Challenges of anomaly detection with LHC data

Inês Ochoa

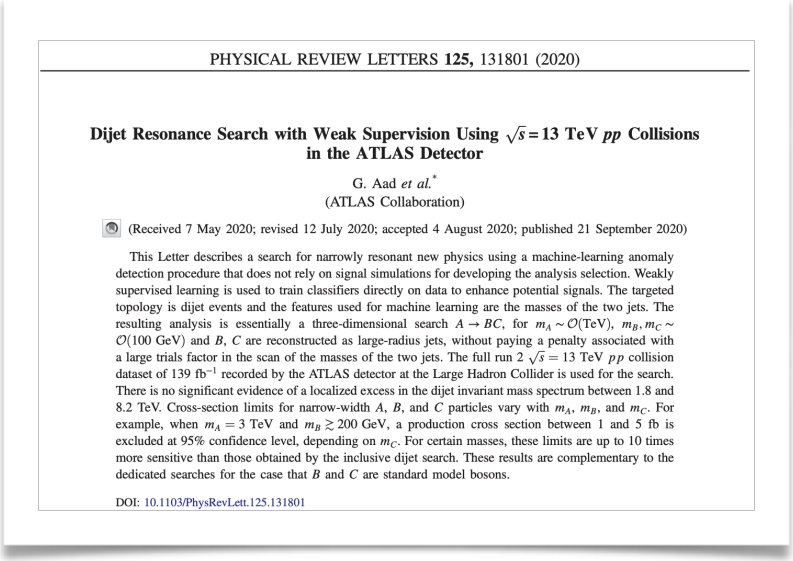
PHYSTAT-Anomalies
May 25th, 2022

Overview

- The goal of this talk is to discuss some of the practical challenges, limitations and assumptions when doing anomaly detection with *actual LHC data*.

- I will consider the *dijet resonance search via weak supervision* by ATLAS to highlight these challenges.

- See talks by Ben, Gregor and Sasha for a wider coverage of anomaly detection methodology.



PHYSICAL REVIEW LETTERS **125,** 131801 (2020)

Dijet Resonance Search with Weak Supervision Using $\sqrt{s} = 13$ TeV pp Collisions in the ATLAS Detector

G. Aad *et al.*[*]
(ATLAS Collaboration)

(Received 7 May 2020; revised 12 July 2020; accepted 4 August 2020; published 21 September 2020)

This Letter describes a search for narrowly resonant new physics using a machine-learning anomaly detection procedure that does not rely on signal simulations for developing the analysis selection. Weakly supervised learning is used to train classifiers directly on data to enhance potential signals. The targeted topology is dijet events and the features used for machine learning are the masses of the two jets. The resulting analysis is essentially a three-dimensional search $A \to BC$, for $m_A \sim \mathcal{O}(\text{TeV})$, $m_B, m_C \sim \mathcal{O}(100\ \text{GeV})$ and B, C are reconstructed as large-radius jets, without paying a penalty associated with a large trials factor in the scan of the masses of the two jets. The full run 2 $\sqrt{s} = 13$ TeV pp collision dataset of 139 fb^{-1} recorded by the ATLAS detector at the Large Hadron Collider is used for the search. There is no significant evidence of a localized excess in the dijet invariant mass spectrum between 1.8 and 8.2 TeV. Cross-section limits for narrow-width A, B, and C particles vary with m_A, m_B, and m_C. For example, when $m_A = 3$ TeV and $m_B \gtrsim 200$ GeV, a production cross section between 1 and 5 fb is excluded at 95% confidence level, depending on m_C. For certain masses, these limits are up to 10 times more sensitive than those obtained by the inclusive dijet search. These results are complementary to the dedicated searches for the case that B and C are standard model bosons.

DOI: 10.1103/PhysRevLett.125.131801

ATLAS Collaboration, Phys.Rev.Lett. 125 (2020) 13, 131801

Outline

- Learning from data
- Classification without labels (CWoLa)
- ATLAS dijet search:
  - Bump-hunting with CWoLa
- Challenges and methodologies
- Final remarks

Most plots from:
ATLAS paper
A. Cukierman's EP-IT Data science seminar

Why learn directly from *data**?

1. Avoid imperfect simulations of physics processes and particle interactions.
   - Minimising background-model dependence, which leads to sub-optimal performance of trained algorithms on data.
2. In searches for new physics, avoid tuning analyses to specific final states or beyond-the-Standard-Model scenarios.
   - Therefore minimising biases or blind-spots in our physics coverage.

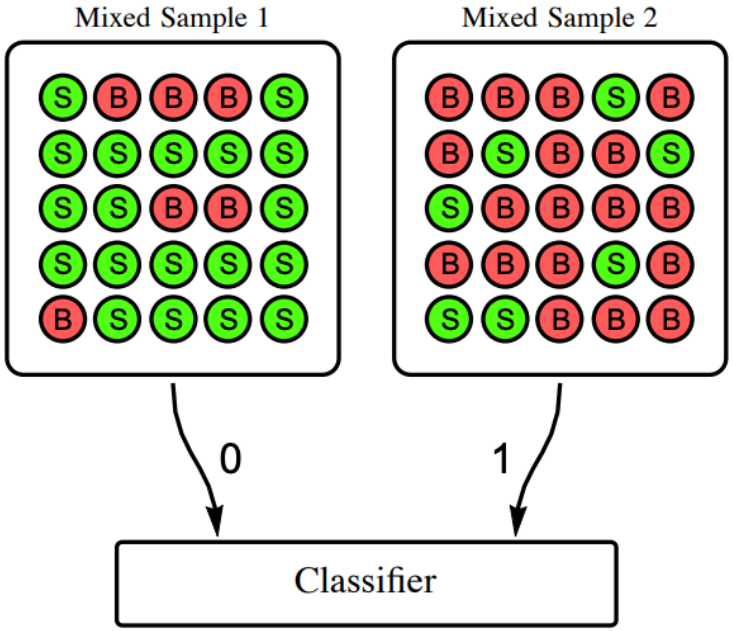
- One obvious drawback: there are no background and signal labels in data.
  - This is where **unsupervised** or **weakly-supervised** learning methods enter.

* With minimal use of simulation.

Classification without labels (CWoLa)

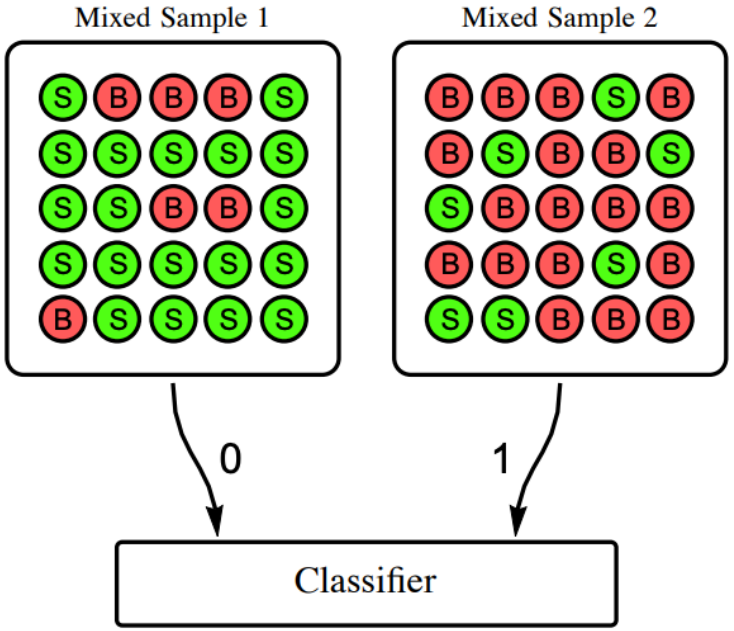
Metodiev, E.M., Nachman, B. & Thaler, J. High Energ. Phys. 2017, 174 (2017)

CWoLa: Classification Without Labels (I)



- Weak supervision: noisy labels.
- Start with two mixed samples which contain both signal and background.
- No knowledge of signal and background labels nor of their fractions in each sample is needed.
- Train a (supervised) classifier to distinguish between samples 1 and 2.

CWoLa: Classification Without Labels (II)



$$\frac{p_1(x)}{p_2(x)} = \frac{f_1 p_S + (1-f_1)p_B}{f_2 p_S + (1-f_2)p_B} = \frac{f_1 \frac{p_S}{p_B} + (1-f_1)}{f_2 \frac{p_S}{p_B} + (1-f_2)}$$

- For $f_2 \ll 1$:

Signal enriched sample

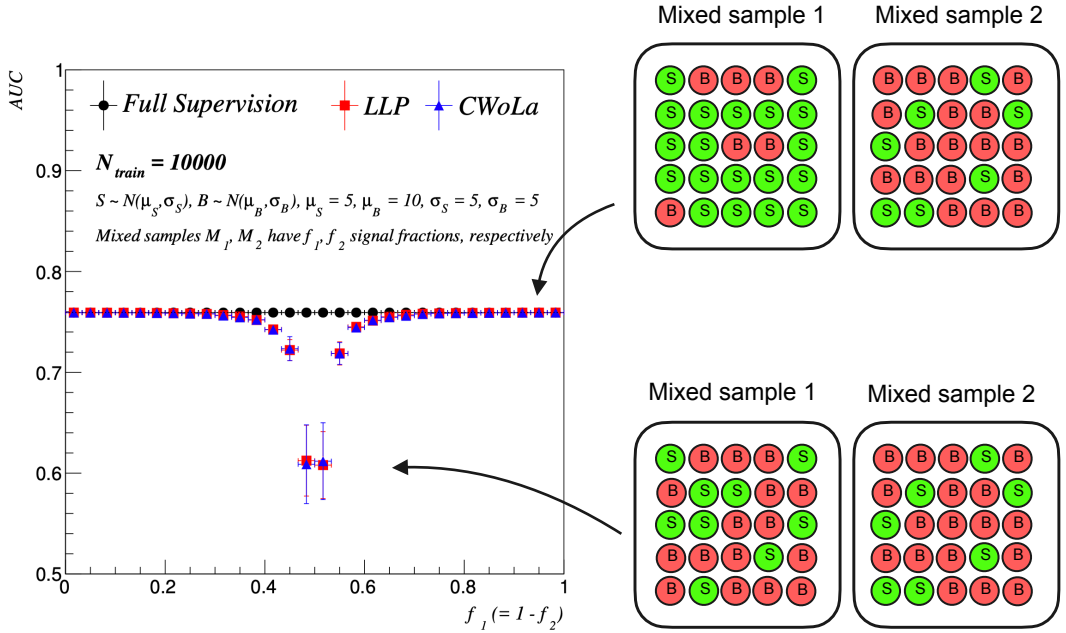
$$\frac{p_1(x)}{p_2(x)} = (1-f_1) + f_1 \frac{p_S}{p_B}$$

Reference: background dominated sample

f_1: signal fraction in sample 1
f_2: signal fraction in sample 2

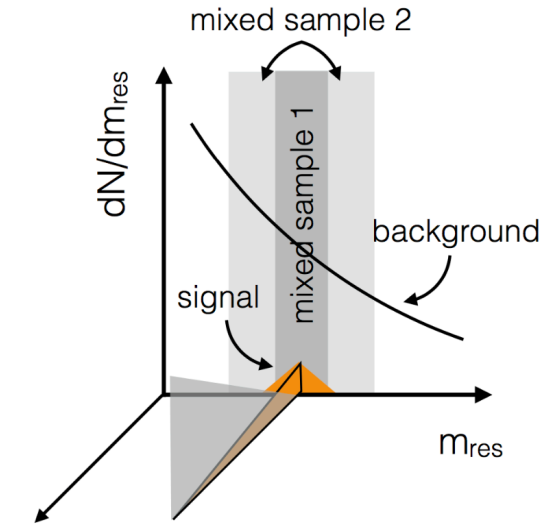
CWoLa: Classification Without Labels (III)

- Assumes no (large) differences between B and S events in samples 1 and 2.
- Does not require any knowledge of f_1 and f_2 for training.
- Requires fractions f_1 and f_2 to be sufficiently different.



CWoLa: Classification Without Labels (IV)

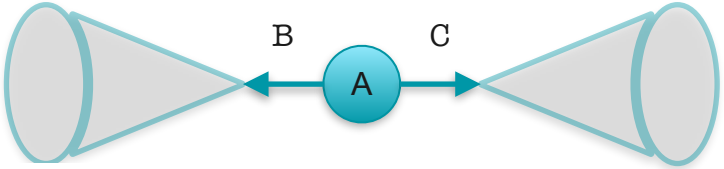
- Classifier trained on feature(s) Y that can increase signal purity.
  - No assumptions on Y other than ~same distribution for background events in the two mixed samples.
  - Confirmed via simulation, theory or control regions.
- In the presence of signal, classifier learns systematic correlations between the two mixed samples and Y.
- In the presence of background-only, classifier should select randomly.

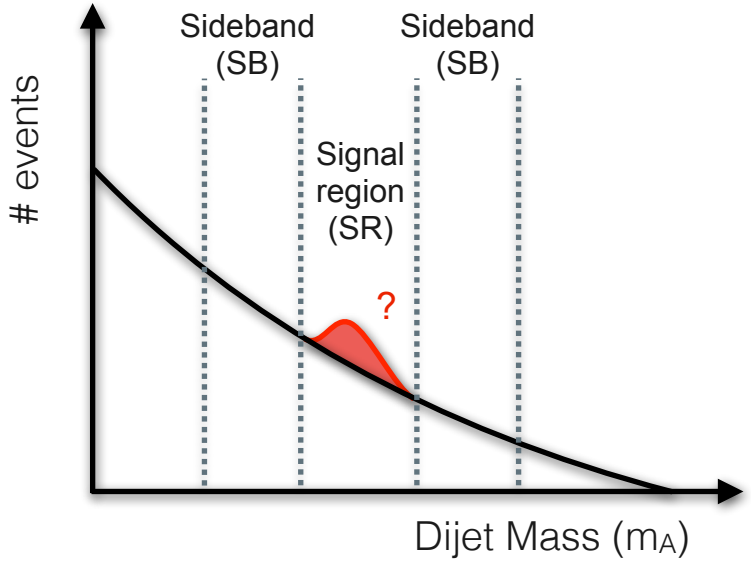


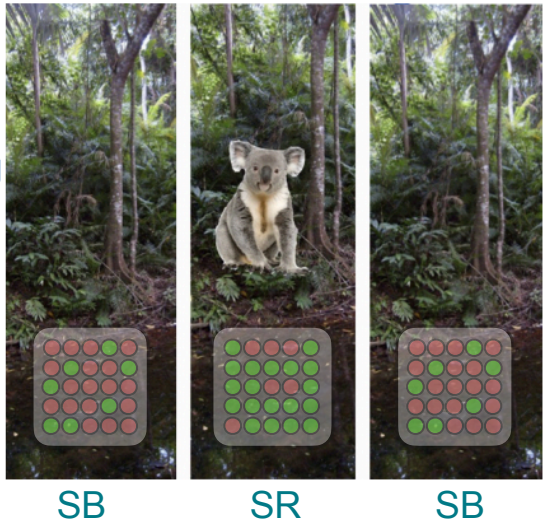
ATLAS dijet search

ATLAS Collaboration, Phys.Rev.Lett. 125 (2020) 13, 131801

Bump-hunting with CWoLa (I)

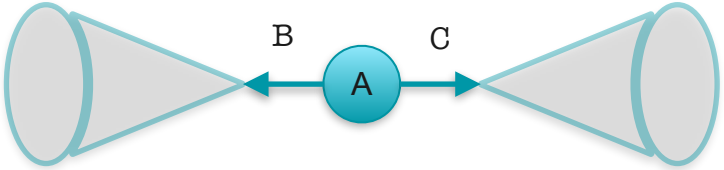


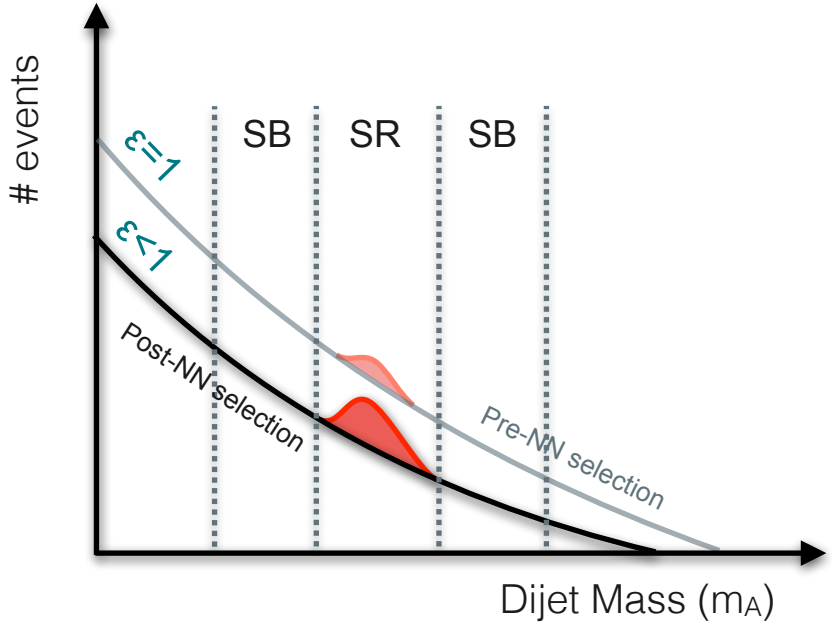




- Signal well-localised in 1 dimension: mass of the dijet system, m_A. ✓
- Features to provide S vs B discrimination: jet masses m_B and m_C. ✓
- Two classes: multijet and signal.

Bump-hunting with CWoLa (II)



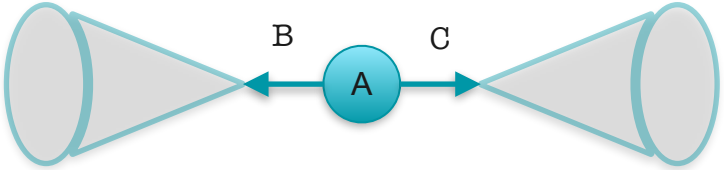


Two main steps:

1. Sensitivity to signal: Train a NN to distinguish between SR and SBs and use it to build a signal-enriched region.
2. Statistical analysis: Fit m_A distribution under the background-only hypothesis.

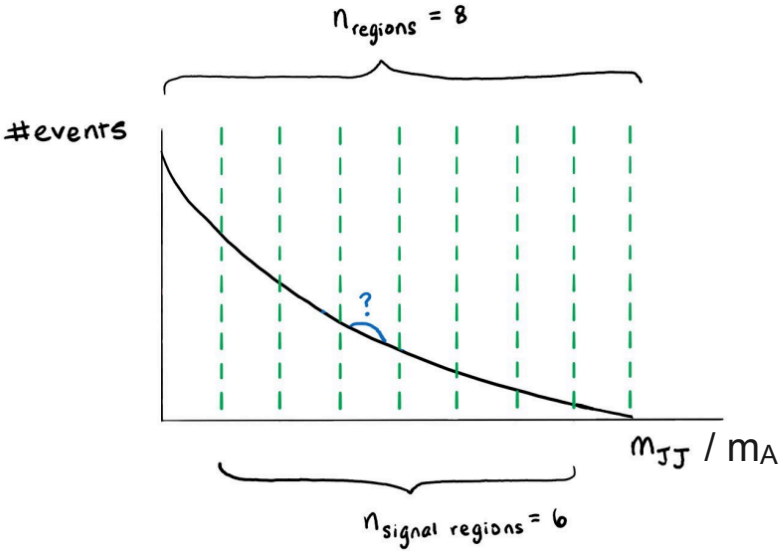
➡ Repeat for different definitions of SR and SB: scan of m_A.

Bump-hunting with CWoLa (III)

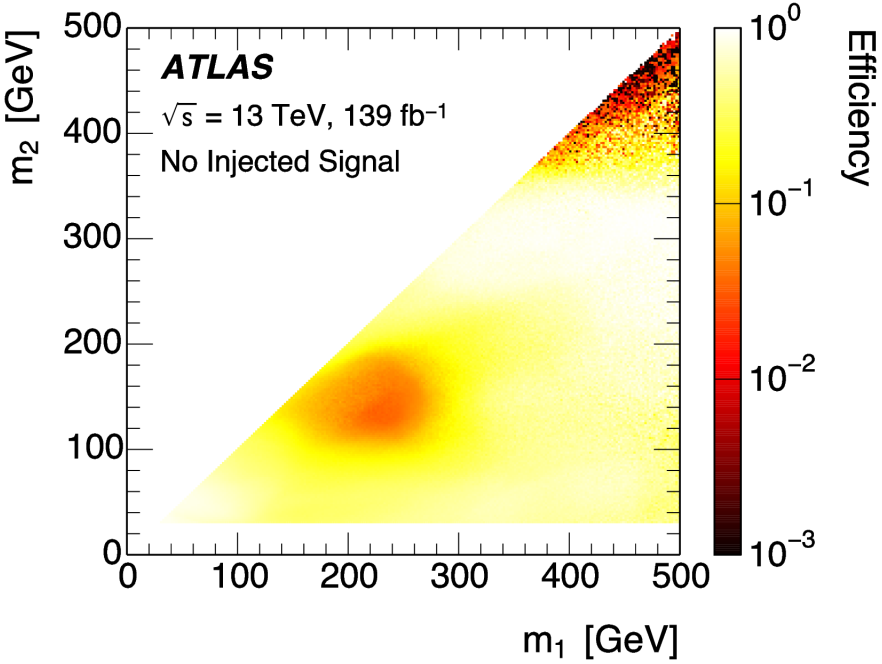


- Dijet mass split into 6 signal regions:
  - Bump-hunt range 2.28-6.81 TeV (fit range: 1.8-8.2 TeV)
  - Window size of 20% m_A (driven by detector resolution for narrow resonances).

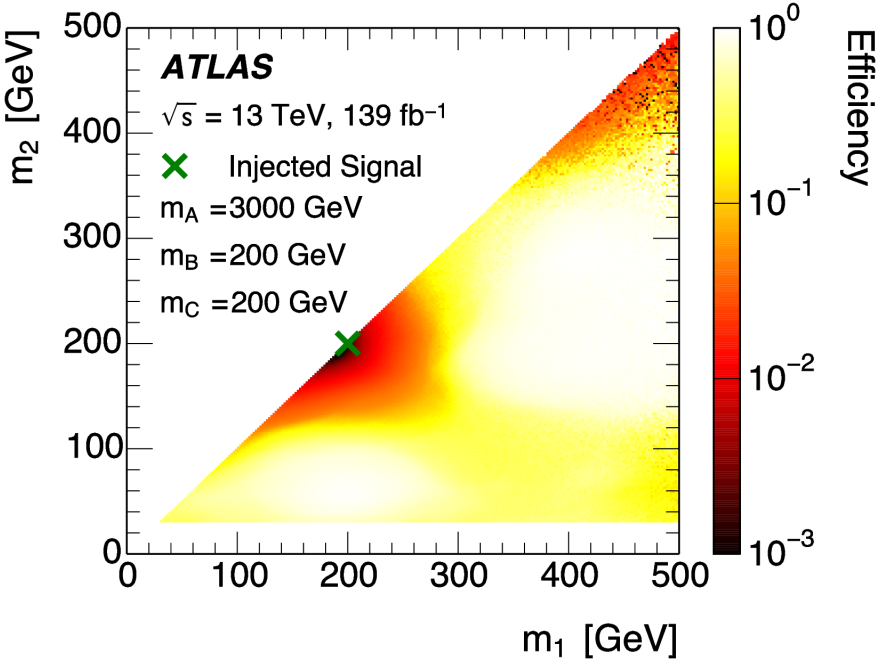
- The efficiency of the NN cut is not optimised, but two fixed signal selections are used:
  - ε = 0.01, 0.1



Bump-hunting with CWoLa (IV)



NN output training directly on data



NN output with injected signal at **X**

Challenges and methodologies

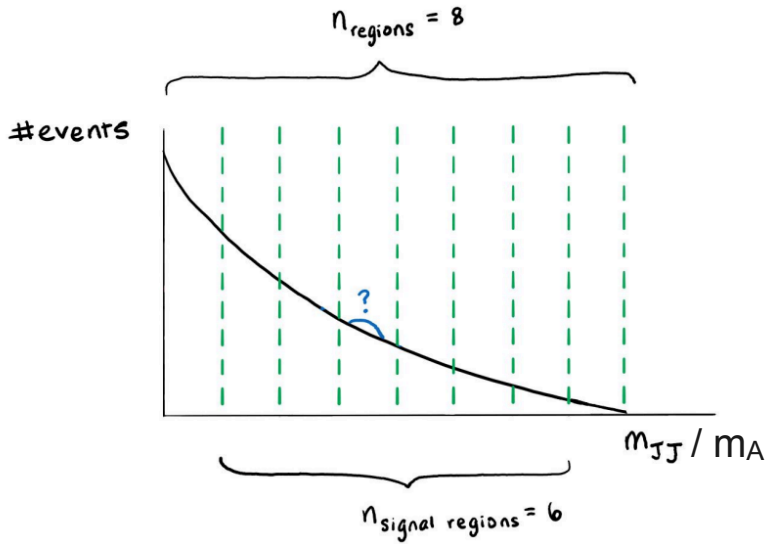
ATLAS Collaboration, Phys.Rev.Lett. 125 (2020) 13, 131801

Collins, J. et al, Phys.Rev.D 99 (2019) 1, 014038

Look elsewhere effect (I)

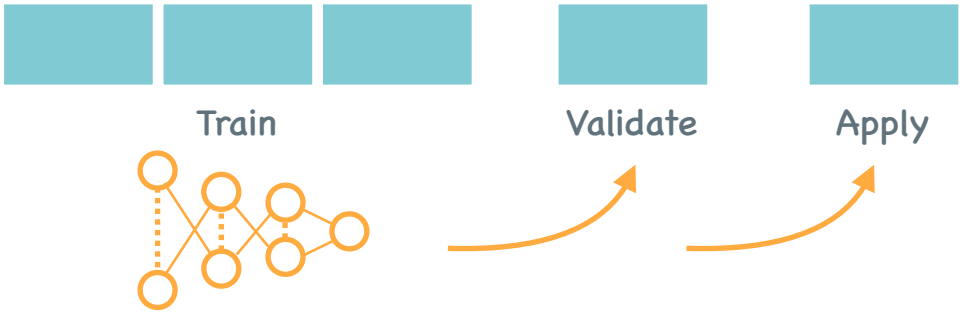
- Trials factors: for a “classic” 3D scan in m_A, m_B, m_C, the trials factor could be very large.
  - Large LEE from scanning over feature space: addressed as described in the next slide.
  - LEE for scan in m_A *not* avoided.
    - Regions are defined ahead of time and are non-overlapping.

- An additional (smaller) factor could come from scanning different thresholds in the NN efficiency ε.
  - Here, two regions with efficiency thresholds (10%, 1%) are sufficiently distant to be considered independent.



Look elsewhere effect (II)

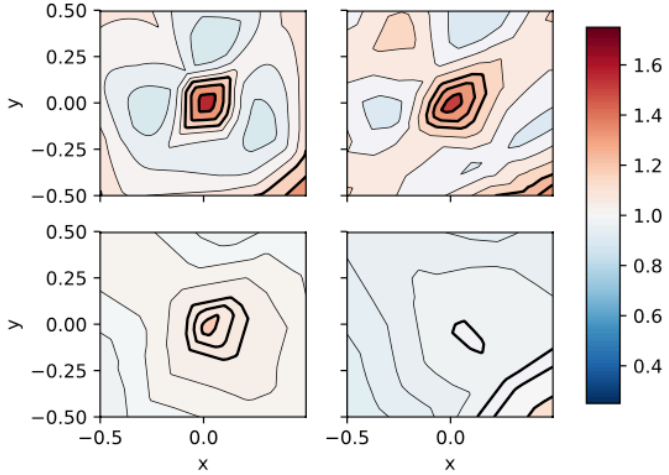
- In order to remove a large LEE from the scan in m_B, m_C, avoid training and evaluating in the same data.
  - Split into *train* and *test* set such that no event is selected with a NN it was trained with.
  - Applying a cut on the NN output is equivalent to selecting the most signal-like 2D bins.
- In the ATLAS dijet analysis, this is addressed with k-fold cross-validation method:

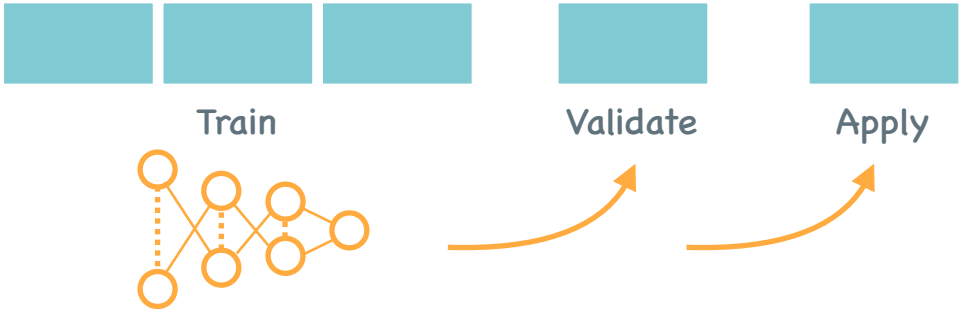


Look elsewhere effect (III)

- If only background is present, any statistical fluctuation in the train dataset is uncorrelated from those in test.
- If a real signal exists, an excess in the train dataset should also be present in the test dataset.
- Training + ensembling multiple classifiers helps mitigate impact of *overfitting* on statistical fluctuations.

4 independent training runs on same data:



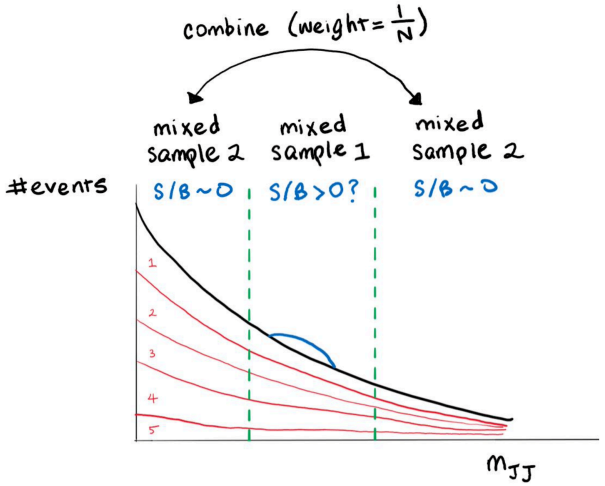


5 x 4 x 3 (random state initialisations) = 60 NNs

Choice of features: decorrelation

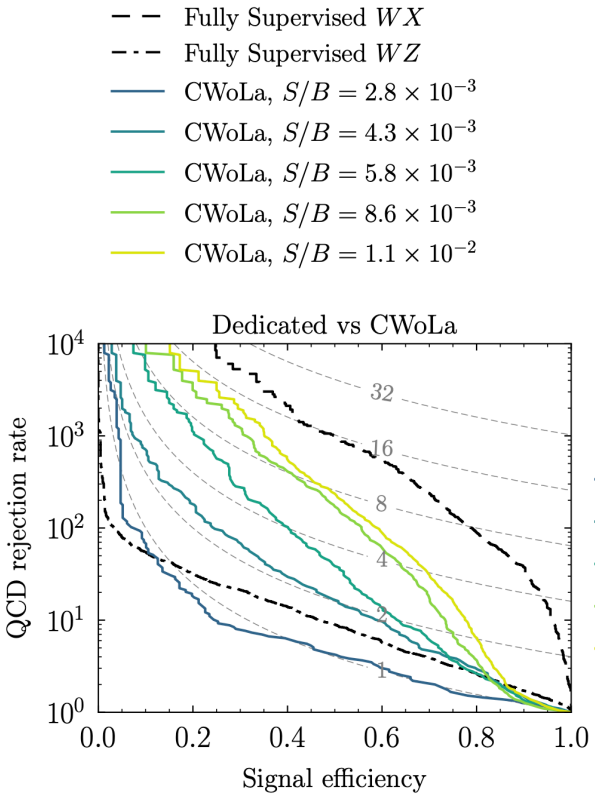
- Method relies on there being no significant differences between background in sidebands and background in the signal region.
  - No fake bumps: if no signal, m_A spectrum should remain smooth after tagging.
  - Features need to vary slowly with m_A: true for m_B and m_C.

- Any correlations are further reduced by:
  - Scaling of 1D $m_J = \{m_B, m_C\}$ distribution in sidebands to the signal region.
  - Restricting m_B, m_C ranges to 30-500 GeV.
  - Combining sidebands and assigning same total weight to each.



Training statistics and S/B

- Difficulty set by relative size of S in the mixed samples and total number of events available for training.
  - Weakly-supervised NN more powerful when local S/B is high.
  - Performance of unsupervised approaches independent of S/B.
- Trivial: limited B statistics impact training performance.
- Choice of SR vs location of the peak:
  - In ATLAS search, signal efficiency unaffected by shifted peak location in most of the mass range.



$pp \to W' \to WX, X \to WW$

Fitting procedure (I)

ATLAS Collaboration, JHEP 09 (2019) 091

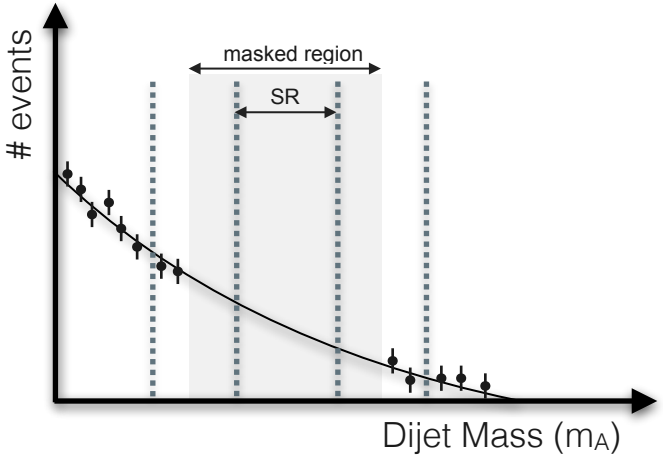
ATLAS Collaboration, Phys.Rev.Lett. 125 (2020) 13, 131801

UA2 Collaboration, Z.Phys.C 49 (1991) 17-28

- Fit m_A spectrum with a parametric function for evaluating B-only hypothesis.
  - Model-independent results: p-value in m_A for each signal region and ε cut.
- Iterative procedure until χ^2 p > 0.05 in sidebands only:

1. $dn/dx = p_1(1-x)^{p_2-\xi_1 p_3} x^{-p_3}$
2. $dn/dx = p_1(1-x)^{p_2-\xi_1 p_3} x^{-p_3+(p_4-\xi_2 p_3-\xi_3 p_2)\log(x)}$
3. $dn/dx = p_1 x^{p_2-\xi_3} e^{-p3x+(p_4-\xi_2 p_3-\xi_3 p_2)x^2}$
4. Sidebands reduced by 400 GeV on both sides, repeat.

- Future challenge: fit with more data or higher ε cuts.
  - Will require non-parametric approaches.



Fit range: 1.8-8.2 TeV

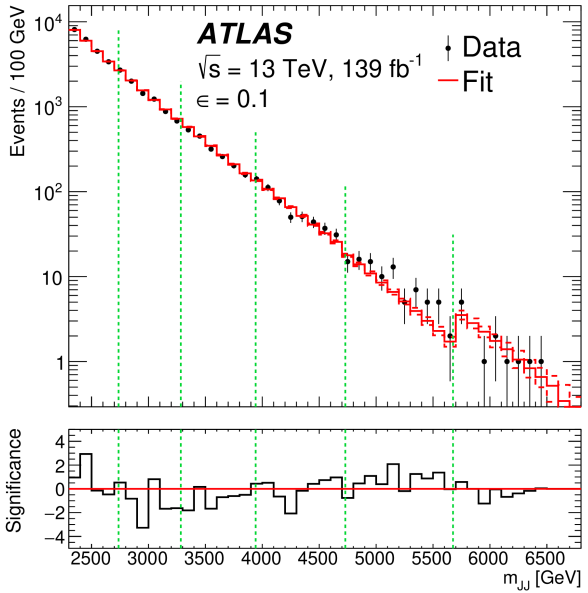
Fitting procedure (II)

ATLAS Collaboration, JHEP 09 (2019) 091

ATLAS Collaboration, Phys.Rev.Lett. 125 (2020) 13, 131801

UA2 Collaboration, Z.Phys.C 49 (1991) 17-28

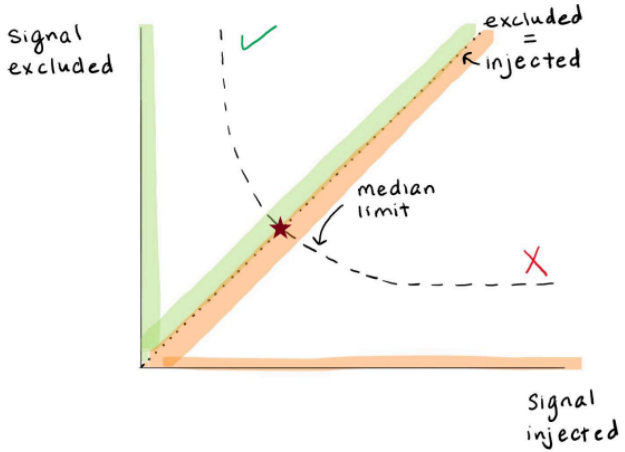
- Fit m_A spectrum with a parametric function for evaluating B-only hypothesis.
  - Model-independent results: p-value in m_A for each signal region and ε cut.
- Iterative procedure until χ^2 p > 0.05 in sidebands only:
  1. $dn/dx = p_1(1-x)^{p_2-\xi_1 p_3} x^{-p_3}$
  2. $dn/dx = p_1(1-x)^{p_2-\xi_1 p_3} x^{-p_3+(p_4-\xi_2 p_3-\xi_3 p_2)\log(x)}$
  3. $dn/dx = p_1 x^{p_2-\xi_3} e^{-p3x+(p_4-\xi_2 p_3-\xi_3 p_2)x^2}$
  4. Sidebands reduced by 400 GeV on both sides, repeat.
- Future challenge: fit with more data or higher ε cuts.
  - Will require non-parametric approaches.



Largest positive deviation around 2500 GeV

Setting limits (I)

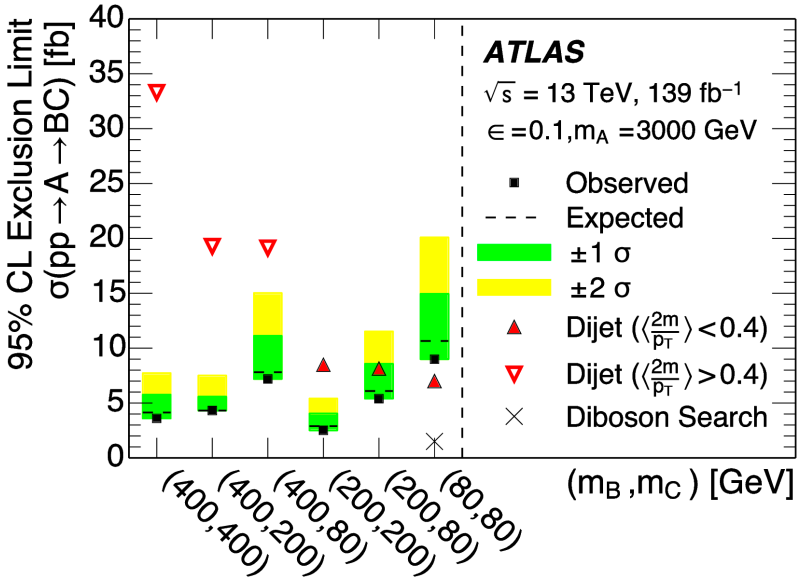
- The classifier's performance depends on the data it sees:
  - Limit depends on the injected signal strength.
  - The learning procedure must be repeated for a new signal and a new cross-section.



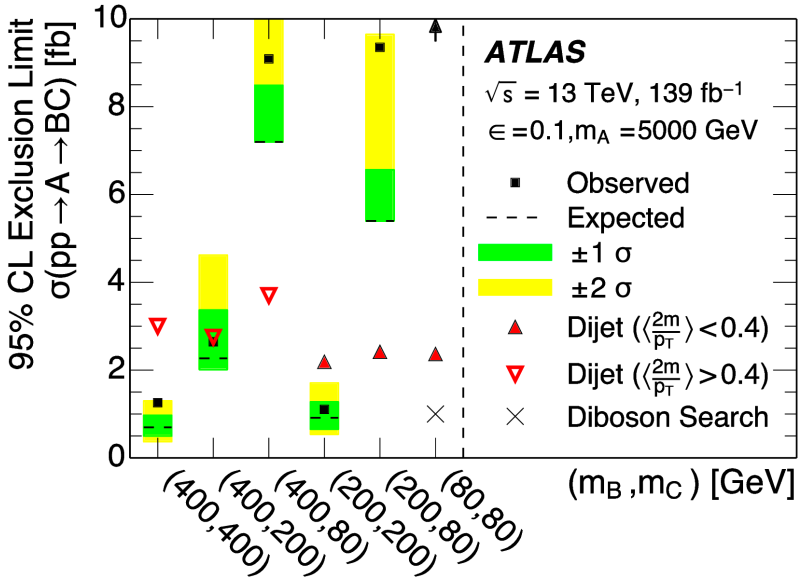
1. Perform coarse scan over injected signal strengths μ.
2. For a given μ, limit is $\max(\sigma_{CL},\sigma_{injected})$:
   - The NN's performance may not be as good if there was less signal than injected.
3. For a given signal, limit is $\min_{\mu}(\max(\sigma_{CL(\mu)},\sigma_{injected(\mu)}))$

For one signal region, 10 injected μ x 5 random samplings of the signal simulation ≃ 3000 NNs

Setting limits (II)



m_A = 3000 GeV



m_A = 5000 GeV

Validation

- Lack of good *control regions* to validate method and assumptions:
  - Whatever the NN learns and we select on depends on the data.

- This search relies on:
  - Simulation.
  - *Validation region* in data, using events with large absolute rapidity difference between the jets.
    - Where S/B ratio is expected to be much lower (true for s-channel resonances).

- More generally, some anomaly detection methods may be suitable to be validated with SM processes.

Computing resources

- Resource intensive: for this result, O(10k) neural networks were trained.

- Additional resources if:
  - Finer grid of signal strength injections for limit setting.
  - More complex scans of m_A or of NN efficiency thresholds.
  - Performing further re-interpretation of results in absence of an excess:
    - RECASTing requires access to data for retraining with injected signals.

Final remarks

Final remarks

- We always need minimal assumptions regarding what new physics is.
- For this method, the key physics starting points are:
  - New physics is a (narrow) resonance:
    - Localized over-density / bump in a given dimension.
  - The background process is smooth in this dimension.
    - Allows us to define signal-enriched and signal-depleted regions.

- Uncovered here:
  - Methods that don't rely on decorrelation between features and m_A (e.g. SALAD, CATHODE, …)
  - Methods using simulation for background model.
  - Non-resonant physics or wide resonances.

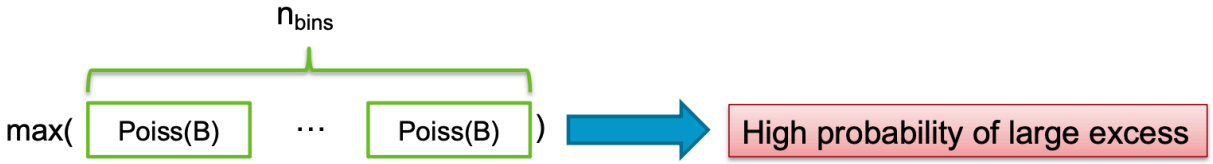
- **Anomaly detection at the LHC will require a combination of methods to fully exploit the data.**

Backup

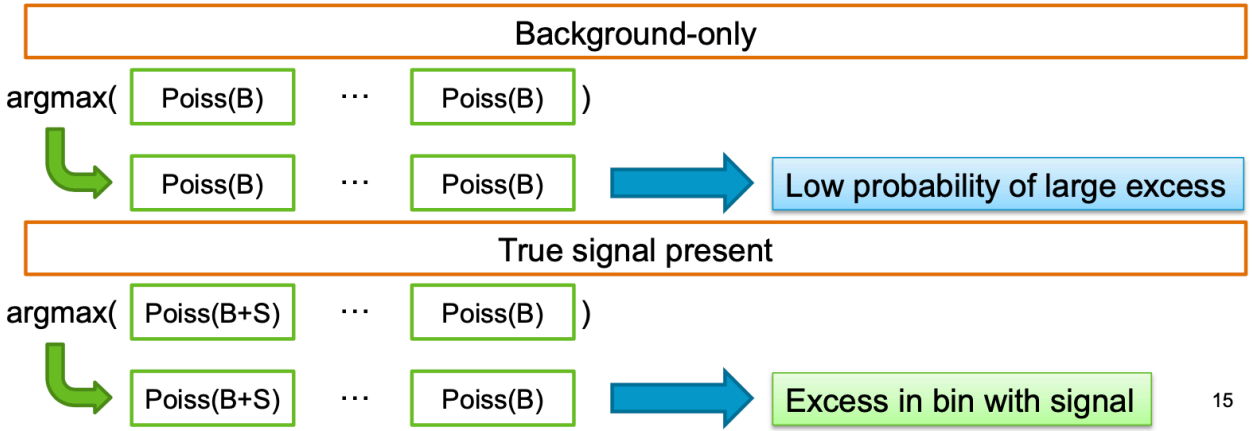
Trials Factors

SLAC

- Trials factor for discovery potential with large numbers of bins



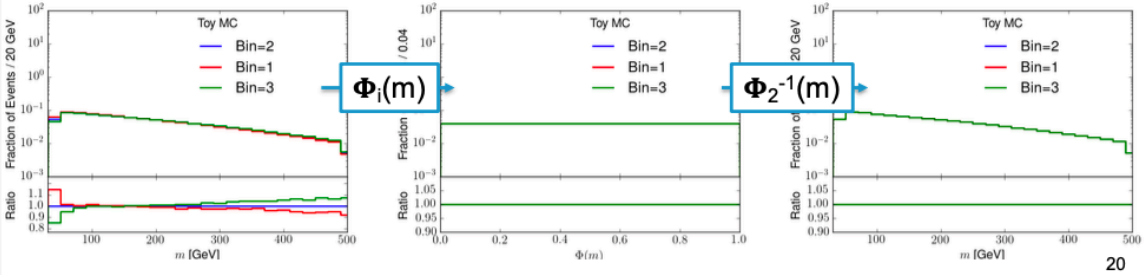
- In 3D m_A,m_B,m_C space, n_{bins} >> 1
- CWoLa hunting (for fixed m_A):

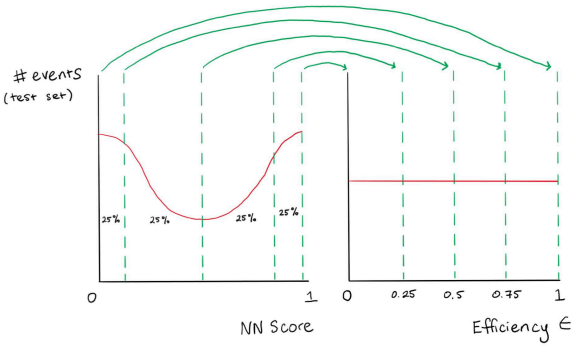


15

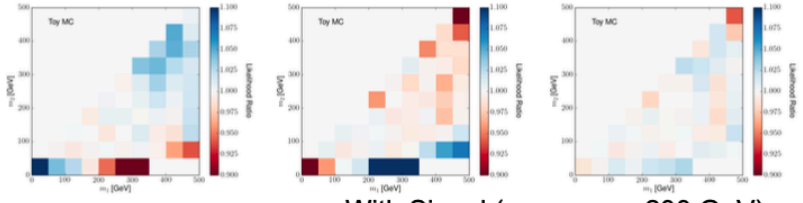
- Decorrelate 1D $m_J = \{m_1, m_2\}$ distribution by percentile scaling
  - Use empirical distribution function
  - $\Phi_i(x)$ = (# of samples in bin i <= x)/(# of samples in bin i)
    - Uniform by definition

A. Cukierman's EP-IT Data science seminar

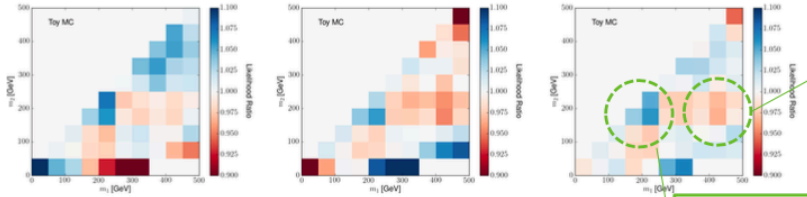




Vs Lower Sideband | Vs Upper Sideband | Vs Comb. Sidebands



With Signal ($m_B = m_C = 200$ GeV):



Deficit – sidebands slightly biased by 1D correction to sig. region

Excess! NN should learn to target