Boosting mono-jet searches with model agnostic machine learning

arXiv2204.11889 TF, Michael Krämer, Maximilian Lipp and Alexander Mück

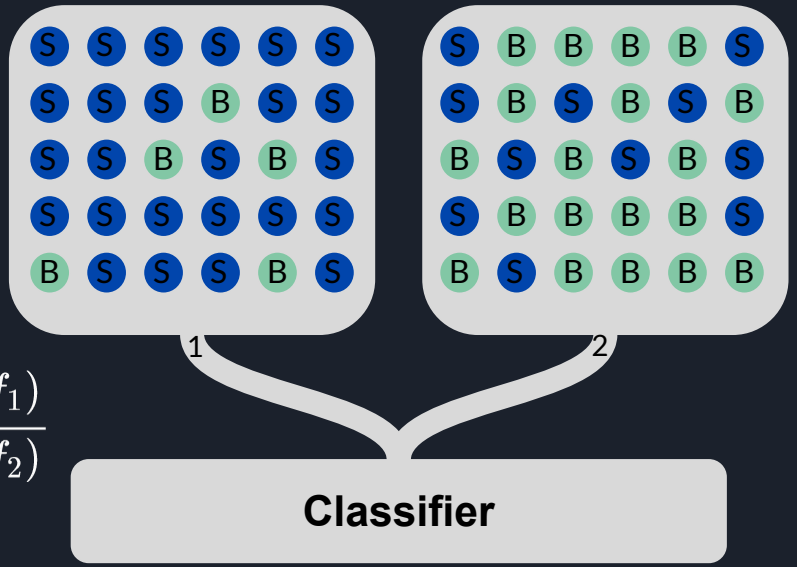
Thorben Finke (finke@physik.rwth-aachen.de)
RWTH Aachen University
Institute for Theoretical Particle Physics and Cosmology

Classification without labels (CWoLa) [arXiv1708.02949]

- Two samples M_1 and M_2 with signal fractions f_1 and f_2 with $f_1 > f_2$
- Optimal classifier for M_1 and M_2 also optimal for signal (S) and background (B)

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1 p_S + (1-f_1) p_B}{f_2 p_S + (1-f_2) p_B} = \frac{f_1 L_{S/B} + (1-f_1)}{f_2 L_{S/B} + (1-f_2)}$$

$$\partial_{L_{S/B}} L_{M_1/M_2} = \frac{(f_1 - f_2)}{\left(f_2 L_{S/B} + 1 - f_2\right)^2} > 0$$



What is our data?

In general: jets as signatures from particle collisions
In particular: a list of properties for every jet constituent

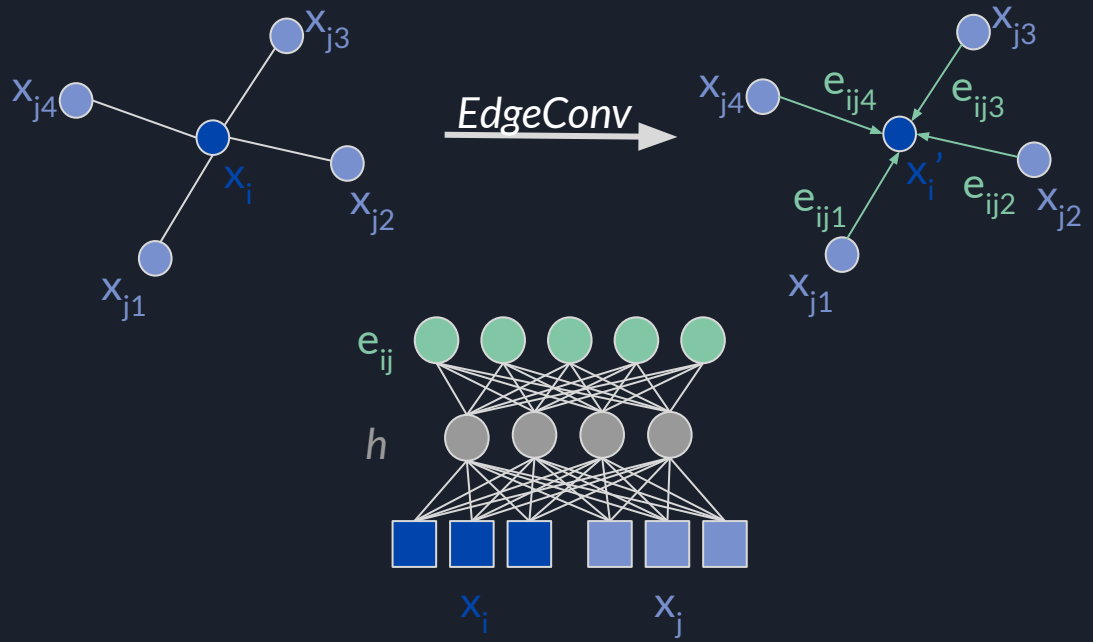
- *M1*: the signal region of the ATLAS mono-jet search (varying f1)
- *M2*: corresponding control regions (f2=0)
  - For simplicity we use the background simulated in the signal region

- Signal: semi-visible jets
  - Produced by a strongly interacting dark sector coupling to the SM
- Background: SM jets that fall into the signal region

DGCNN (ParticleNet) as classifier

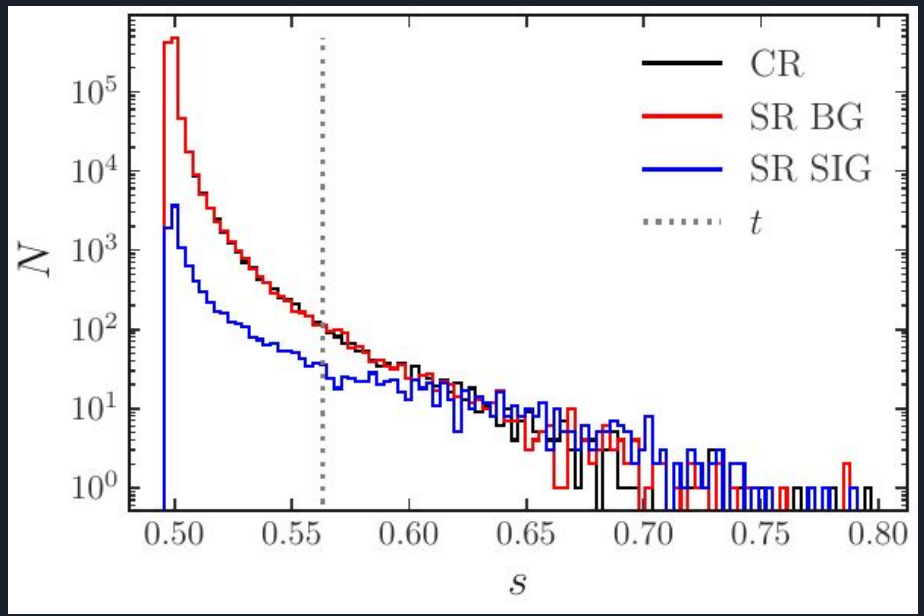
arXiv1902.08570

- Construct knn graph
- Use edge function *h* to calculate edge features e_{ij} for all edges
- Aggregate edge features to obtain node x'
- Repeat for all nodes



Classifier output (f_1 = 1 %)

- Peak at ~0.5
  - Expected from indistinguishable background
- Background in signal and control region follow same distribution
- Choose a threshold based on control region
  - Arbitrarily set to keep 0.1 %
- Beyond threshold similar amount of signal and background

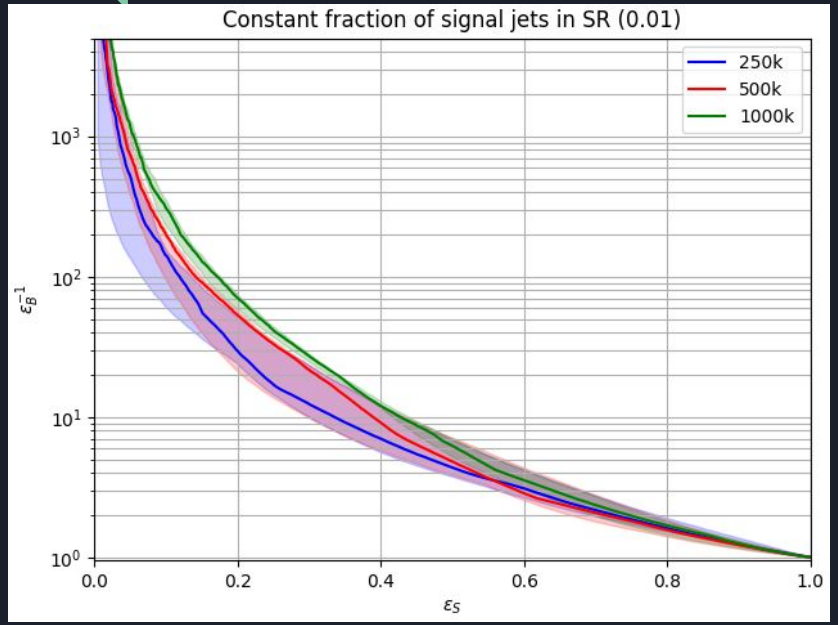


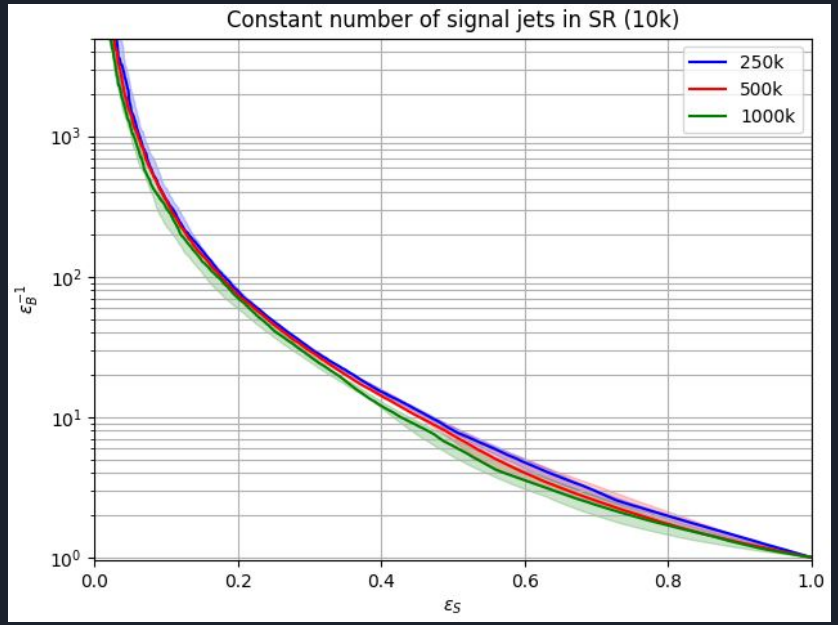
Results using only main background

- Does not introduce fake signal
- High sensitivity beyond current ATLAS limits (<40k events at 95 % CL)

f_1	n^{SR}	n^{SIG}	stat. sign.
0 %	1048	0	1.07
0.6 %	1306	247	6.84
1 %	1666	625	14.89

What limits performance?





- The total number of signal events sets the performance
  - Statistics needed in high dimensions for the signal to stand out of noise

Conclusion

- The CWoLa method allows for enhanced sensitivity without using truth level information
- Sensitive to any difference in control and signal region
  - Can be used to check validity of the control region
- The method will benefit from more statistics as the total number of signal events is more crucial than the fraction
- To improve on this method one needs to make the classifier robust to small signal fractions

BACKUP

The ATLAS mono-jet search

Selection cuts:

- $E_T^{miss} > 200$ GeV
- leading AK4 jet with $p_T > 150$ GeV and $|\eta| < 2.4$
- < 4 additional jets with $p_T > 30$ GeV and $|\eta| < 2.8$
- $\Delta\phi(p_T^{jet}, E_T^{miss}) > 0.4$)
- lepton veto

SM backgrounds:

- Z+jet production with invisibly decaying Z (61 %)
- W+jet production with leptonically decaying W and non-identification of the charged lepton (31 %)
- Top quark production (3.5 %)
- Di-boson production (2 %)

Resulting in $O(10^6)$ background events and a model agnostic limit of 40k additional events at 95 % CL

Results using also additional backgrounds

r_{tt}^{CR}	r_{VV}^{CR}	n^{SR}	n^{DM}
0 %	0 %	4383	223
2.8 %	1.6 %	1465	456
3.5 %	2.0 %	1686	633

- Added 3.5 % top and 2 % di-boson background to 1 % signal in signal region
- Ignoring additional backgrounds in control region leads to wrong signal
- Matching the background perfectly recovers performance from before
- Not matching the background perfectly decreases performance, but does not spoil it completely