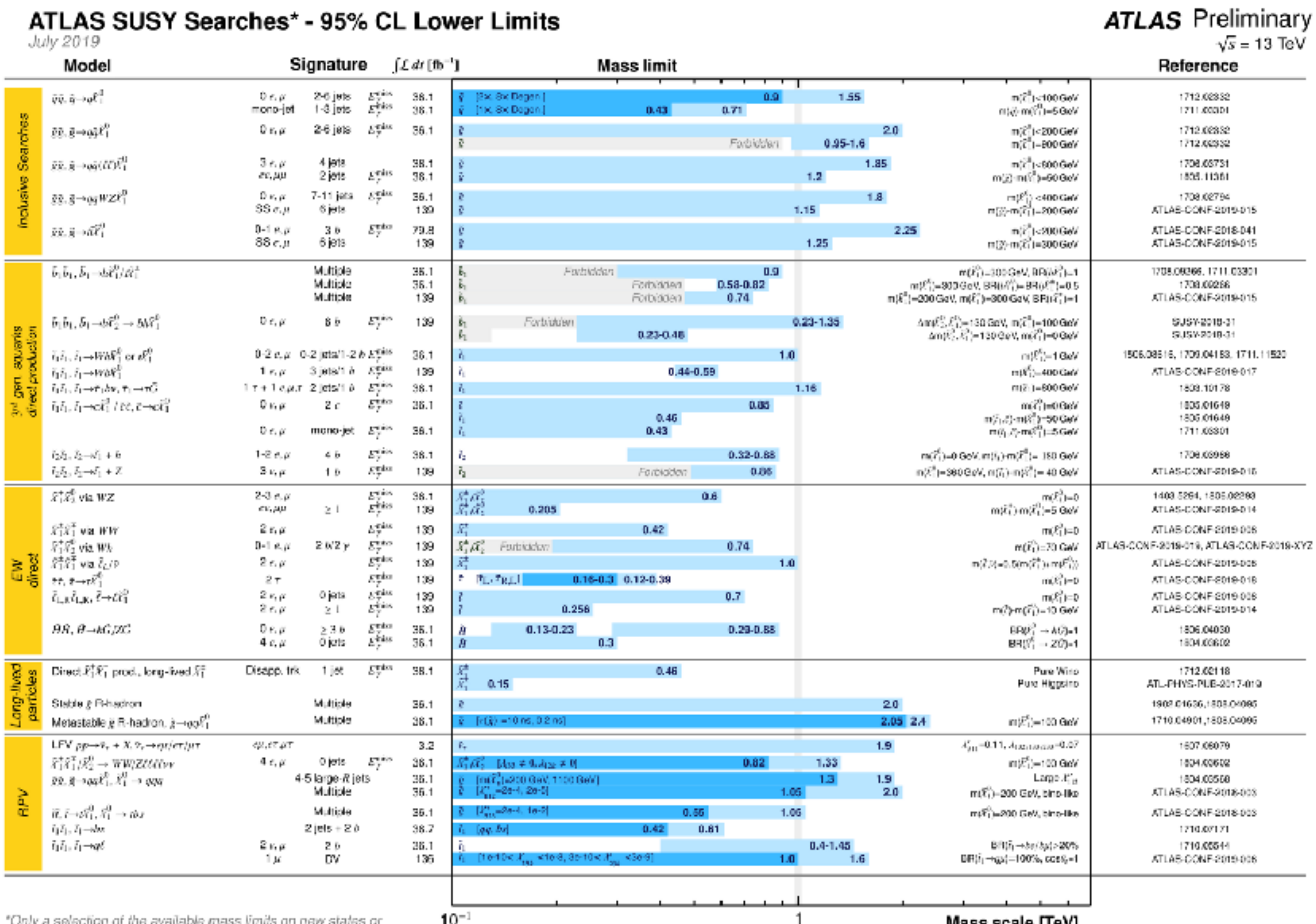
Searching for the Unexpected with Unsupervised Machine Learning

David Shih

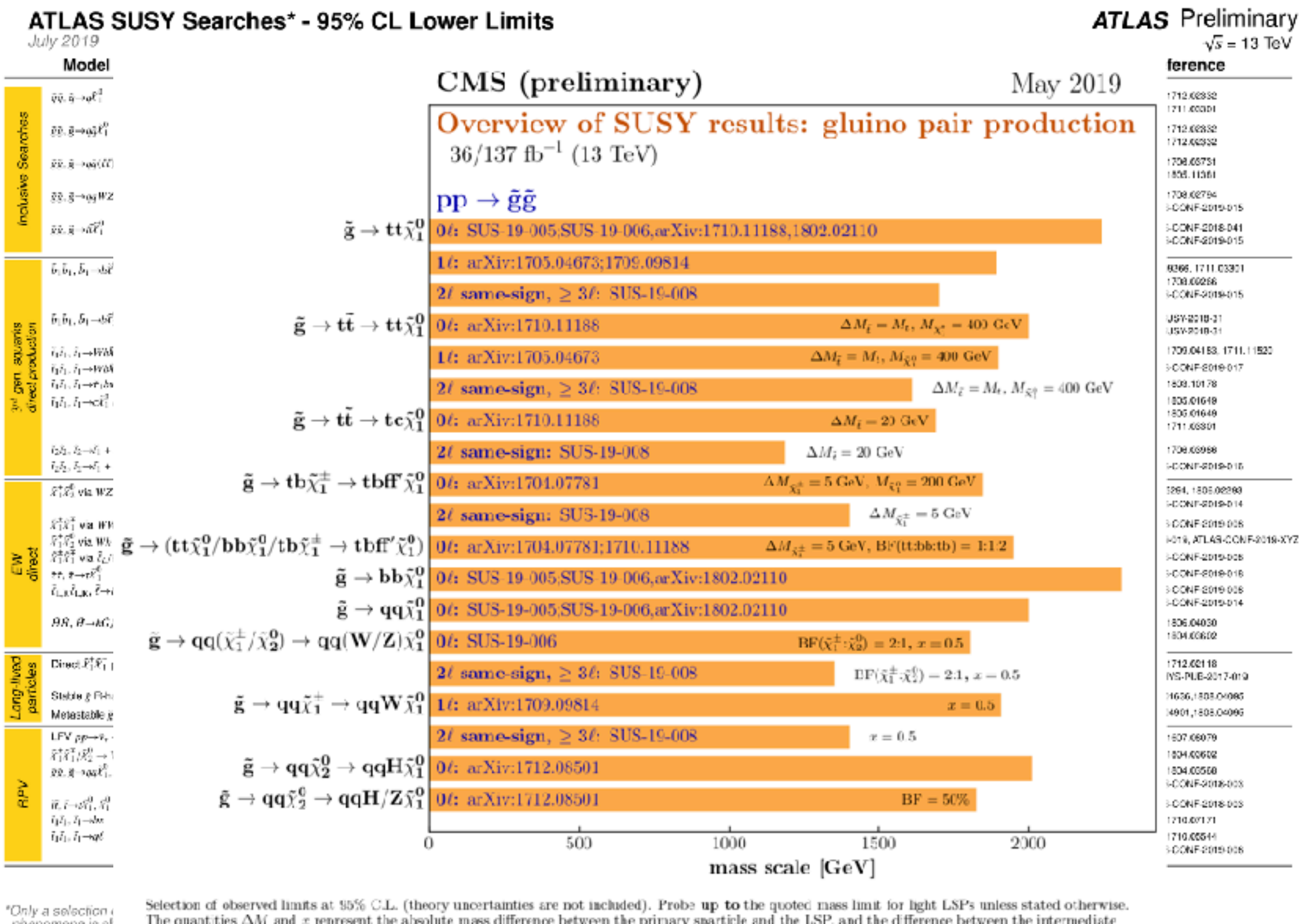
Bay Area Particle Theory Seminar

October 25, 2019

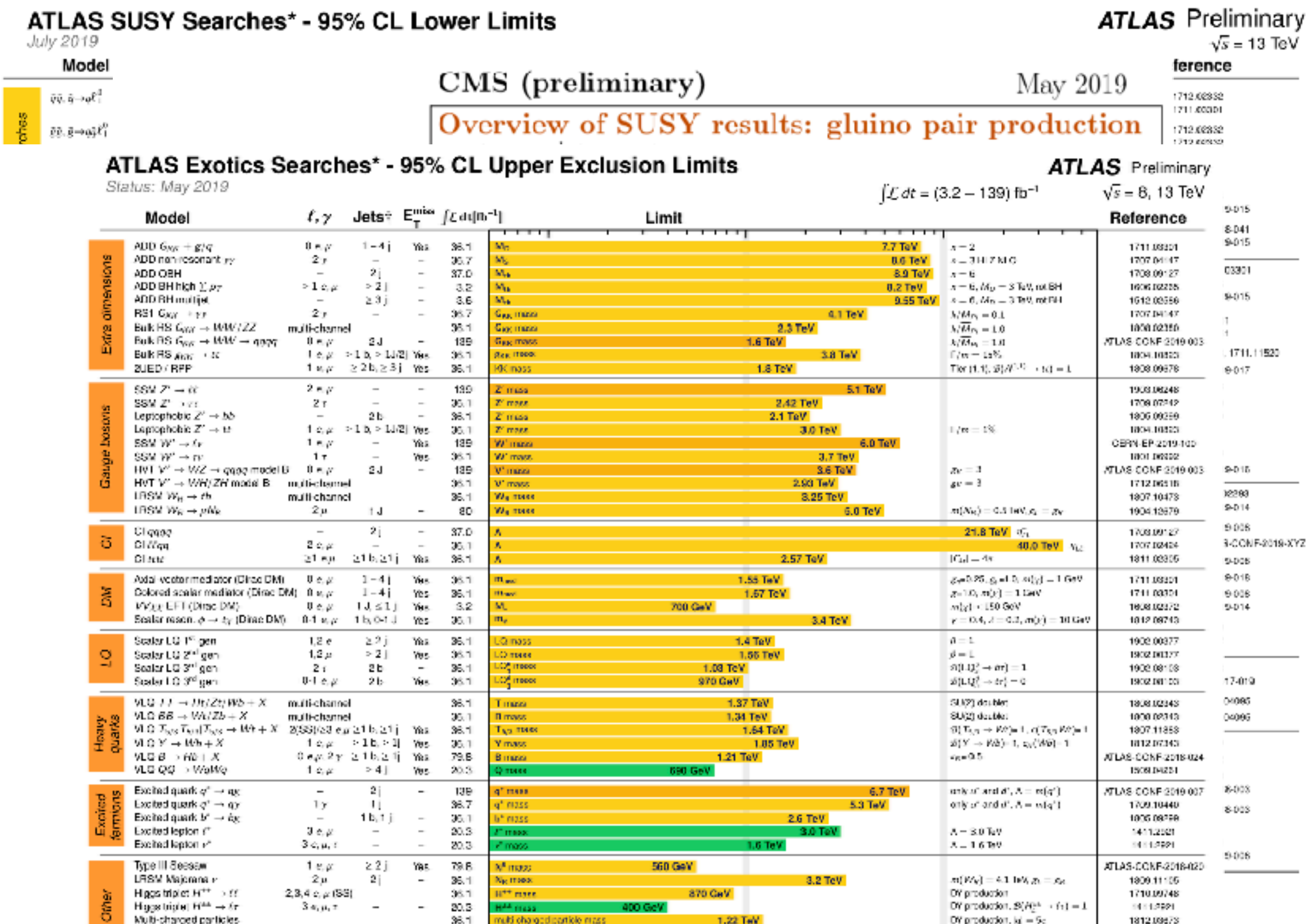
Current Status of NP Searches @ LHC



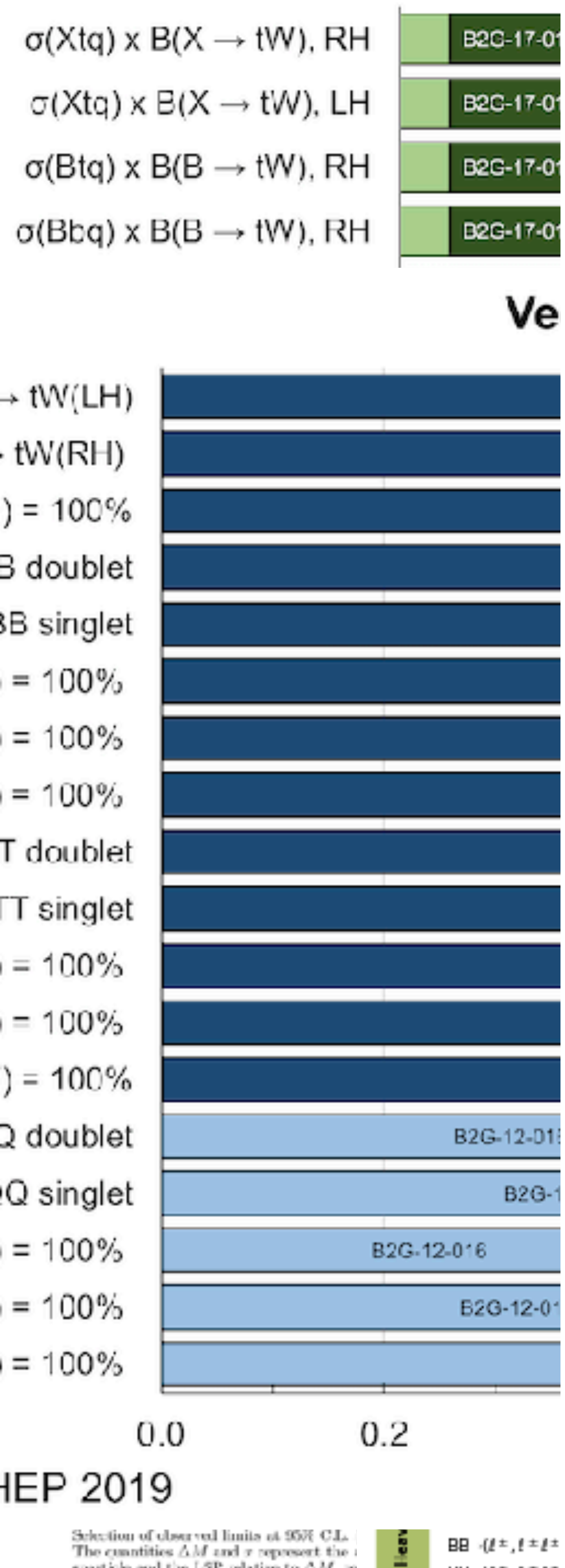
Current Status of NP Searches @ LHC

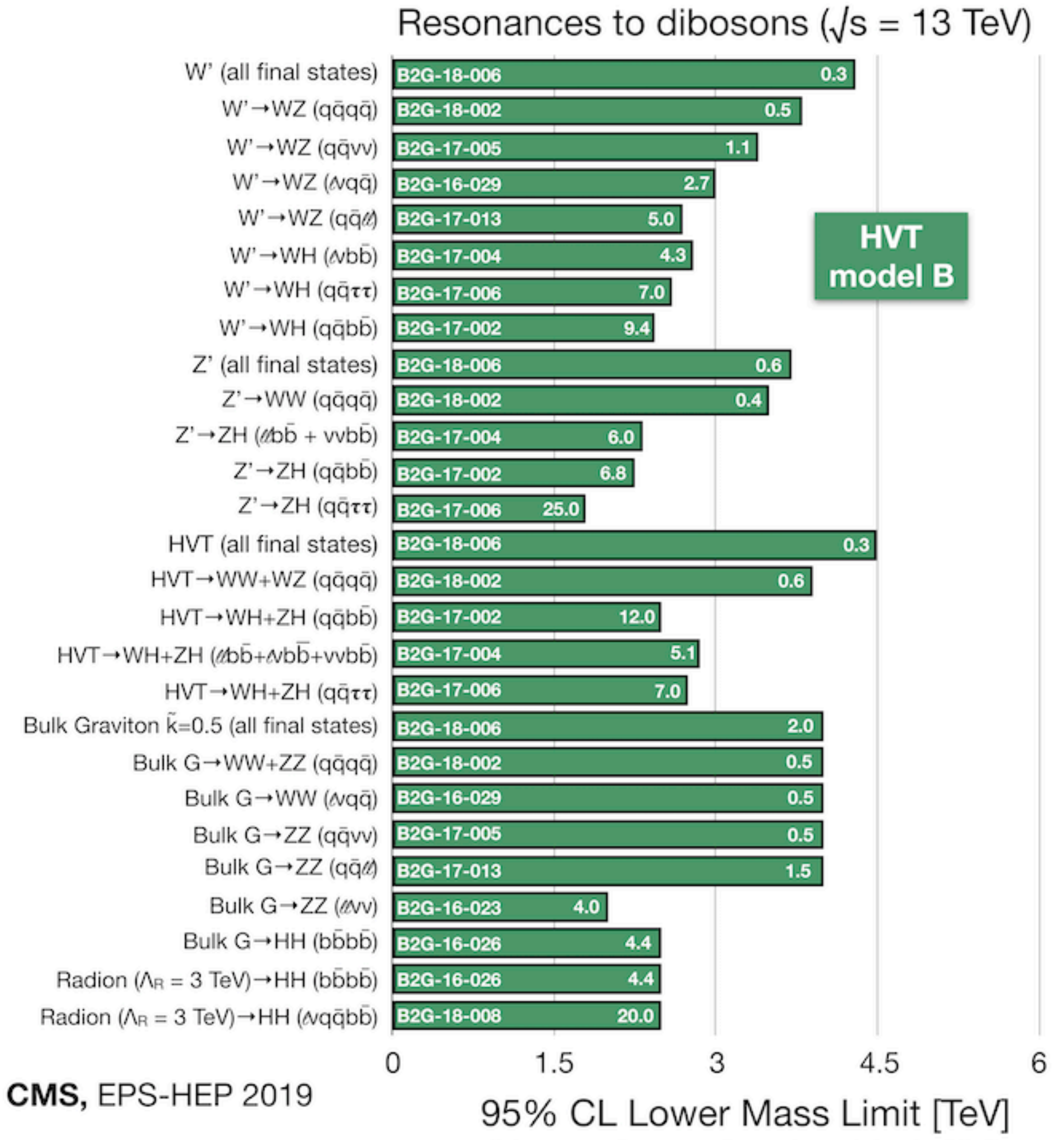


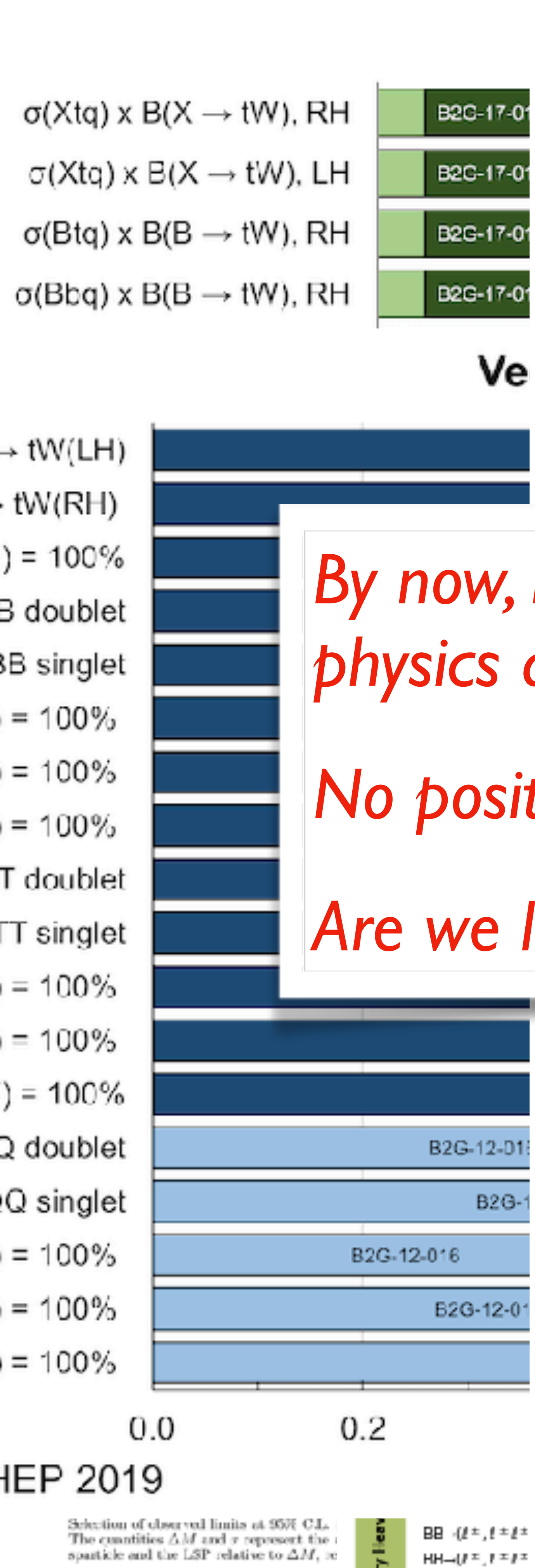
Current Status of NP Searches @ LHC

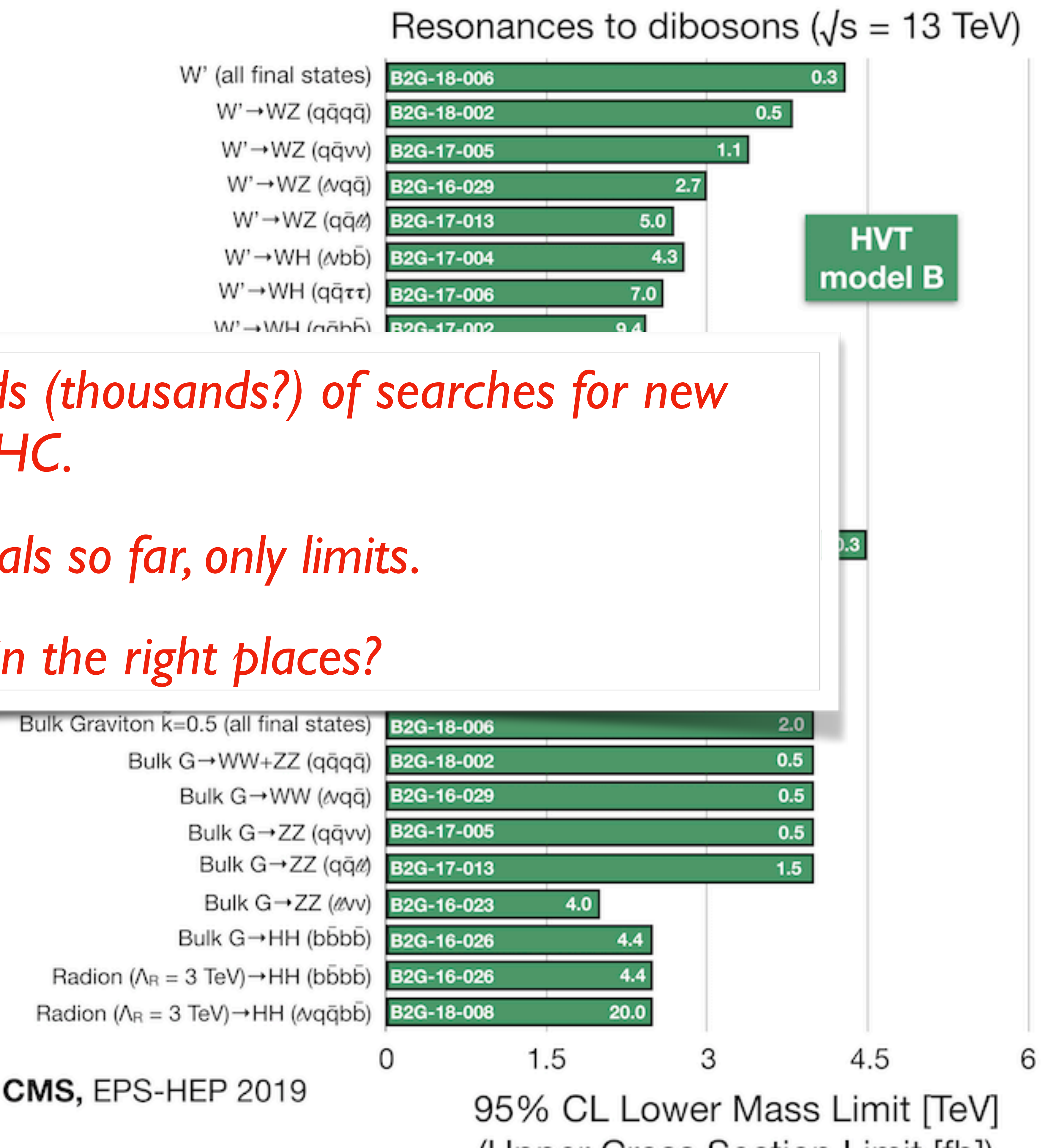


Vector-like quark single production









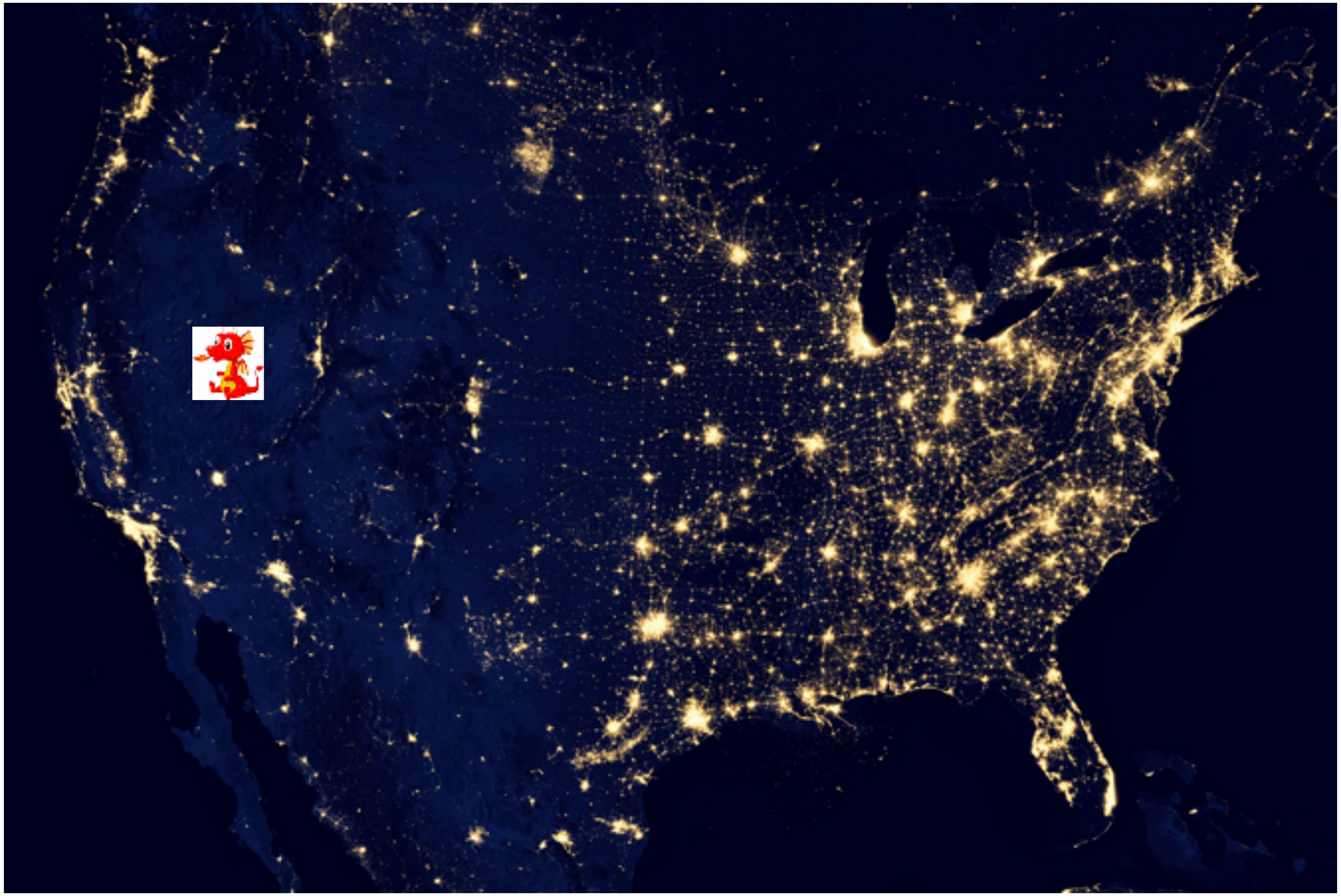
By now, hundreds (thousands?) of searches for new physics at the LHC.

No positive signals so far, only limits.

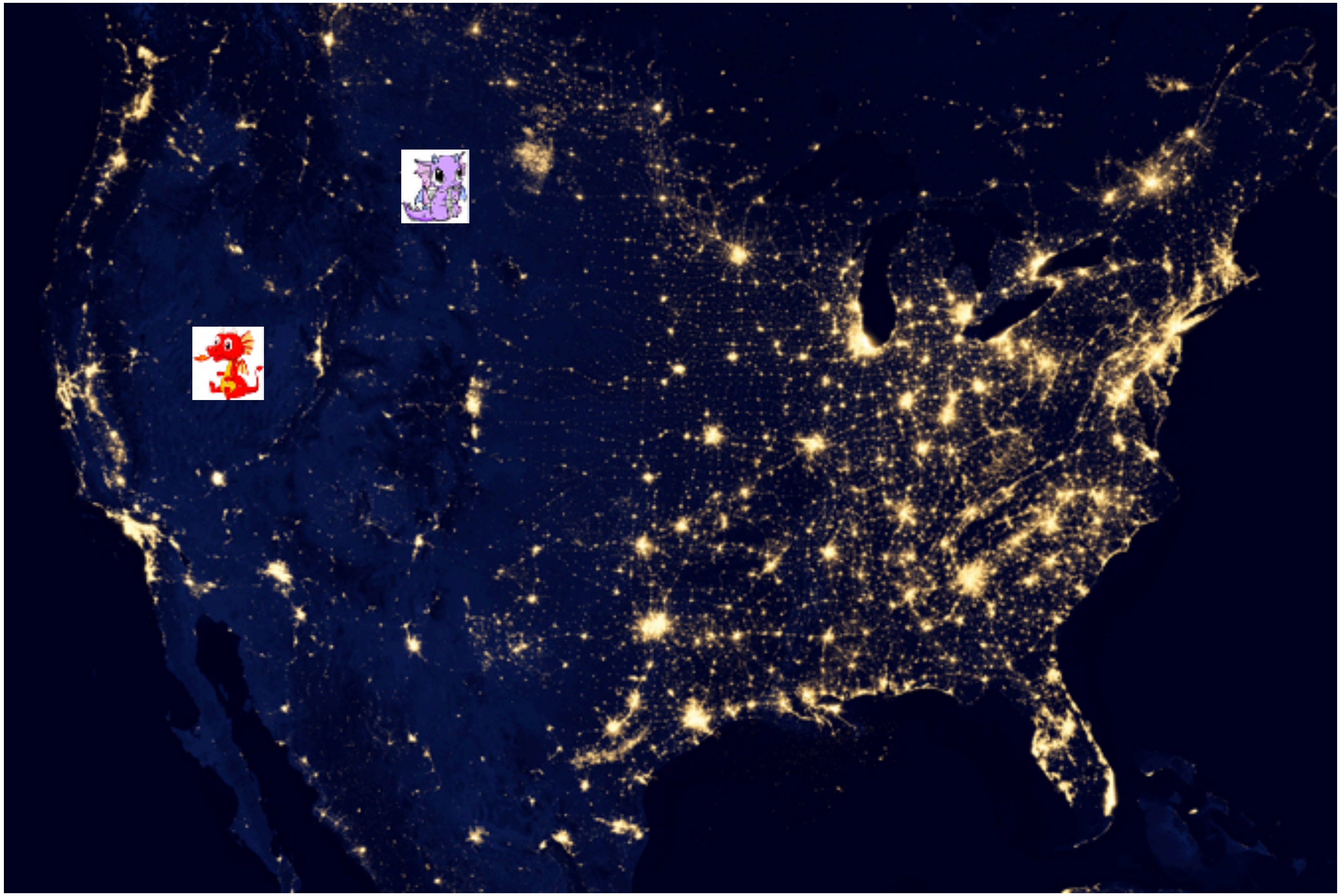
Are we looking in the right places?

Our current coverage

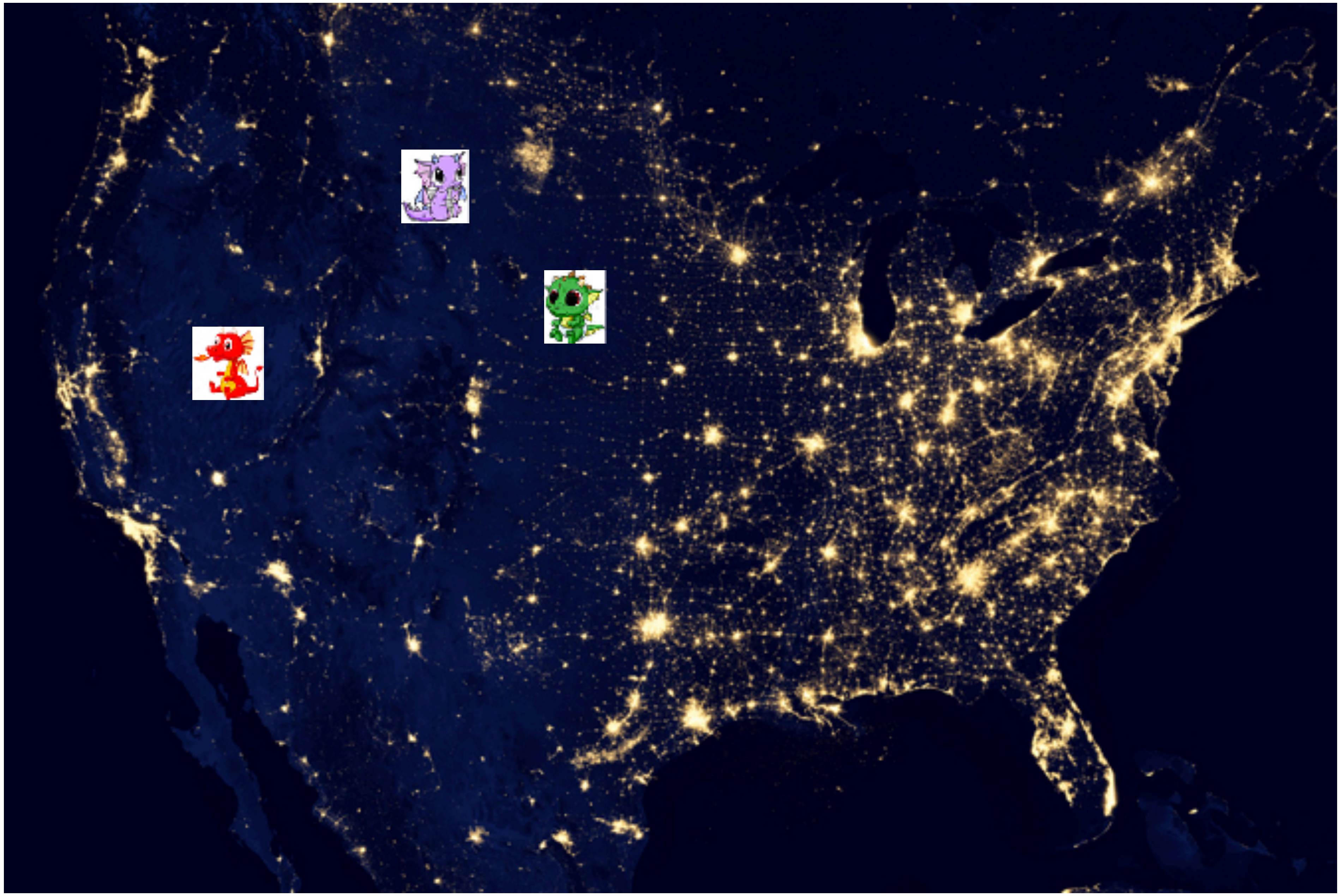
Our current coverage



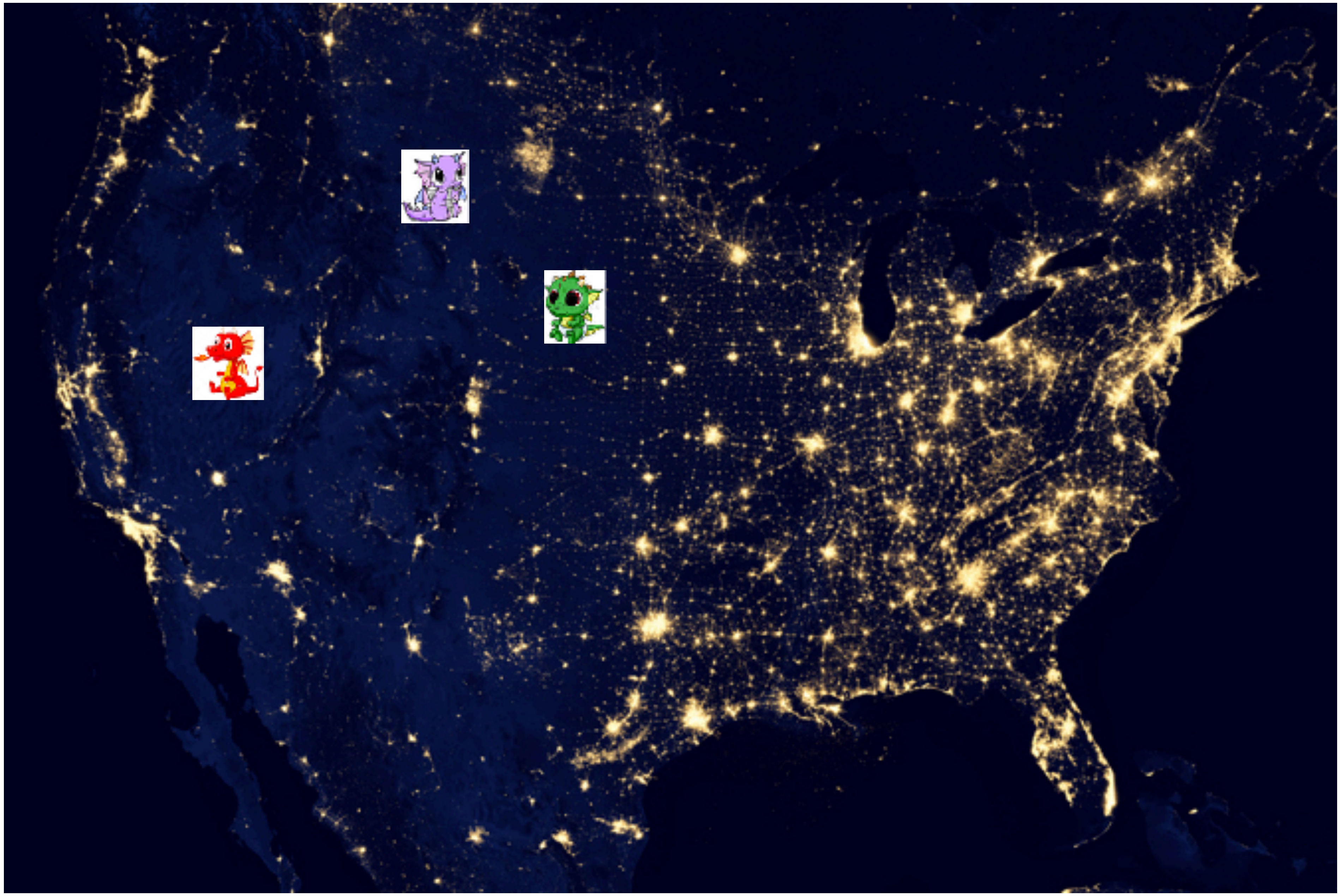
Our current coverage



Our current coverage



Our current coverage



How can we make sure we don't miss anything?

Model Independent Searches

Can we search for new physics without a model in mind?

Model Independent Searches

<u>Can we search for new physics without a model in mind?</u>

Yes?? 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)

Model independent searches

The general approach behind all of these:

- Bin the data into exclusive final states,
- Compare a zillion 1D histograms between data and ***simulation*** (or just look at high pT tails),
- Look for discrepancies.

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

Major drawbacks

The general approach behind all of these:

- Bin the data into exclusive final states,
- Compare a zillion 1D histograms between data and ***simulation*** (or just look at high pT tails),
- Look for discrepancies.

Major drawbacks

Enormous look elsewhere effect

The general approach behind all of these:

- Bin the data into exclusive final states,
- Compare a zillion 1D histograms between data and ***simulation*** (or just look at high pT tails),
- Look for discrepancies.

Major drawbacks

Enormous look elsewhere effect

The general approach behind all of these:

- Bin the data into exclusive final states,
- Compare a zillion 1D histograms between data and ***simulation*** (or just look at high pT tails),
- Look for discrepancies.

Sub-optimal signal / background discrimination

Major drawbacks

Enormous look elsewhere effect

The general approach behind all of these:

- Bin the data into exclusive final states,
- Compare a zillion 1D histograms between data and ***simulation*** (or just look at high pT tails),
- Look for discrepancies.

Sub-optimal signal / background discrimination

Over-reliance on simulation for background prediction

An example of what is found

CDF Run II (927 pb^{-1})

Sleuth@CDFII
result

(top 5)

fraction of pseudo experiments in this final state as interesting as CDF data

SLEUTH Final State	$\mathcal{P}$
$b\bar{b}$	0.0055
$j\,\not{p}$	0.0092
$\ell^+\ell'^+\,\not{p}jj$	0.011
$\ell^+\ell'^+\,\not{p}$	0.016
$\tau\,\not{p}$	0.016

fraction of pseudo experiments in any final state as interesting as CDF data

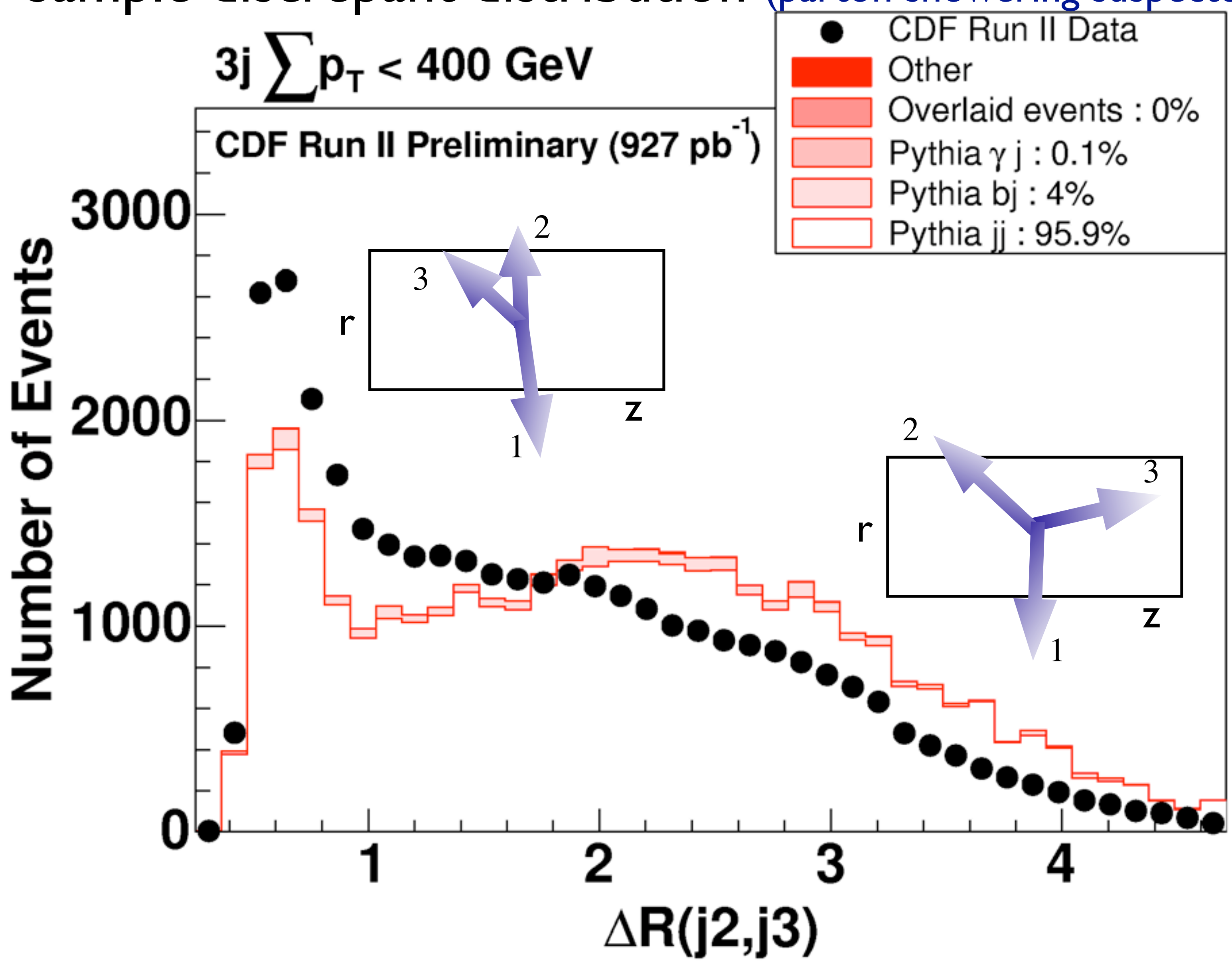
$$\tilde{\mathcal{P}} = 0.46$$

- 46% of pseudo experiments are expected to be as interesting
- Sleuth finds no significant excess in CDF Run II high-p_T data
- This does not prove there is no new physics present

From B. Knuteson talk at UMich (2008)

An example of what is found

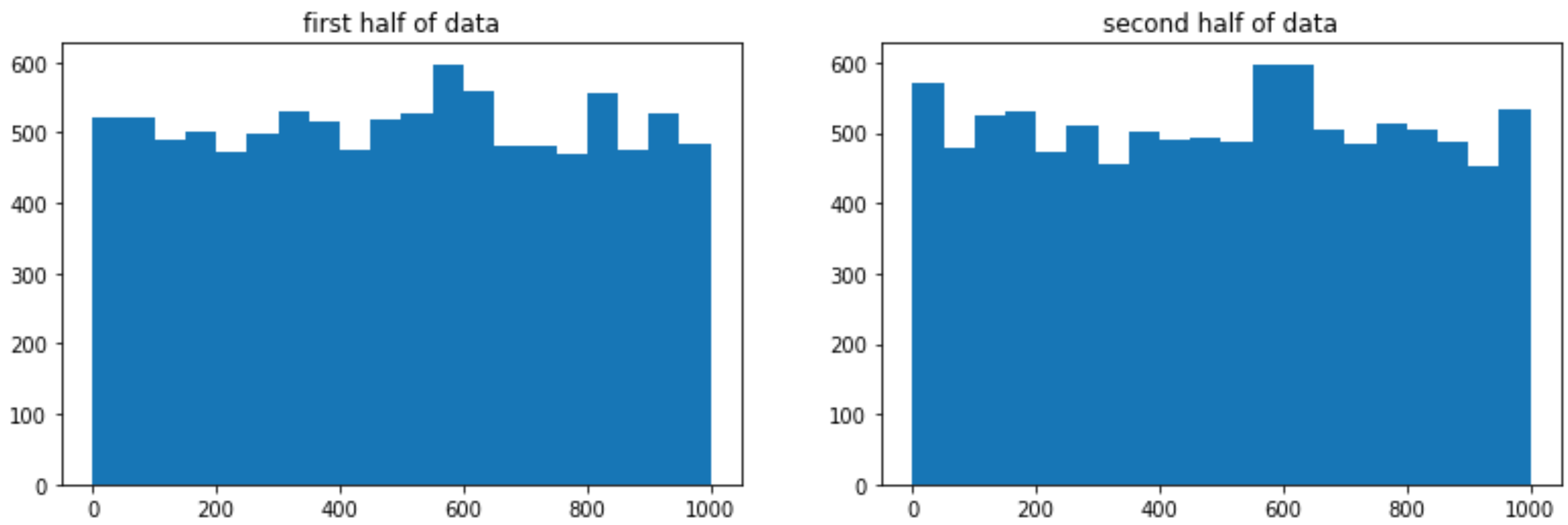
Sample discrepant distribution (parton showering suspected)



From B. Knuteson talk at UMich (2008)

Look elsewhere effect

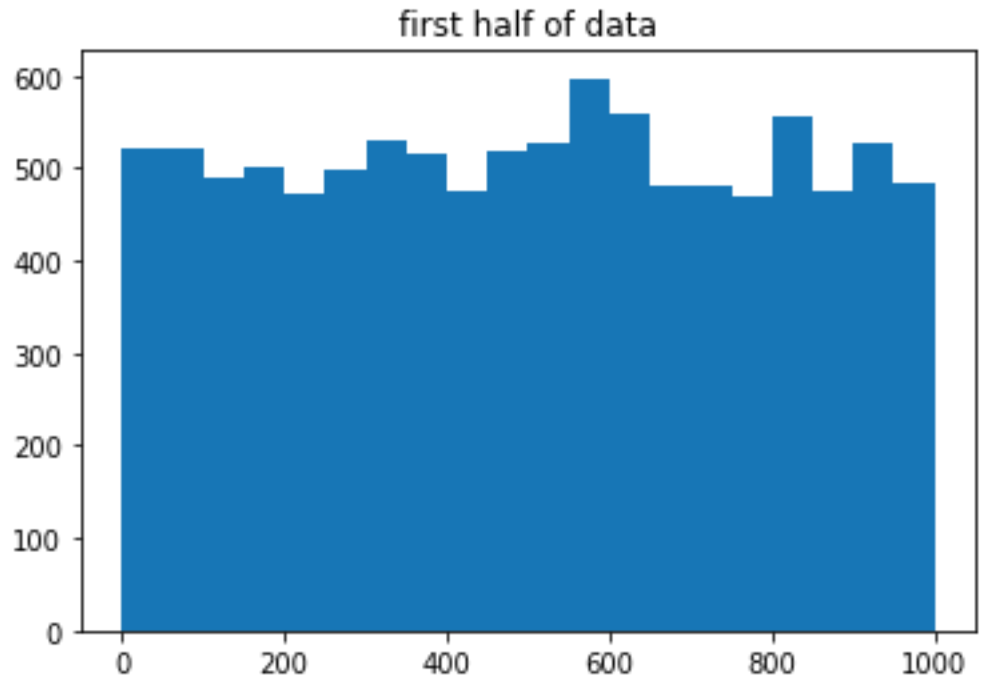
Look elsewhere effect

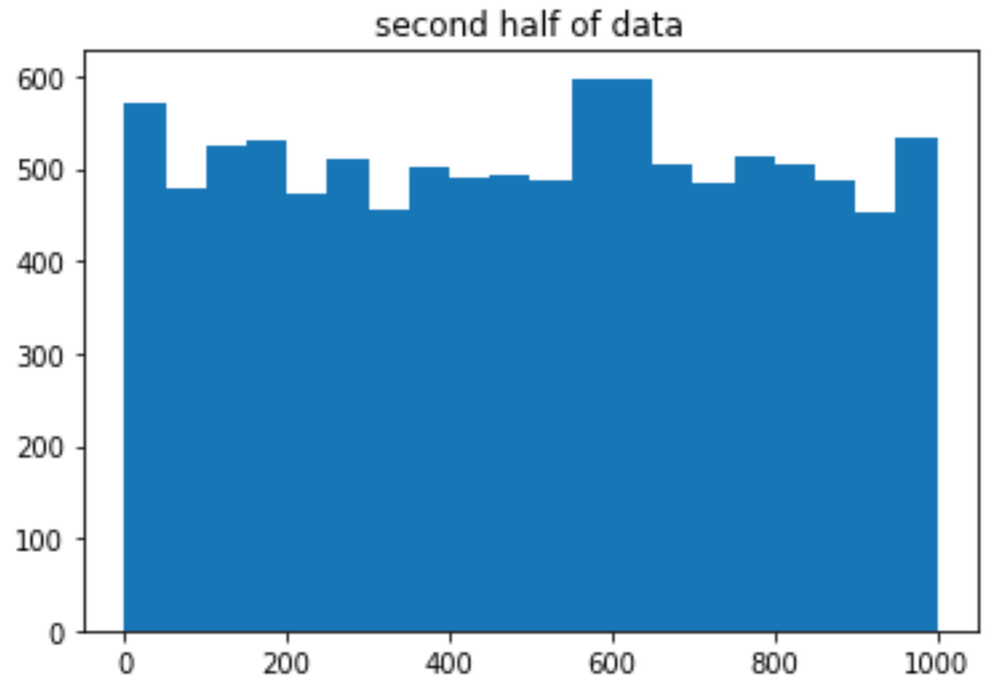


Simple way to mitigate LEE:

- Divide data in two

Look elsewhere effect

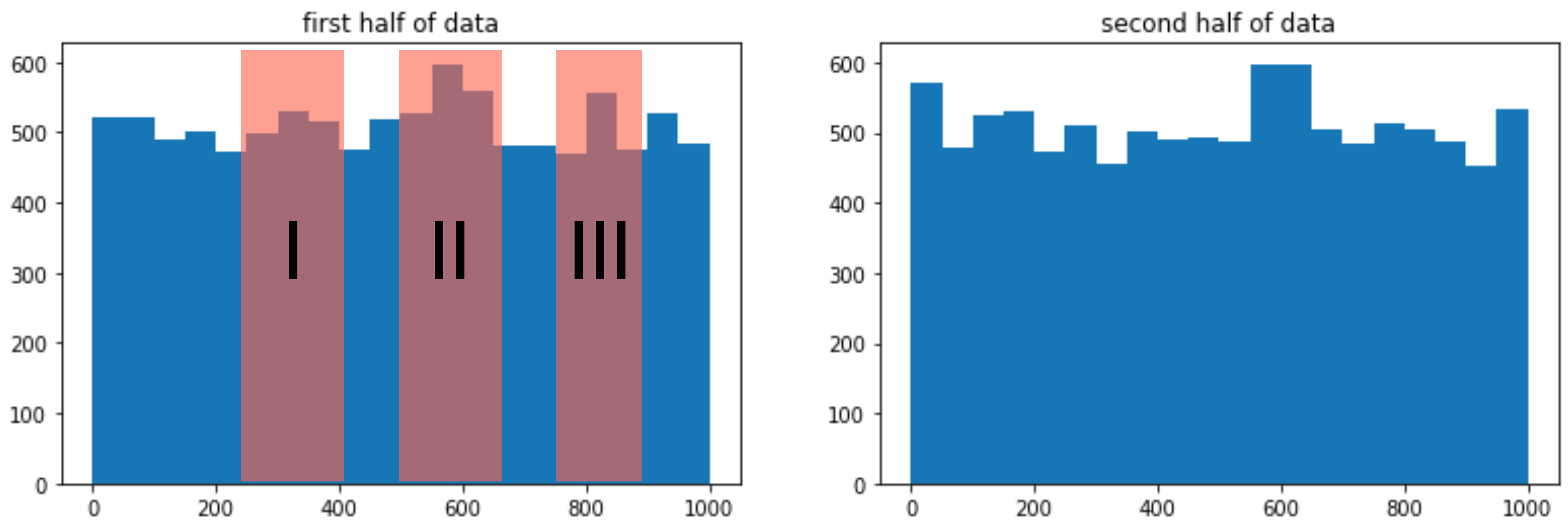




Simple way to mitigate LEE:

- Divide data in two
- Look for discrepancies in first half

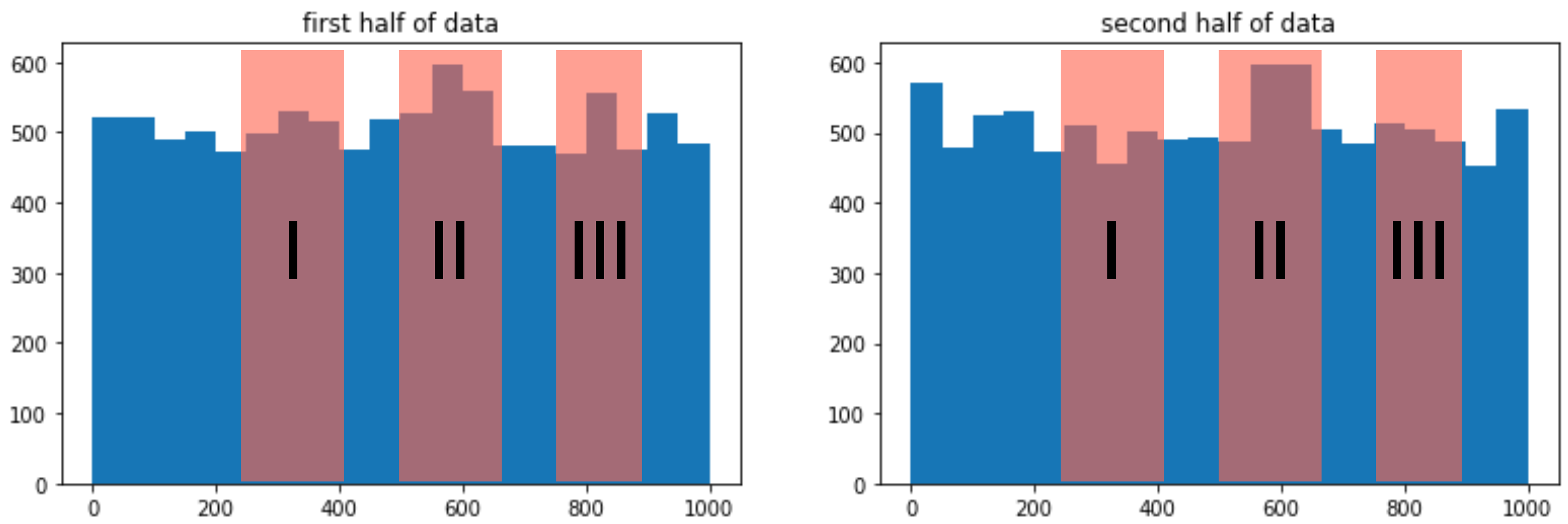
Look elsewhere effect



Simple way to mitigate LEE:

- Divide data in two
- Look for discrepancies in first half
- Fix regions of interest around these discrepancies and restrict search to these in the second half

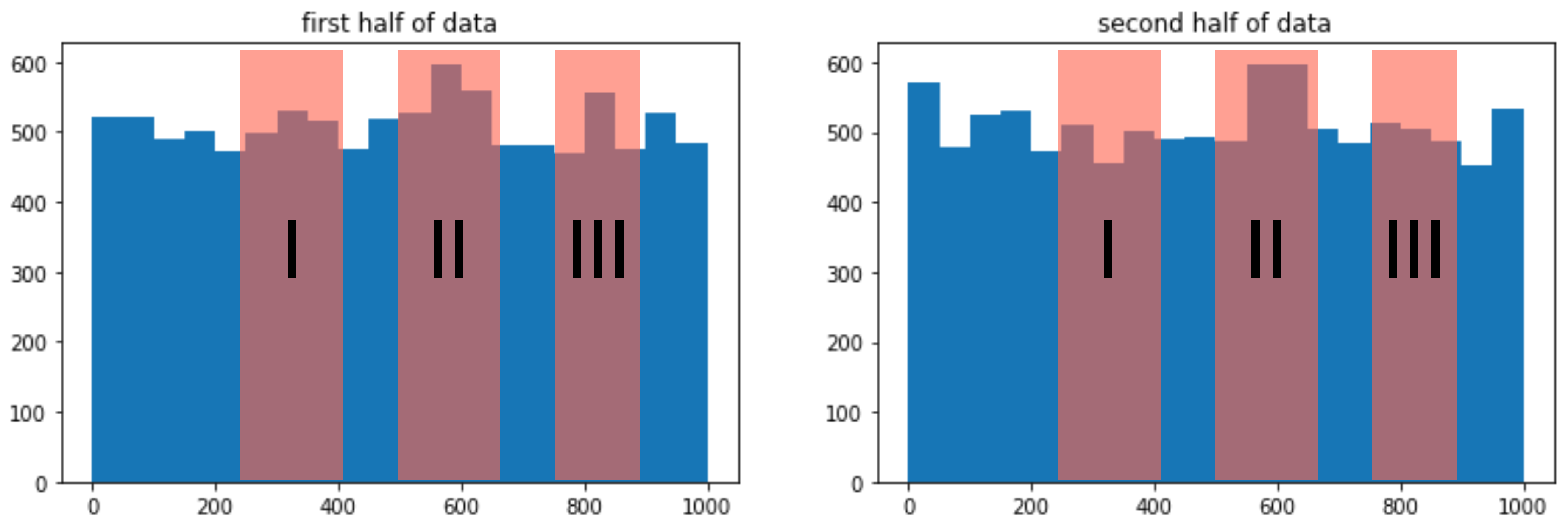
Look elsewhere effect



Simple way to mitigate LEE:

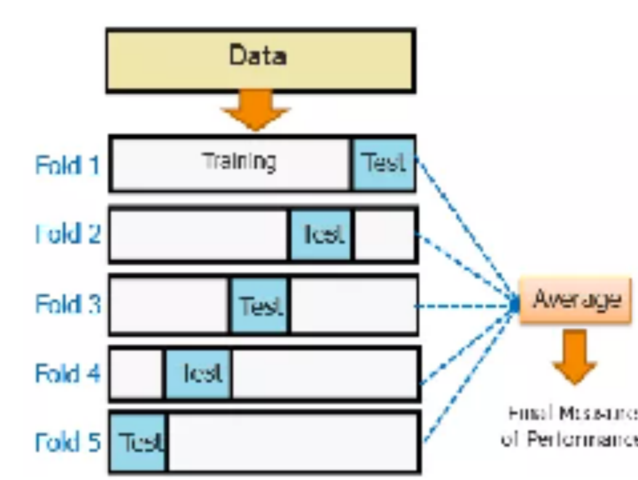
- Divide data in two
- Look for discrepancies in first half
- Fix regions of interest around these discrepancies and restrict search to these in the second half

Look elsewhere effect



Simple way to mitigate LEE:

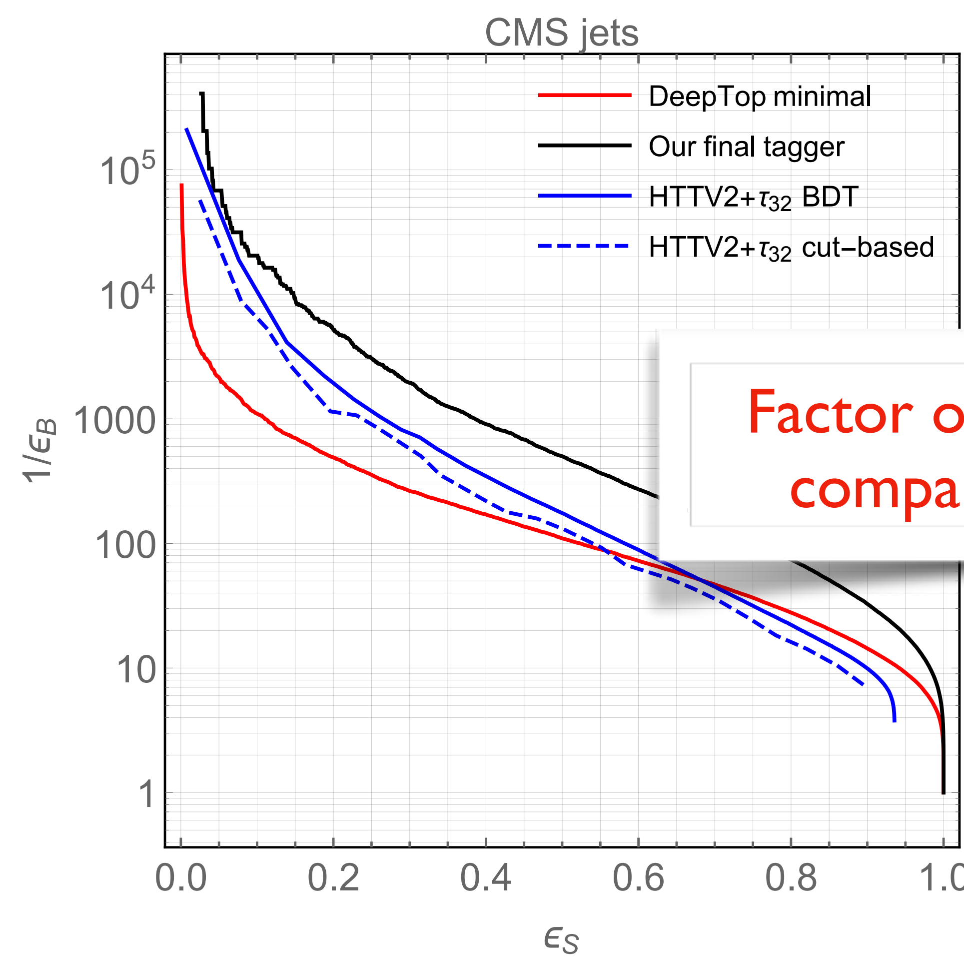
- Divide data in two
- Look for discrepancies in first half
- Fix regions of interest around these discrepancies and restrict search to these in the second half
- An even better approach: k-fold cross validation, see e.g. 1805.02664



Sub-optimal signal/background discrimination

Can do much better than comparing 1D histograms.

Using **deep learning**, can leverage the entire high-dimensional phase space to separate signals from background!



Factor of 3-5 improved performance compared to cut-based methods!

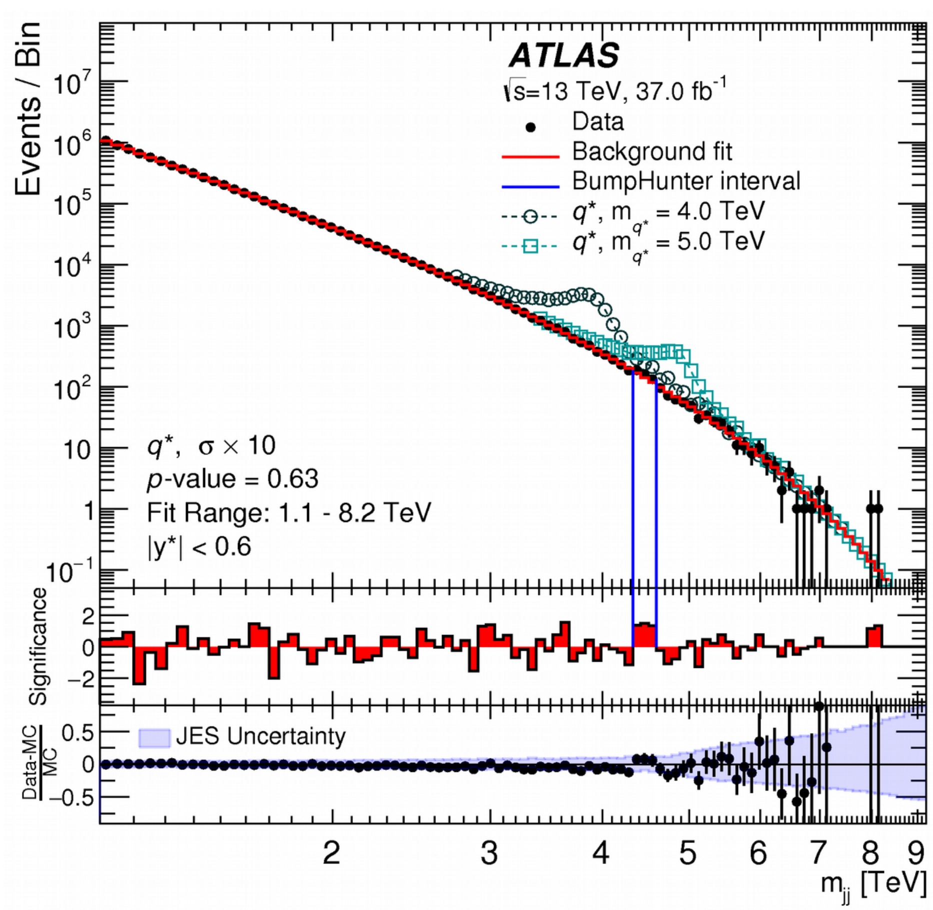
From Macaluso & DS 1803.00107 top tagging with CNNs

See D'Agnolo & Wulzer 1806.02350: train DNN to distinguish data from bg MC. Optimal version of existing general searches

Over-reliance on simulation for backgrounds

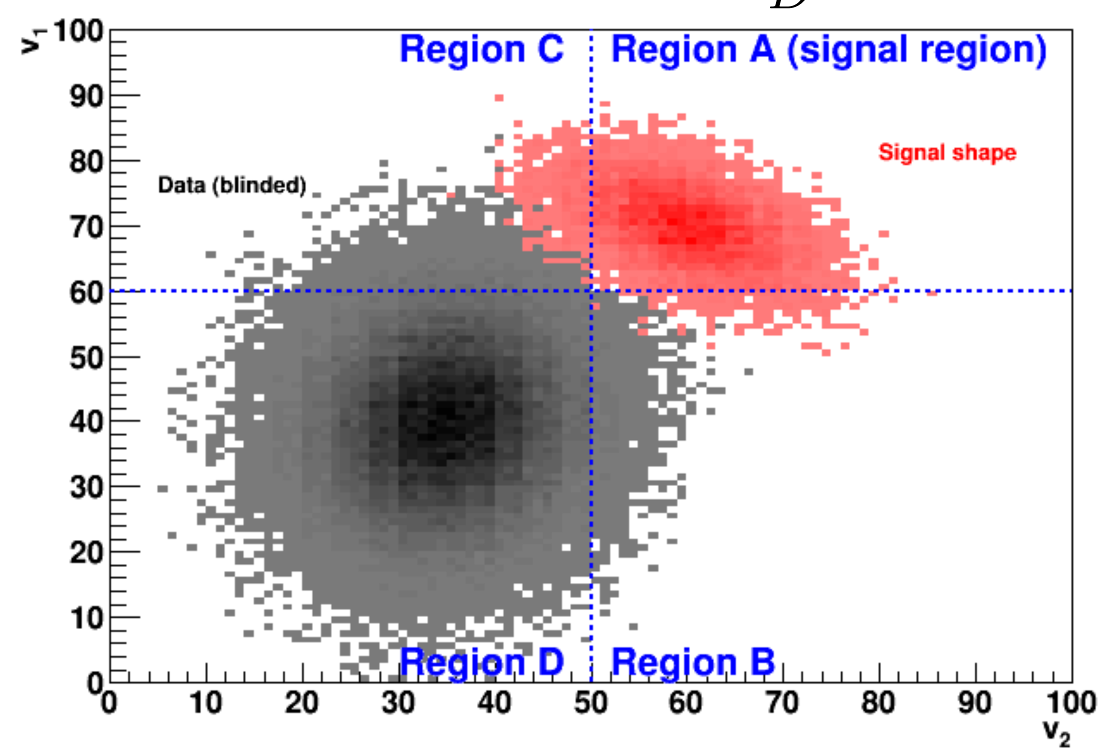
Existing approaches are largely signal model independent, but not background model independent.

To achieve full model independence, we would like to predict backgrounds directly from the data.



Bump hunt with sideband fit

$$N_A = N_B \times \frac{N_C}{N_D}$$



ABCD method

Unsupervised ML for NP searches

Unsupervised ML for NP searches

We want to search for new physics in the data in a signal **and** background independent way.

Cannot use simulation or ground-truth labels. We need **unsupervised ML.**

Two main approaches proposed so far:

Unsupervised ML for NP searches

We want to search for new physics in the data in a signal **and** background independent way.

Cannot use simulation or ground-truth labels. We need **unsupervised ML.**

Two main approaches proposed so far:

- CWoLa hunting
  Collins, Howe & Nachman 1805.02664, 1902.02634

Unsupervised ML for NP searches

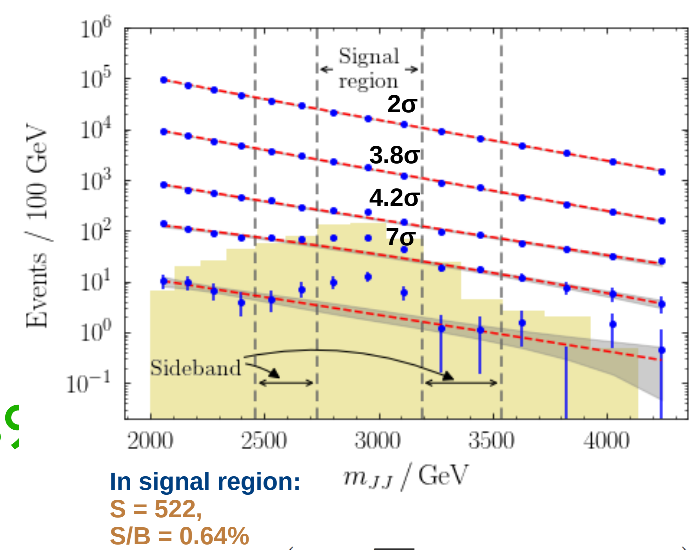
We want to search for new physics in the data in a signal **and** background independent way.

Cannot use simulation or ground-truth labels. We need **unsupervised ML.**

Two main approaches proposed so far:

- CWoLa hunting
  Collins, Howe & Nachman 1805.02664, 1902.02634

- Autoencoders
  Farina, Nakai & DS 1808.08992; Heimel et al 1808.08



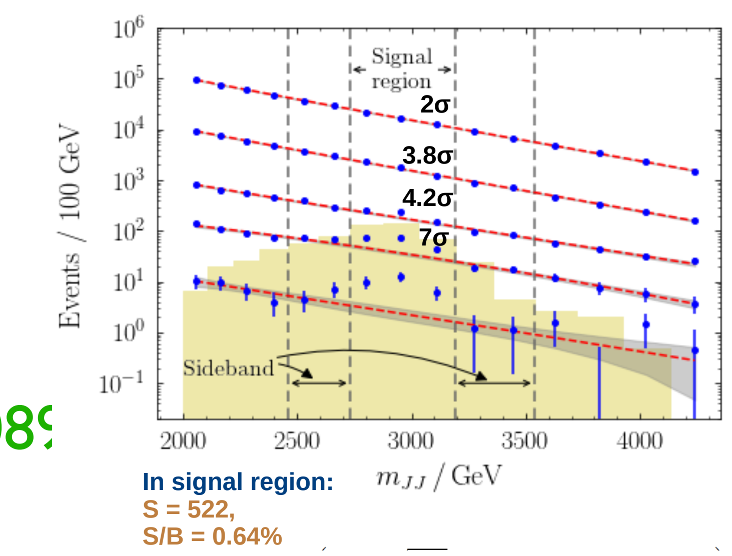
Unsupervised ML for NP searches

We want to search for new physics in the data in a signal **and** background independent way.

Cannot use simulation or ground-truth labels. We need **unsupervised ML.**

Two main approaches proposed so far:

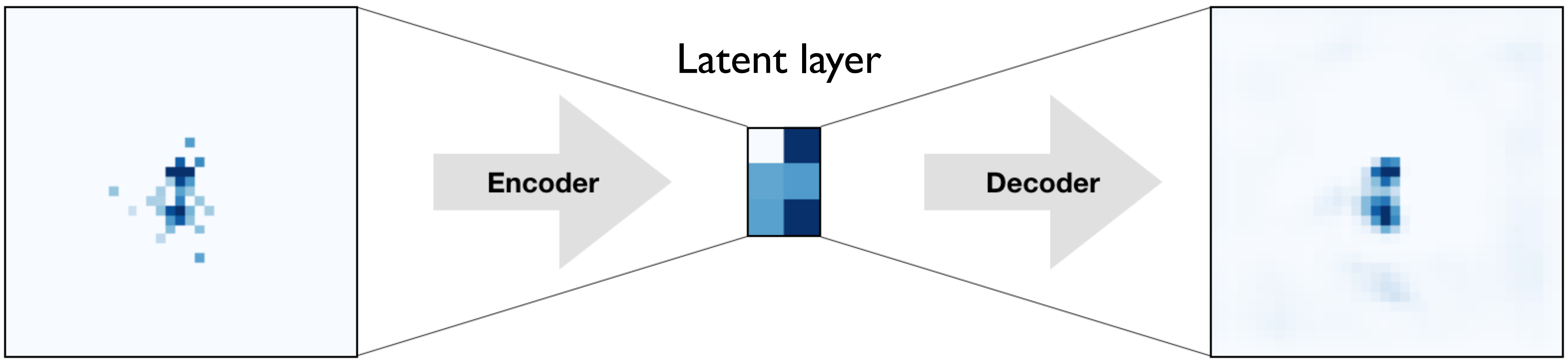
- CWoLa hunting
  Collins, Howe & Nachman 1805.02664, 1902.02634
- Autoencoders
  Farina, Nakai & DS 1808.08992; Heimel et al 1808.08



There are surely more methods waiting to be invented!

Searching for NP with deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992



An autoencoder maps an input into a "latent representation" and then attempts to reconstruct the original input from it.

The encoding is lossy, so the reconstruction is not perfect.

Many real world applications of autoencoders, including anomaly detection, fraud detection, denoising, compression, generation, density estimation

See also:
Hajer et al "Novelty Detection Meets Collider Physics" 1807.10261
Cerri et al "Variational Autoencoders for New Physics Mining at the Large Hadron Collider" 1811.10276

Searching for NP with deep autoencoders

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992

Loss function for autoencoder:

"reconstruction error"

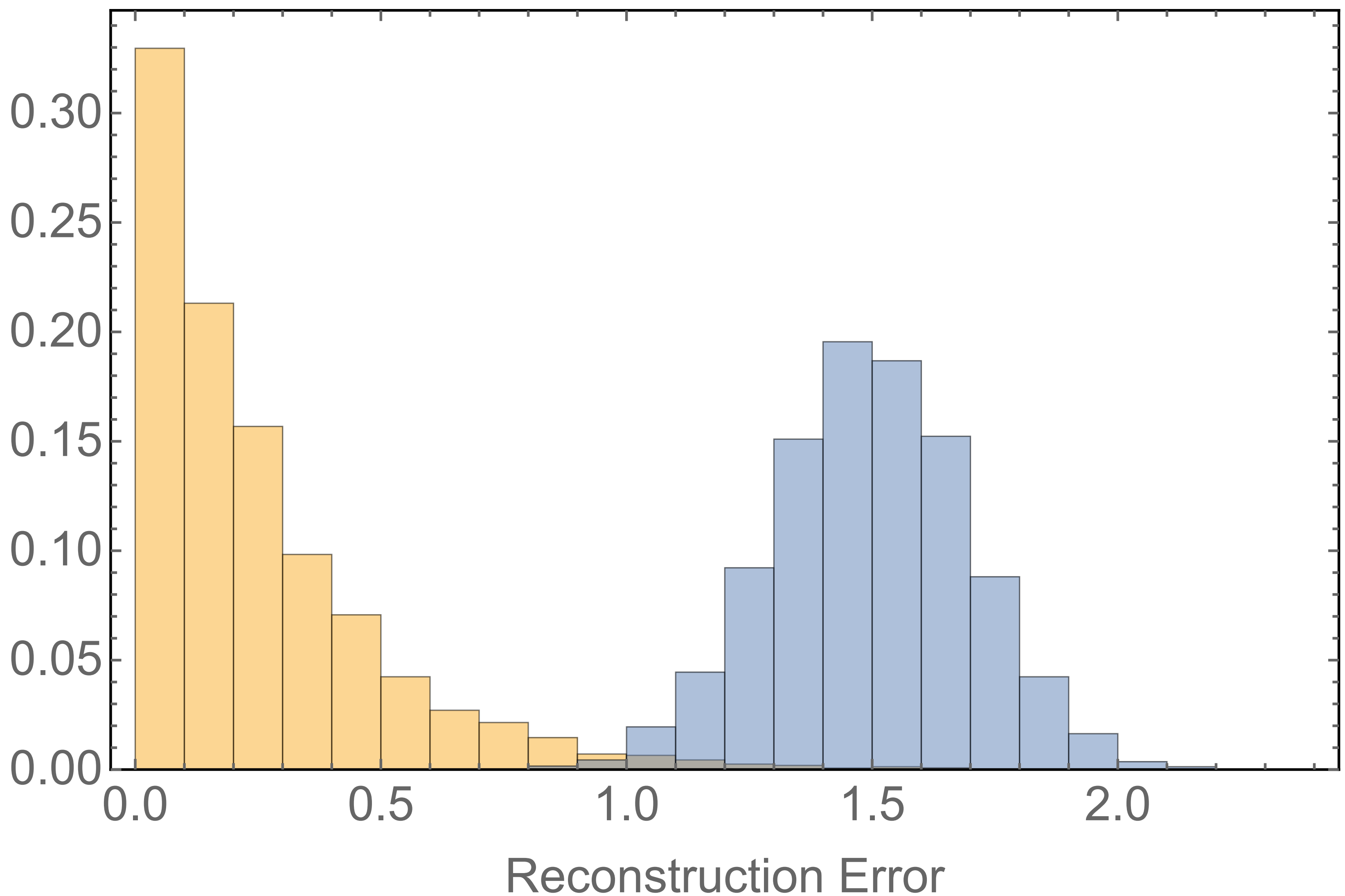
$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

Searching for NP with deep autoencoders

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992

Loss function for autoencoder: $$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

"reconstruction error"

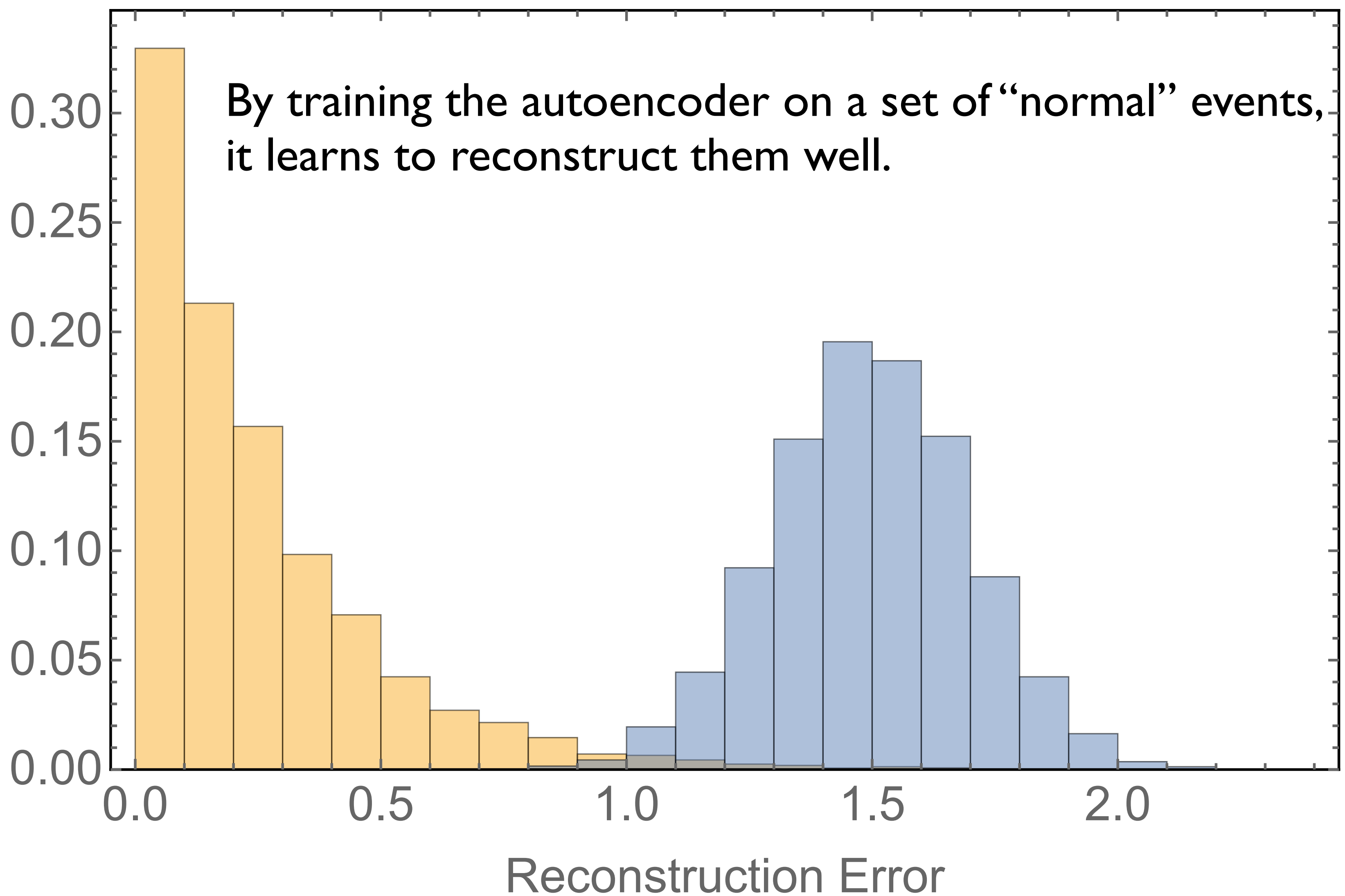


Searching for NP with deep autoencoders

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992

Loss function for autoencoder: $$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

"reconstruction error"

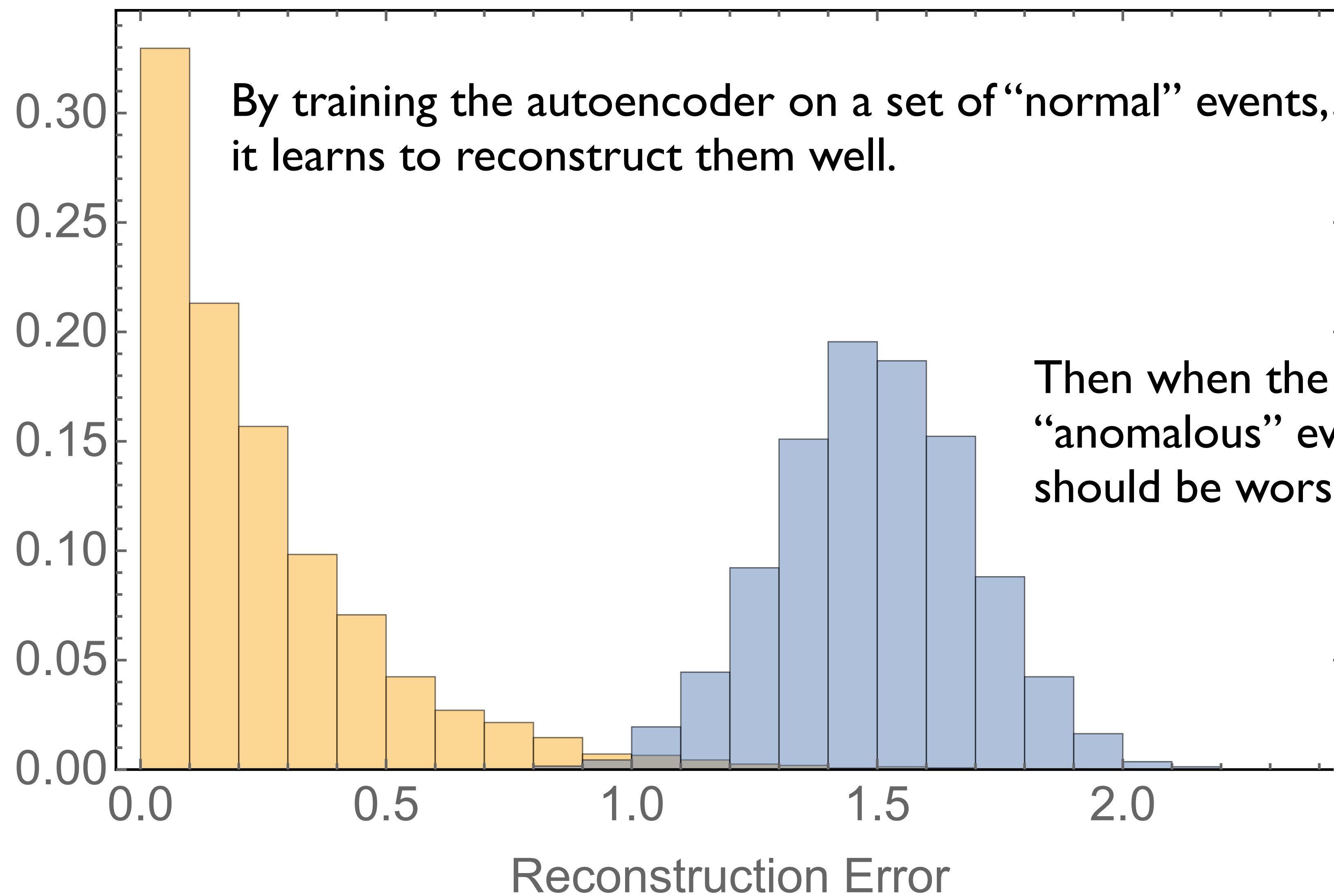


Searching for NP with deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992

Loss function for autoencoder: $$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

"reconstruction error"



By training the autoencoder on a set of "normal" events, it learns to reconstruct them well.

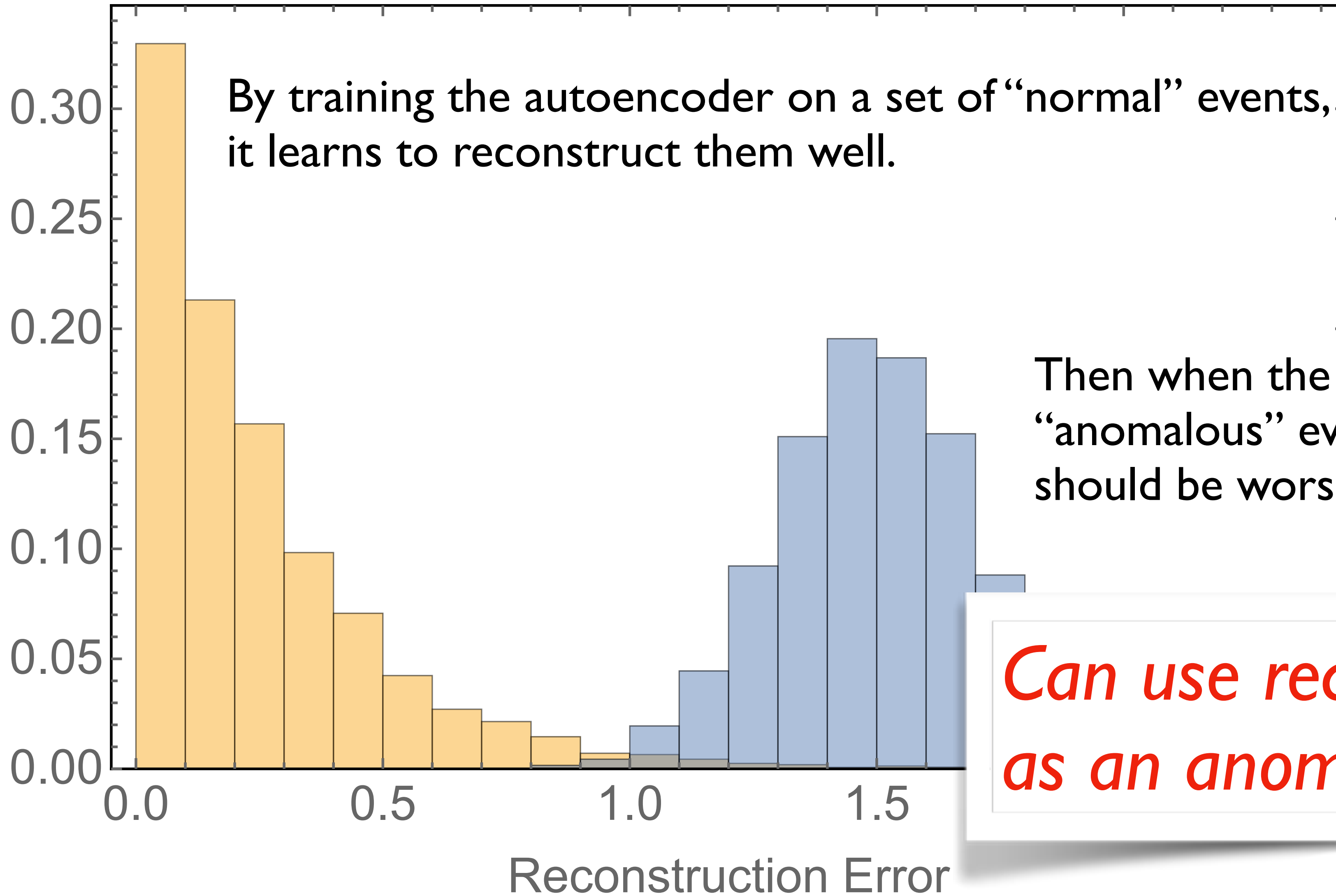
Then when the autoencoder encounters "anomalous" events, its performance should be worse.

Searching for NP with deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992

Loss function for autoencoder: $$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

"reconstruction error"



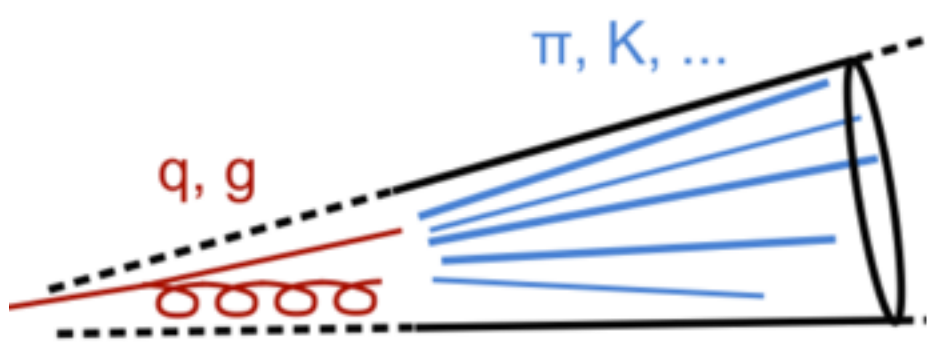
By training the autoencoder on a set of "normal" events, it learns to reconstruct them well.

Then when the autoencoder encounters "anomalous" events, its performance should be worse.

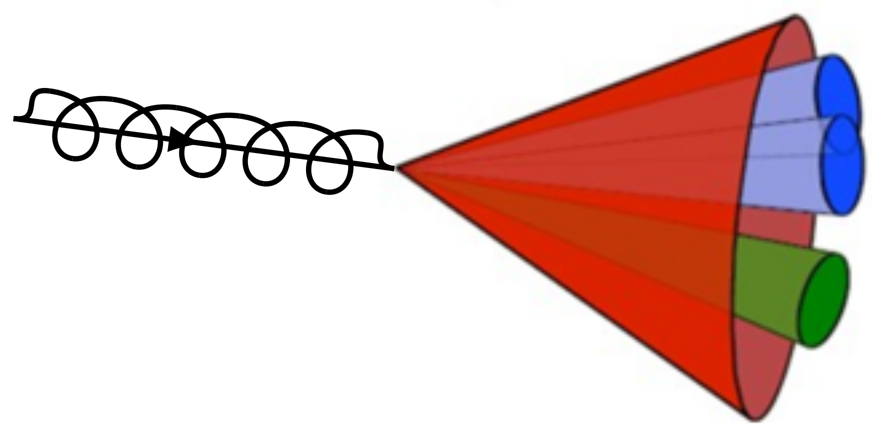
Can use reconstruction error as an anomaly threshold!

Sample definitions

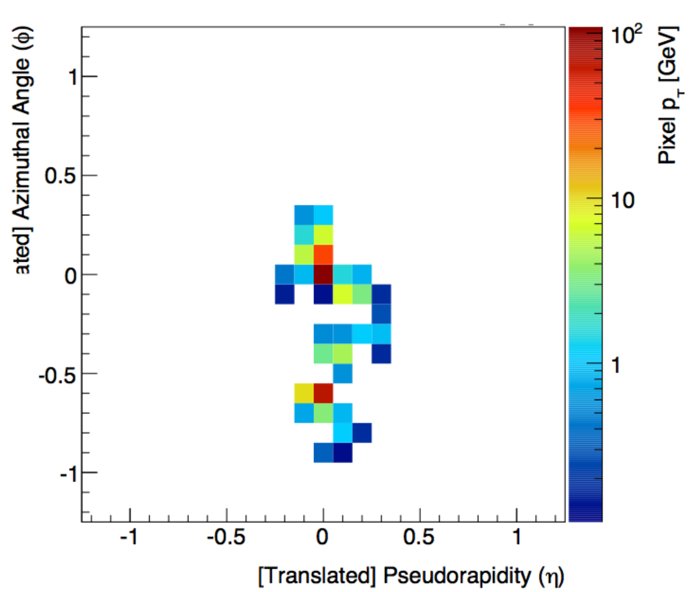
- Background: QCD jets
  (p_T: 800-900 GeV, $|\eta|<1$, anti-kt R=1)



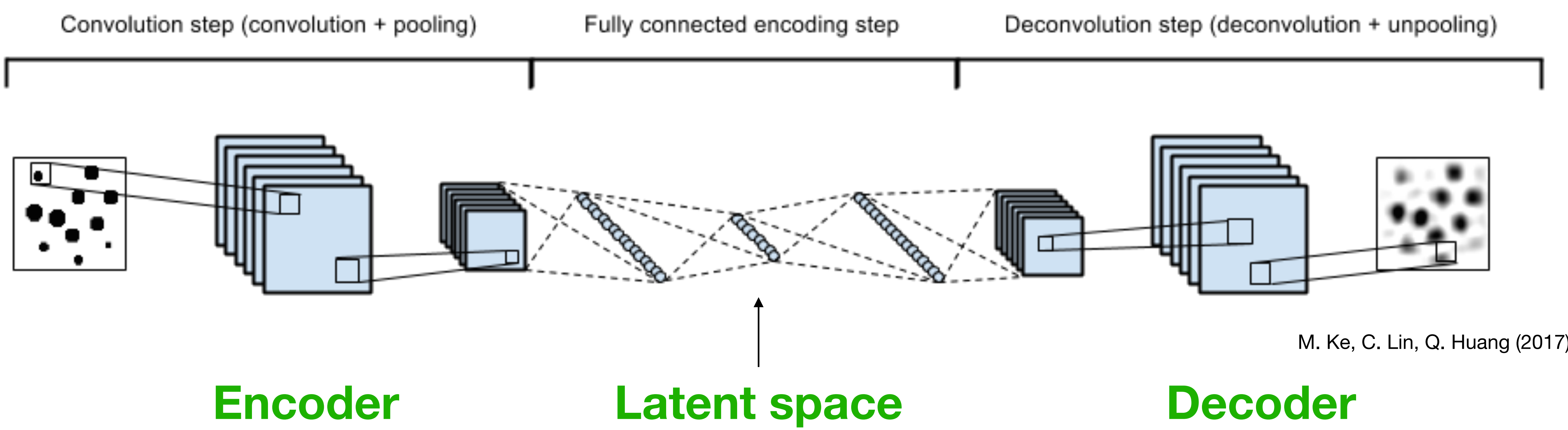
- Signals:
  - All-hadronic tops
  - 400 GeV gluinos decaying via RPV



- We formed jet images in η and ϕ with a pixel resolution and intensity given by the calorimeter towers.



Autoencoder architecture



128C3-MP2-128C3-MP2-128C3-32N-6N-32N-12800N-128C3-US2-128C3-US2-1C3

Our primary autoencoder used convolutional neural networks (CNNs) for encoding and decoding the jet images.

We also considered autoencoders based on PCA and simple DNNs.

Many more architectures are possible.

Choosing the latent dimension

Should choose the latent dimension in an unsupervised manner (ie without optimizing on a specific signal)

- d too large → autoencoder becomes identity transform
- d too small → autoencoder cannot learn all the features

Choosing the latent dimension

Should choose the latent dimension in an unsupervised manner (ie without optimizing on a specific signal)

- d too large → autoencoder becomes identity transform
- d too small → autoencoder cannot learn all the features

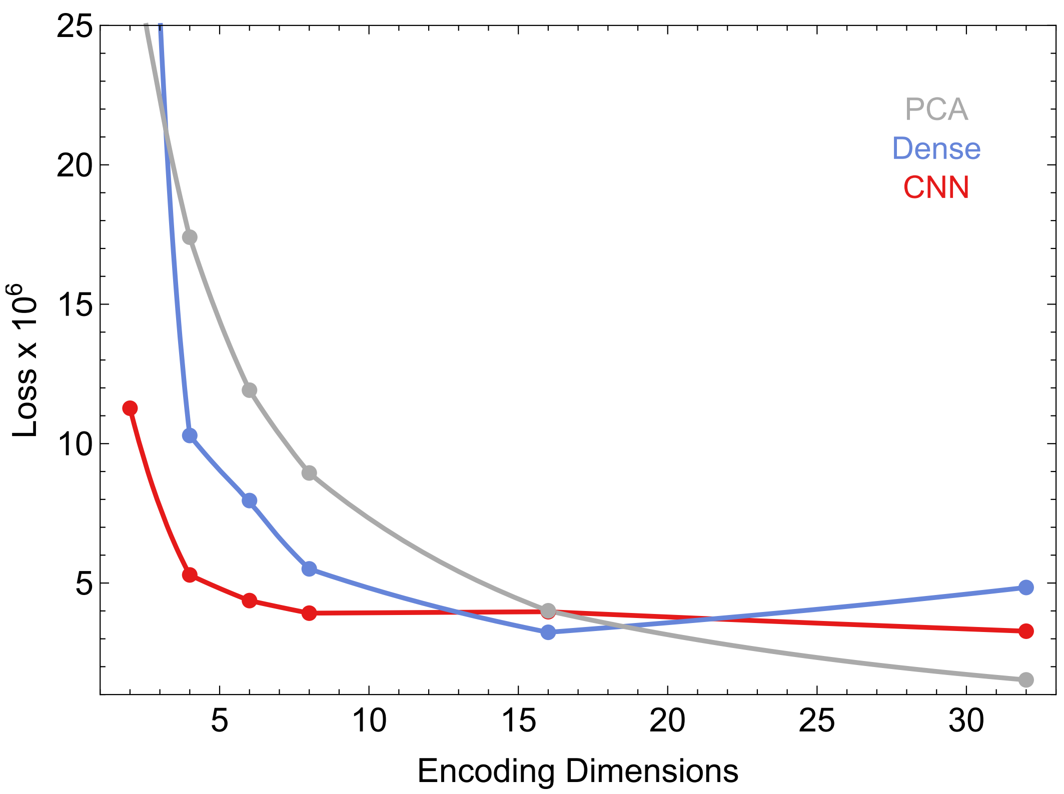
→ *idea: look for saturation point as d is increased*

Choosing the latent dimension

Should choose the latent dimension in an unsupervised manner (ie without optimizing on a specific signal)

- d too large → autoencoder becomes identity transform
- d too small → autoencoder cannot learn all the features

idea: look for saturation point as d is increased

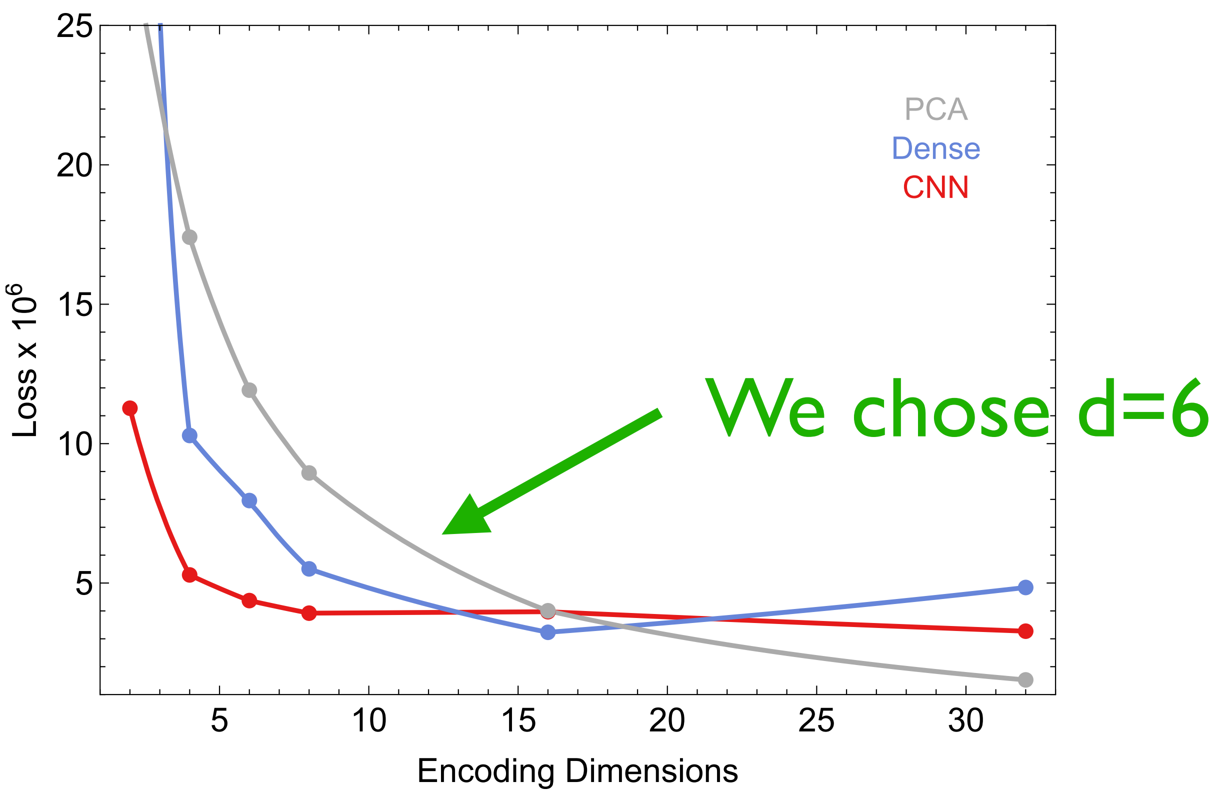


Choosing the latent dimension

Should choose the latent dimension in an unsupervised manner (ie without optimizing on a specific signal)

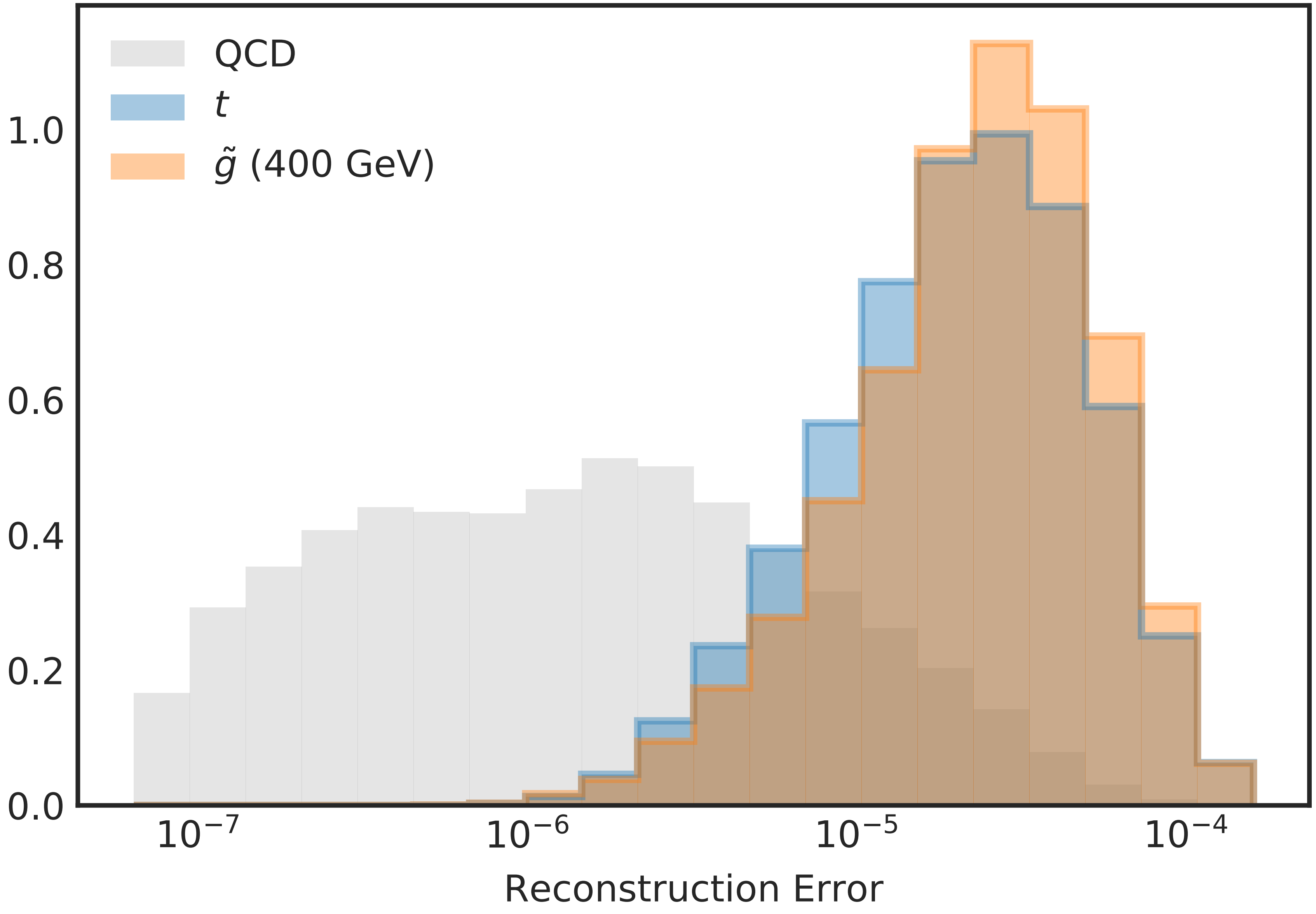
- d too large → autoencoder becomes identity transform
- d too small → autoencoder cannot learn all the features

→ *idea: look for saturation point as d is increased*



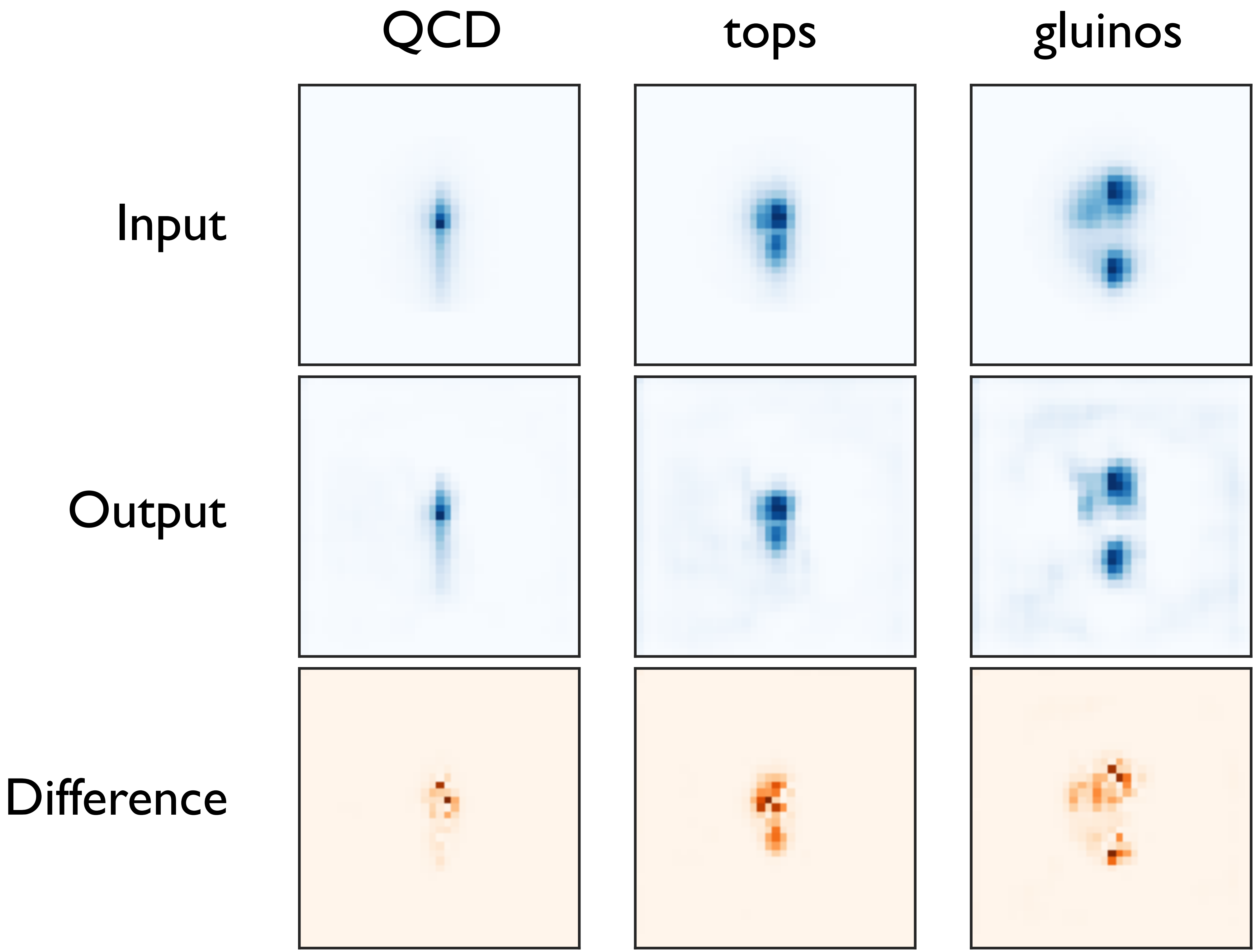
Performance: weakly supervised mode

Train the AE on QCD backgrounds only.



It works as an anomaly detector!

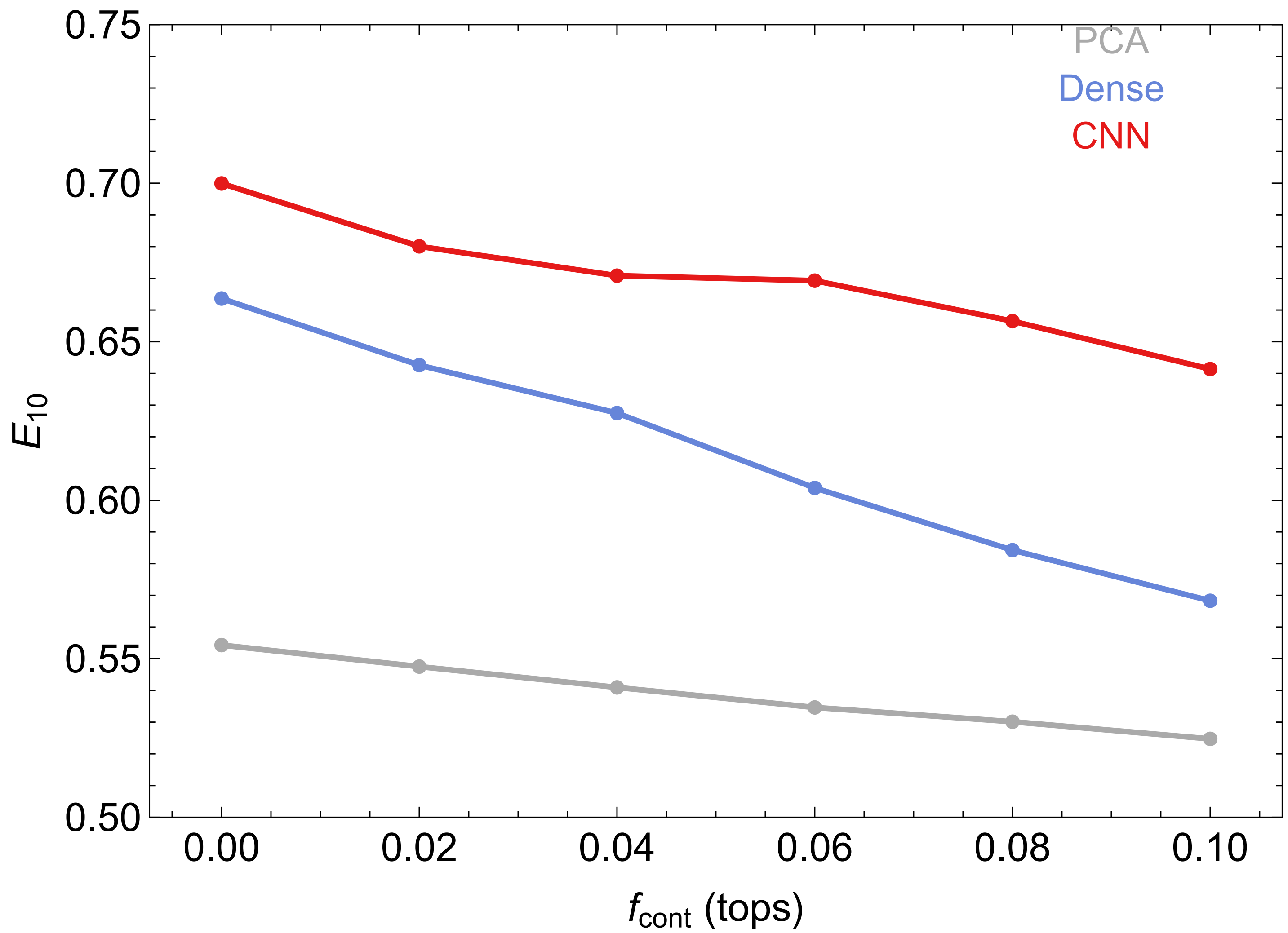
Performance: weakly supervised mode



Fully unsupervised mode

Can also train on QCD background "contaminated" with a small fraction of signal. This could be representative of actual data.

(E_{10} = signal efficiency at 10% bg efficiency)



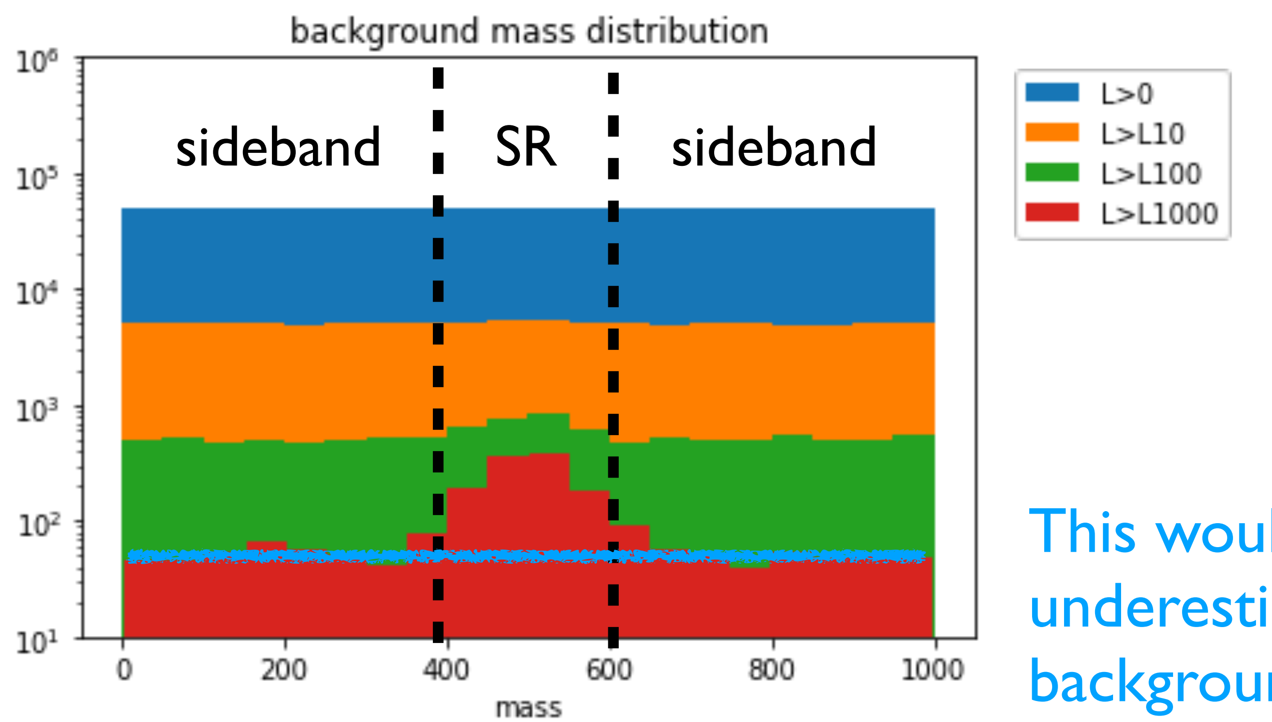
Performance of AE robust even up to 10% contamination!

Background estimation

Finally, background estimation. Remember: unlike past attempts at model-independent searches, we want it to be data-driven!

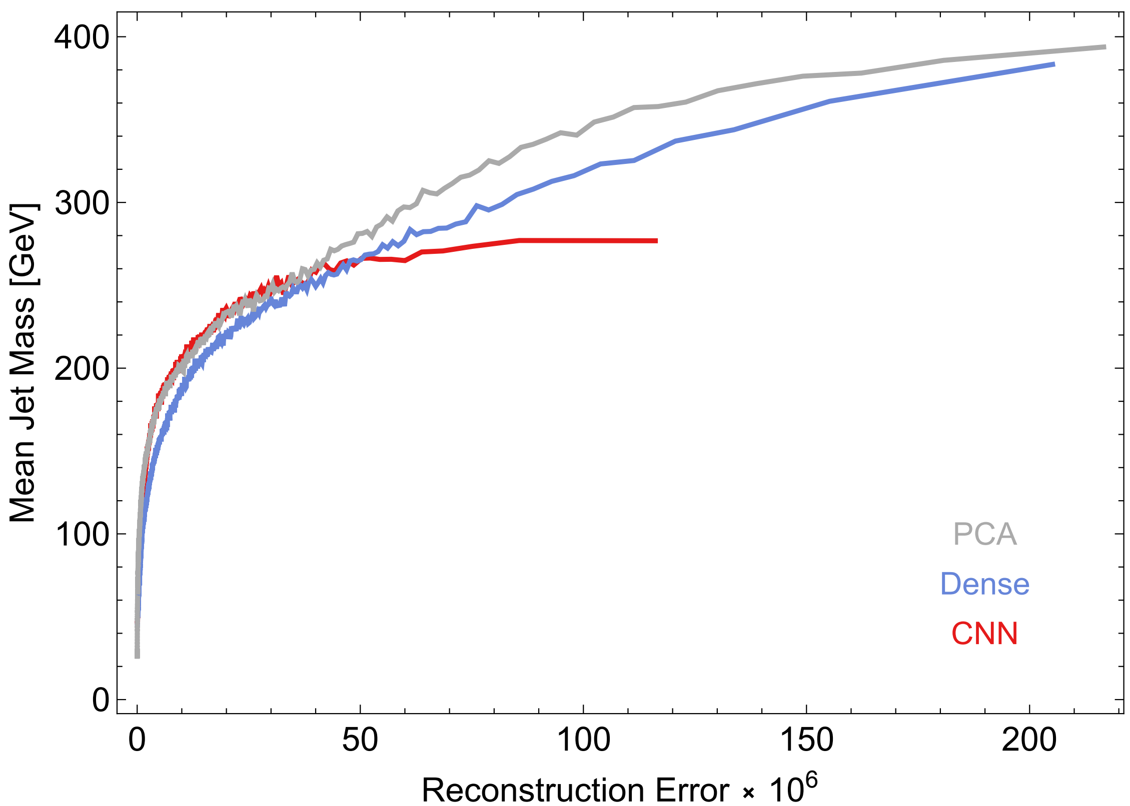
One idea: combine autoencoder with a bump hunt in jet mass. Estimate backgrounds using sidebands in mass.

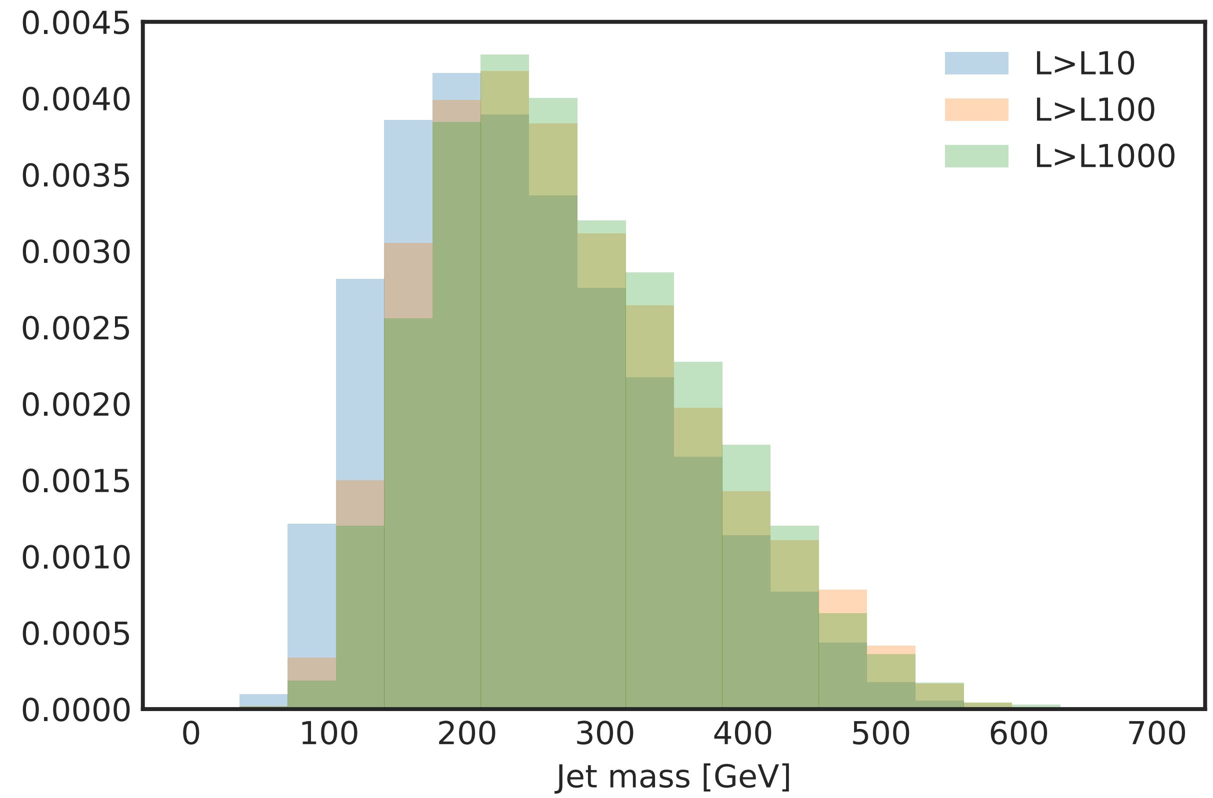
Only works if cutting on reconstruction error does not sculpt the mass distribution of the background!



This would greatly underestimate the background in the SR

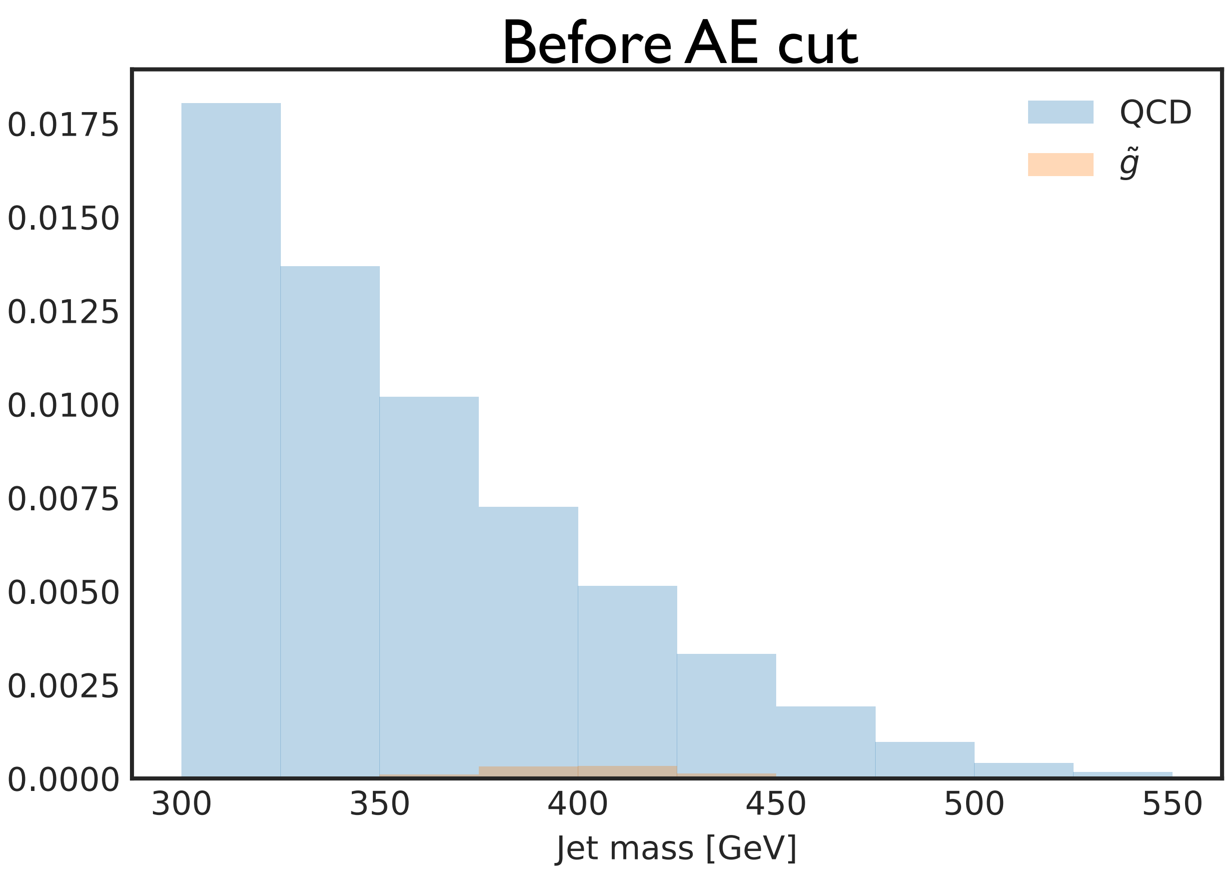
Bump hunt with autoencoder

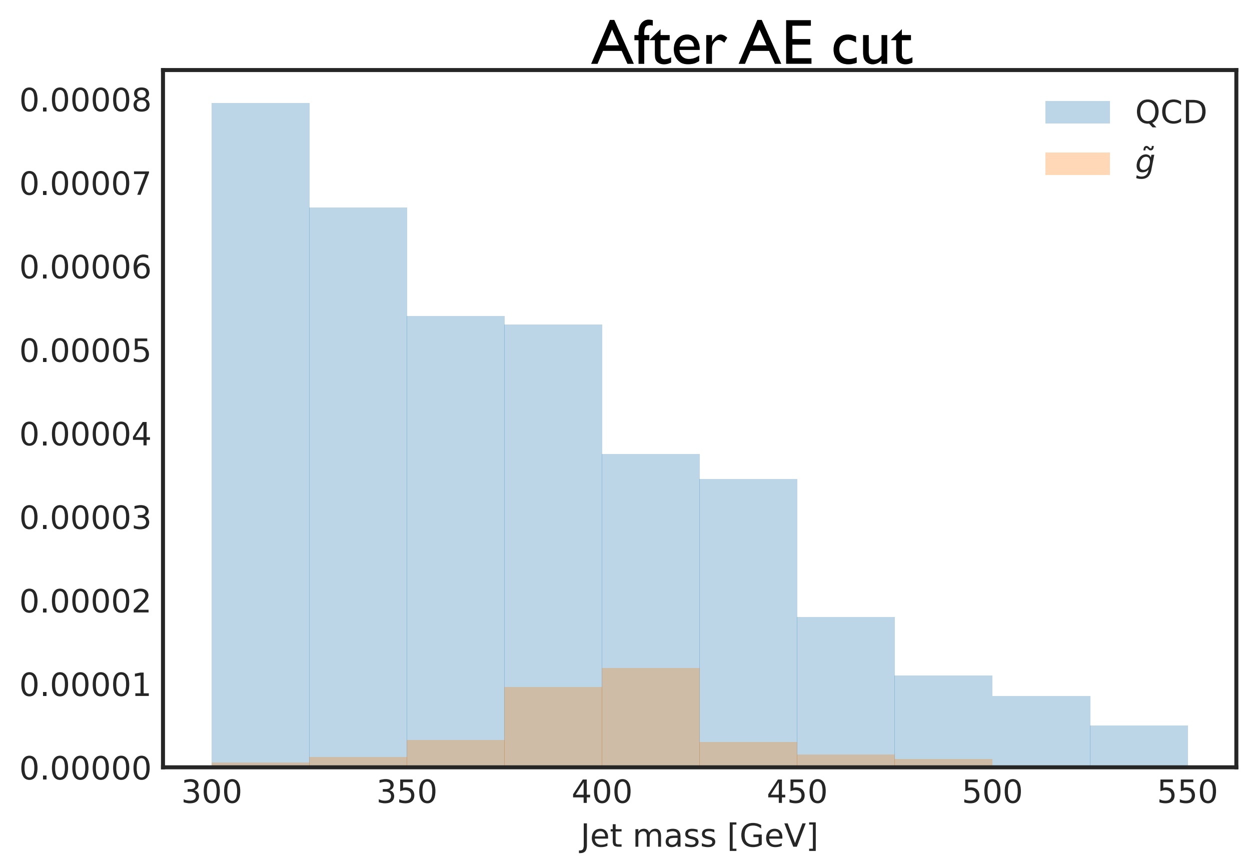




We find empirically that the background jet mass distribution is fairly stable against cuts on CNN AE reconstruction loss above ~250 GeV.

Bump hunt with autoencoder





Train directly on data that contains 400 GeV gluinos.

Use the AE to clean away QCD jets.

Enhance the significance of the bump hunt! (improve S/B by factor of ~6)

Could really discover new physics this way!

Autoencoder with explicit decorrelation

A more controlled approach to mass decorrelation would be to explicitly penalize correlations in the training of the autoencoder.

One promising method: autoencoder with adversarial decorrelation (Heimel et al 1808.08979; based on 1611.01046, 1703.03507 and the idea behind GANs)

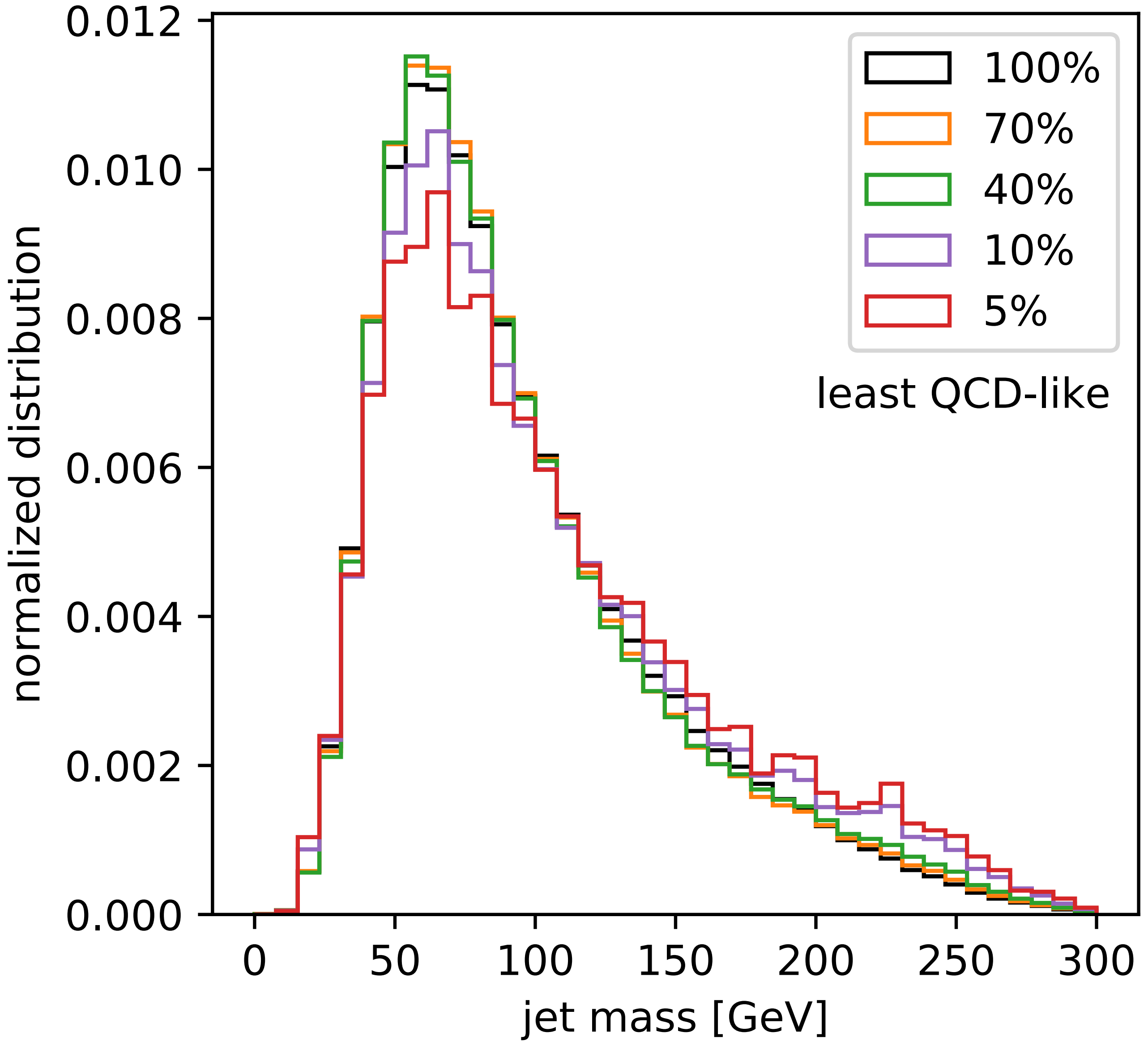
- Introduce a second NN, the adversary, that tries to predict the mass from the reconstruction loss.
- Penalize the total loss function when the adversary does well.

$$L_{adv} = \sum_i (f_{adv}(L_{AE}(x_i)) - m_i)^2$$

$$L_{tot} = L_{AE} - \lambda L_{adv}$$

Autoencoder with explicit decorrelation

Heimel et al 1808.08979



Current implementation needs QCD backgrounds for decorrelation.
Can generalize to fully unsupervised case?

Alternatives to adversaries

Adversaries are notoriously **tricky to train** — saddle point optimization

$$\min_{\theta_{\rm clf}} \max_{\theta_{\rm adv}} L_{\rm clf}(y(\theta_{\rm clf})) - \lambda L_{\rm adv}(y(\theta_{\rm clf}), m; \theta_{\rm adv})$$

Would be great if we could achieve the same performance but with a convex regularizer term

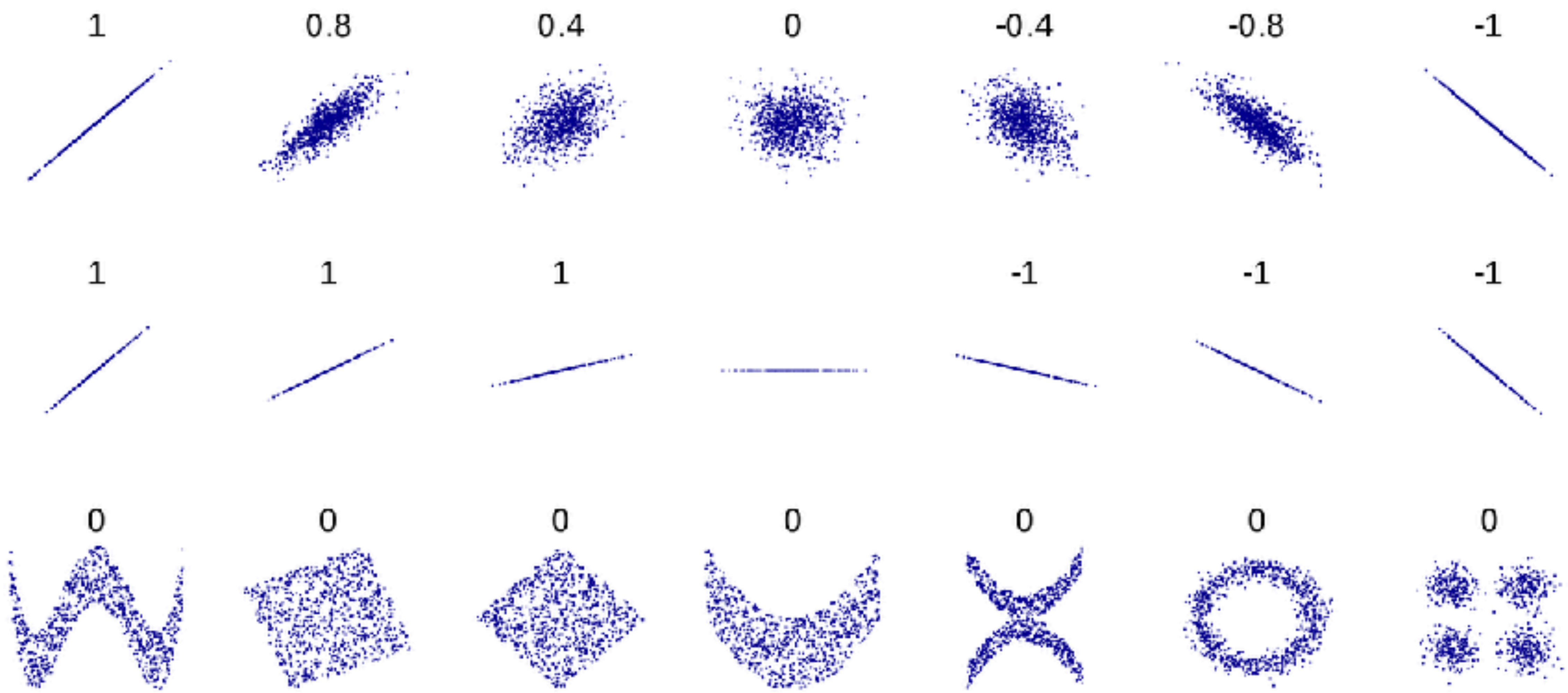
$$\min_{\theta_{\rm clf}} L_{\rm clf}(y(\theta_{\rm clf})) + \lambda C_{\rm reg}(y(\theta_{\rm clf}), m)$$

First idea: can we just use Pearson correlation coefficient?

$$C_{\rm reg} = R(y, m) \propto \sum_i y_i m_i$$

Problem: this only measures linear correlations

Pearson correlation



y and m can be highly correlated yet R=0

Distance correlation ("DisCo")

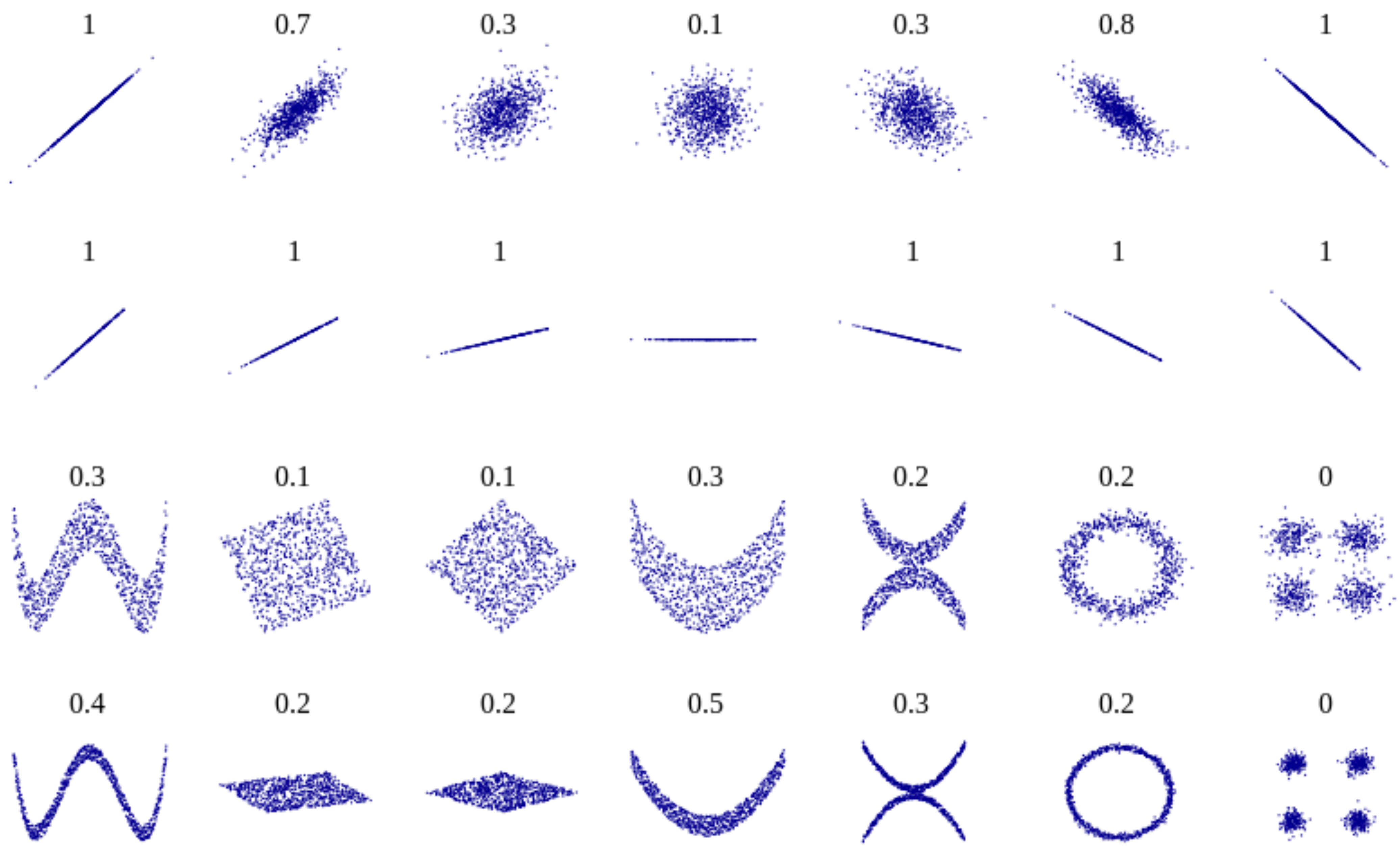
Work in progress with Gregor Kasieczka

Promising idea: "distance correlation"
(Szekely, Rizzo, Bakirov 2007; Szekely & Rizzo 2009)

$$\mathrm{dCov}^2(X,Y) = \langle |X-X'||Y-Y'|\rangle + \langle |X-X'|\rangle\langle |Y-Y'|\rangle - 2\langle |X-X'||Y-Y''|\rangle$$

- Zero iff X,Y are independent; positive otherwise
- Computationally tractable
- Straightforward sample definition — doesn't require binning

Distance correlation



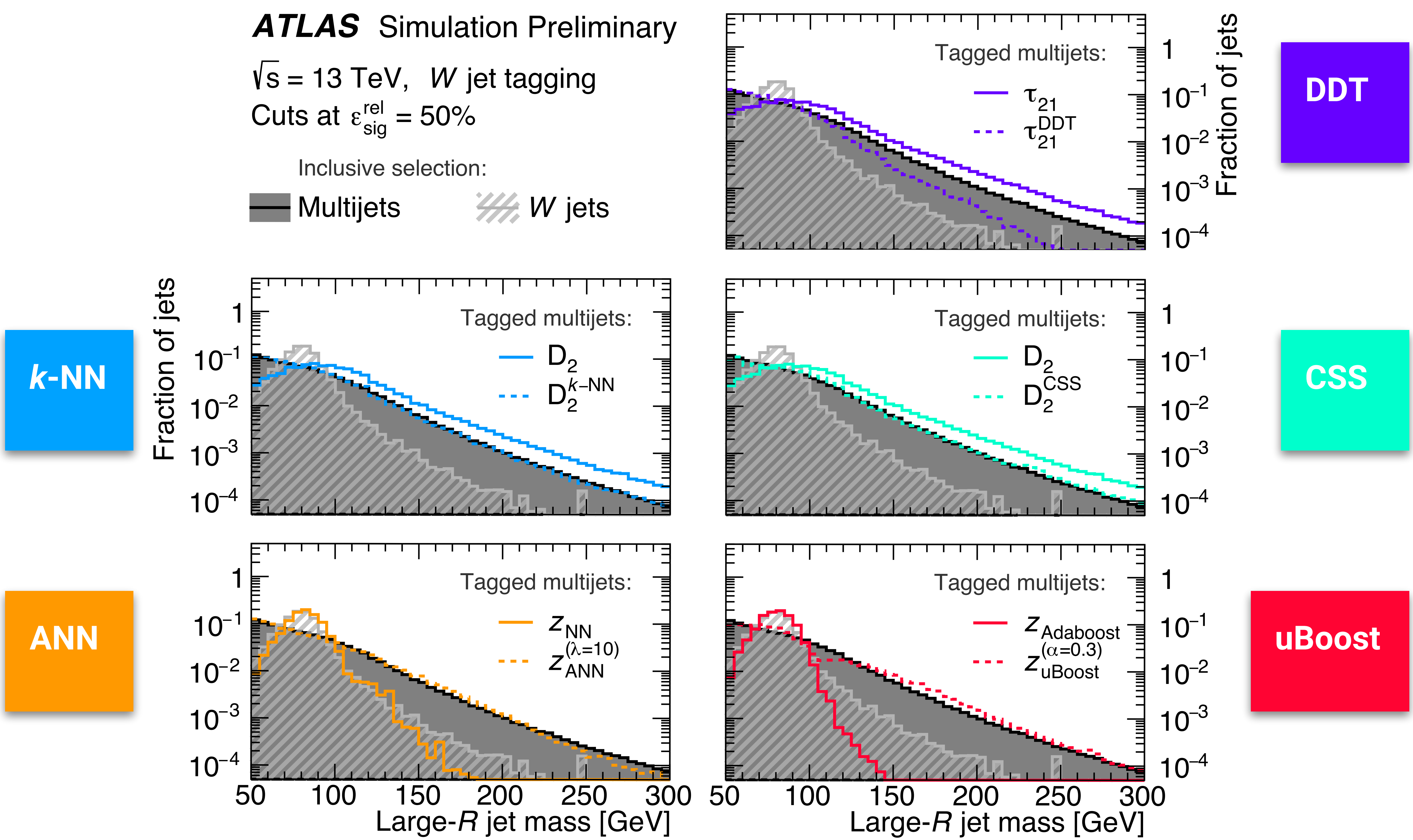
Disco is sensitive to nonlinear correlations!

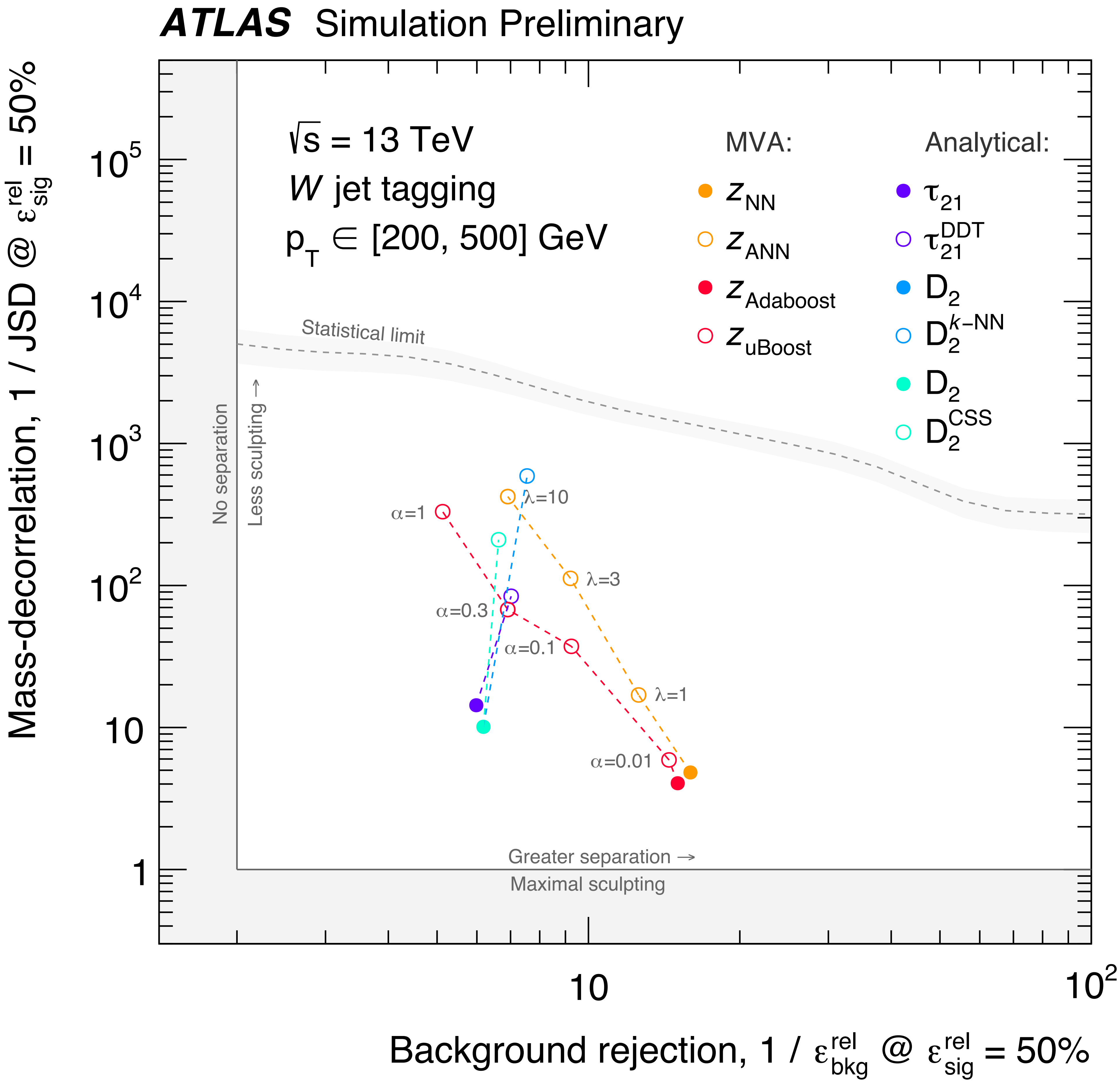
State of the art: ATLAS study of various decorrelation methods in context of boosted W-tagging.

ATL-PHYS-PUB-2018-014

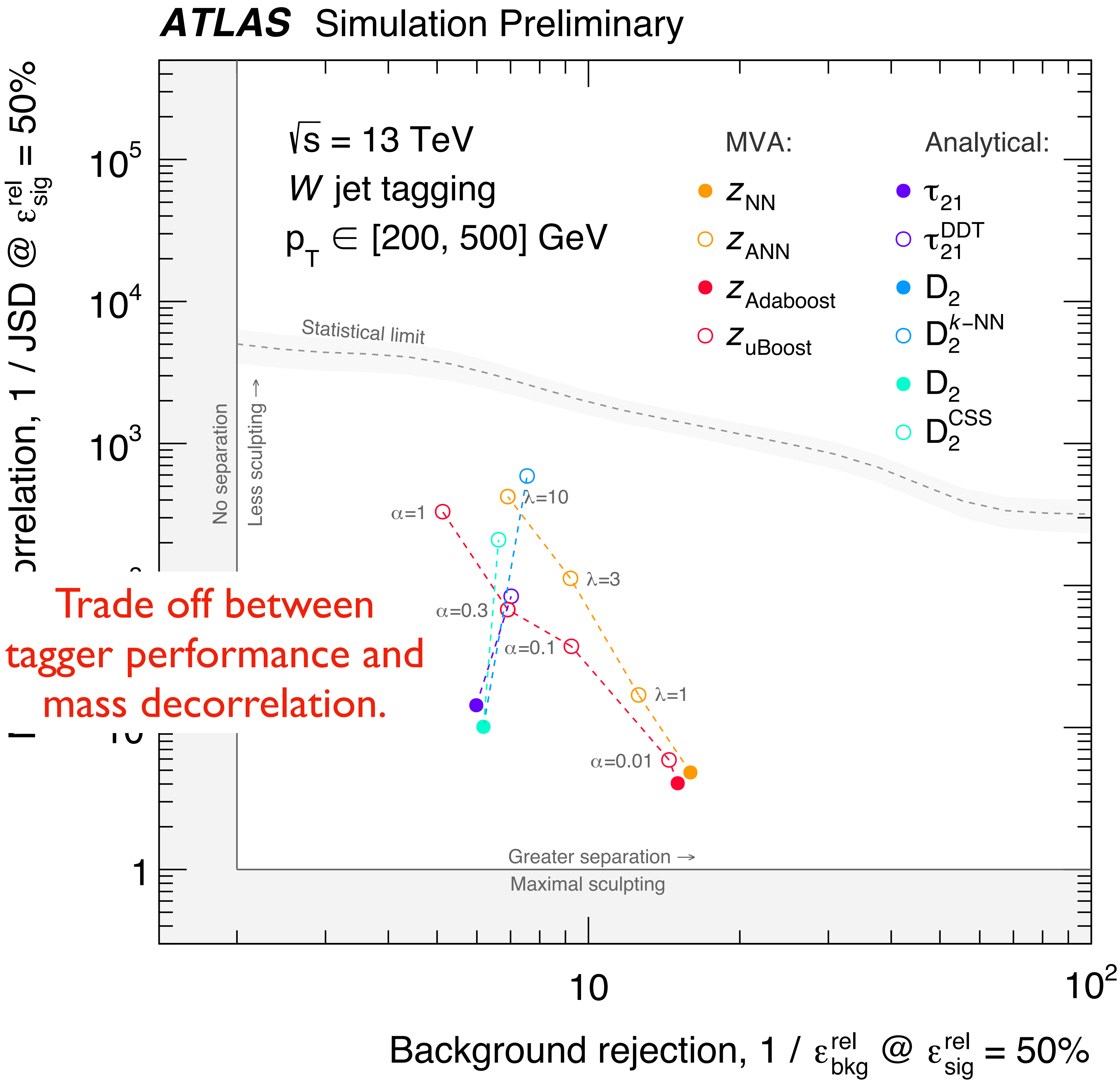
Analytical / single-variable

Multivariate (MVA)

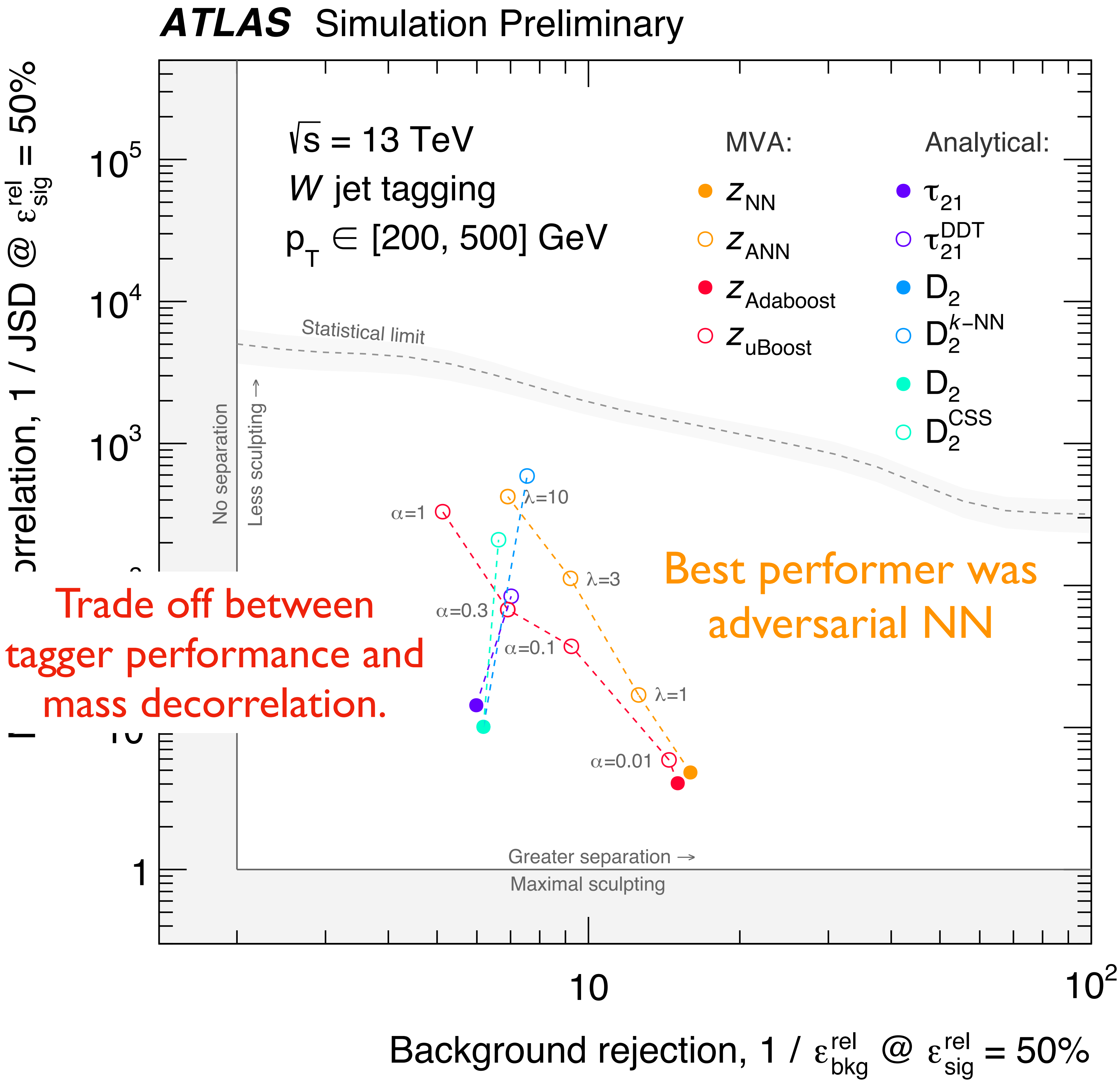




ATLAS Simulation Preliminary
√s = 13 TeV
W jet tagging
$p_T \in [200, 500]$ GeV
MVA:
z_{NN}
z_{ANN}
$z_{Adaboost}$
z_{uBoost}
Analytical:
τ_{21}
τ_{21}^{DDT}
D_2
D_2^{k-NN}
D_2
D_2^{CSS}
Statistical limit
No separation
Less sculpting →
Greater separation →
Maximal sculpting
λ=10
λ=3
λ=1
α=1
α=0.3
α=0.1
α=0.01
Mass-decorrelation, 1 / JSD @ ε_{sig}^{rel} = 50%
Background rejection, 1 / ε_{bkg}^{rel} @ ε_{sig}^{rel} = 50%

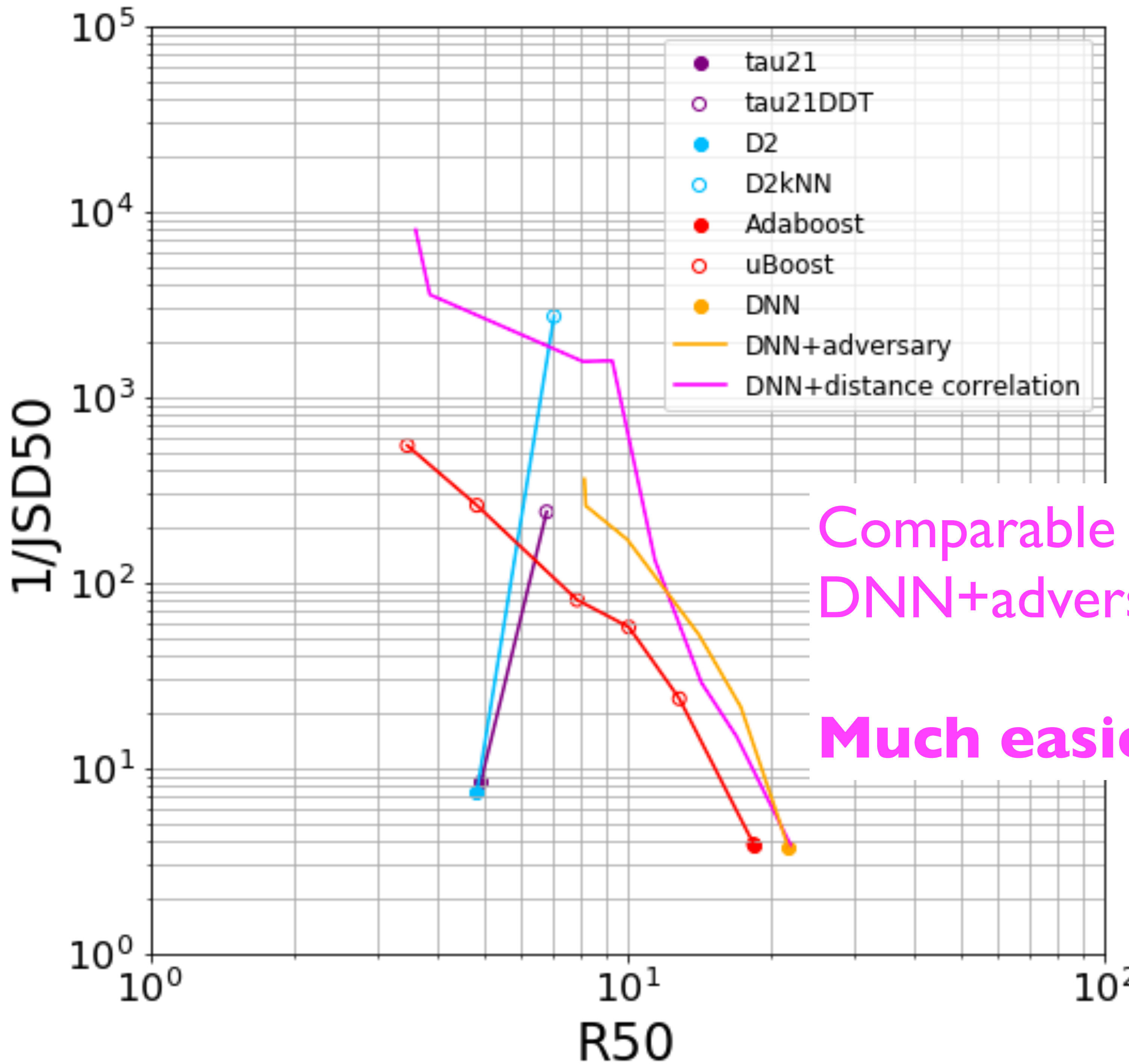


Trade off between tagger performance and mass decorrelation.


ATLAS Simulation Preliminary
$\sqrt{s}$ = 13 TeV
W jet tagging
$p_T \in$ [200, 500] GeV
MVA:
z_{NN}
z_{ANN}
$z_{Adaboost}$
z_{uBoost}
Analytical:
τ_{21}
τ_{21}^{DDT}
D_2
D_2^{k-NN}
D_2
D_2^{CSS}
Statistical limit
No separation
Less sculpting →
Greater separation →
Maximal sculpting
α=1
α=0.3
α=0.1
α=0.01
λ=10
λ=3
λ=1
Background rejection, $1 / \varepsilon_{bkg}^{rel}$ @ ε_{sig}^{rel} = 50%
1 / JSD @ ε_{sig}^{rel} = 50%
Best performer was adversarial NN
Trade off between tagger performance and mass decorrelation.

Distance correlation ("DisCo")

Work in progress with Gregor Kasieczka



Comparable performance to DNN+adversary.

Much easier to train.

Summary/Outlook

There is increasingly strong motivation to perform model-independent searches at the LHC. We need to ensure that we have not missed anything in the data.

However, attempts at model-independent searches have suffered from major drawbacks: an enormous trials factor, simplistic signal/background discrimination, and an over-reliance on simulation. Together, they have given the philosophy of model-independent searches a bad name.

In this talk we have explored new ideas for model-independent searches inspired by recent breakthroughs in unsupervised deep learning.

We have seen that:

- Deep autoencoders can find subtle signals in the data in a model agnostic way.
- Decorrelation is essential in combining data-driven background estimation with deep learning.
- A new method "DisCo" can more easily achieve state-of-the-art decorrelation in NN training.

Outlook

I think there is an exciting future for model-independent searches @ LHC!

- Comparison of different approaches to model-independent searches (AE, CWoLa, … — work in progress with Pablo Martin & Ben Nachman)
- Careful study of the LEE in these different approaches
- More applications of DisCo
  - e.g. using it to improve the ABCD method ("Double DisCo" — work in progress with Gregor Kasieczka, Ben Nachman & Matt Schwartz)
- Real life applications of decorrelation?
  - E.g. making less discriminatory AIs for hiring/admissions/bail decisions/sentencing/… ?
- **New ideas for model-independent searches**
  - e.g. using recent breakthroughs in density estimation to search for anomalies
  - many more…?!

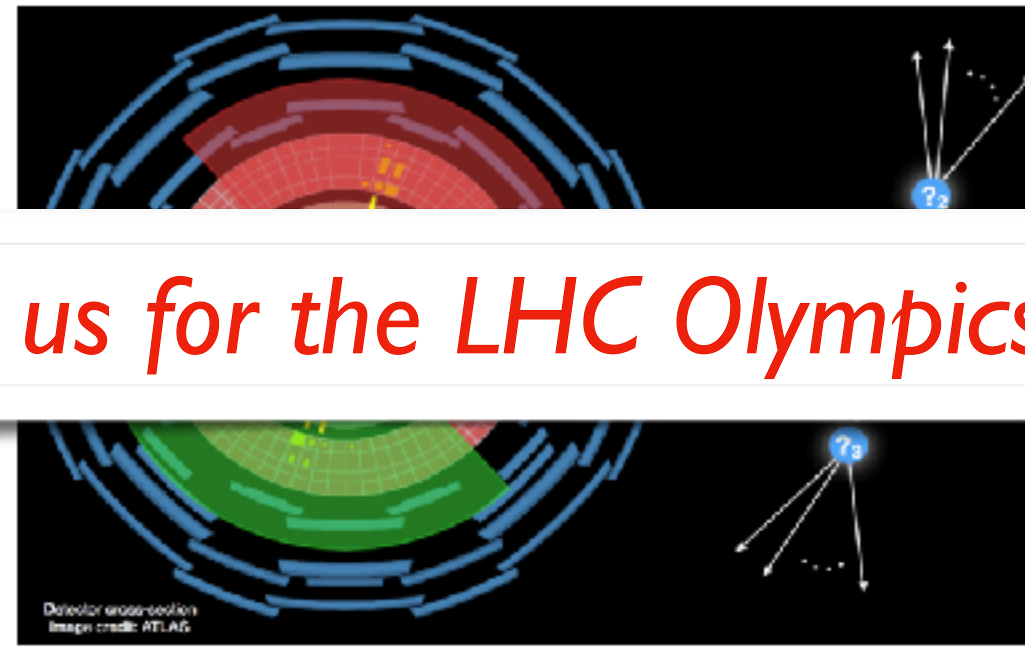
ML4Jets2020

15-17 January 2020
Europe/Zurich timezone

Search...

- Overview
- Timetable
- Participant List
- LHCOlympics2020
- Slack channel

LHCOlympics2020



Come join us for the LHC Olympics 2020!

Despite an impressive and extensive effort by the LHC collaborations, there is currently no convincing evidence for new particles produced in high-energy collisions. At the same time, there has been a growing interest in machine learning techniques to enhance potential signals using all of the available information.

In the spirit of the first LHC Olympics (circa 2005-2006) [1st, 2nd, 3rd, 4th], we are organizing the 2020 LHC Olympics. Our goal is to ensure that the LHC search program is sufficiently well-rounded to capture "all" rare and complex signals. The final state for this olympics will be focused (generic dijet events) but the observable phase space and potential BSM parameter space(s) are large: all hadrons in the event can be used for learning (be it "cuts", supervised machine learning, or unsupervised machine learning).

For setting up, developing, and validating your methods, we provide background events and a benchmark signal model. You can download these from this page. To help get you started, we have also prepared simple python scripts to read in the data and do some basic processing.

The final test will happen 2 weeks before the ML4Jets2020 workshop. We will release a new dataset where the "background" will be similar to but not identical to the one in the development set (as is true in real data!). The goal of the challenge is to see who can "best" identify BSM (yes/no, what mass, what cross-section) in the dataset. There are many ways to quantify "best" and we will use all of the submissions to explore the pros/cons of the various approaches.

Thanks for your attention!