Dark Machines – Unsupervised LHC data

ATLAS Group Meeting - 08/06/21

Julien Donini - LPC/Université Clermont Auvergne

Dark Machines: https://darkmachines.org/

- Collective of physicists and data scientists
- Several projects: astrophysics, collider physics, ML …

Unsupervised searches at LHC

- Set of **simulated MC data** for unsupervised BSM searches
- Original **description** of the challenge [2002.12220]
- Dark Machine anomaly score **challenge** [2105.14027]

Generated samples

- **13 TeV** pp collisions: MG5_aMCNLO + Pythia + Delphes (ATLAS)
- **SM** data + several **BSM** scenarios
- About **10/fb** of data, available in csv files (yes...)

Available MC samples

SM processes			
Physics process	Process ID	σ (pb)	N_{tot} ($N_{10\,\text{fb}^{-1}}$)
$pp \to jj(+2j)$	njets	$19718_{H_T>600\text{GeV}}$	415331302 (197179140)
$pp \to l^{\pm}\nu_l(+2j)$	w_jets	$10537_{H_T>100\text{GeV}}$	135692164 (105366237)
$pp \to \gamma j(+2j)$	gam_jets	$7927_{H_T>100\text{GeV}}$	123709226 (79268824)
$pp \to l^+l^-(+2j)$	z_jets	$3753_{H_T>100\text{GeV}}$	60076409 (37529592)
$pp \to t\bar{t}(+2j)$	ttbar	541	13590811 (5412187)
$pp \to t+\text{jets}(+2j)$	single_top	130	7223883 (1297142)
$pp \to \bar{t}+\text{jets}(+2j)$	single_topbar	112	7179922 (1116396)
$pp \to W^+W^-(+2j)$	ww	82.1	17740278 (821354)
$pp \to W^{\pm}t(+2j)$	wtop	57.8	5252172 (577541)
$pp \to W^{\pm}\bar{t}(+2j)$	wtopbar	57.8	4723206 (577541)
$pp \to \gamma\gamma(+2j)$	2gam	47.1	17464818 (470656)
$pp \to W^{\pm}\gamma(+2j)$	Wgam	45.1	18633683 (450672)
$pp \to ZW^{\pm}(+2j)$	zw	31.6	13847321 (315781)
$pp \to Z\gamma(+2j)$	Zgam	29.9	15909980 (299439)
$pp \to ZZ(+2j)$	zz	9.91	7118820 (99092)
$pp \to h(+2j)$	single_higgs	1.94	2596158 (19383)
$pp \to t\bar{t}\gamma(+2j)$	ttbarGam	1.55	95217 (15471)
$pp \to t\bar{t}Z$	ttbarZ	0.59	300000 (5874)
$pp \to t\bar{t}h(+1j)$	ttbarHiggs	0.46	200476 (4568)
$pp \to \gamma t(+2j)$	atop	0.39	2776166 (3947)
$pp \to t\bar{t}W^{\pm}$	ttbarW	0.35	279365 (3495)
$pp \to \gamma\bar{t}(+2j)$	atopbar	0.27	4770857 (2707)
$pp \to Zt(+2j)$	ztop	0.26	3213475 (2554)
$pp \to Z\bar{t}(+2j)$	ztopbar	0.15	2741276 (1524)
$pp \to t\bar{t}t\bar{t}$	4top	0.0097	399999 (96)
$pp \to t\bar{t}W^+W^-$	ttbarWW	0.0085	150000 (85)

BSM process
Z' + monojet
$Z' + W/Z$
Z' + single top
Z' in lepton-violating $U(1)_{L_\mu - L_\tau}$
$\not{R}$-SUSY stop-stop
$\not{R}$-SUSY squark-squark
SUSY gluino-gluino
SUSY stop-stop
SUSY squark-squark
SUSY chargino-neutralino
SUSY chargino-chargino

Tested algorithms

Abbreviation	Algorithm	Section	Hyperparameters	# Submitted
SimpleAE	Autoencoders	4.1	Tab. 6	1
VAEs	Variational Autoencoders	4.2	Tab. 7	140
DeepSetVAE	Deep Set Variational Autoencoders	4.3	Tab. 8	4
ConvVAE (NoF)	Convolutional Variational Autoencoders	4.4	Tab. 9	1
Planar	ConvVAE+Planar Flows	4.5.1	Tab. 10	1
SNF	ConvVAE+Sylvester Normalizing Flows	4.5.2	Tab. 11	3
IAF	ConvVAE+Inverse Autoregressive Flows	4.5.3	Tab. 12	1
ConvF	ConvVAE+Convolutional Normalizing Flows	4.5.4	Tab. 13	1
CNN	Convolutional (β)VAE	4.6		2
KDE	Kernel Density Estimation	4.7	Tab. 14	36
Flow	Spline autoregressive flow	4.8	Tab. 15	2
Deep SVDD	Deep SVDD	4.9	Tab. 16 & 17	80
Combined (Deep SVDD & Flow)	Spline autoregressive flow with Deep SVDD	4.10		8
DAGMM	Deep Autoencoding Gaussian Mixture Model	4.11	Tab. 19	384
ALAD	Adversarial Anomaly Detection	4.12	Tab. 21	96
Latent	Anomaly Detection in the Latent Space	4.13	Tab. 22	288

See for details: https://arxiv.org/abs/2105.14027

Samples and data format

Datasets: 3 different sets

- Individual SM and BSM files: **original** dataset (a bit outdated)
- Pre-processed w/ **4 types** of selections: **Hackathon dataset**
- **Secret dataset**

Format: csv files for all datasets containing for each event a line

```
event ID; process ID; event weight; MET; METphi; obj1, E1, pt1, eta1, phi1;
obj2, E2, pt2, eta2, phi2; ...
```

Particle **types**: (b)-jets, electron, muons, gamma

Example: this is a ttbar+γ event with 1 b-jets, 2-jets, 1 gamma

57;ttbarGam;1;66814.7;0.820827;**j**,807565,119079,2.6017,3.02583;**b**,204221,49828.9,2.0879,
-0.469806; **j**,120426,45285.9,1.6327,-1.18;**g**,52303.8,22421.9,1.4907,-0.156513;

Variable event **size** and **structure** (i.e line length) for each events

Parsing files

Parsing can be ~easily performed using **regular expressions**

- See this very nice tutorial: https://docs.python.org/3/howto/regex.html

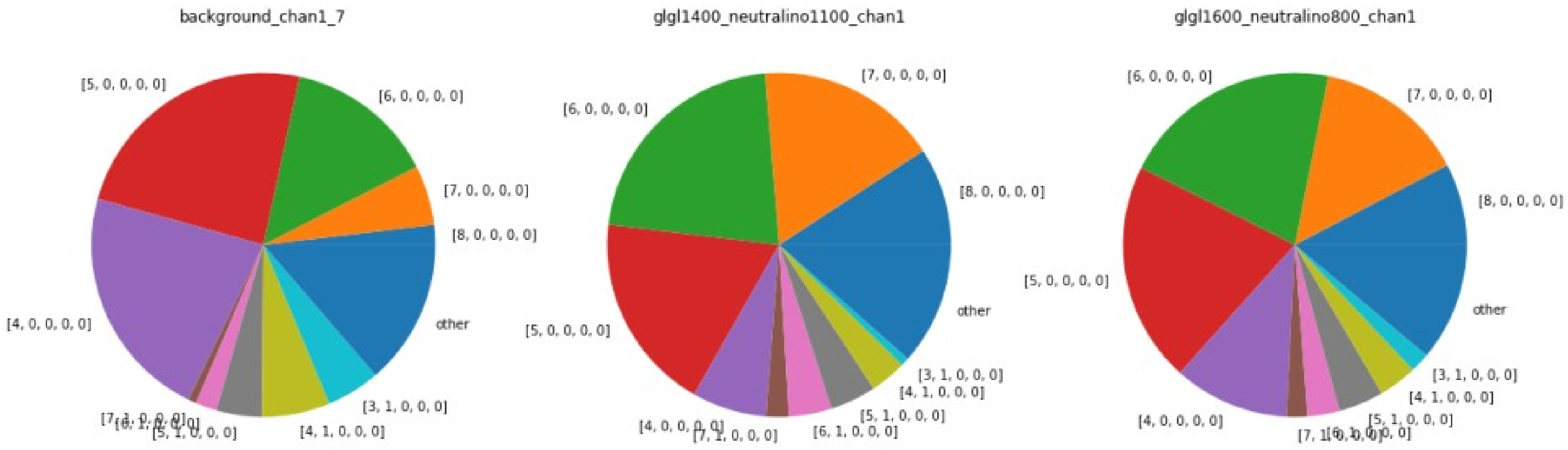
```
# use https://regexper.com to visualise these if required
rx_dict = {
    #'jets': re.compile(r'j,(?P<jets>.*,.*,.*,.*;)\n')
    'header': re.compile(r'(?P<id>\d+);(?P<name>.*?);(?P<weight>.*?);(?P<MET>.*?);(?P<METphi>.*?);'),
    'jets': re.compile('j,(.*?),(.*?),(.*?),(.*?);'),
    'b-jets': re.compile('b,(.*?),(.*?),(.*?),(.*?);'),
    'elec': re.compile('e.,(.*?),(.*?),(.*?),(.*?);'),
    'muons': re.compile('m.,(.*?),(.*?),(.*?),(.*?);'),
    'gamma': re.compile('g,(.*?),(.*?),(.*?),(.*?);')
}
#rx_dict

def get_header(name,line):
    key = rx_dict.get(name)
    if key:
        header = rx_dict[name].match(line)
        header_new = [int(header.group('id')),header.group('name'),float(header.group('weight')),float(header.group
        return header_new
    else:
        print("Warning: '%s' name not found in dictionary" % name )
    return None


def get_particles(name,line):
    key = rx_dict.get(name)
    npart = 0
    if key:
        particles = rx_dict[name].findall(line)
        if particles:
            part = np.array(particles).astype(np.float)
            npart = part.shape[0]
            return part, npart
    else:
        print("Warning: '%s' name not found in dictionary" % name )
    return None, npart
```

This is how I did it – there are probably smarter ways

Signature selection

Select **any** signatures in data: [jets, b-jets, elec, muons, gamma]

- Dump **4-vectors** for all selected particles
- High level variables with python equivalent of **TLorentzvector** class
  https://github.com/scikit-hep/scikit-hep/blob/master/skhep/math/vectors.py

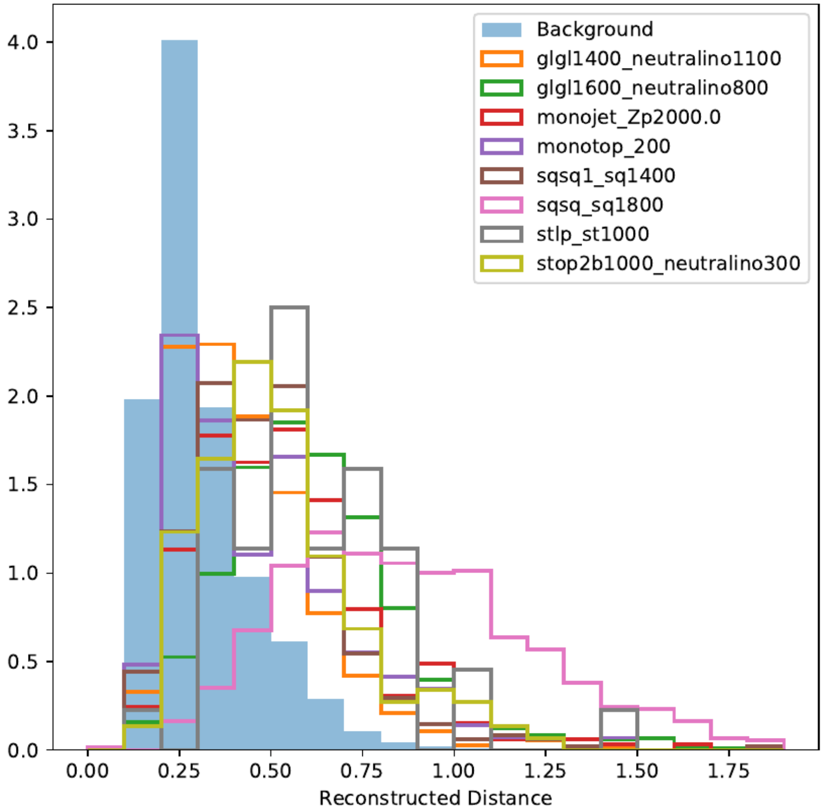


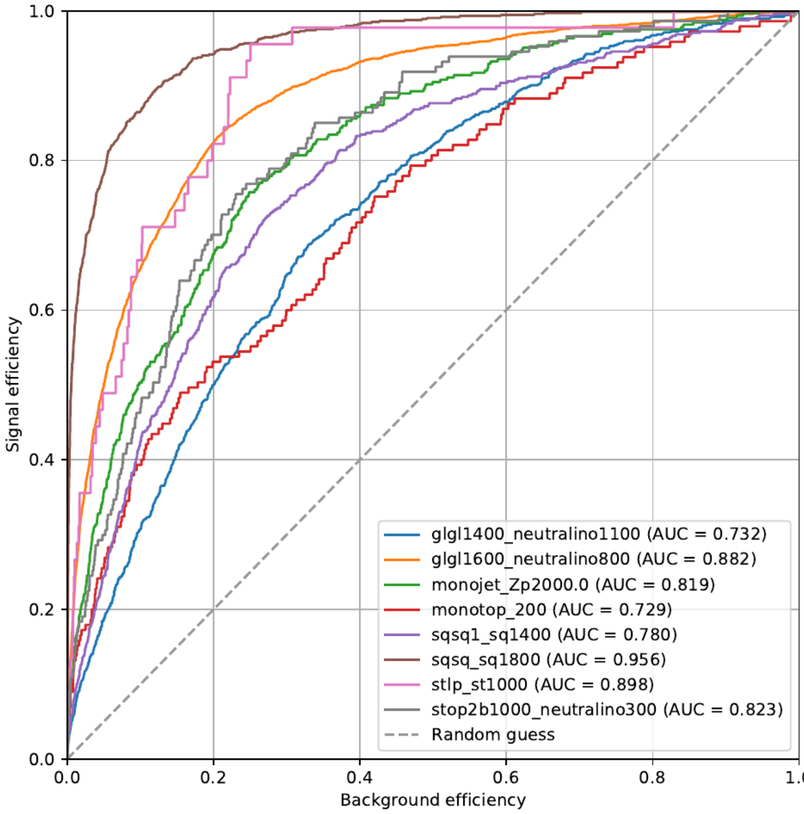
Example of multijet selection for
background and two signal models

Example of approach

Simple **Autoencoder**, trained on **background** events

- Data corresponding to channel 1 of **Hackathon** dataset
- **Preselection**: high H_T, MET and jet multiplicity ($\geq$ 4 jets)



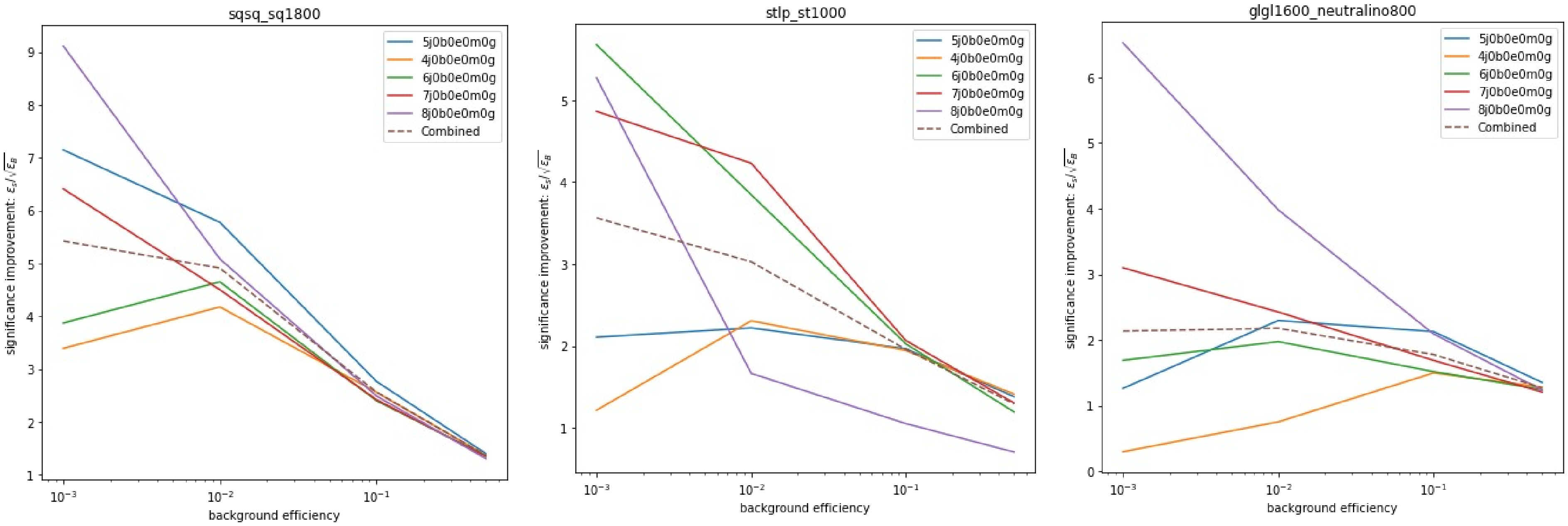


5-jet channel

Significance improvement

Signal **significance improvement** $\frac{\epsilon_S}{\sqrt{\epsilon_B}}$ vs background efficiency

Here for 3 different signals and $\epsilon_B =$ 0.1%, 1%, 10%, 50%



5-8 jets channels and combined channels

Conclusions

Dark Machine LHC is a rich and complete dataset for unsupervised learning

- First **explorations** with several ML approaches described in 2105.14027

Started to look at this datasets (parsing, signature selections, simple AE…)

Still some stones left **unturned**:

- No **data-driven** approach
- No real **multi-channel** combination
- Simplistic significance calculation
- ...
- Good **playground** for Louis and Ioan's methods ;-)

BSM processes

BSM process	Channel 1	Channel 2a	Channel 2b	Channel 3
Z' + monojet	$\times$	$\times$		$\times$
$Z' + W/Z$				$\times$
Z' + single top	$\times$			$\times$
Z' in lepton-violating $U(1)_{L_\mu - L_\tau}$		$\times$	$\times$	
$\not{R}$-SUSY stop-stop	$\times$		$\times$	$\times$
$\not{R}$-SUSY squark-squark	$\times$			$\times$
SUSY gluino-gluino	$\times$	$\times$	$\times$	$\times$
SUSY stop-stop	$\times$			$\times$
SUSY squark-squark	$\times$			$\times$
SUSY chargino-neutralino		$\times$	$\times$	
SUSY chargino-chargino			$\times$	

- Channel 1:

$$H_T \geq 600 \text{ GeV}, \quad E_T^{\text{miss}} \geq 200 \text{ GeV}, \quad E_T^{\text{miss}}/H_T \geq 0.2, \tag{2.2}$$

with at least four (b)-jets with $p_T > 50$ GeV, and one (b)-jet with $p_T > 200$ GeV.

- Channel 2a:

$$E_T^{\text{miss}} \geq 50 \text{ GeV}, \tag{2.3}$$

and at least 3 muons/electrons with $p_T > 15$ GeV.

- Channel 2b:

$$E_T^{\text{miss}} \geq 50 \text{ GeV}, \quad H_T \geq 50 \text{ GeV},$$

and at least 2 muons/electrons with $p_T > 15$ GeV.

- Channel 3:

$$H_T \geq 600 \text{ GeV}, \quad E_T^{\text{miss}} > 100 \text{ GeV}.$$