Model agnostic selections for new Physics searches

Olmo Cerri, Javier Duarte, Jennifer Ngadiuba, Thong Q. Nguyen, Maurizio Pierini, Maria Spiropulu, Jean-Roch Vlimant, **Kinga Anna Wozniak**

Chep 2019, Adelaide

ANOMALY DETECTION WITH ML

Problem Statement:

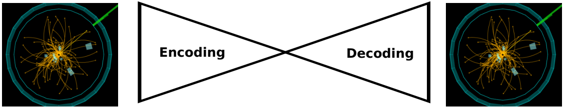
- Input dataset to analysis: large share of BG, small share of SIG
- Enhance SIG over BG ratio
- Look for ”anomalies” in the dataset, discard others

ML Solutions:

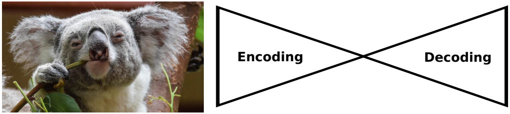
- Supervised algorithms: New physics search based on a model
- Unsupervised algorithm: allows for model independence, just flag / veto

AUTOENCODER

- Unsupervised learning how BG event looks like
- By mapping input (set of QCD events) onto itself



- Loss: $\Delta(\text{Input}, \text{Output})$
- Inference: When input is "unusual"
  $\rightarrow \Delta$ is large $\rightarrow$ raise flag for anomaly

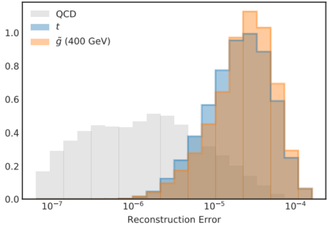


?

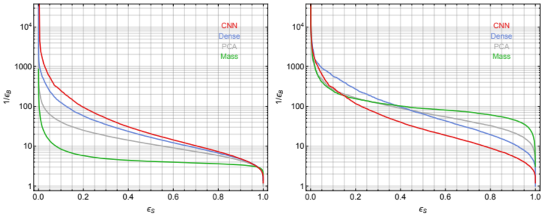
Related Work: Event Selection

Convolutional autoencoder selection performance of QCD versus boosted top jets and RPV gluino jets as signal

Farina et al.: arXiv:1808.08992, Heimel et al.: arXiv:1808.08979: Already applied to anomaly detection (with plain autoencoders)

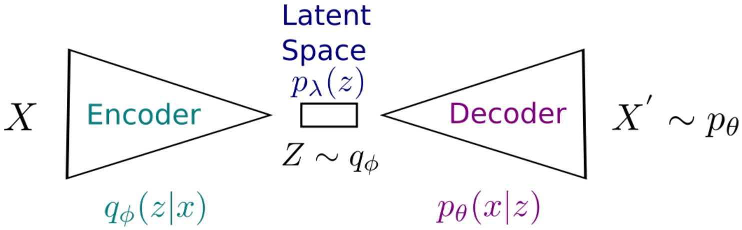


QCD background (gray) and two signals: tops (blue) and 400 GeV gluinos (orange).



ROC of signal tagging efficiency vs background rejection (1/bg-efficiency) computed on top jets (left) and gluino jets (right)

VARIATIONAL AUTOENCODER



- $X \in \mathbb{R}^N, \quad Z \in \mathbb{R}^M, \quad M \ll N$
- Stochastic Encoder & Decoder (outputs = parameters of probability distributions)
- Prior on latent space $\rightarrow$ 2 Loss terms: reconstruction + divergence in latent space

Kingma et al.: arXiv:1312.6114,

An et al.: `http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf`

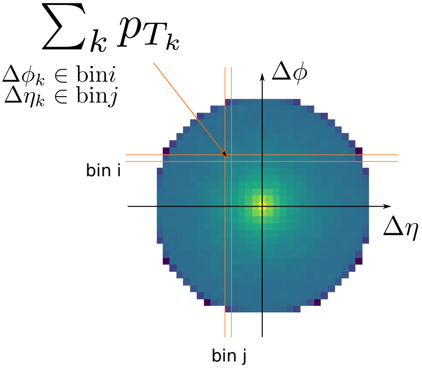
VAE for Jet Images

Dataset:

- 10 fb^{-1} of QCD simulation
- Pythia + Delphes
- Clustered particle-flow Jets (=2) with anti-kt (R=0.8)
- $\rightarrow$ Set of Dijet events

Set of constituents in each jet

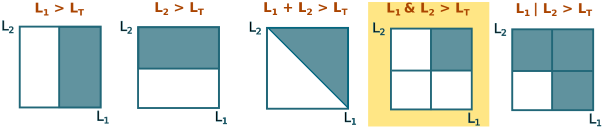
- Momentum & 2 angles
- $\rightarrow$ transform to angle-binned p_T-image



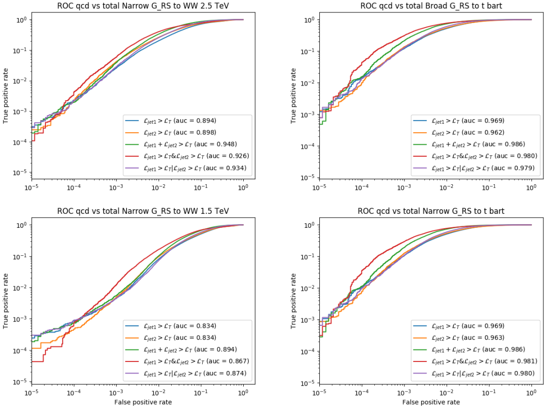
$\rightarrow$ **Train on Set of QCD Images**

De Oliveira et al.: arXiv:1511.05190 (Jet images)

From Anomalous Jets to Anomalous Events



- Given events with N jets
- Combine loss from N jets to one event loss
- Different combinations possible (performance)
- Select events cutting on event loss
- Pick best performing on set of benchmark models (RS gravitons to different final states)

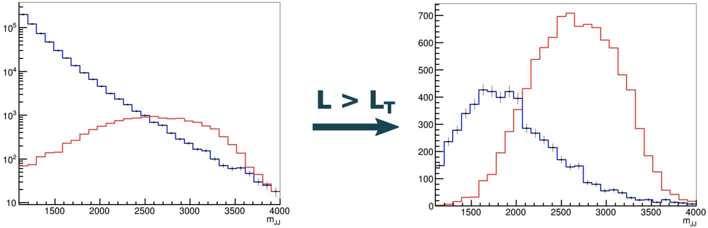


VAE to boost Supervised Searches on tails

Select events $\begin{cases} \text{Dijets } p_T > 40\,\text{GeV} \\ |\eta| < 2.4 \\ m_{JJ} > 1100\ \text{GeV} \end{cases}$

CMS analysis for selection: arXiv:1806.00843

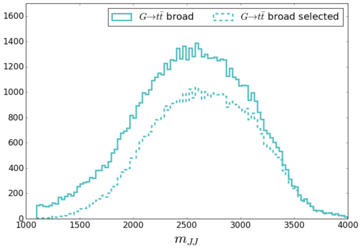
- Train VAE on data sideband: $|\Delta\eta| \geq 1.4$
- Apply VAE to signal region: $|\Delta\eta| < 1.4$
- Select events with **Loss** $>$ **threshold** s.t. some fraction of events (here 1%) is kept

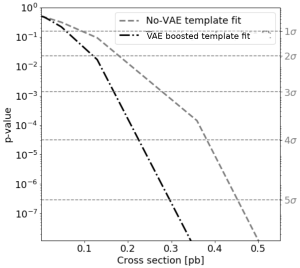


- PROS: enhance SIG (here: $G_{RS} \rightarrow t\bar{t}$ broad with xsec 10 pb)
- CONS: shape the BG, in a way that could be dangerous for a signal in that mass range
- Can still run bump hunt if 'excess' is not on bulk of the distribution

VAE to boost Supervised Searches on tails

- Consider broad RS Graviton $\rightarrow t\bar{t}$ with 3.5 TeV mass 'Data' cocktail (pythia simulation of QCD + G_{RS} at different cross sections)
  - Traditional approach: Bump hunt
  - VAE-boosted approach: Bump hunt after VAE loss cut
- NOTE: here running simple template fit, assuming (for both) that we know BG shape $\rightarrow$ overestimated significance but comparison meaningful



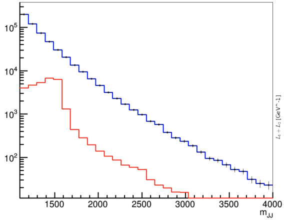


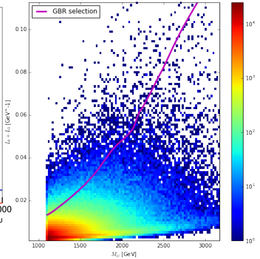
VAE TO BOOST SUPERVISED SEARCHES ON BULK

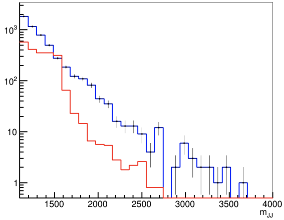
Select events $\begin{cases} \text{Dijets } p_T > 40\text{ GeV} \\ |\eta| < 2.4 \\ m_{JJ} > 1100\text{GeV} \end{cases}$

Ref: CMS analysis for selection

- Train VAE on data sideband: $|\Delta\eta| \geq 1.4$.
- Apply VAE to signal region: $|\Delta\eta| < 1.4$
- Select events with M_{JJ} dependent cut (quantile regression keeping 1% of events)



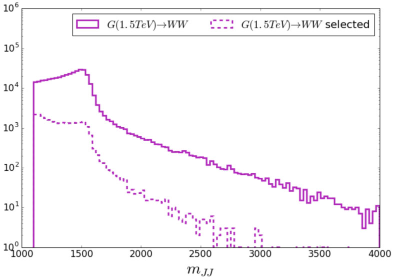


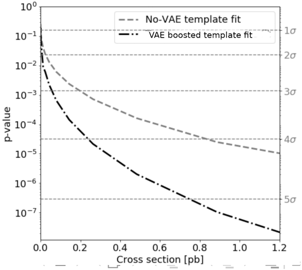


- PROS: keeps background unbiased
- CONS: reshape the signal here: $G_{RS} \rightarrow WW$ with xsec 40 pb) in unfavourable way (penalise tail)
- Can still run a bump hunt if core of the distribution

VAE to boost Supervised Searches on Bulk

- Consider RS Graviton $\rightarrow$ WW with 1.5 TeV mass Data cocktail (pythia simulation of QCD + G_{RS} at different cross sections)
  - Traditional approach: Bump hunt
  - VAE-boosted approach: Bump hunt after VAE loss cut
- NOTE: here running simple template fit, assuming (for both) that we know BG shape $\rightarrow$ overestimated significance but comparison meaningful





Conclusions & Outlook

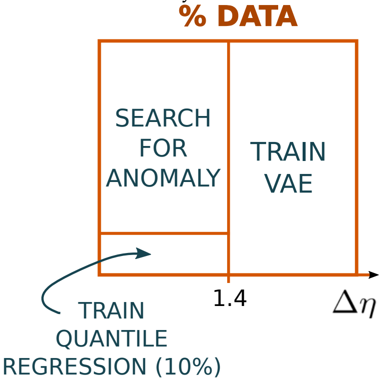
- Unsupervised Machine Learning can be used to enrich the selected datasets with new physics anomalies offline, to define SM veto
  - Good sensitivity on the tail, but strong shape bias on the bulk (might affect analysis strategy)
  - Recover sensitivity with bkg-unbiased selection: reshape signal but still improves sensitivity
- Next steps:
  - Repeat the study in a fully-realistic setup (e.g., with full shape fit)
  - From model-independent selection to model-independent analysis

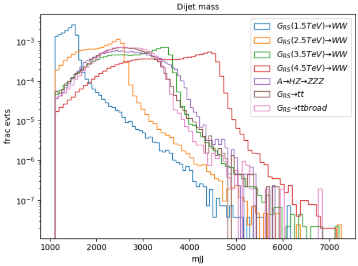
Appendix

STRATEGY

Select events $\begin{cases} \text{Dijets } p_T > 40\,\text{GeV} \\ |\eta| < 2.4 \\ m_{JJ} > 1100\text{GeV} \end{cases}$

Ref: CMS analysis for selection



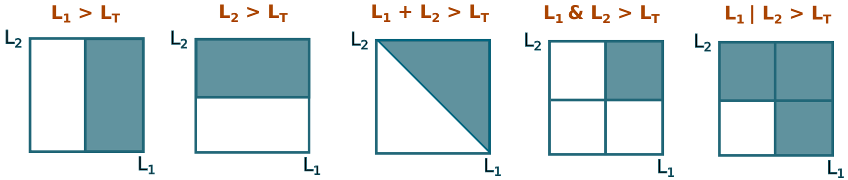


Benchmark Models

- RS Graviton $G_{RS} \to WW$ ($m_{JJ} = 1.5, 2.5, 3.5, 4.5$ TeV)
- $A \to ZH(ZZ)$ ($m_{JJ} = 13$ TeV)
- $G_{RS} \to tt$ ($m_{JJ} = 13$ TeV)
- Broad $G_{RS} \to tt$ ($m_{JJ} = 13$ TeV)

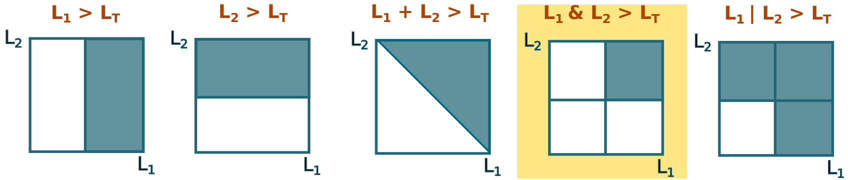
From Anomalous Jets to Anomalous Events

- Trained on control region
- Performance evaluated on signal region
- Use different event anomaly definitions, based on the vae score of the two jets



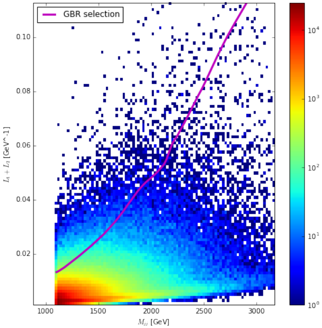
From Anomalous Jets to Anomalous Events

- Trained on control region
- Performance evaluated on signal region
- Use different event anomaly definitions, based on the vae score of the two jets



QUANTILE REGRESSION & BSM EFFICIENCY

Quantile regression with DNN (Keras, 6 Dense Layers, ReLU, Input: m_{JJ}, Output: percentile loss cut)



Sample	Trg evts	Eff. trg [%]	VAE sel. evts	Eff VAE [%]
qcdSigExt	1000038	1.5	9822	0.98
GtoTTBroad	64822	66.1	7000	10.80
GtoTTNarr	63572	66.2	6892	10.84
GtoWW4	585624	60.4	65423	11.17
AtoHZ	63128	64.4	6390	10.12
GtoWW1	371292	37.5	26539	7.15
GtoWW2	474653	51.6	48743	10.27
GtoWW3	518954	56.4	56786	10.94

VAE FOR LHC

ROC with and without $A \to 4\mu$ contamination

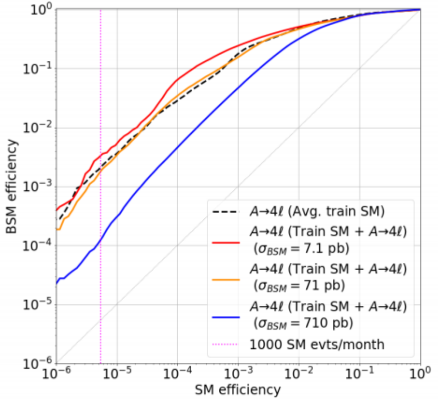
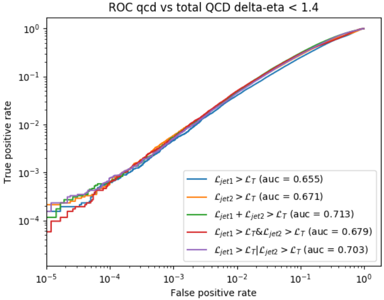


Figure: $\sigma = 7.15$ *pb*: 0.02% contamination ($\tilde{1}00$ events per month), $\sigma = 71.5$ *pb*: 0.19% contamination, $\sigma = 715$ *pb*: 1.89% contamination of the training sample

Efficiency Baseline QCD signalregion



Training

Loss Function: Reconstruction Loss + KL-Divergence

Reconstruction Loss, min Log Likelihood
Decoder Model: $p_\theta(x|z) = exp(k(z))$:

$$
\begin{aligned}
-log\, p(X|k(z)) &= -\sum_i log(k_i(z)e^{-k_i(z)x_i}) \\
&= \sum k_i(z)x_i - \sum_i log(k_i(z))
\end{aligned}
\tag{1}
$$

where x_i is pixel i of jet image X

Training

Loss Function: Reconstruction Loss + KL-Divergence

KL Divergence = relative entropy

$$D_{KL}(q_\phi(z|x)||p_\lambda(z)) = \int p(z) \log \frac{q(z|x)}{p(z)} dz \tag{2}$$

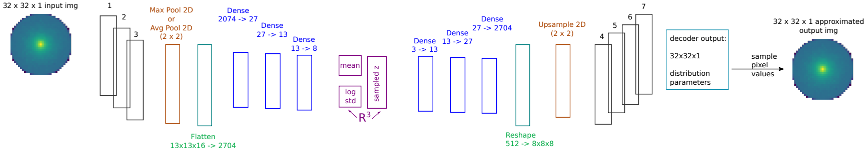
can be solved analytically if prior and approximate posterior are Gaussian

Architecture

$$z \in \mathbb{R}^5$$

Layers:

1: Conv2D, 8 filter, 3x3 filter-size, no stide, padding: valid, RELU
2: Conv2D, 12 filter, 3x3 filter-size, no stride, padding: valid, RELU
3: Conv2D, 16 filter, 3x3 filter-size, no stride, padding: valid, RELU
4: Conv2DTranspose, 16 filter, 3x3 filter-size, padding: same, RELU
5: Conv2DTranspose, 12 filter, 3x3 filter-size, padding: same, RELU
6: Conv2DTranspose, 8 filter, 3x3 filter-size, padding: same, RELU
7: Conv2DTranspose, 1 filter, 3x3 filter-size, padding:same, sigmoid (beta in [2,5])



RUNTIME PARAMETER

- Training: 120K QCD sideband events
- Decoder Model: Exponential
- z-dimension: $\mathbb{R}^5$