Anomaly detection with not-so-dense(ity) estimators at ATLAS

Jack Harrison[1]

[1] *Institut de Física d'Altes Energies, Barcelona*

Cross-Collserola Meeting in Astrophysics, Cosmology and Particles

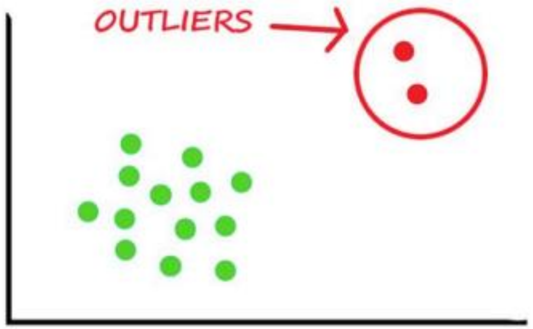
What is anomaly detection?

- Machine learning generally falls under three categories:

1. Supervised learning
   - Distinct labels for all examples
2. Semi-supervised learning
   - A small number of examples are labelled
3. **Unsupervised learning**
   - No labels

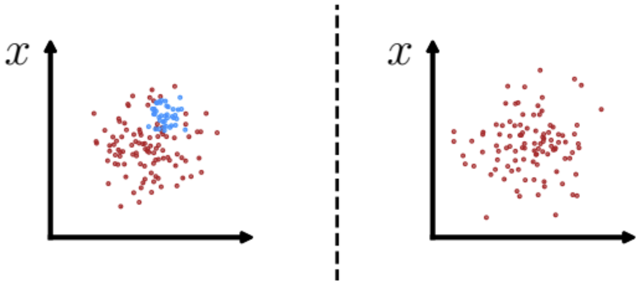
- Anomaly detection is a branch of unsupervised learning, where we try and learn directly from the distribution in order to detect outliers. It's generally split into two themes:

Outlier detection



low $p(x)$

Over-density detection



high $\frac{p_{data}(x)}{p_{bg}(x)}$

Shameless grab from [David Shih's talk](#).

Why do we need anomaly detection?

TABLE I. Existing two-body exclusive final state resonance searches at $\sqrt{s} = 8$ TeV. The ∅ symbol indicates no existing search at the LHC.

	e	μ	τ	γ	j	b	t	W	Z	h
e	±∓[4],±±[5]	±±[5, 6] ±∓[6, 7]	[7]	∅	∅	∅	∅	∅	∅	∅
μ		±∓[4],±±[5]	[7]	∅	∅	∅	∅	∅	∅	∅
τ			[8]	∅	∅	∅	[9]	∅	∅	∅
γ				[10]	[11–13]	∅	∅	[14]	[14]	∅
j					[15]	[16]	[17]	[18]	[18]	∅
b						[16]	[19]	∅	∅	∅
t							[20]	[21]	∅	∅
W								[22–25]	[23, 24, 26, 27]	[28–30]
Z									[23, 25, 31]	[28, 30, 32, 33]
h										[34–37]

arXiv:1610.09392

The picture gets worse when including BSM resonances:

		e	μ	τ	q/g	b	t	γ	Z/W	H
	e	[37, 38]	[39, 40]	[39]	∅	∅	∅	[41]	[42]	∅
	μ		[37, 38]	[39]	∅	∅	∅	[41]	[42]	∅
	τ			[45, 46]	∅	[47]	∅	∅	∅	∅
	q/g				[29, 30, 50, 51]	[52]	∅	[53, 54]	[55]	∅
	b					[29, 52, 56]	[57]	[54]	[58]	[59]
	t						[61]	∅	[62]	[63]
	γ							[65, 66]	[67–69]	[68, 70]
	Z/W								[71]	[71]
	H									[72, 73]
BSM → SM$_1$ × SM$_1$	q/g									
	γ/π^0's									
	b									
	⋮									
BSM → SM$_1$ × SM$_2$	tZ/H									
	bH									
	⋮									
BSM → complex	$\tau qq'$									
	eqq'									
	$\mu qq'$									
	⋮									

		BSM → SM$_1$ × SM$_1$				BSM → SM$_1$ × SM$_2$				BSM → complex			
		q/g	γ/π^0's	b	⋯	tZ/H	bH		⋯	$\tau qq'$	eqq'	$\mu qq'$	⋯
	e	∅	∅	∅		∅	∅	∅		∅	[43, 44]	∅	
	μ	∅	∅	∅		∅	∅	∅		∅	∅	[43, 44]	
	τ	∅	∅	∅		∅	∅	∅		[48, 49]	∅	∅	
	q/g	∅	∅	∅		∅	∅	∅		∅	∅	∅	
	b	∅	∅	∅		[60]	∅	∅		∅	∅	∅	
	t	∅	∅	∅		[64]	[60]	∅		∅	∅	∅	
	γ	∅	∅	∅		∅	∅	∅		∅	∅	∅	
	Z/W	∅	∅	∅		∅	∅	∅		∅	∅	∅	
	H	[74]	∅	∅		∅	∅	∅		∅	∅	∅	
BSM → SM$_1$ × SM$_1$	q/g	∅	∅	∅		∅	∅	∅		∅	∅	∅	
	γ/π^0's		[75]	∅		∅	∅	∅		∅	∅	∅	
	b			[76, 77]		∅	∅	∅		∅	∅	∅	
	⋮												
BSM → SM$_1$ × SM$_2$	tZ/H												
	bH												
	⋮												
BSM → complex	$\tau qq'$												
	eqq'												
	$\mu qq'$												
	⋮												

arXiv:1907.06659

- The phase-space of potential BSM models is too large to be covered by dedicated analyses.
- We need a way of casting a broad net, that is sensitive to a range of new physics.

A standard search

Pick a signal model

Generate it with your chosen masses + couplings

Define a specific signal region, where signal populates.
Define orthogonal control regions, where signal mustn't populate

Check modelling in control regions and assume it holds in signal regions
Compare the number of recorded events in your signal region, with the number you expect from your simulation

Problems:

What if you pick wrong?

What if your signal populates your control regions, or only part of your signal regions?

What if your background modelling doesn't hold across orthogonal regions?

Normalising flows

- We want some model that can take an input $z \sim p_z$, which is easy to sample, and can generate our complicated distribution p_x. We do this by learning an invertible mapping $x = f_\theta(z)$ and $z = f_\theta^{-1}(x)$.

- We rely on the change of variables formula for probability densities;

$$p_x(x;\theta) = p_z\left(f_\theta^{-1}(x)\right)|\det(\frac{\partial f_\theta^{-1}(x)}{\partial x})|$$

- In order to use this, we require;
  1. Input and output dimensionality is the same
  2. The learnt mapping must be invertible
  3. The determinant of the Jacobian needs to be tractable (and efficient).

- There are several ways to ensure 1-3, here we focus on the Real Non-Volume Preserving (RealNVP) model where we separate z into two disjoint subsets, z_1 and z_2 and then apply two neural networks (s_θ, m_θ):

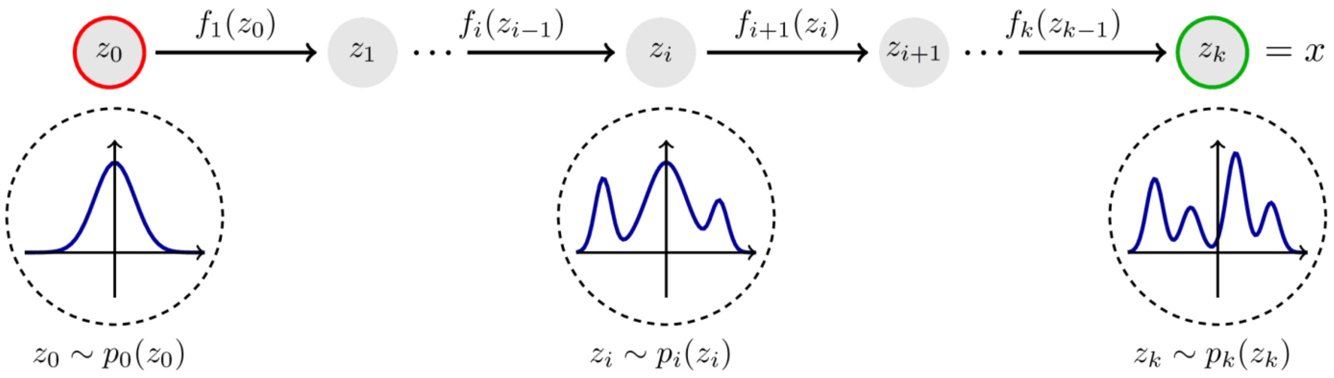
$$x_1 = z_1,$$
$$x_2 = e^{s_\theta(z_1)} + m_\theta(z_1)$$

Normalising flows

› Stacking layers of transforms leads us to a loss function that is allows us to explicitly minimise the negative log-likelihood of our input data (D):

$$\log p_x(x;\theta) = \log p_z(z_0) - \sum_{i=1}^{k} \log \left| \det \frac{df_i}{dz_{i-1}} \right|,$$

$$\mathcal{L}(D) = -\frac{1}{D} \sum \log p_x$$



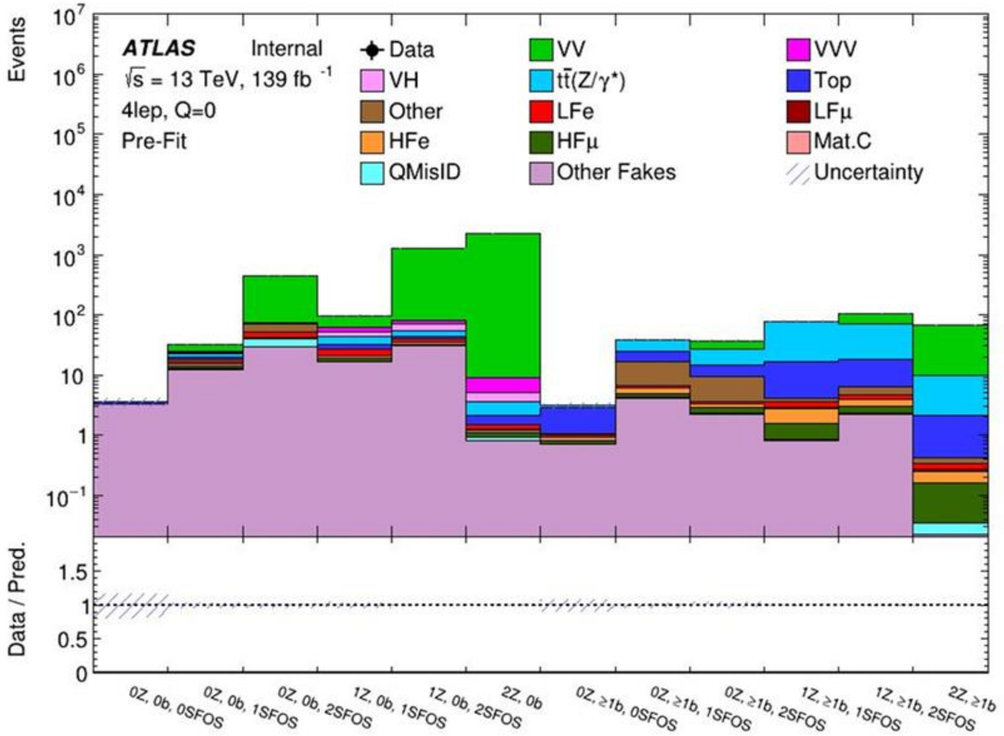
Base density

Transformed density

Further reading here.

Model-independent multi-lepton analysis

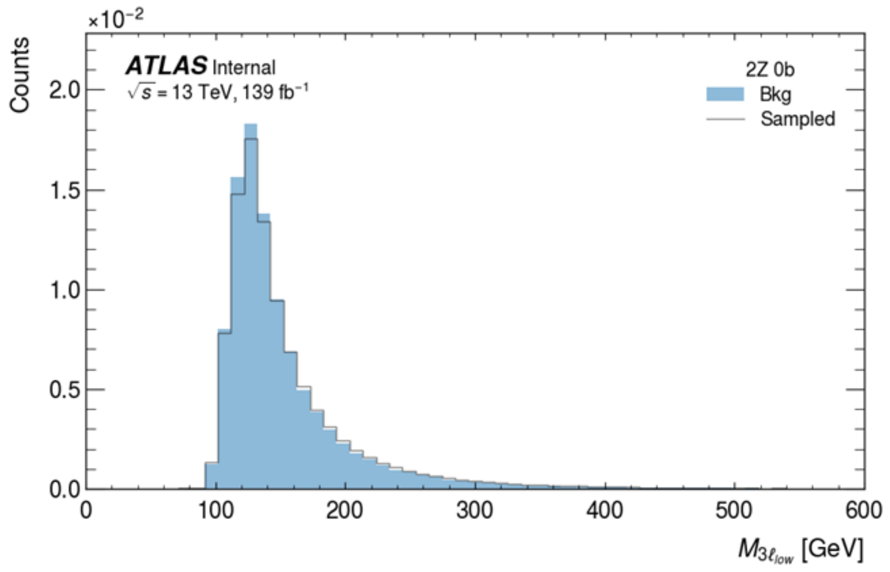
› Search for new physics in events with ≥4 light leptons (e, μ);



› Gives us a large scope for potential BSM models, as heavy resonances can easily decay through chains that produce high lepton multiplicities.

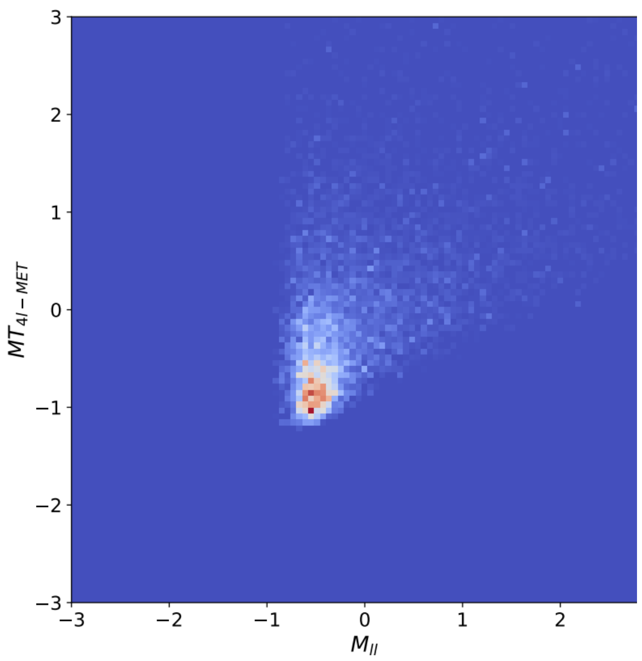
Using normalising flows

› We then train one normalising flow per region, training on our simulated MC background only, before evaluating on signal and background.

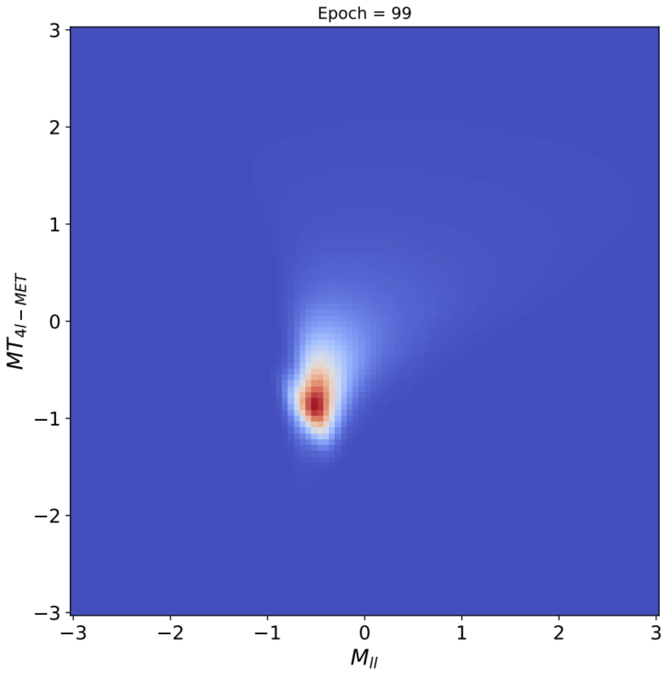


Since the normalising flow is a generative model, we can check the learnt distribution by sampling.

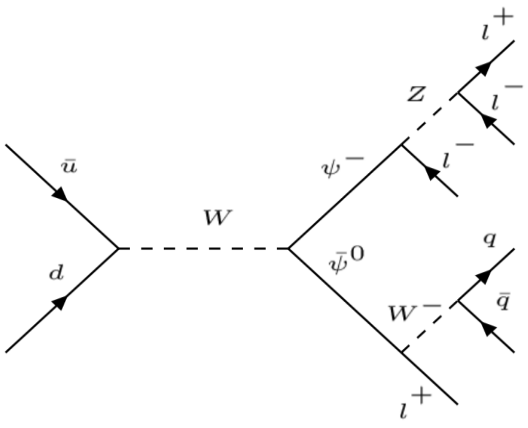
Target density



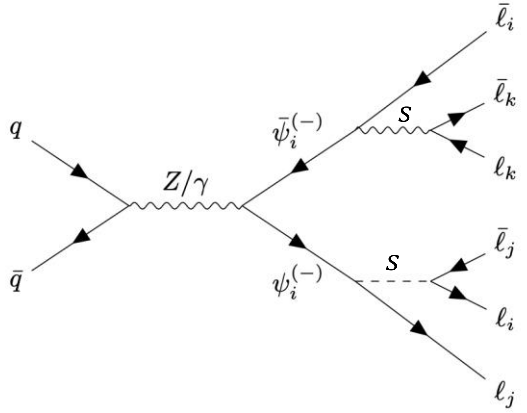
Learned density



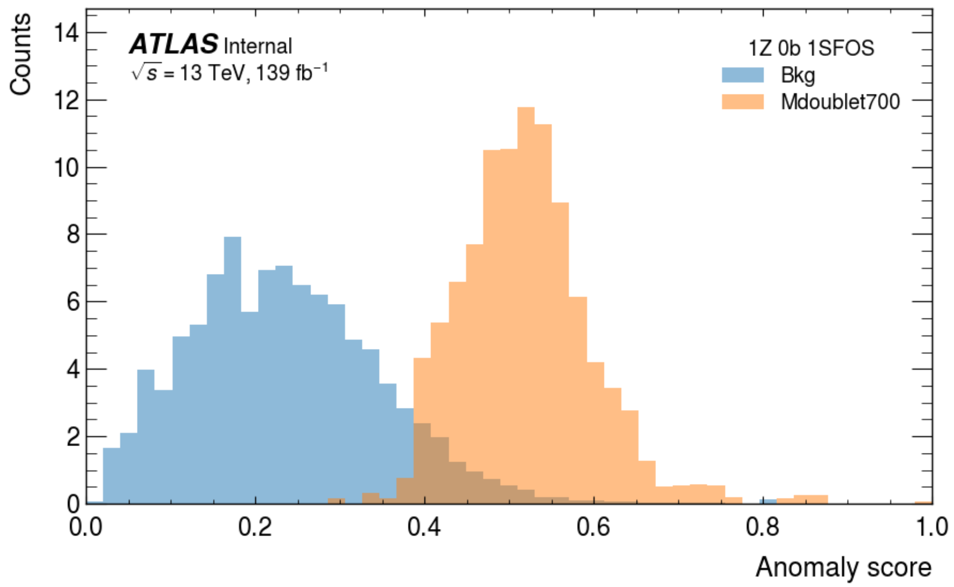
Example signal models:

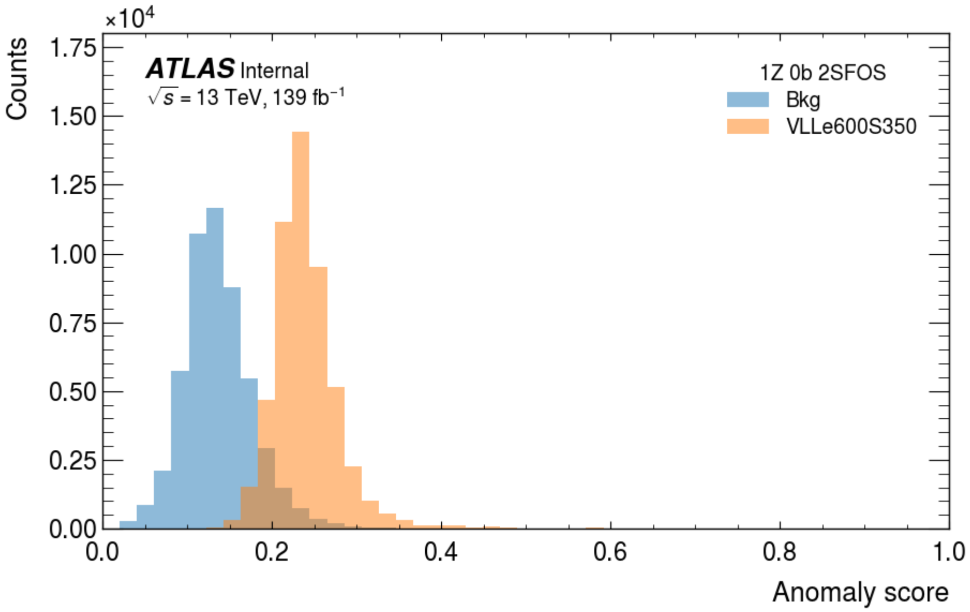


VLL decaying through W/Z/h.



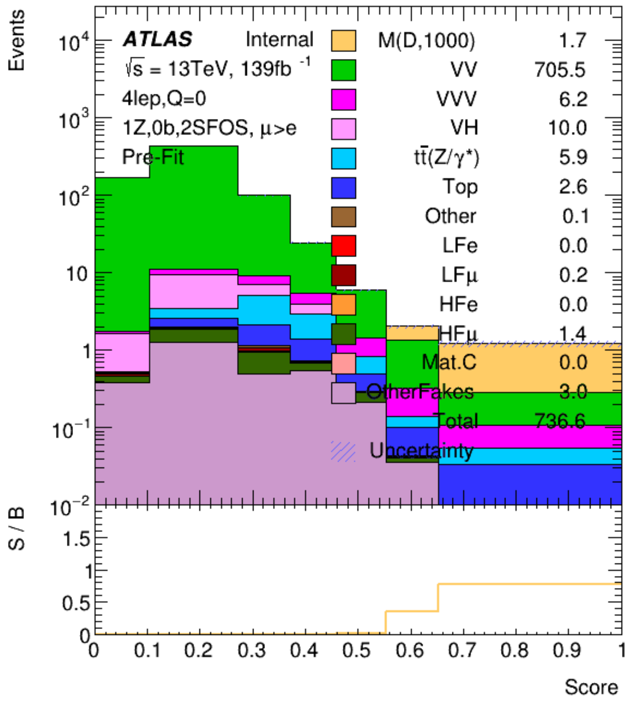
VLL decaying through BSM scalar S.

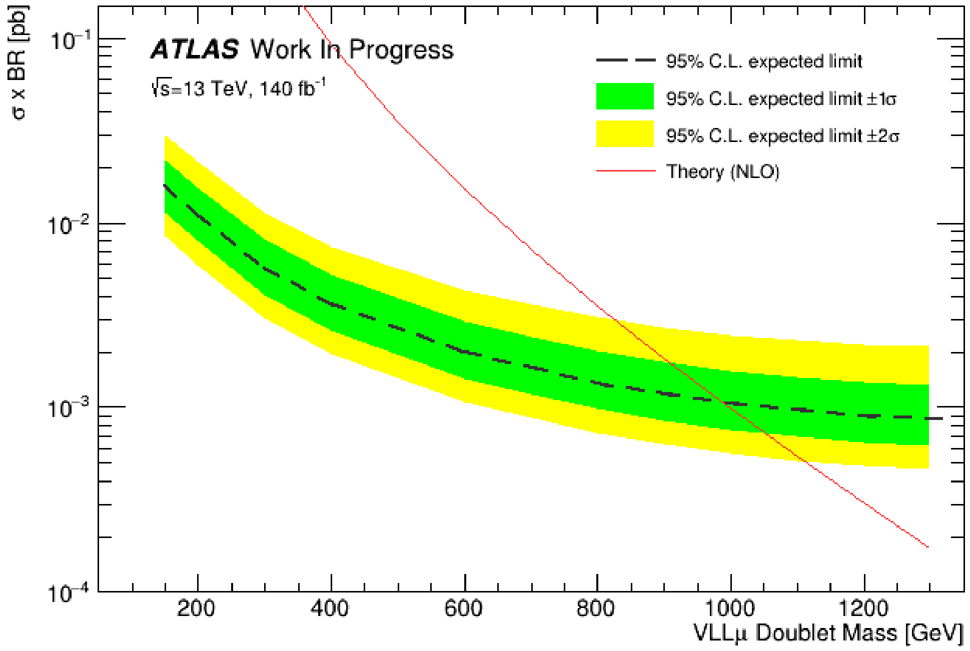




Sensitivity plots

› We can then use the anomaly score as a discriminating variable, which we can fit to and produce model-dependent limits.

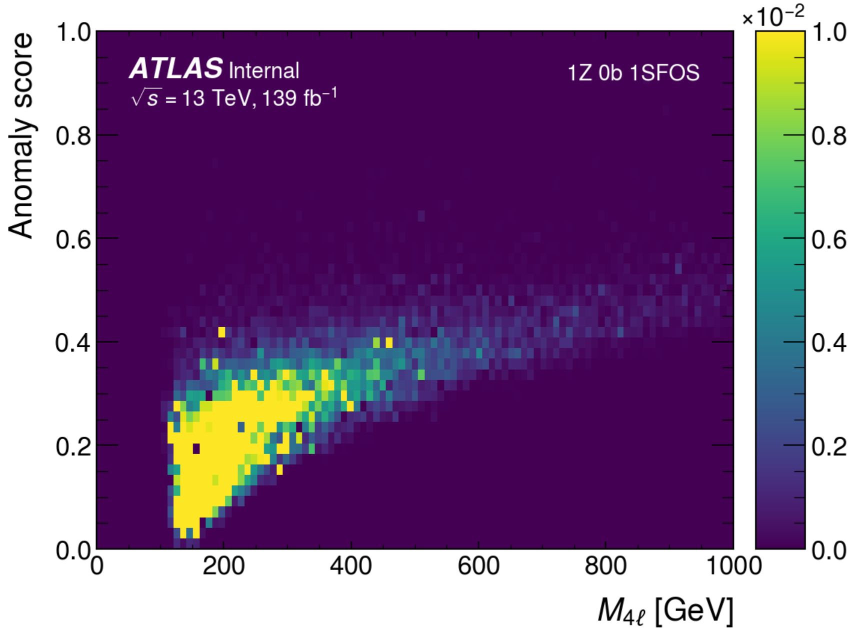


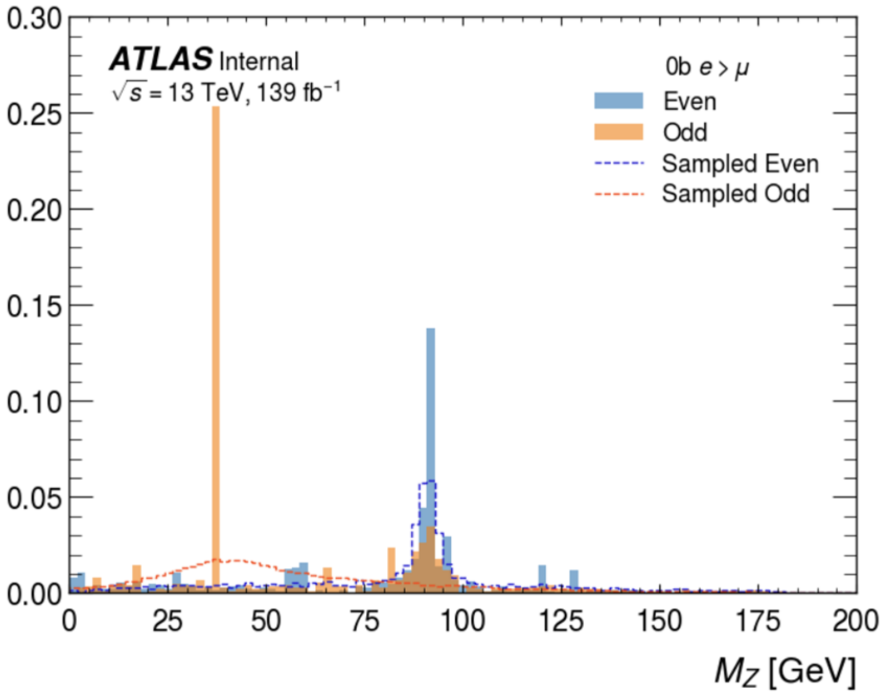


› We can drastically reduce the background count in a model-independent way, and produce competitive limits with the model-dependent search (although not better).

Potential challenges

- Only searches for ‘low-probability’ models, what about excesses of background-like events?
- Reduced expressive-ness, and trouble with discrete inputs.
- Overfitting to low-statistics:
- Difficult to explain the model’s choice of low-probability. Motivates our choices of using physically-motivated input variables.





Summary

- Anomaly detection provides a way to cover the vast phase space of potential new BSM models.

- We can use normalising flows to explicitly learn the probability density of the background.

- We can use them to produce sensitive searches with a wide scope, significantly reducing the background level while remaining signal-agnostic.

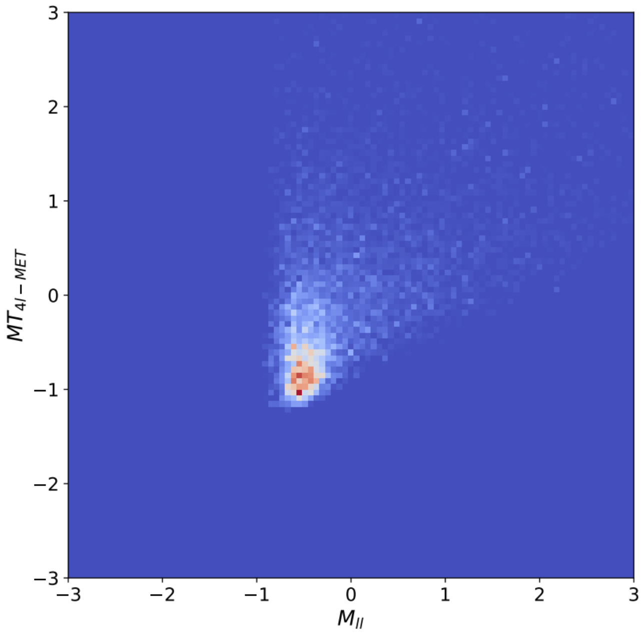
- Several challenges remain with this technique, and the sensitivity is still lower than the ideal scenario, meaning further improvements to model-agnostic searches are still out there.

Backup

IFAE Institut de Física d'Altes Energies
EXCELENCIA SEVERO OCHOA
BIST Barcelona Institute of Science and Technology
ATLAS EXPERIMENT
CERN

Normalising flow trained on background

Target density



Learned density