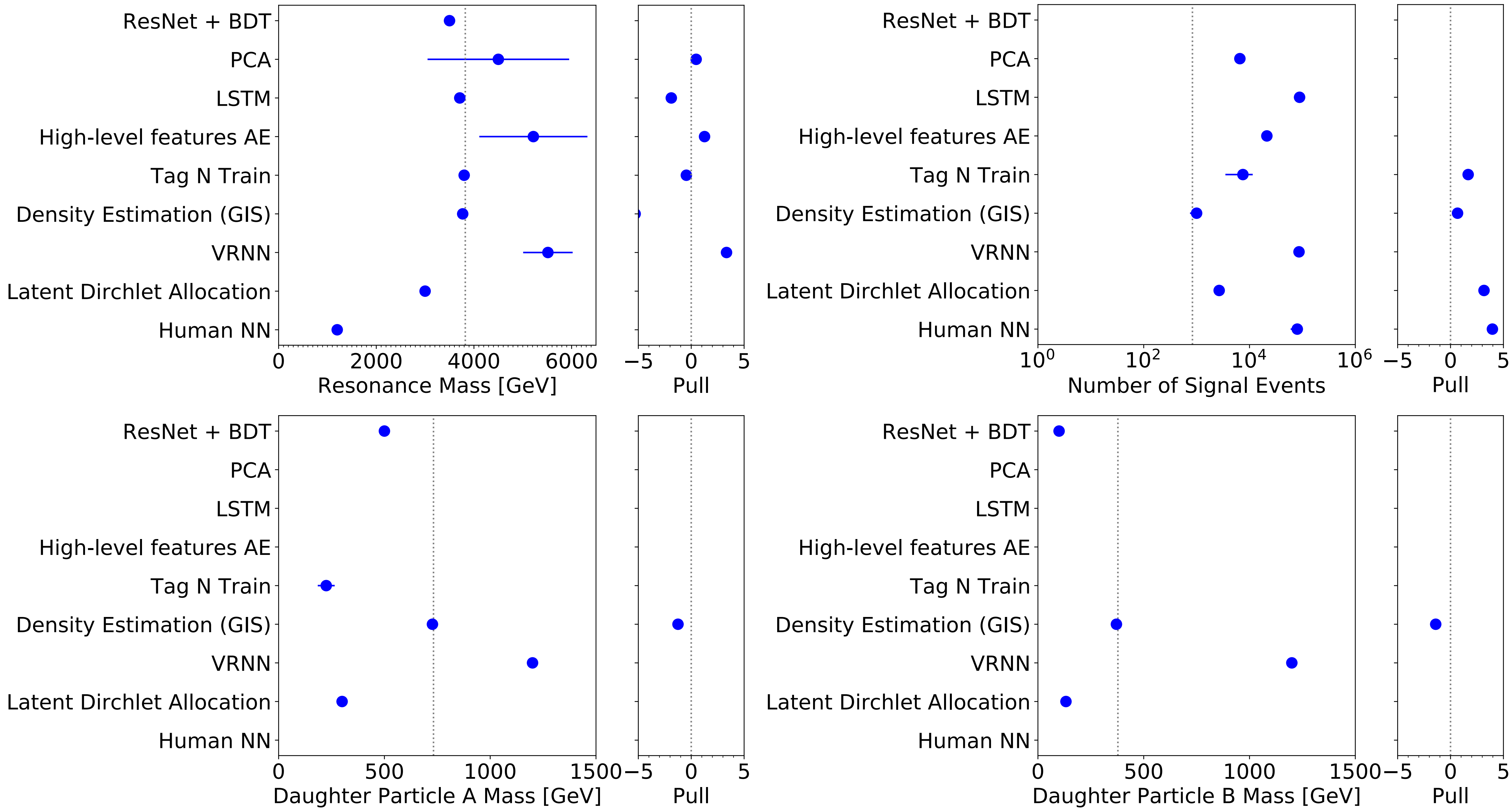
mY = 378 GeV




ResNet + BDT
PCA
LSTM
High-level features AE
Tag N Train
Density Estimation (GIS)
VRNN
Latent Dirchlet Allocation
Human NN
Resonance Mass [GeV]
Pull
Number of Signal Events
Daughter Particle A Mass [GeV]
Daughter Particle B Mass [GeV]



Two approaches clearly stood out:



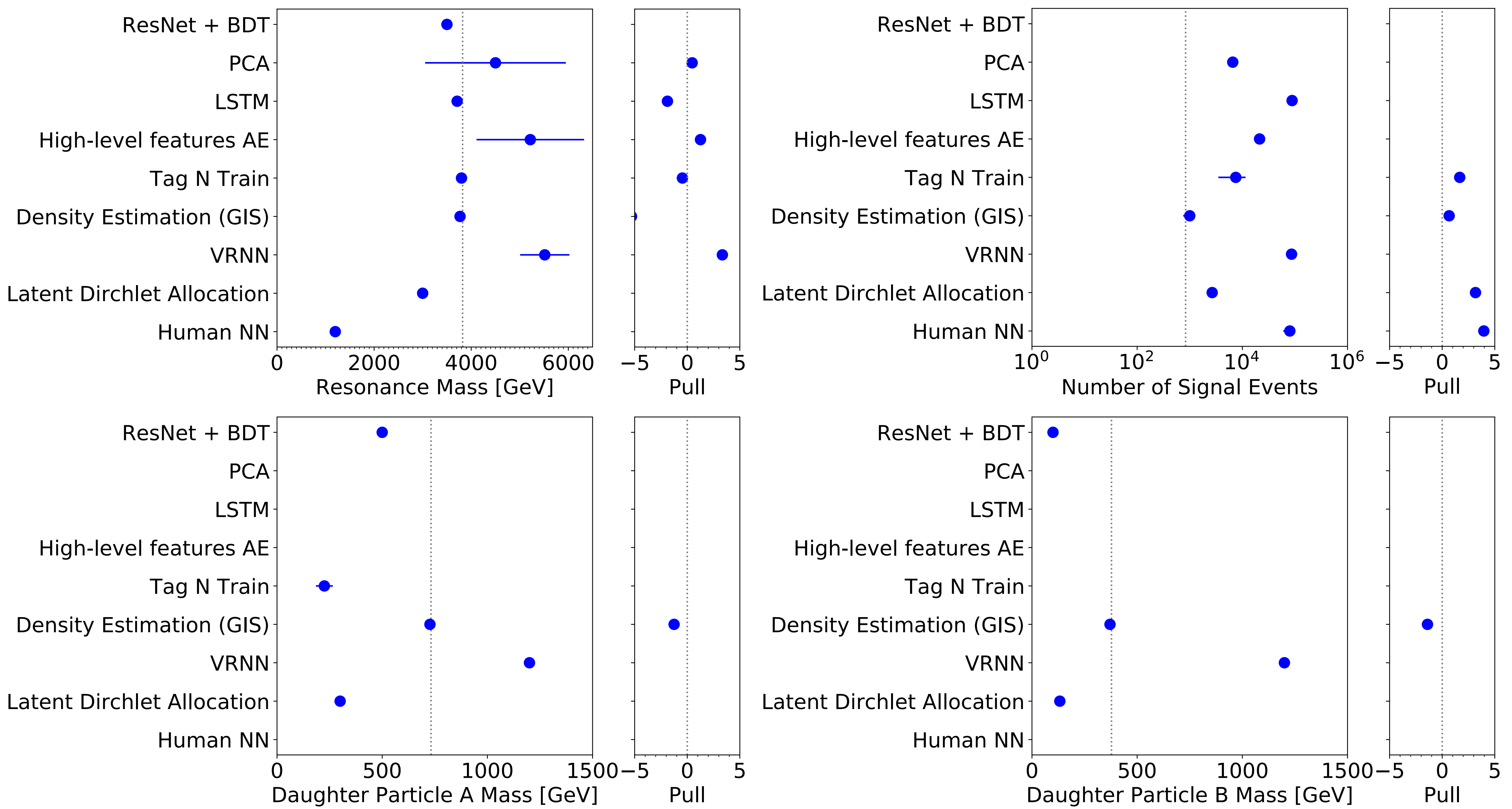
Conditional density estimation for anomaly detection

George Stein, Uros Seljak, Biwei Dai, He Jia

BERKELEY CENTER *for* COSMOLOGICAL PHYSICS

Two approaches clearly stood out:

Used the ANODE method with a novel density estimator!



Two approaches clearly stood out:

Tag N' Train

Oz Amram & Cristina Mantilla Suarez (Johns Hopkins)

JOHNS HOPKINS UNIVERSITY

Used a combination of autoencoders and CWoLa hunting

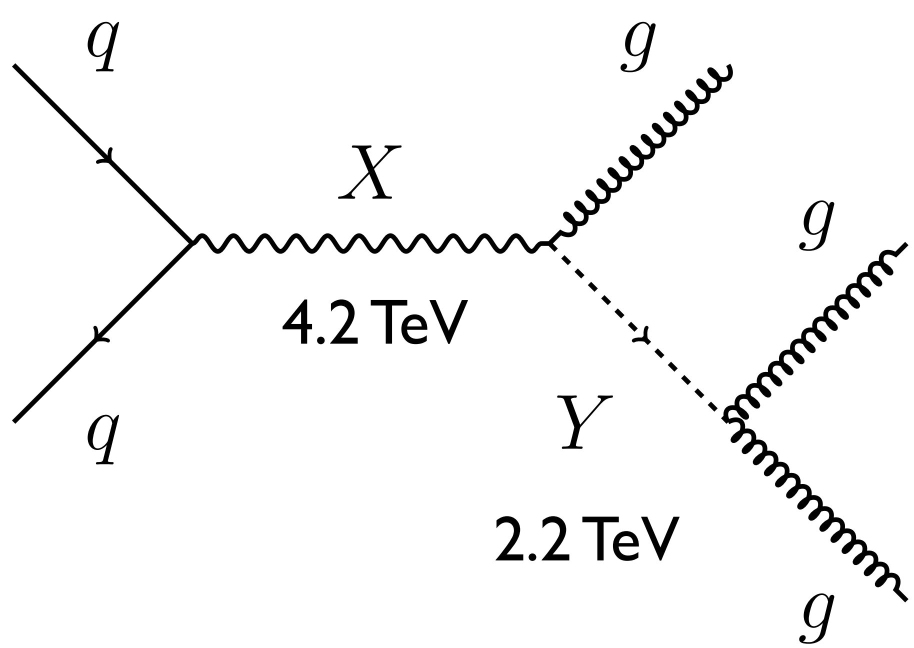
Box 2

No signal! QCD background only.

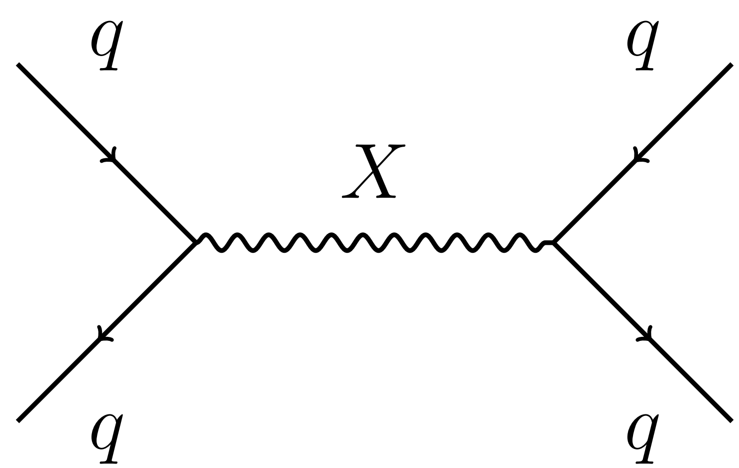
4 of the 5 submissions found false positives…

Clearly a matter of concern / area of future improvement for anomaly detection approaches!

Box 3



2000 events



1200 events

No jet substructure.

Two decay modes of X resonance. Need to combine to reach discovery significance.

No approach succeeded in finding the signal.

ANODE: Anomaly Detection with Density Estimation

Nachman & DS 2001.04990

Example of a new approach inspired by LHCO2020.

(See Ben's talk for additional new approaches!)

Use **neural density estimation** to directly learn the conditional probability densities from the data

$$\mathcal{P}(x|data;\ m_{JJ} \in SR)$$

$$\mathcal{P}(x|data;\ m_{JJ} \notin SR) = \mathcal{P}(x|bg;\ m_{JJ} \notin SR)$$

interpolate in (x,mJJ)

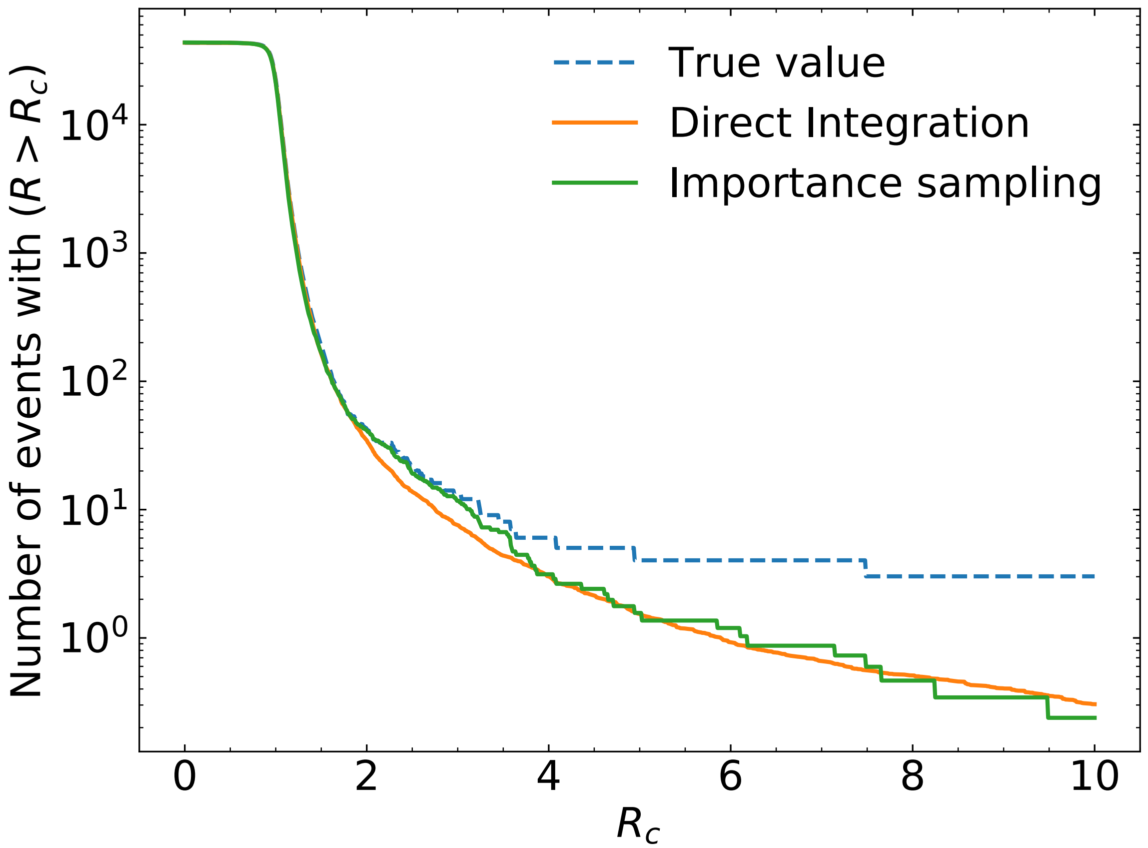
$$\mathcal{P}(x|bg;\ m_{JJ} \in SR)$$

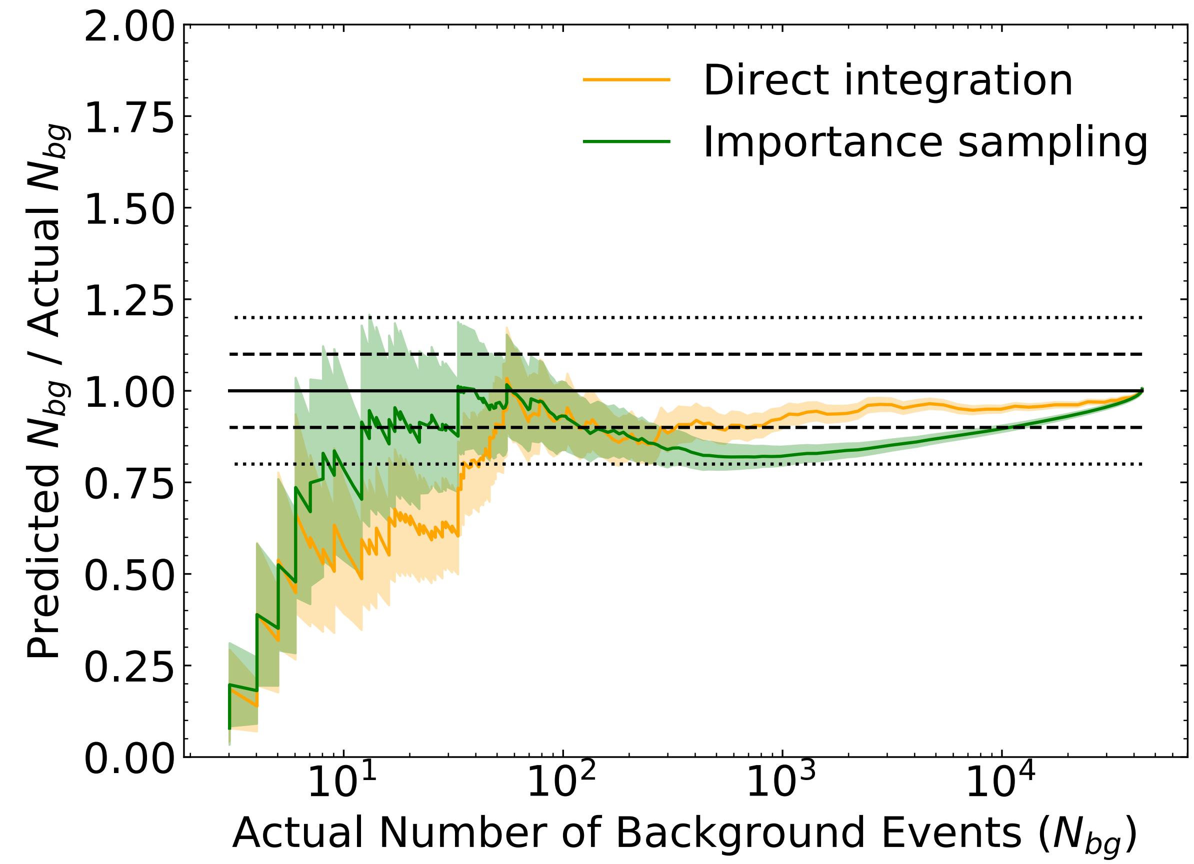
Construct the likelihood ratio: $$R(x) = \frac{\mathcal{P}(x|data;\ m_{JJ} \in SR)}{\mathcal{P}(x|bg;\ m_{JJ} \in SR)}$$

ANODE: Results on LHCO R&D Dataset

Nachman & DS 2001.04990

Novel aspect of ANODE: can estimate backgrounds directly with P(x|bg; m∈SR)

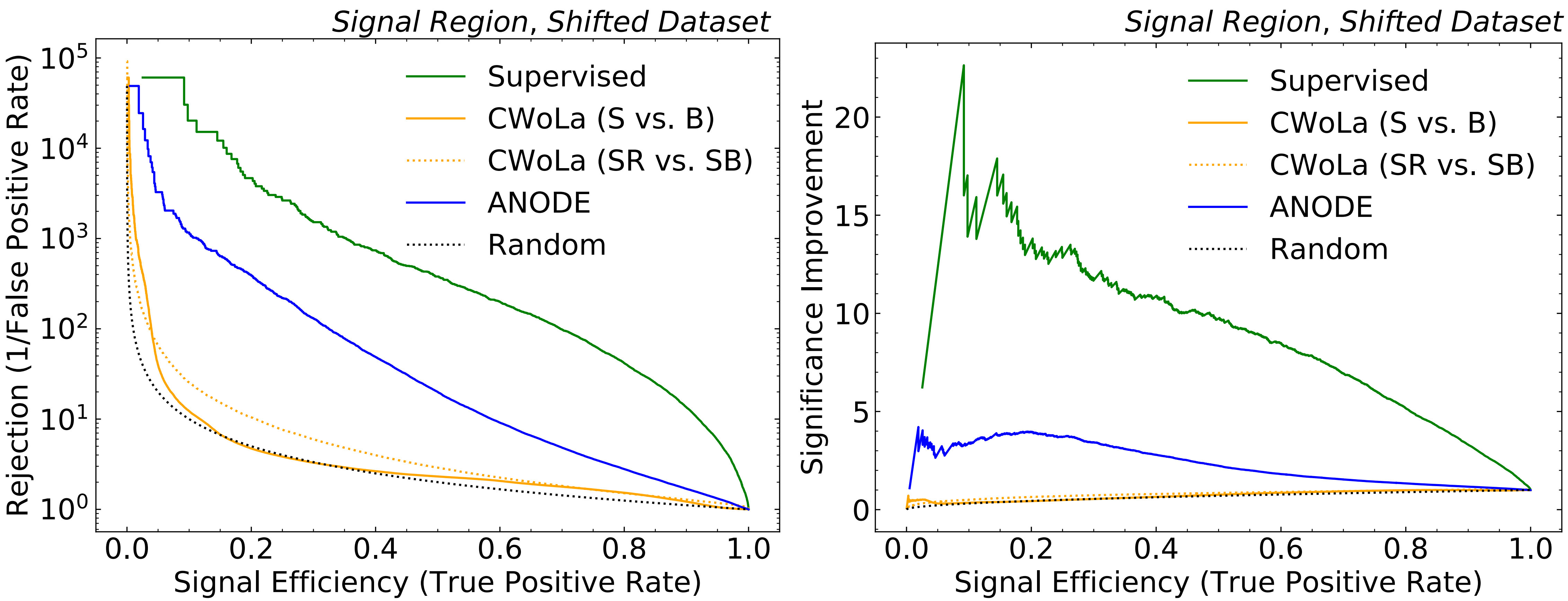




ANODE: Results on LHCO R&D Dataset

Ben Nachman & DS 2001.04990

Can also consider performance on a feature set which is not independent of m. We introduced artificial correlations just as proof of concept: $m_{J_{1,2}} \to m_{J_{1,2}} + c\, m_{JJ}$



ANODE is robust while CWoLa completely fails!