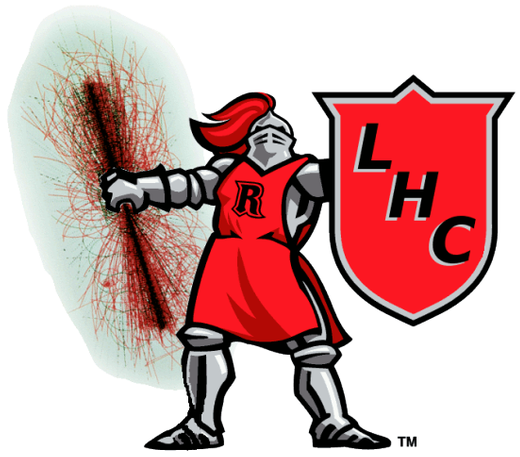
Anomaly Detection and LHCO2020

Introduction and Overview

David Shih

ML4Jets 2020, NYU

January 16, 2020



Excitement about Discovery Potential @ LHC

Then (2007)

History of the LHC Olympics

Prior to the LHC turn-on:

- Four dedicated workshops: July 2005 (CERN), February 2006 (CERN), August 2006 (KITP), March 2007 (Princeton).

- A number of black boxes consisting of reconstructed objects (electrons, photons, jets, MET, etc) were prepared using Madgraph, Pythia & PGS

- Anticipation was high that the LHC would discover ~~new physics~~ supersymmetry right away, so the focus was on characterizing the ~~new physics~~ supersymmetric model

- So almost all the black boxes were signal only and heavily based on supersymmetry

- And the primary goal was signal characterization (measuring masses, spins, branching ratios)

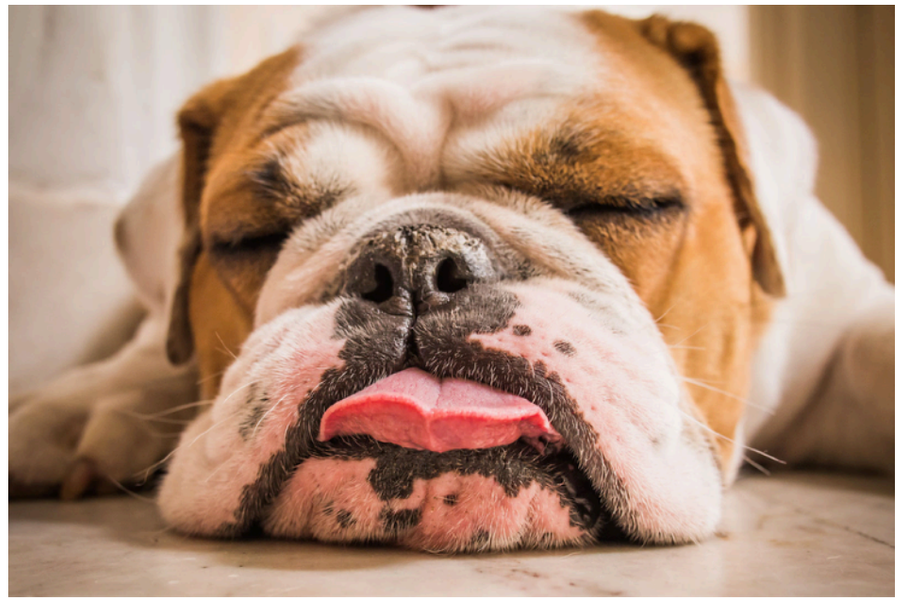
Excitement about Discovery Potential @ LHC

Then

Excitement about Discovery Potential @ LHC

Then

and now



LHC Olympics 2020

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

LHC Olympics 2020

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

The LHC has been operating for ~10 years

- Despite countless model-specific searches for new physics at the LHC, no evidence for new physics yet.
- What if we're not looking in the right places?

LHC Olympics 2020

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

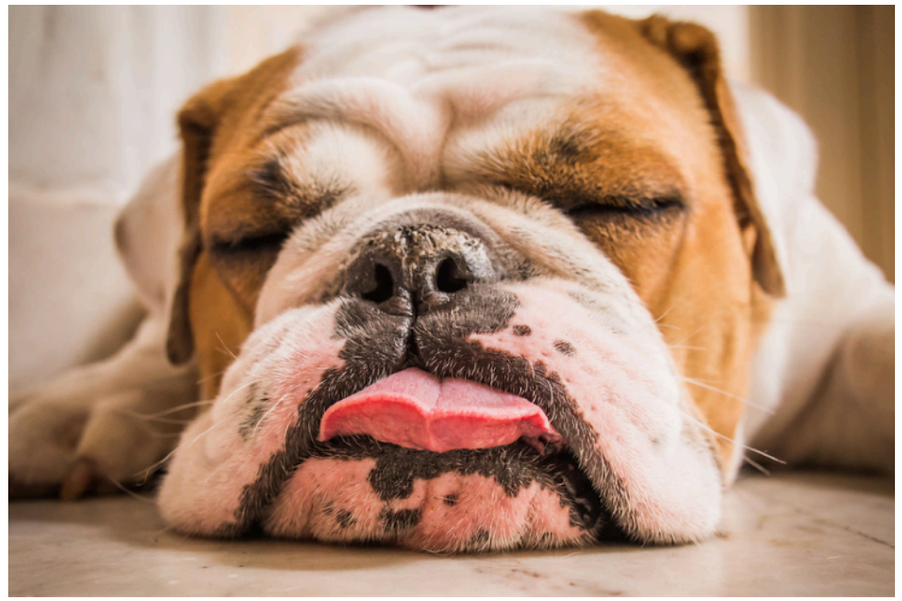
The LHC has been operating for ~10 years

- Despite countless model-specific searches for new physics at the LHC, no evidence for new physics yet.
- What if we're not looking in the right places?
- We need a new data challenge!
- New motivation: To spur the development of innovative, model-independent search methods, especially using deep learning
- And now the primary goal of the challenge is anomaly detection: to find the signal in the data if it is there (and then characterize it)

Excitement about Discovery Potential @ LHC

Then

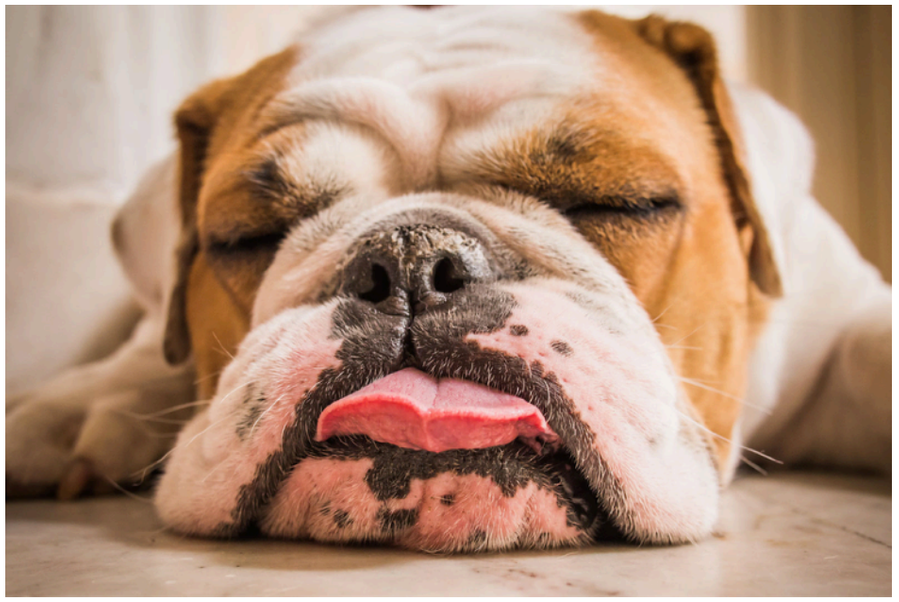
and now



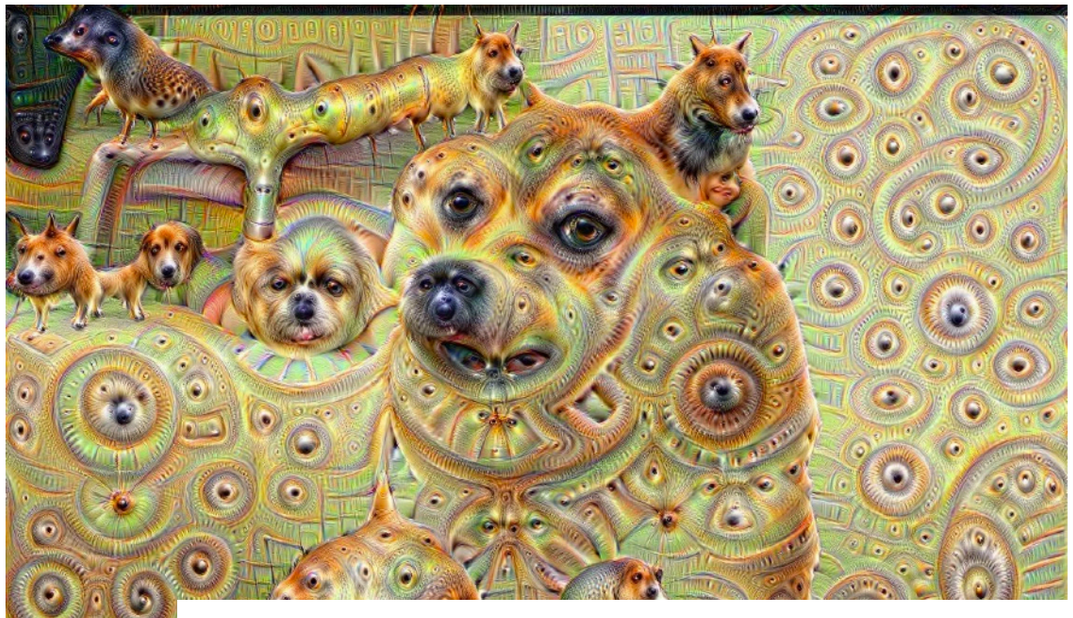
Excitement about Discovery Potential @ LHC

Then

and now



into the future



credit to Google Deep Dream and M. Perelstein

LHC Olympics 2020: Black Boxes

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

Three black boxes of simulated data were prepared:

- 1 million events each
- 4-vectors of every reconstructed particle (hadron) in the event
  - Particle ID, charge, etc not included
- Single R=1 jet trigger pT>1.2 TeV
- Black boxes are meant to be representative of actual data, meaning they are mostly background and may contain signals of new physics

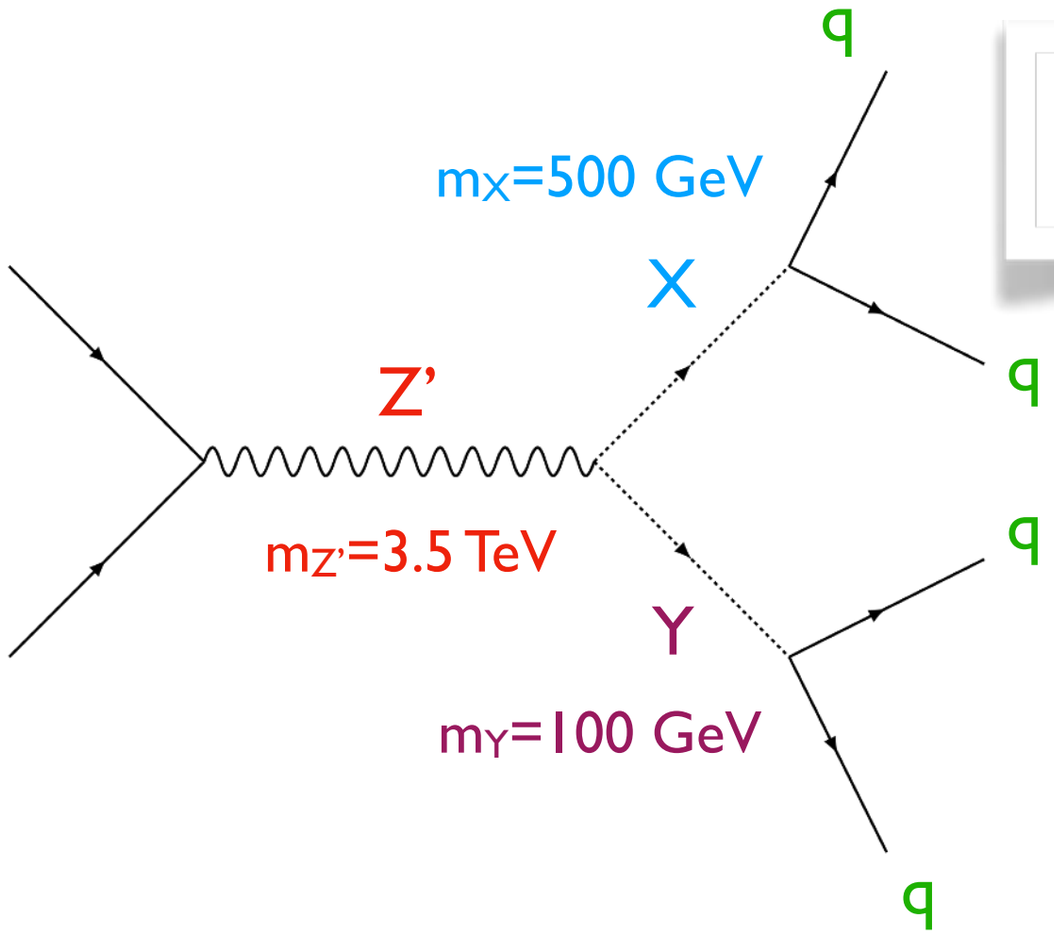
In addition, a sample of 1M QCD dijet events (produced with Pythia8 and Delphes3.4.1) was provided as a background sample.

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

LHC Olympics 2020: R&D Dataset

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

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



No explicit search at the LHC for this scenario!

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

LHC Olympics 2020: Submission format

Organizers: Gregor Kasieczka, Ben Nachman & David Shih

A p-value associated with the dataset having no new particles (null hypothesis).

Short answer text

As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).

Short answer text

How many signal events (+uncertainty) are in the dataset (before any selection criteria).

Short answer text

Please consider submitting plots or a Jupyter notebook! (these will be private and used only for the presentation / documentation at the end)

Add file

https://docs.google.com/forms/d/e/1FAIpQLScw323fa9qpLbdMvGtr2YeqcGTjE5Zm18-umIDiPldi_cWxVA/viewform

Elements of a successful search strategy (?)

What we *thought* it would take to do well in the challenge:

1. A model-agnostic search strategy with broad sensitivity to new physics
2. Accurate method of background estimation

(It is not enough to have a discriminant that is sensitive to new physics. One must also be able to predict the background in the signal region!)

Elements of a successful search strategy (?)

What we *thought* it would take to do well in the challenge:

1. A model-agnostic search strategy with broad sensitivity to new physics
2. Accurate method of background estimation

(It is not enough to have a discriminant that is sensitive to new physics. One must also be able to predict the background in the signal region!)

We'll see to what extent this was true!

Elements of a successful search strategy (?)

What we *thought* it would take to do well in the challenge:

1. A model-agnostic search strategy with broad sensitivity to new physics
2. Accurate method of background estimation

(It is not enough to have a discriminant that is sensitive to new physics. One must also be able to predict the background in the signal region!)

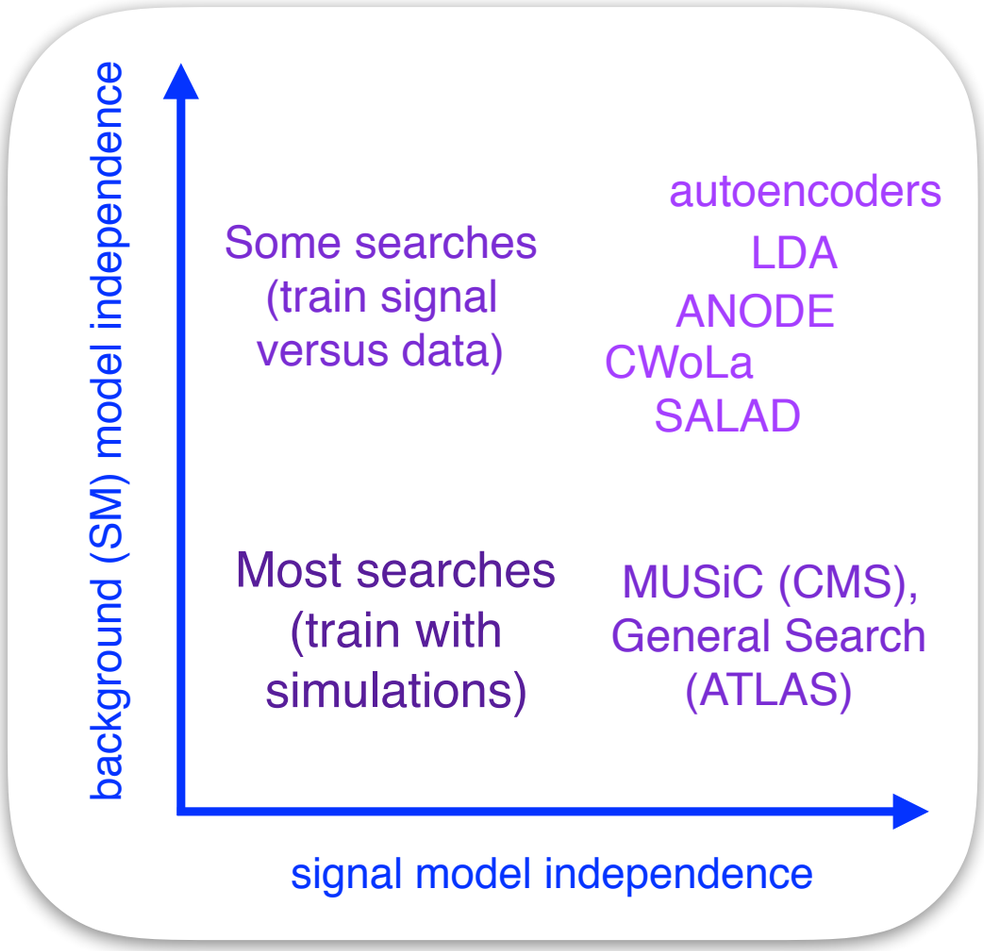
We'll see to what extent this was true!

But first, let's have a look at the landscape of model-independent search strategies…

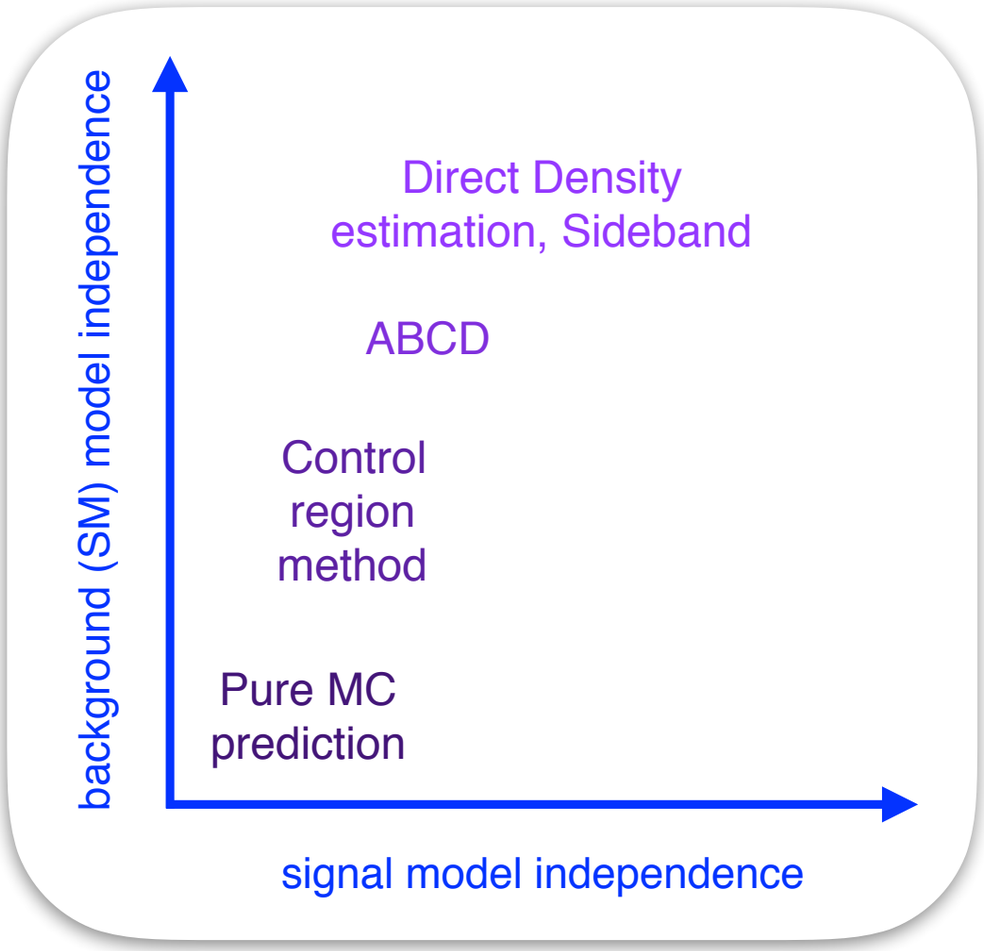
Overview of search strategies

Search strategies vary in their degrees of *signal and background model dependence*

from Nachman & DS 2001.04990



(a) Signal sensitivity



(b) Background specificity

Existing Model Independent Searches

A brief history of model independent searches in HEP:

• **D0**	*"Sleuth"*	PRD 62:092004 (2000) PRD 64:012004 (2001) PRL 86:3712 (2001)
• **H1 (Hera)**	*"General Search"*	PLB 602:14-30 (2004) 0705.3721
• **CDF**	*"Sleuth/Vista"*	0712.1311 PRD 78:012002 (2008) 0712.2534 (submitted to PRL, NEVER PUBLISHED) 0809.3781 PRD 79:011101 (2009)
• **CMS**	*"MUSIC"*	CMS-PAS-EXO-14-016
• **ATLAS**	*"Model independent general search"*	1807.07447 EPJC 79:120 (2019)

Existing Model Independent Searches

The general idea behind **all** of these: ***data vs MC comparison***

From CDF 0712.2534:

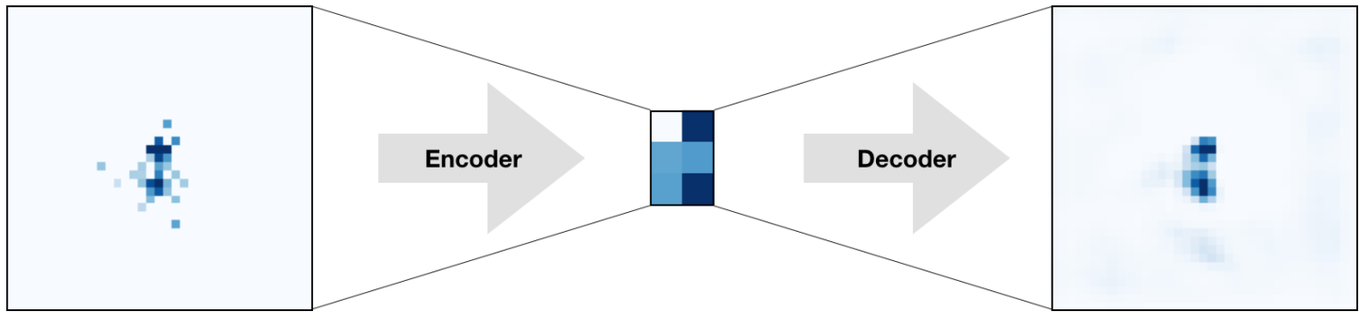
> A global comparison of data to standard model prediction is made in 16,486 kinematic distributions in 344 populated exclusive final states. In each final state, the bottom quark (b), and missing momentum ($\not{p}$). Monte Carlo event generators are used to determine the standard model prediction. VISTA partitions data and Monte

Existing searches compare many 1D histograms.

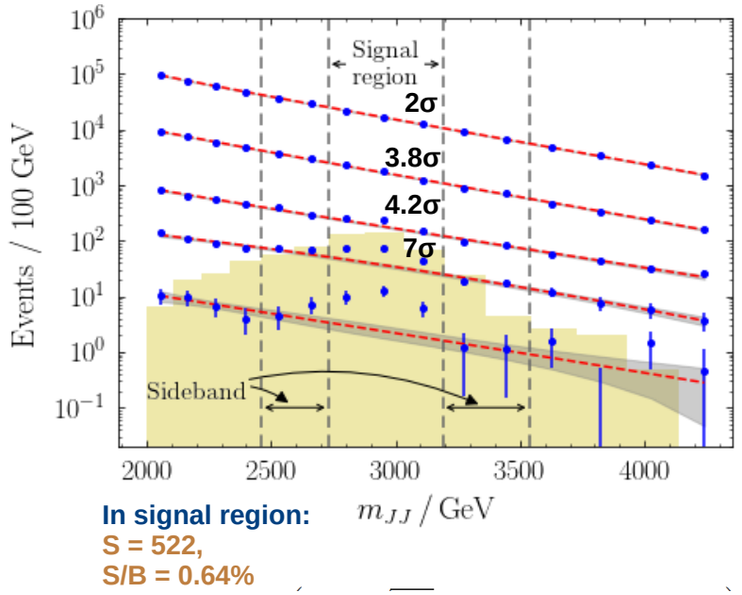
Optimal version would use DNN on full phase space to distinguish data from background MC (D'Agnolo, Wulzer et al 1806.02350, 1912.12155)

Signal model independent but background model dependent

Autoencoders

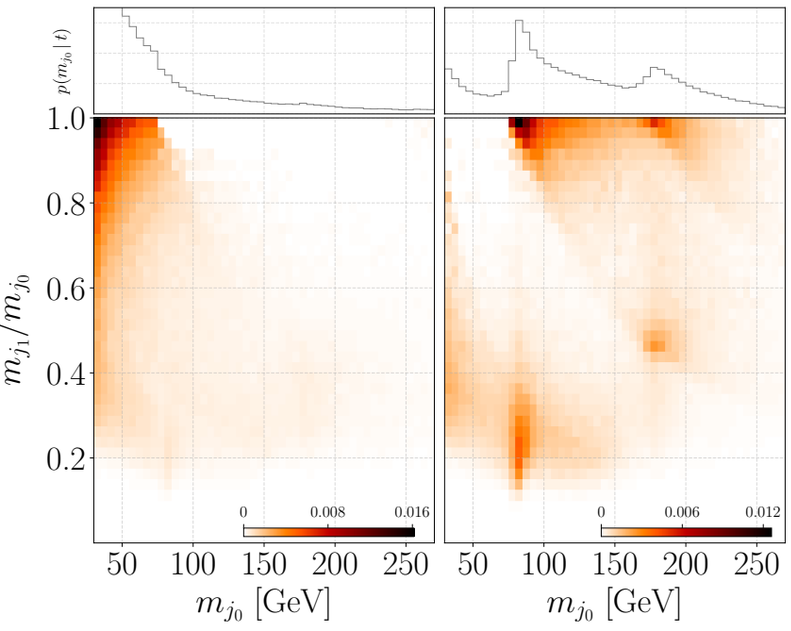


CWoLa

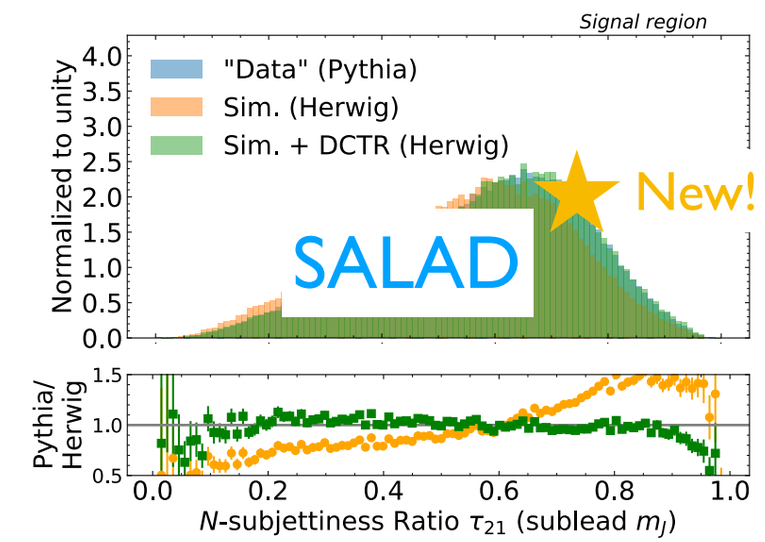


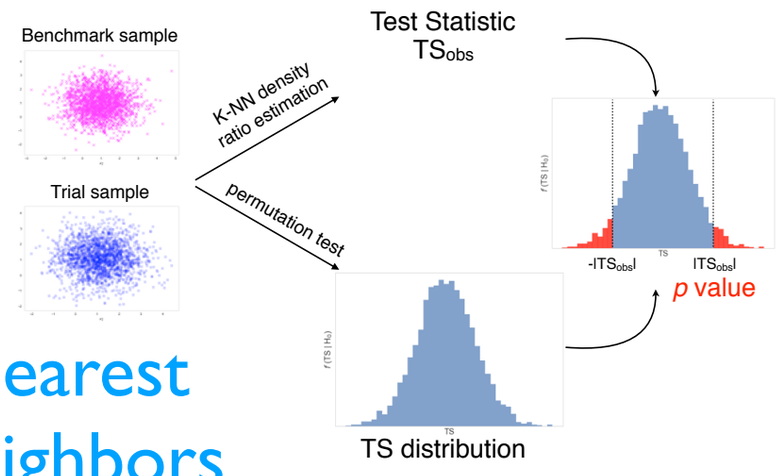
In signal region:
S = 522,
S/B = 0.64%

Probabilistic Modeling

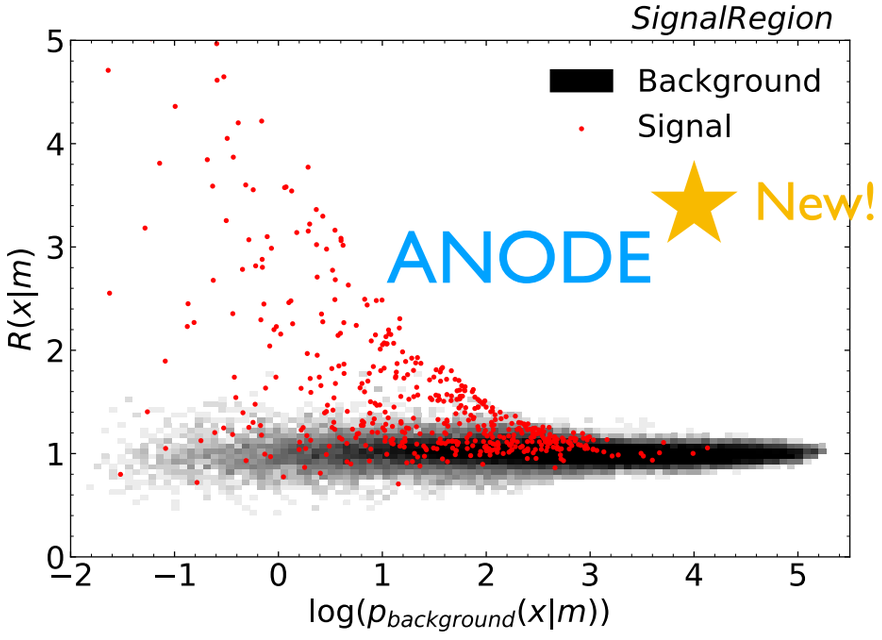


New approaches

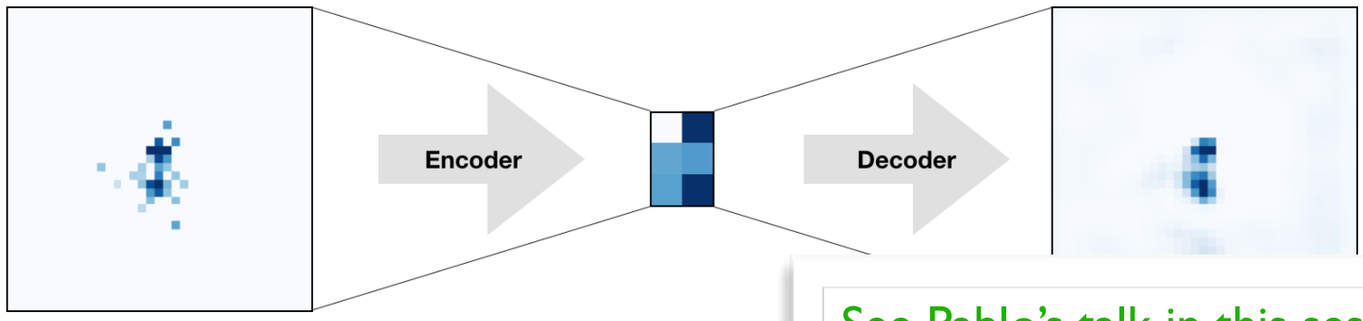




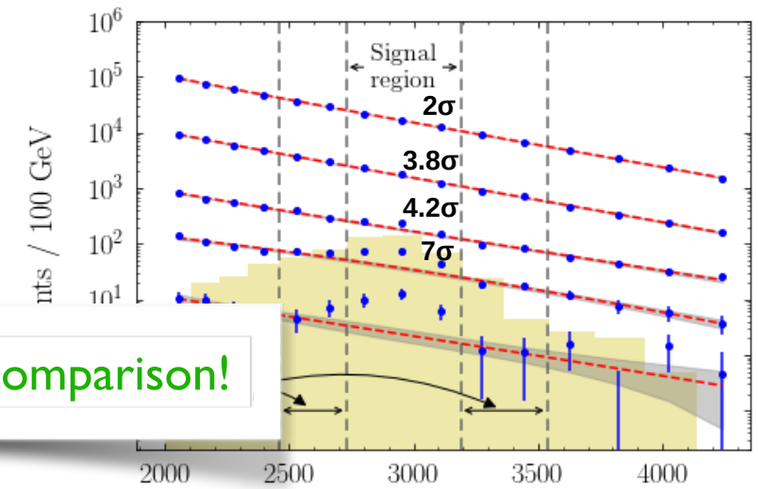
Nearest Neighbors



Autoencoders



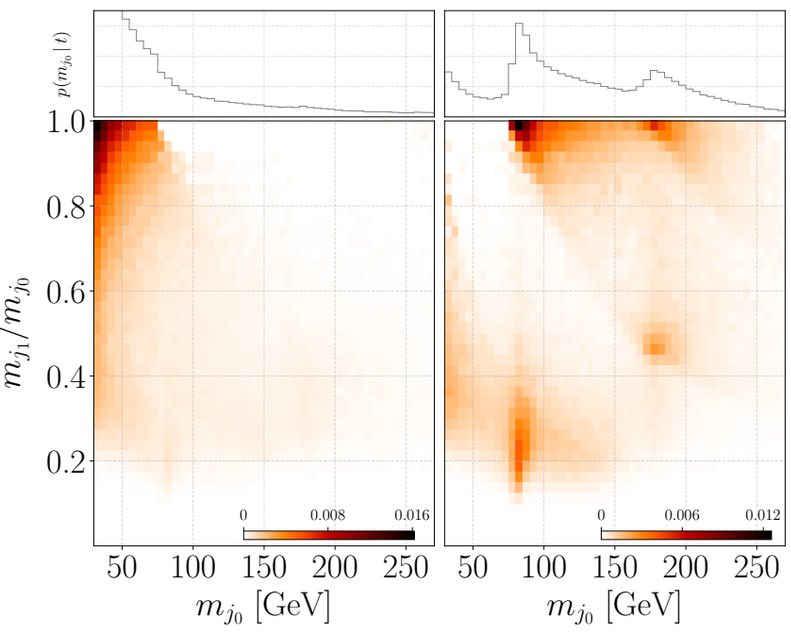
CWoLa



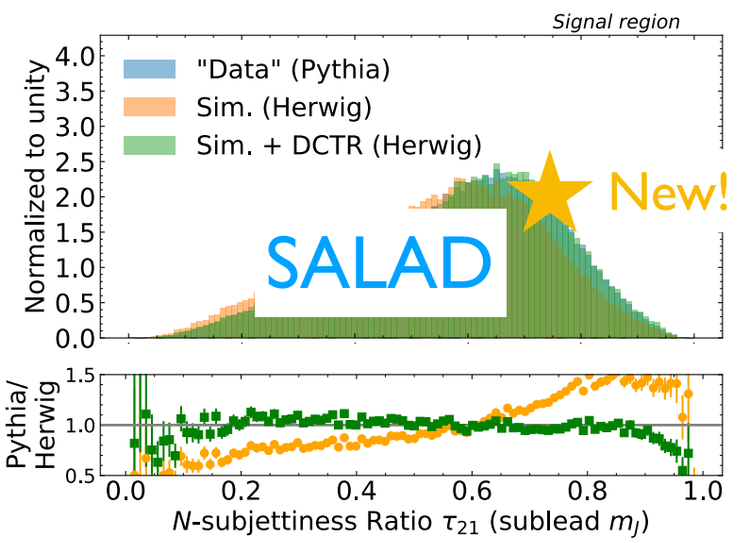
In signal region:
S = 522,
S/B = 0.64%

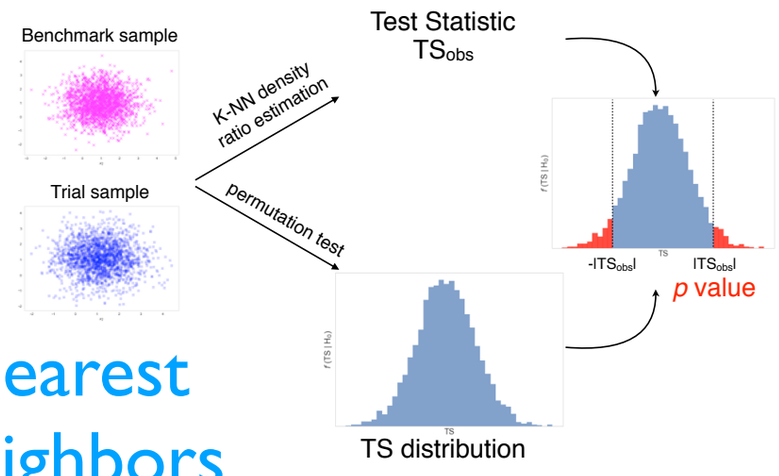
See Pablo's talk in this session for a comparison!

Probabilistic Modeling

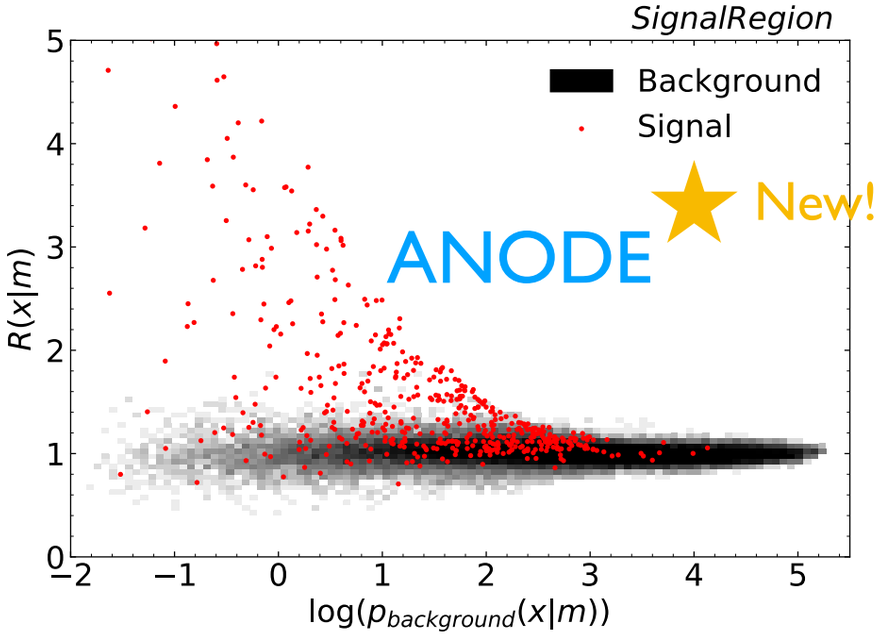


New approaches

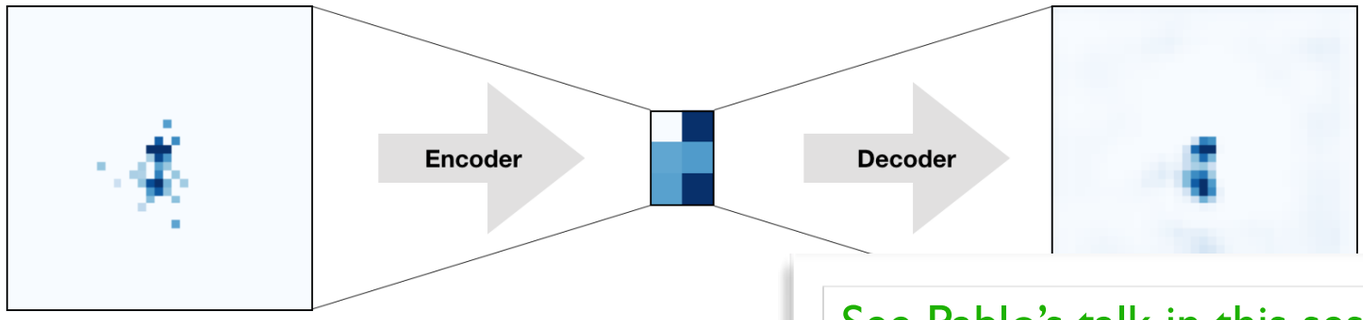




Nearest Neighbors

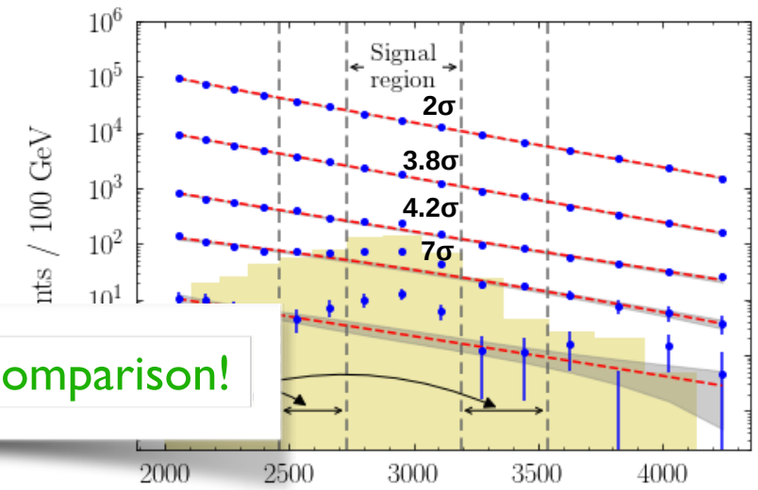


Autoencoders



See Pablo's talk in this session for a comparison!

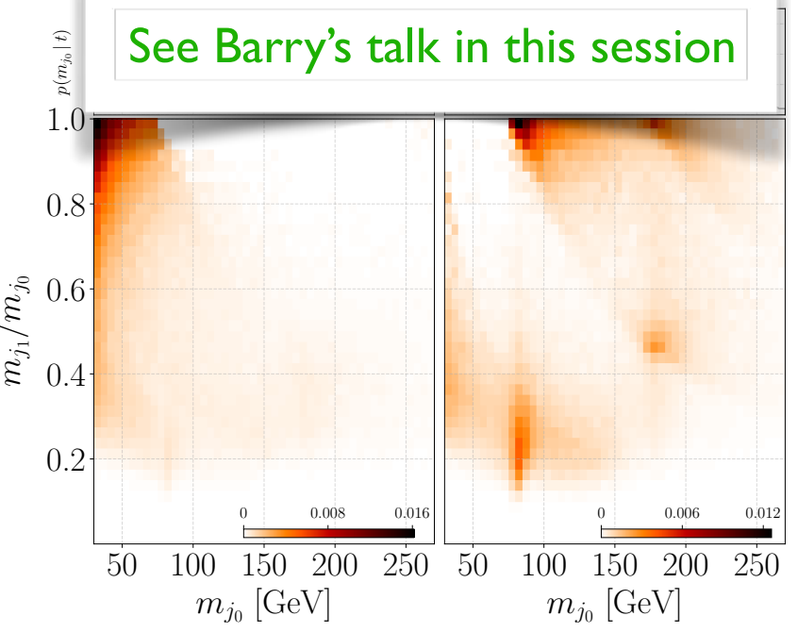
CWoLa



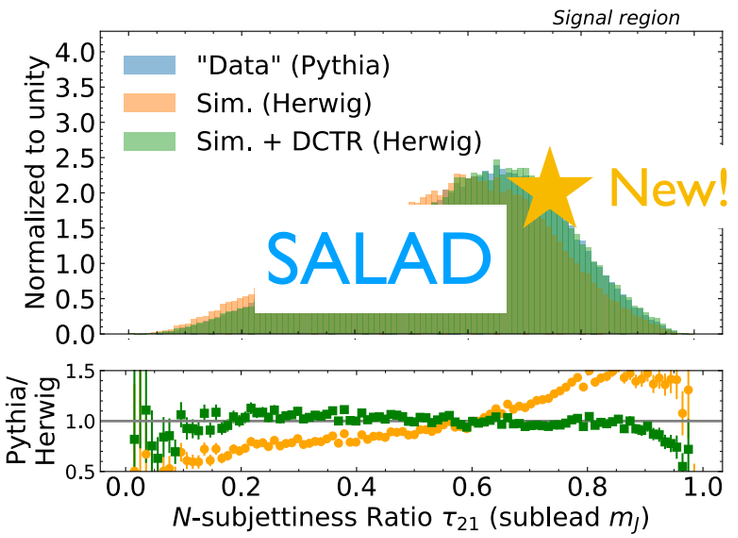
In signal region:
S = 522,
S/B = 0.64%

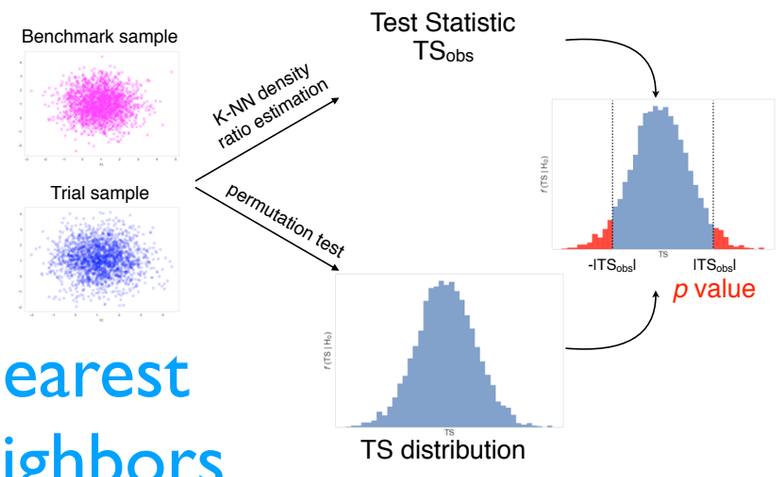
Probabilistic Modeling

See Barry's talk in this session

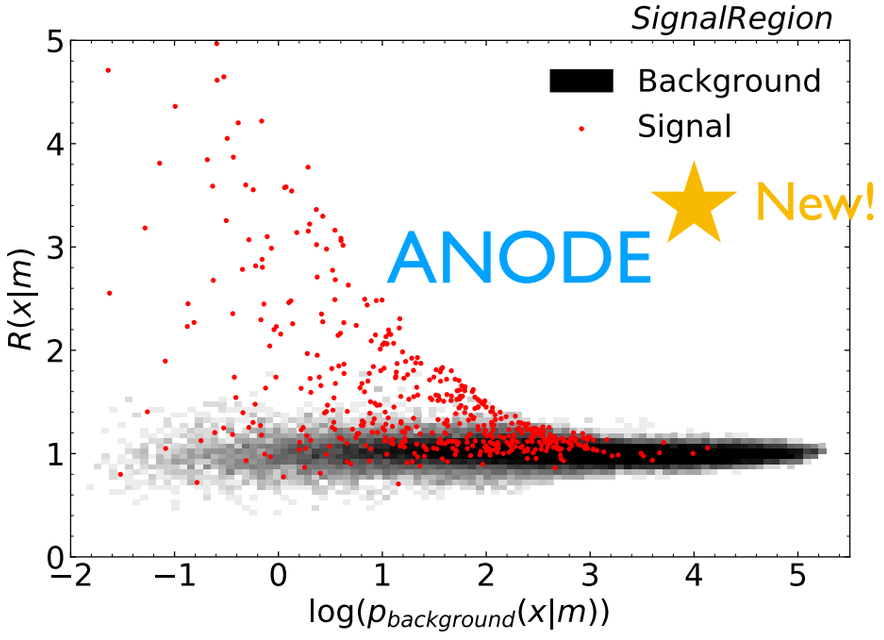


New approaches

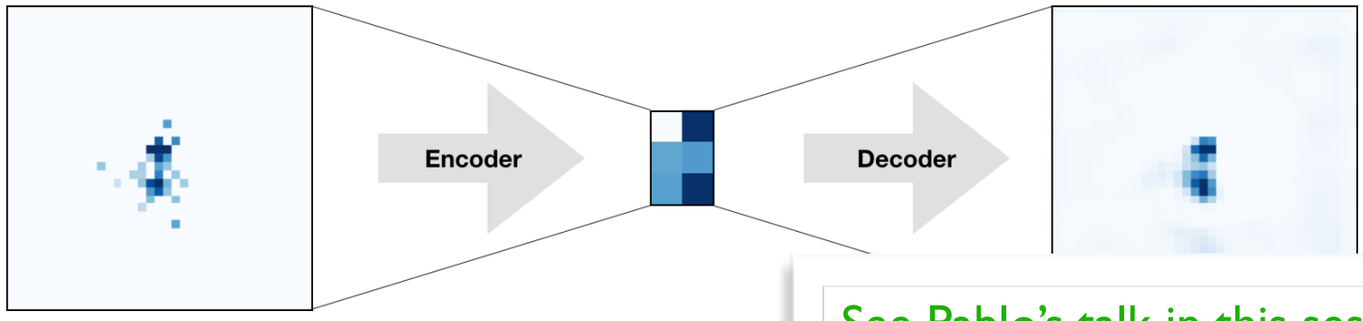




Nearest Neighbors

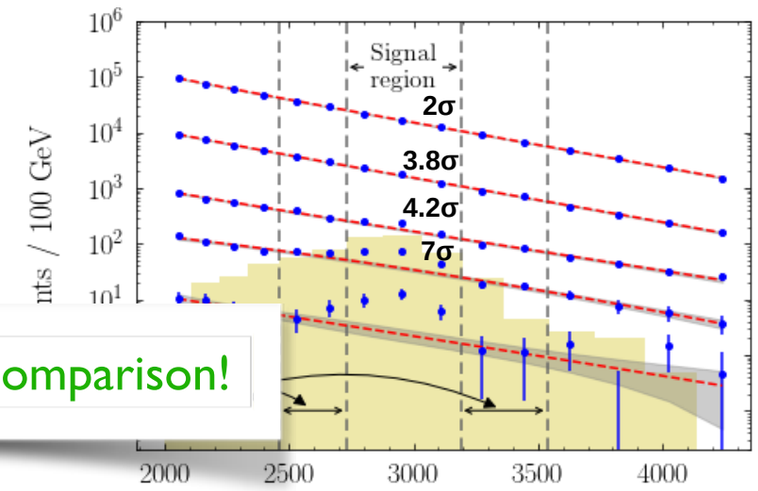


Autoencoders



See Pablo's talk in this session for a comparison!

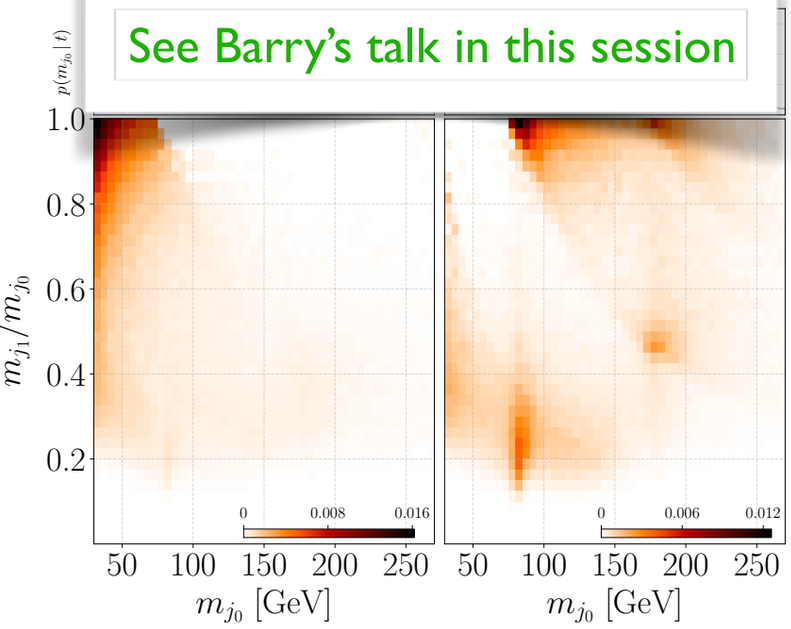
CWoLa



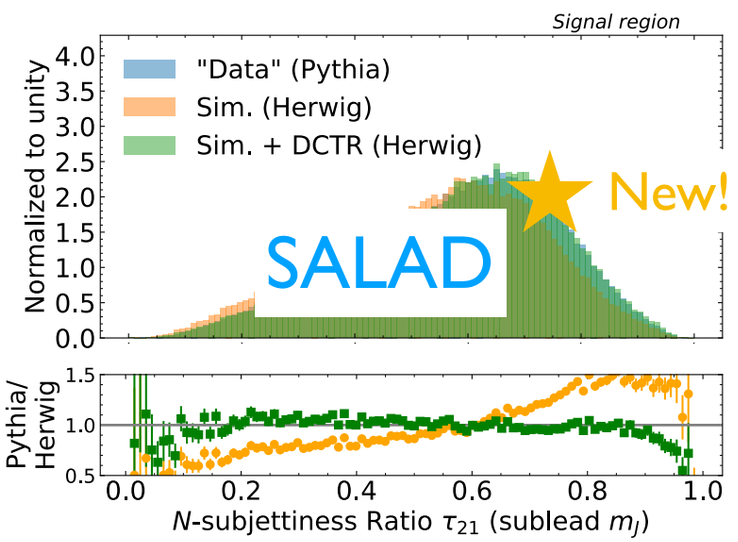
In signal region:
S = 522,
S/B = 0.64%

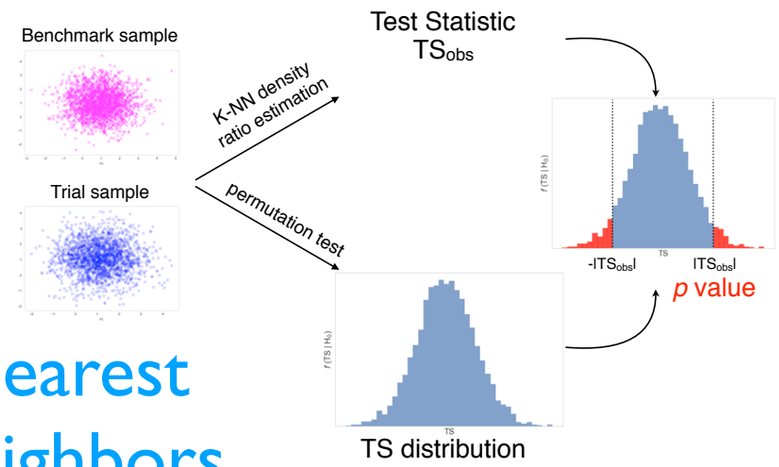
Probabilistic Modeling

See Barry's talk in this session

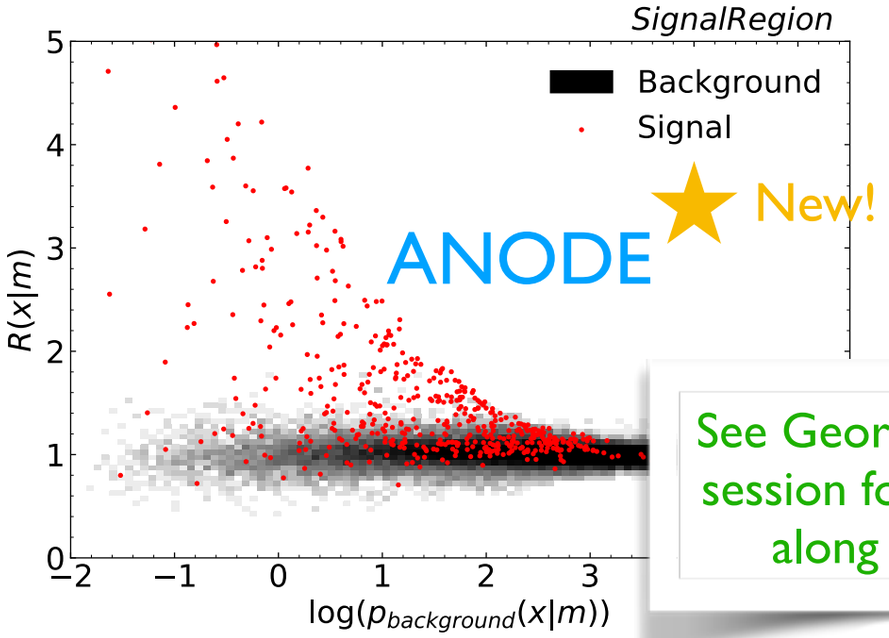


New approaches





Nearest Neighbors



See George's talk in this session for an approach along these lines

Enhancing the bump hunt

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

- Let m be a feature where the signal is assumed to be localized around m=m0 while the background is smooth.
- We can divide the data into the signal region: $m \in (m0-\delta m, m0+\delta m)$ and the sideband region $m \notin (m0-\delta m, m0+\delta m)$

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

- Let m be a feature where the signal is assumed to be localized around m=m0 while the background is smooth.
- We can divide the data into the signal region: $m \in (m0-\delta m, m0+\delta m)$ and the sideband region $m \notin (m0-\delta m, m0+\delta m)$

Traditional bump hunt

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

- Let m be a feature where the signal is assumed to be localized around m=m0 while the background is smooth.
- We can divide the data into the signal region: $m \in (m0-\delta m, m0+\delta m)$ and the sideband region $m \notin (m0-\delta m, m0+\delta m)$

Traditional bump hunt

- Let x be additional discriminating features
- Can we formulate a model-agnostic discriminant R(x) which is broadly sensitive to resonant new physics?

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

- Let m be a feature where the signal is assumed to be localized around m=m0 while the background is smooth.
- We can divide the data into the signal region: m∈(m0-δm,m0+δm) and the sideband region m∉(m0-δm,m0+δm)

Traditional bump hunt

- Let x be additional discriminating features
- Can we formulate a model-agnostic discriminant R(x) which is broadly sensitive to resonant new physics?
- $R(x) = \frac{P_{data}(x|m \in SR)}{P_{bg}(x|m \in SR)}$ would be optimal

Enhancing the bump hunt

A growing number of methods (CWoLa, ANODE, SALAD, …) aim to enhance the bump hunt using additional features:

- Let m be a feature where the signal is assumed to be localized around m=m0 while the background is smooth.
- We can divide the data into the signal region: m∈(m0-δm,m0+δm) and the sideband region m∉(m0-δm,m0+δm)

Traditional bump hunt

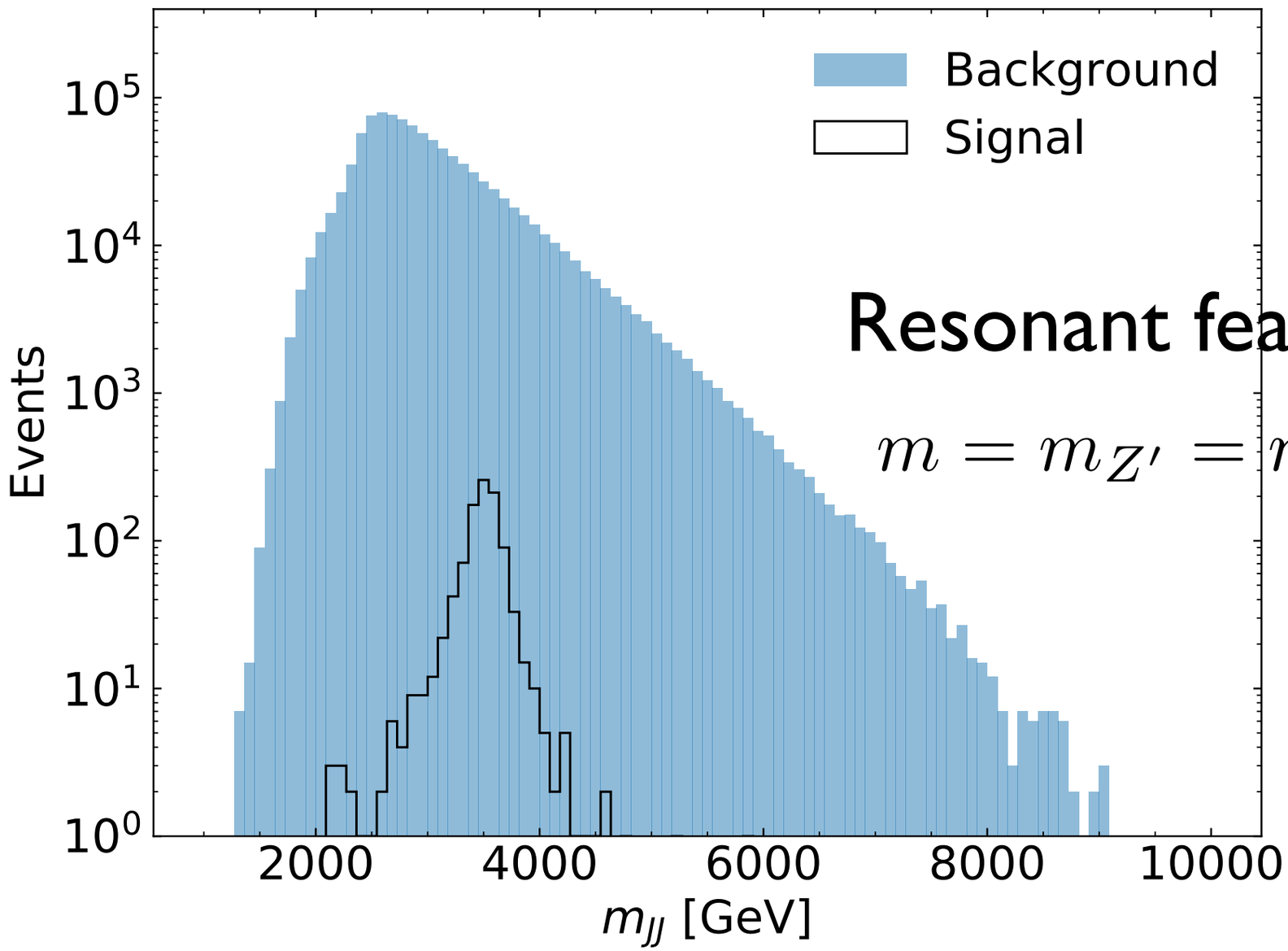
- Let x be additional discriminating features
- Can we formulate a model-agnostic discriminant R(x) which is broadly sensitive to resonant new physics?
- $R(x) = \frac{P_{data}(x|m \in SR)}{P_{bg}(x|m \in SR)}$ would be optimal

Enhanced bump hunt

Signal: $Z' \rightarrow XY$; $X, Y \rightarrow qq$; $m_{Z'}$=3.5 TeV, m_X=500 GeV, m_Y=100 GeV

Background: QCD dijets

Example: LHCO R&D Dataset



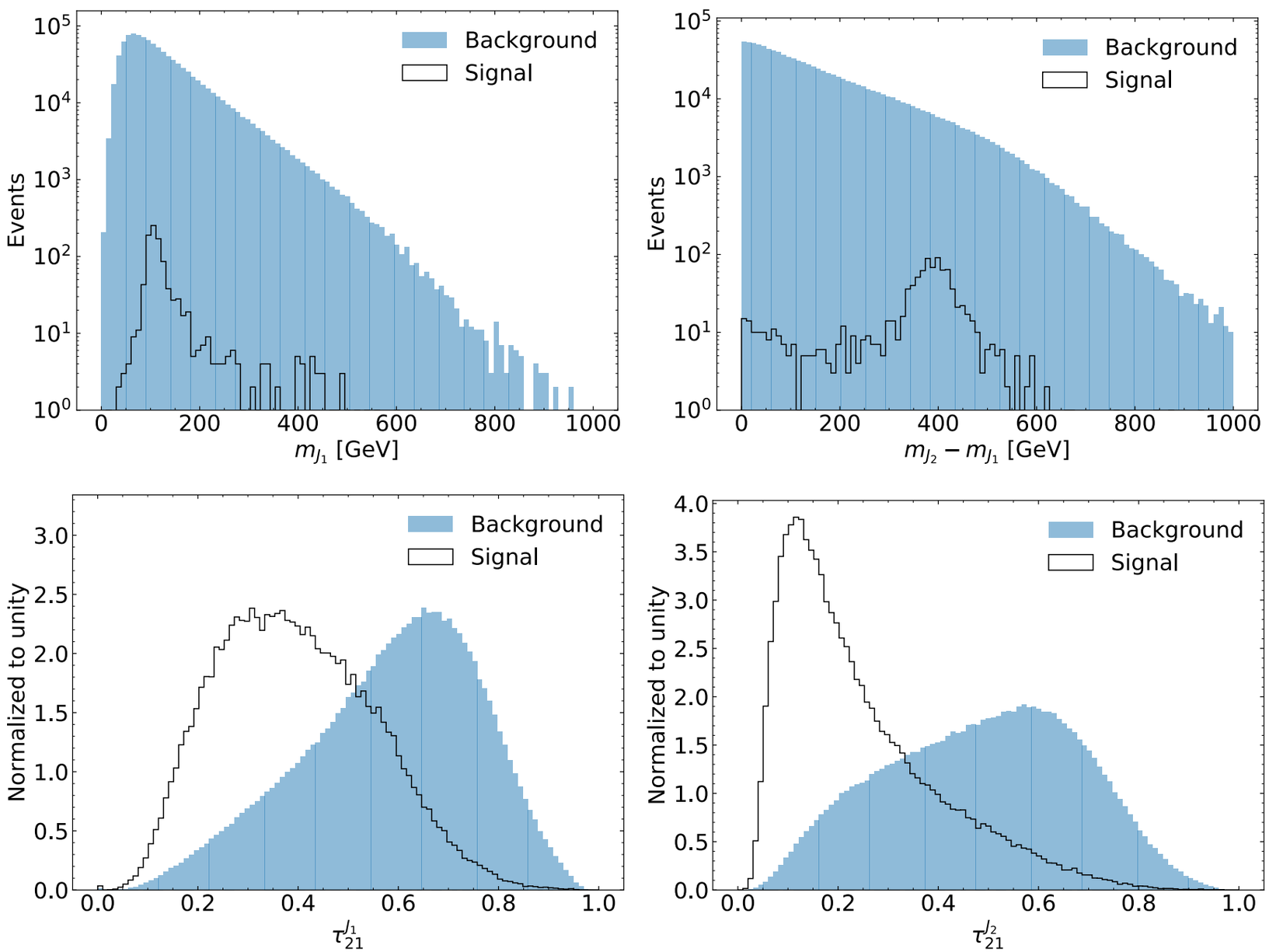
S=500, B=500,000, B_{SR}=61,000

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

Signal: $Z' \to XY$; $X, Y \to qq$; $m_{Z'}$=3.5 TeV, m_X=500 GeV, m_Y=100 GeV

Background: QCD dijets

Example: LHCO R&D Dataset



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

ANODE: ANOmaly detection with Density Estimation

Ben Nachman & DS 2001.04990

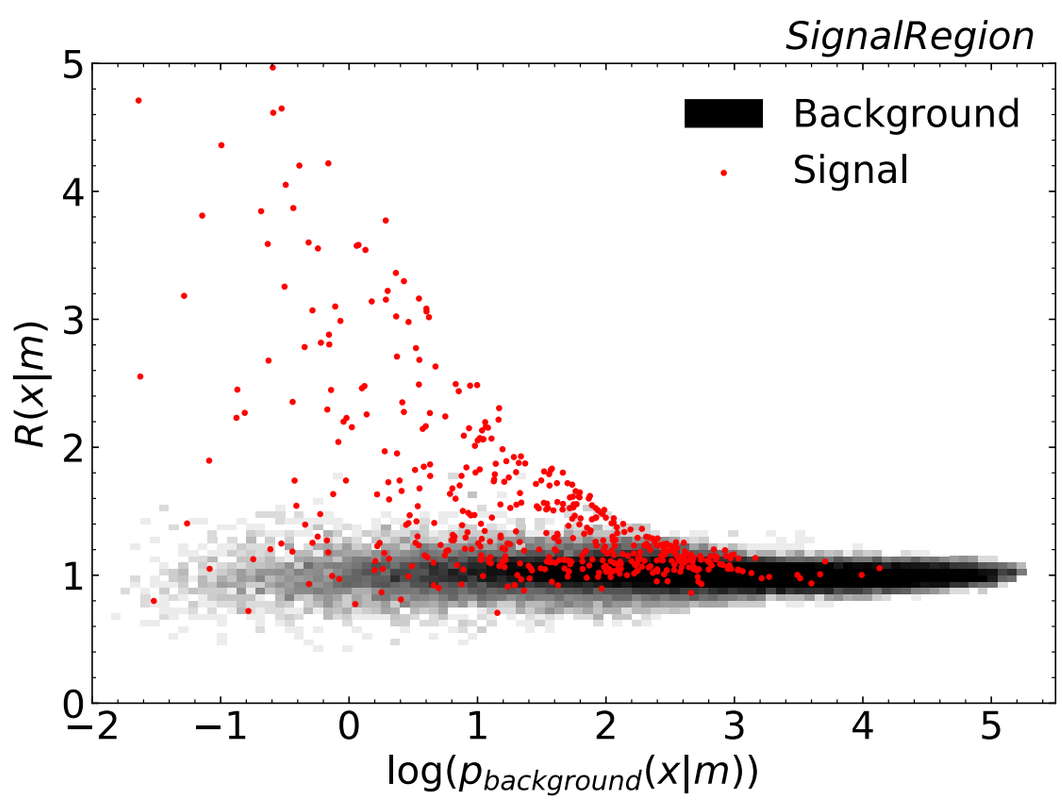
Idea: Leverage recent breakthroughs in high dimensional density estimation to find over-densities in the data that could be from new physics.

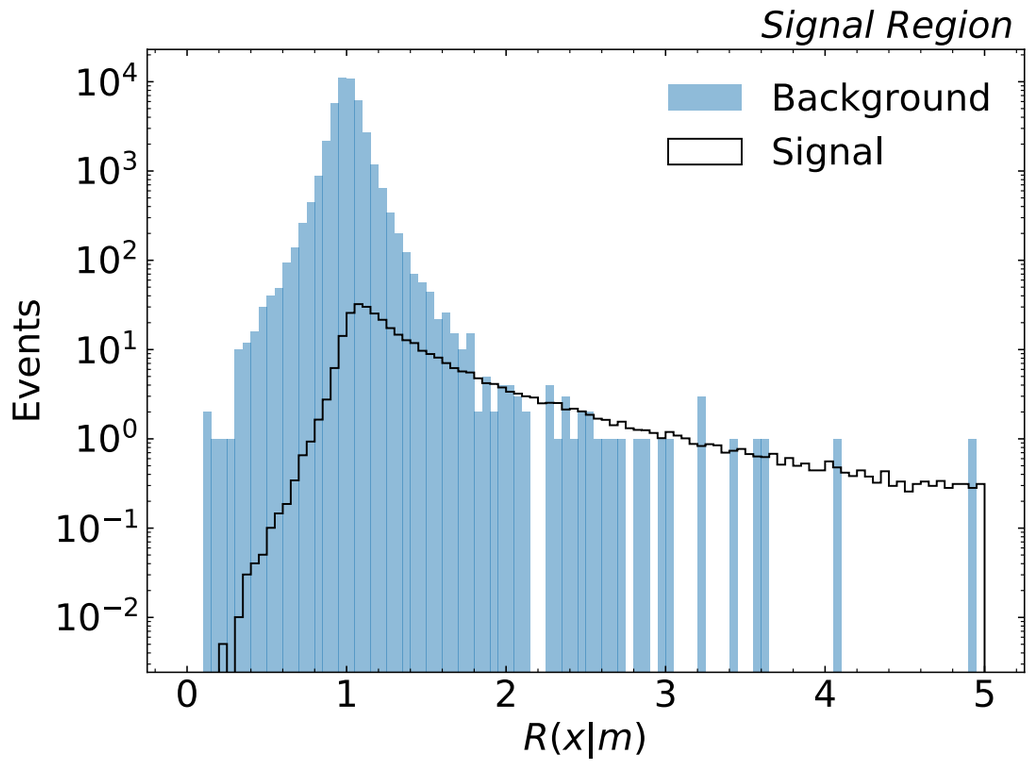
- estimate $P_{data}(x|m \in SR)$ with your favorite method
- estimate $P_{data}(x|m \notin SR)$ with your favorite method
- interpolate $P_{data}(x|m \notin SR)$ into SR to obtain $P_{bg}(x|m \in SR)$.
- Construct likelihood ratio $R(x)=P_{data}(x|m \in SR)/P_{bg}(x|m \in SR)$.

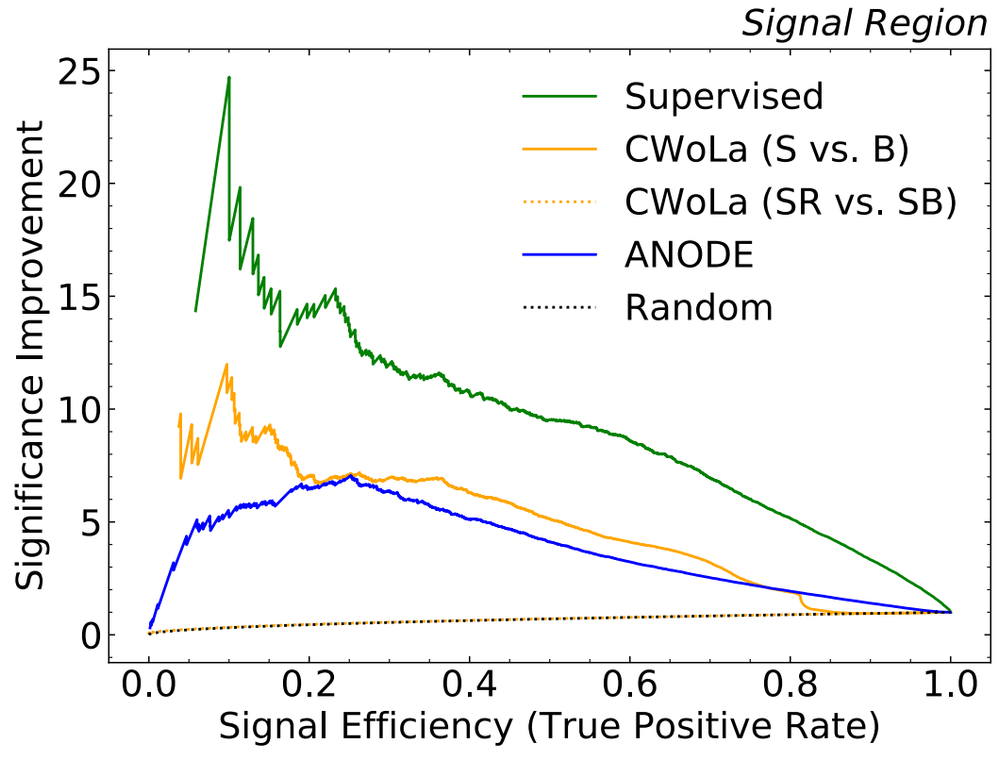
ANODE: Results

Ben Nachman & DS 2001.04990

uses conditional MAF (1705.07057)
for density estimation







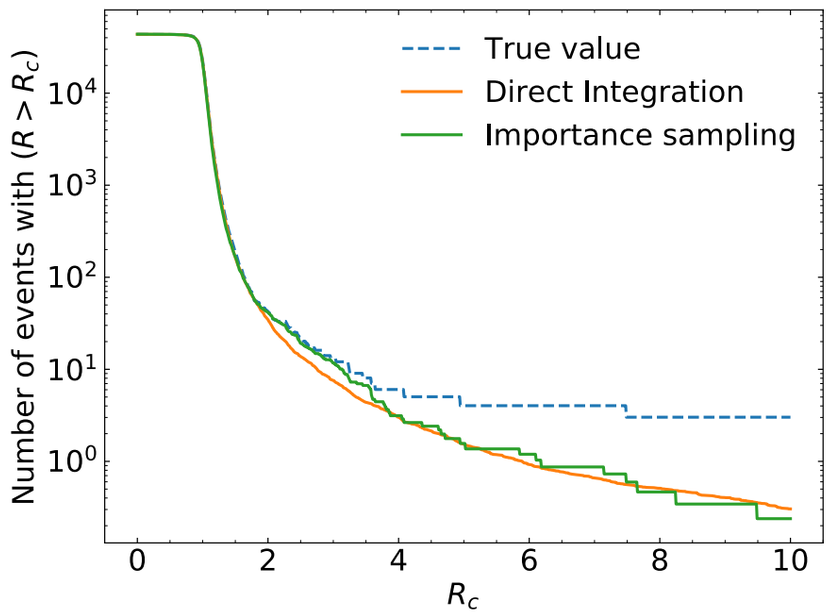
Can construct a very sensitive discriminant this way.

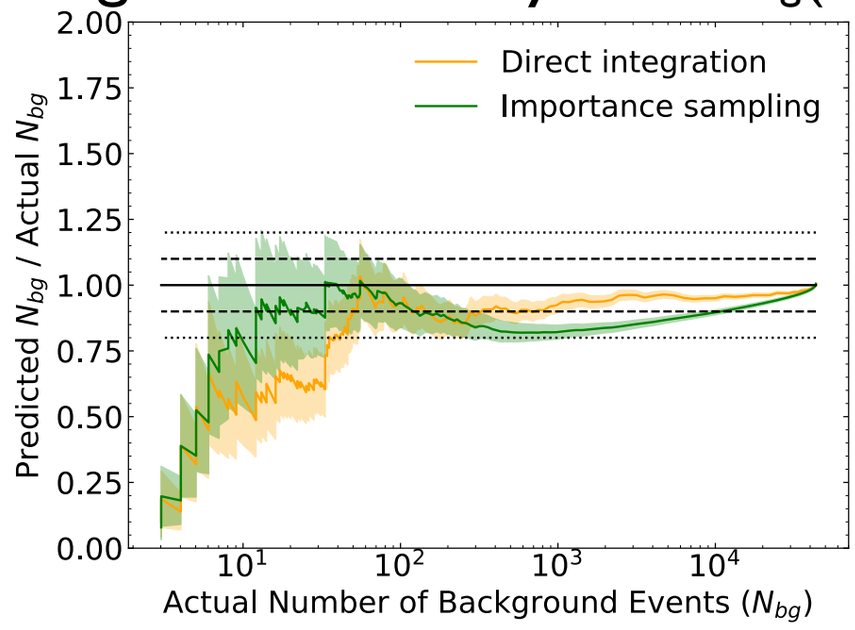
Can enhance the significance of the bump hunt by a factor of up to 7!

ANODE: Results

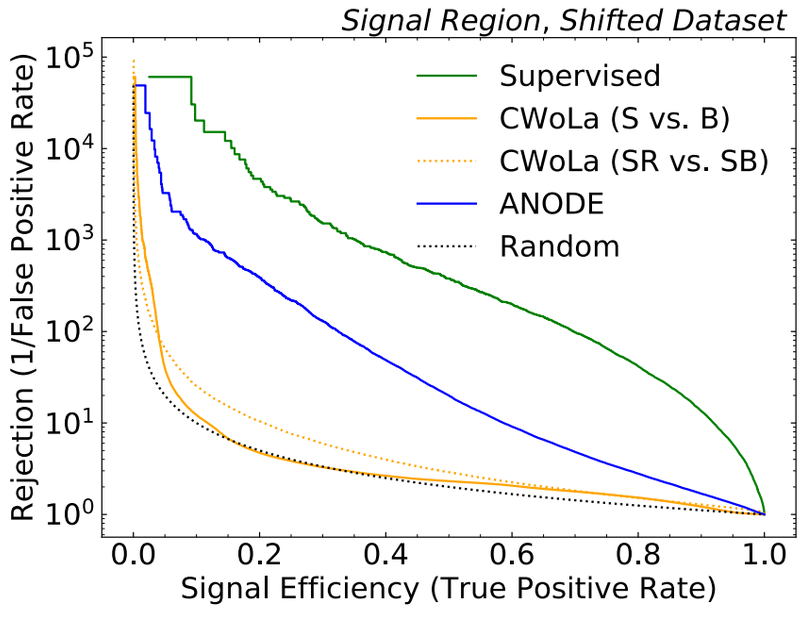
Ben Nachman & DS 2001.04990

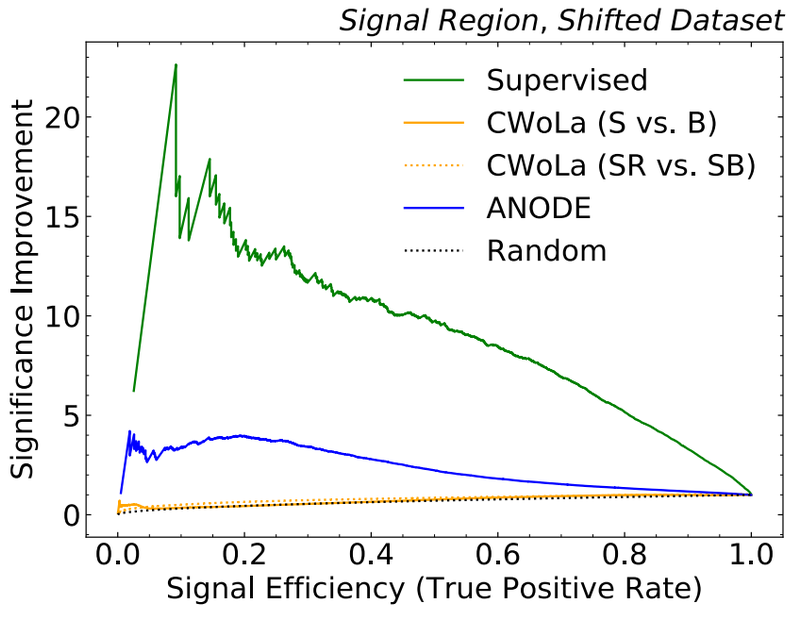
Can estimate backgrounds directly with $P_{bg}(x|m \in SR)$





Robust against correlations in features (eg $m_{J_{1,2}} \to m_{J_{1,2}} + c\, m_{JJ}$)





SALAD: Simulation Assisted Likelihood-free Anomaly Detection

Anders Andreassen, Ben Nachman & DS 2001.05001

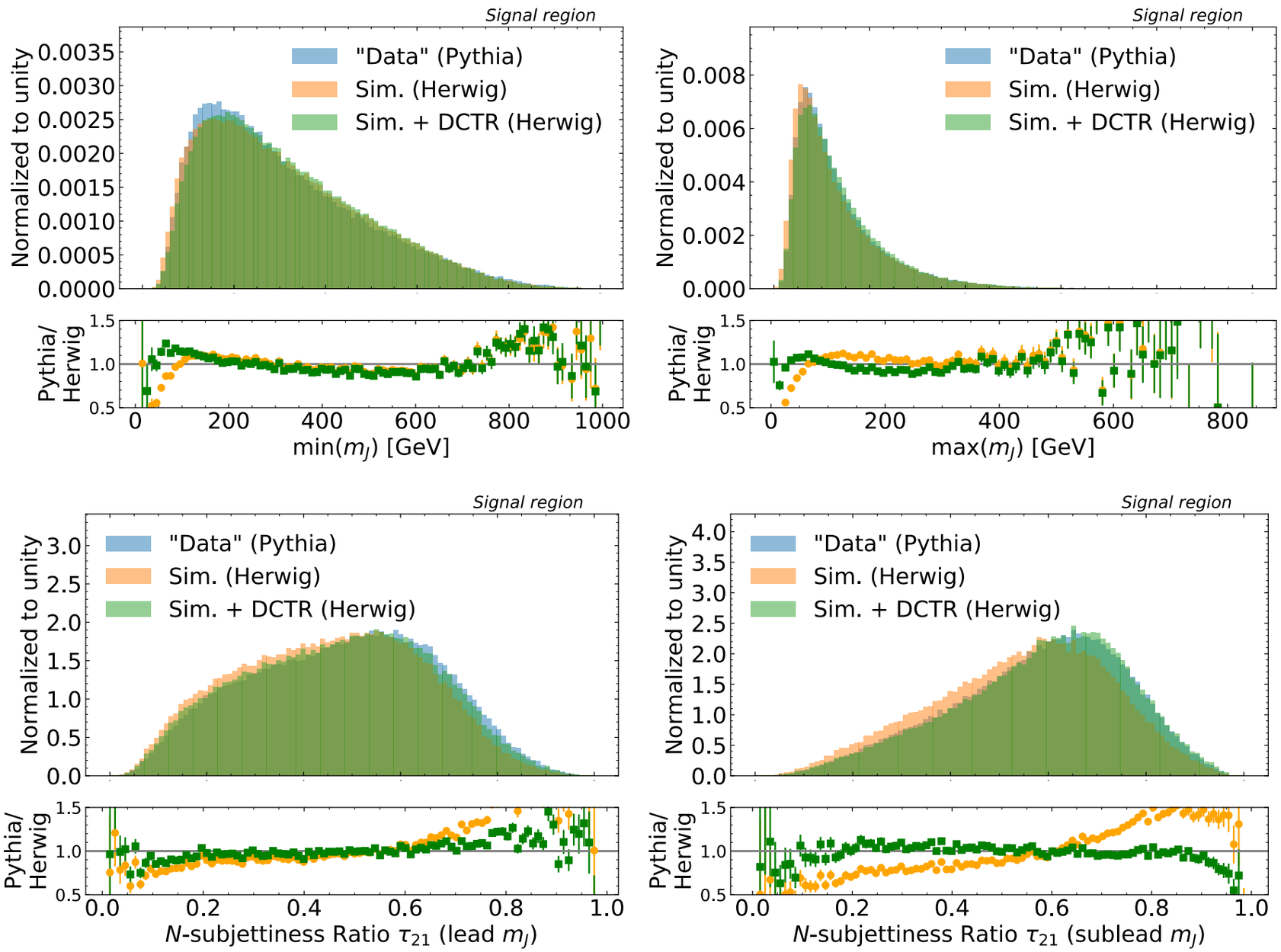
Idea: training data vs. raw simulation may not be sensitive to BSM, limited by quality of simulation. But it's a shame to completely ignore simulations which include a lot of nontrivial physics. What if we could reweight simulation to look like data?

- Use DCTR reweighting method (1907.08209) to reweight background simulation to data in sidebands
- Interpolate into SR
- Using reweighted simulation, generate a sample that follows $P_{bg}(x|m\in SR)$
- Train a classifier to distinguish data from this sample
- Obtain a discriminant that approaches $R(x)=P_{data}(x|m\in SR)/P_{bg}(x|m\in SR)$.

Data: LHCO R&D dataset (same S&B as before)
Simulation: Herwig QCD dijets

SALAD: Results

Anders Andreassen, Ben Nachman & DS 2001.05001

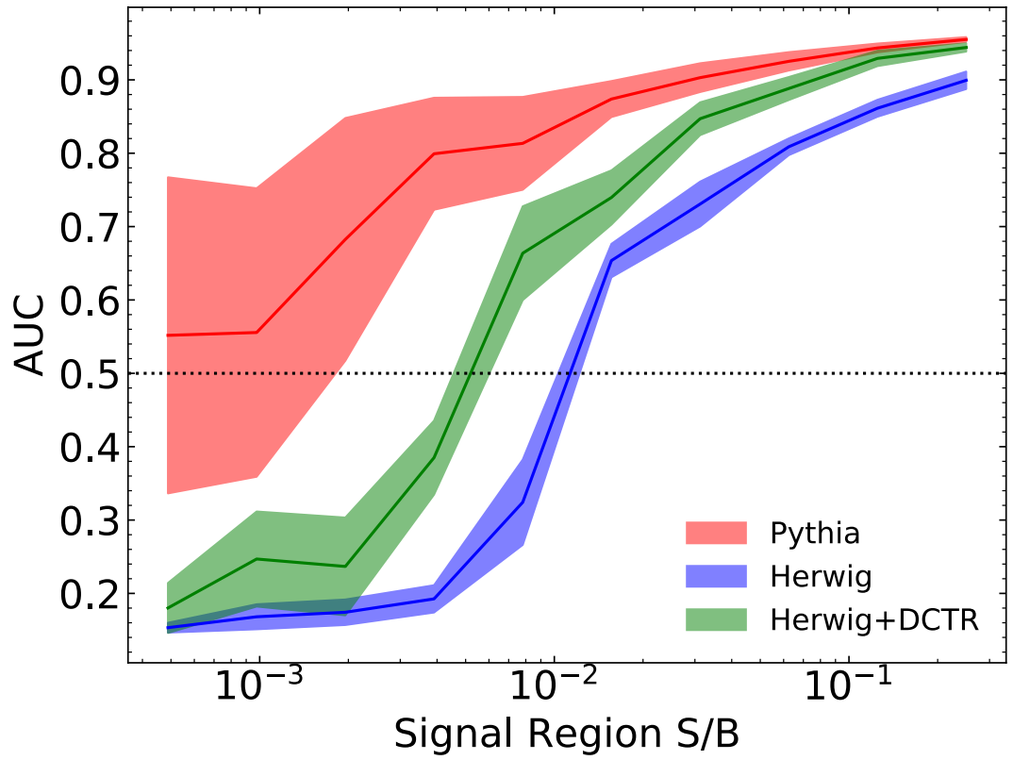


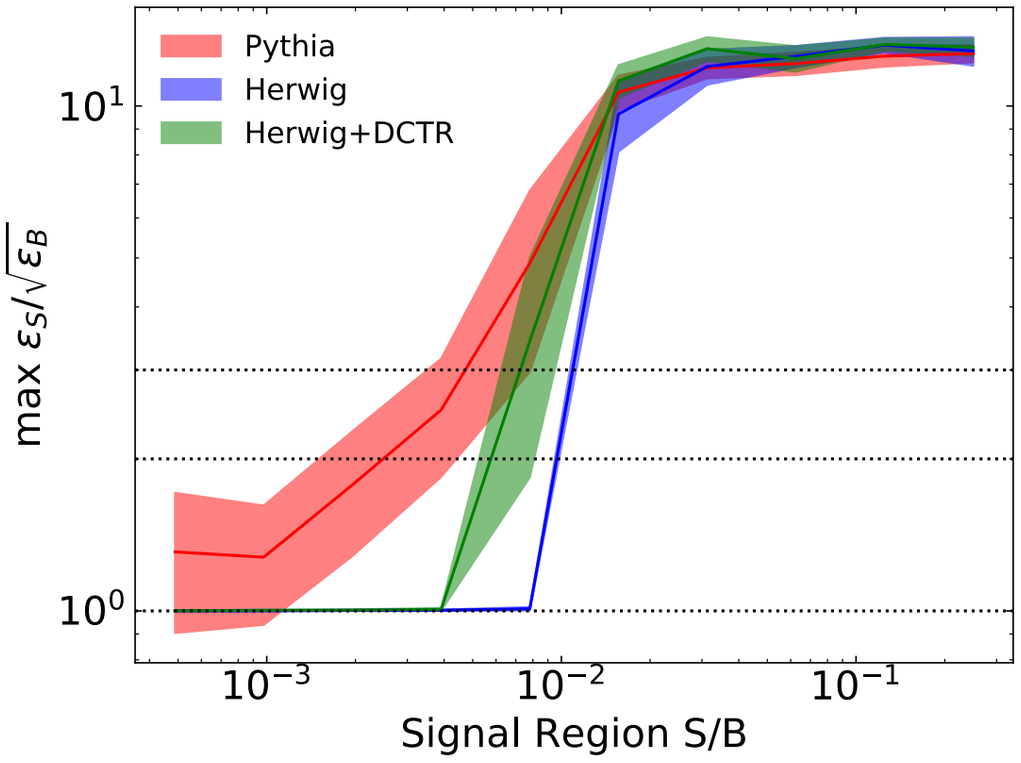
The reweighting+interpolation into the SR works well

Data: LHCO R&D dataset (same S&B as before)
Simulation: Herwig QCD dijets

SALAD: Results

Anders Andreassen, Ben Nachman & DS 2001.05001





Good sensitivity down to S/B ~ 1%

Overview of submissions

- 10 groups submitted results on box 1
- 4 of these groups also submitted results on boxes 2 & 3
- A number of additional groups could not finish the challenge in time but got results on the R&D dataset
- 7 of these groups giving talks in this session about their methods and results

Thanks to all the groups that participated!

Overview of submissions

People tried both supervised and unsupervised methods.

Methods used included

- Autoencoders
- CWoLa hunting
- PCA outlier detection
- LSTM
- CNN+BDT
- variational RNNs for anti-QCD tagging
- density estimation
- biological neural network
- …

Stay tuned for an exciting session!

The results of the LHCO2020 will be discussed in the final talk.