Tag N’ Train

Oz Amram & Cristina Mantilla Suarez (Johns Hopkins)
Paper in Preparation

JOHNS HOPKINS UNIVERSITY

ML4Jets 2020

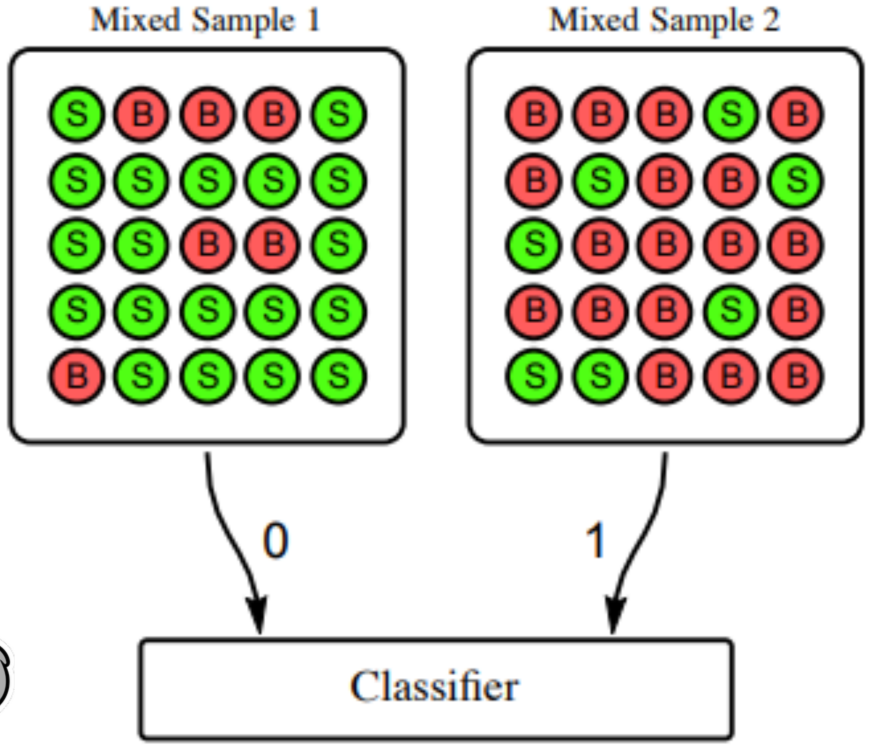
Outline

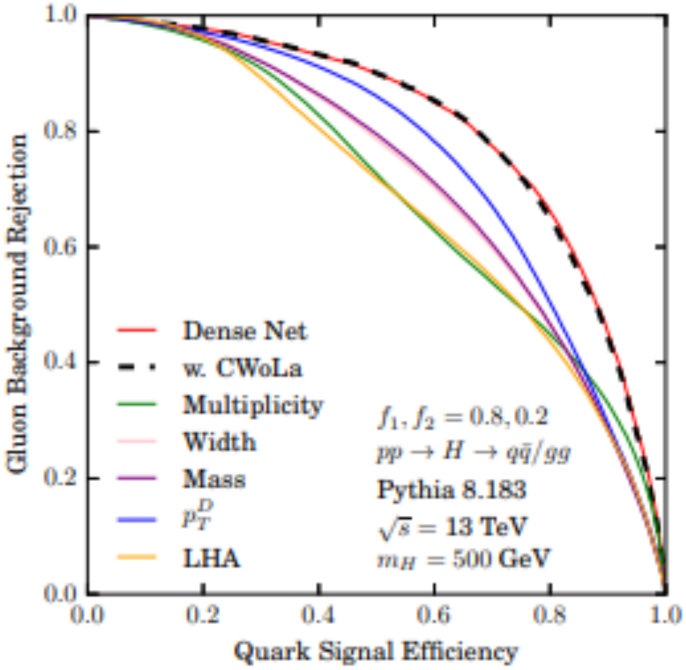
- How to train on data?
- The Tag N' Train algorithm
- Dijet anomaly search
- Future Work

How To Train on Data?

Classification Without Labels

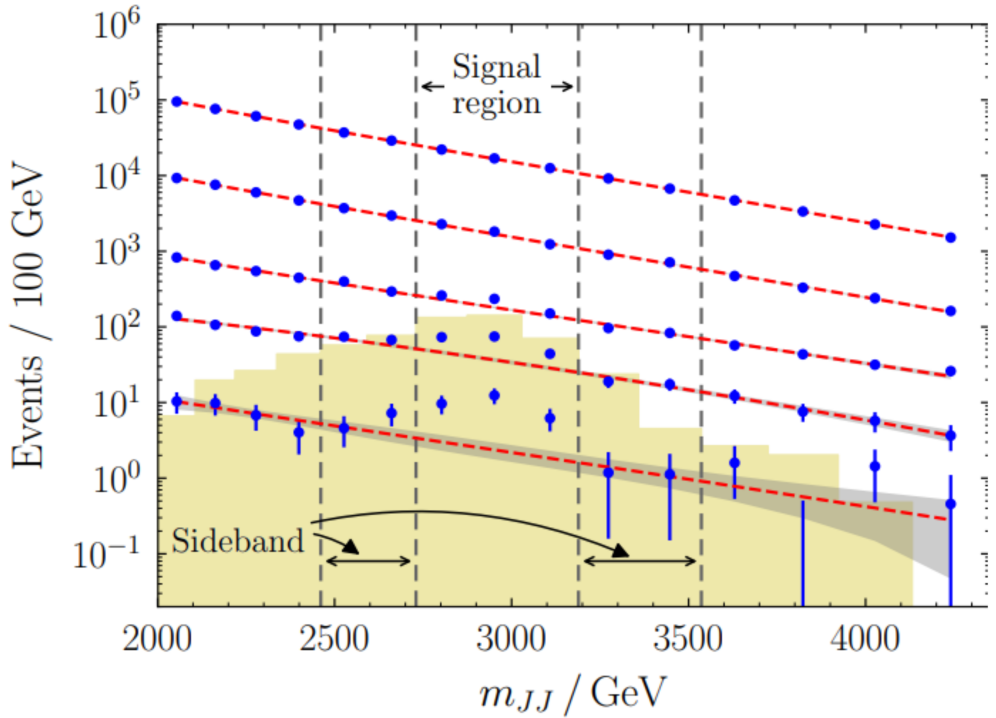
arXiv:1708.02949





CWoLa Hunting

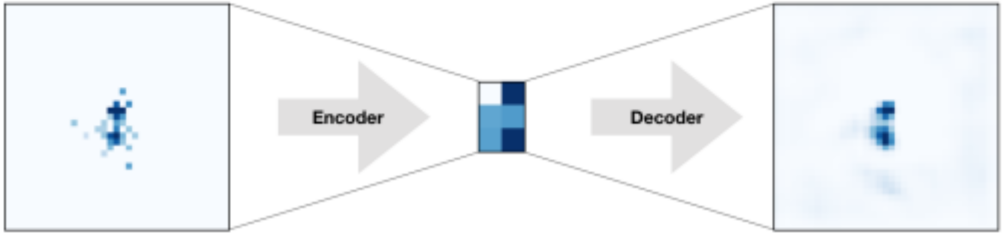
arxiv:1902.02634

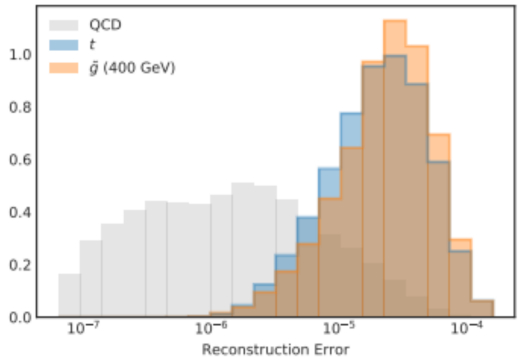


- Signal region = dijet mass window
- Train a classifier on signal region vs. others
- Select events & bump hunt

Anomaly Detection : Autoencoders

arXiv:1808.08992
arXiv:1808.08979





- Train a network to compress and decompress the data
- Can train directly on data, no labels needed
- Anomalous events should have a higher reconstruction loss

Drawbacks

- CWoLa Hunting
  - Worry about sculpting QCD dijet mass distribution
  - Apply to non-resonant signals?

- Autoencoders
  - Only ‘learns’ what QCD looks like
  - Room for improvement as a Sig vs. Bkg classifier

The Tag N' Train Algorithm

How to Combine?

- CWoLa + Autoencoders
- Find samples with enriched signal using autoencoders
- Train better classifiers using these samples

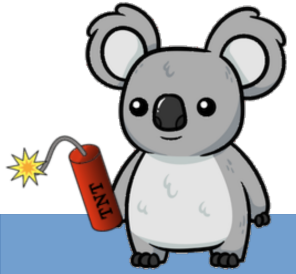
Tag N' Train (TNT)

- A new method of training classifiers on data
- Assumptions:
  - Events have 2 interesting objects in them
  - Correlations between objects not crucial for classification

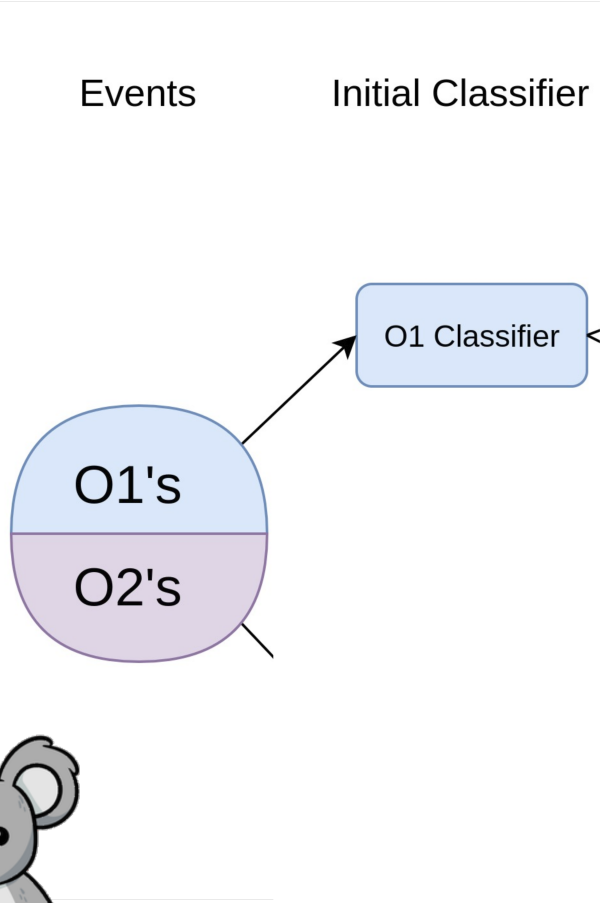
Tag N' Train (TNT)

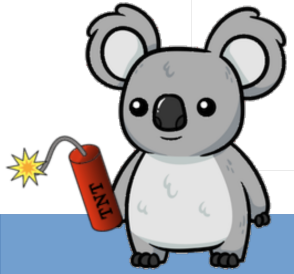
- A new method of training classifiers on data
- Assumptions:
  - Events have 2 interesting objects in them
  - Correlations between objects not crucial for classification

Tag with a weak classifier **N' Train** a better one!

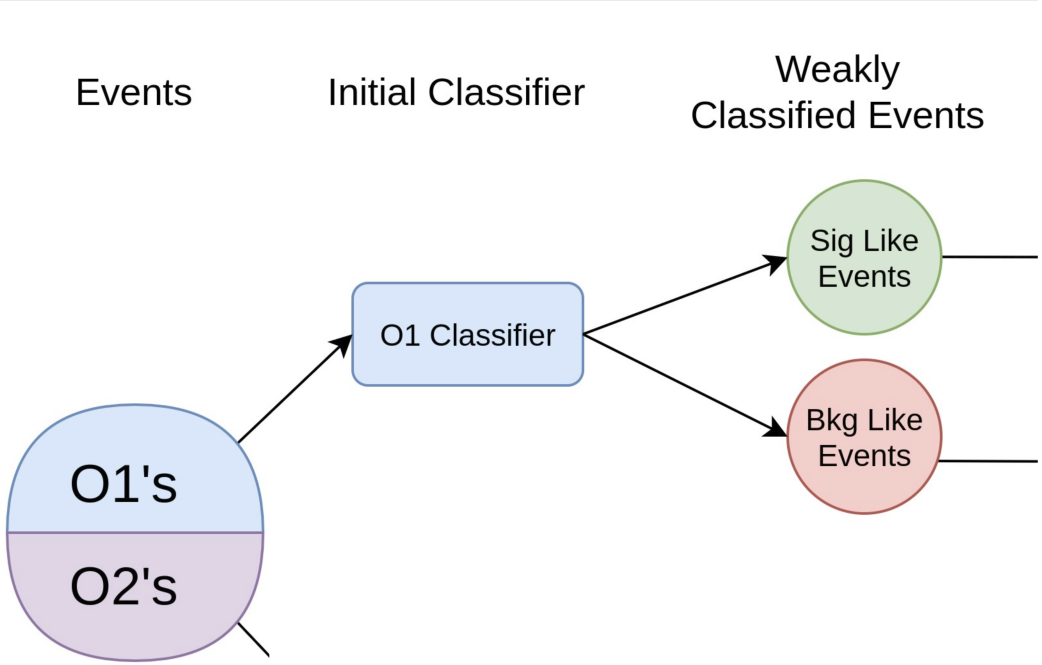


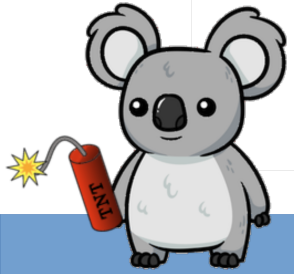
Tag N' Train



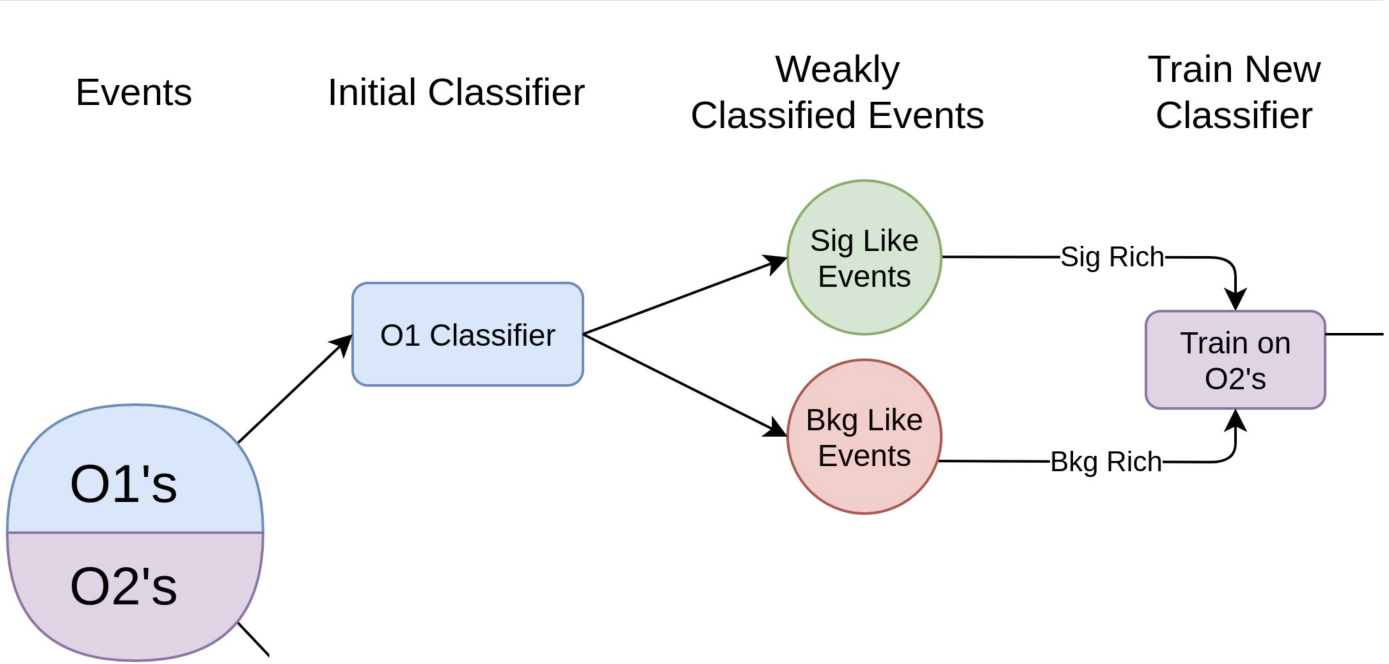


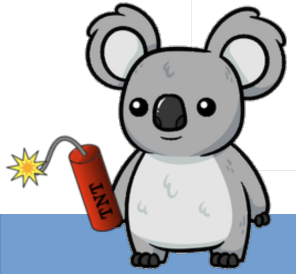
Tag N' Train


Events
Initial Classifier
Weakly Classified Events
O1's
O2's
O1 Classifier
Sig Like Events
Bkg Like Events

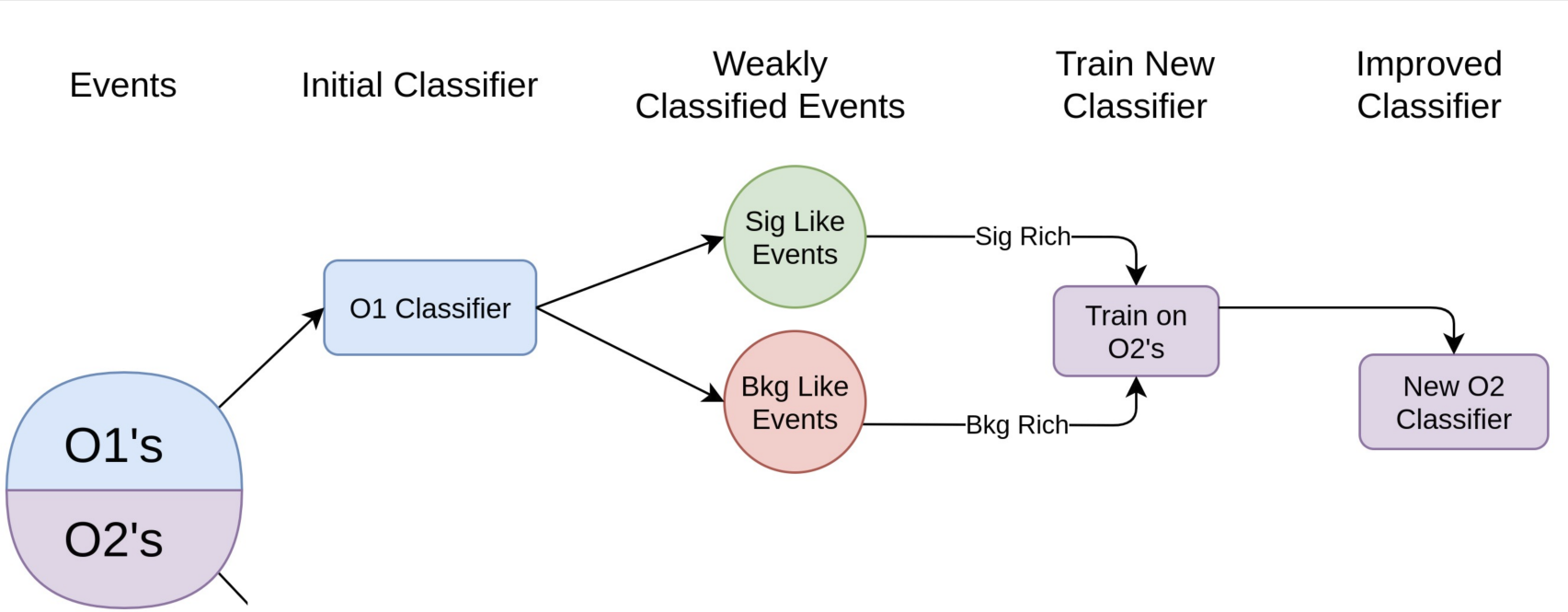

TNT

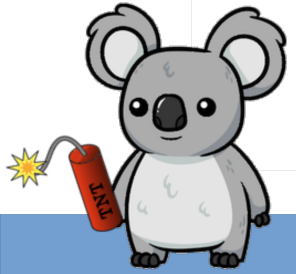
Tag N' Train


Events
Initial Classifier
Weakly Classified Events
Train New Classifier
O1's
O2's
O1 Classifier
Sig Like Events
Bkg Like Events
Sig Rich
Bkg Rich
Train on O2's

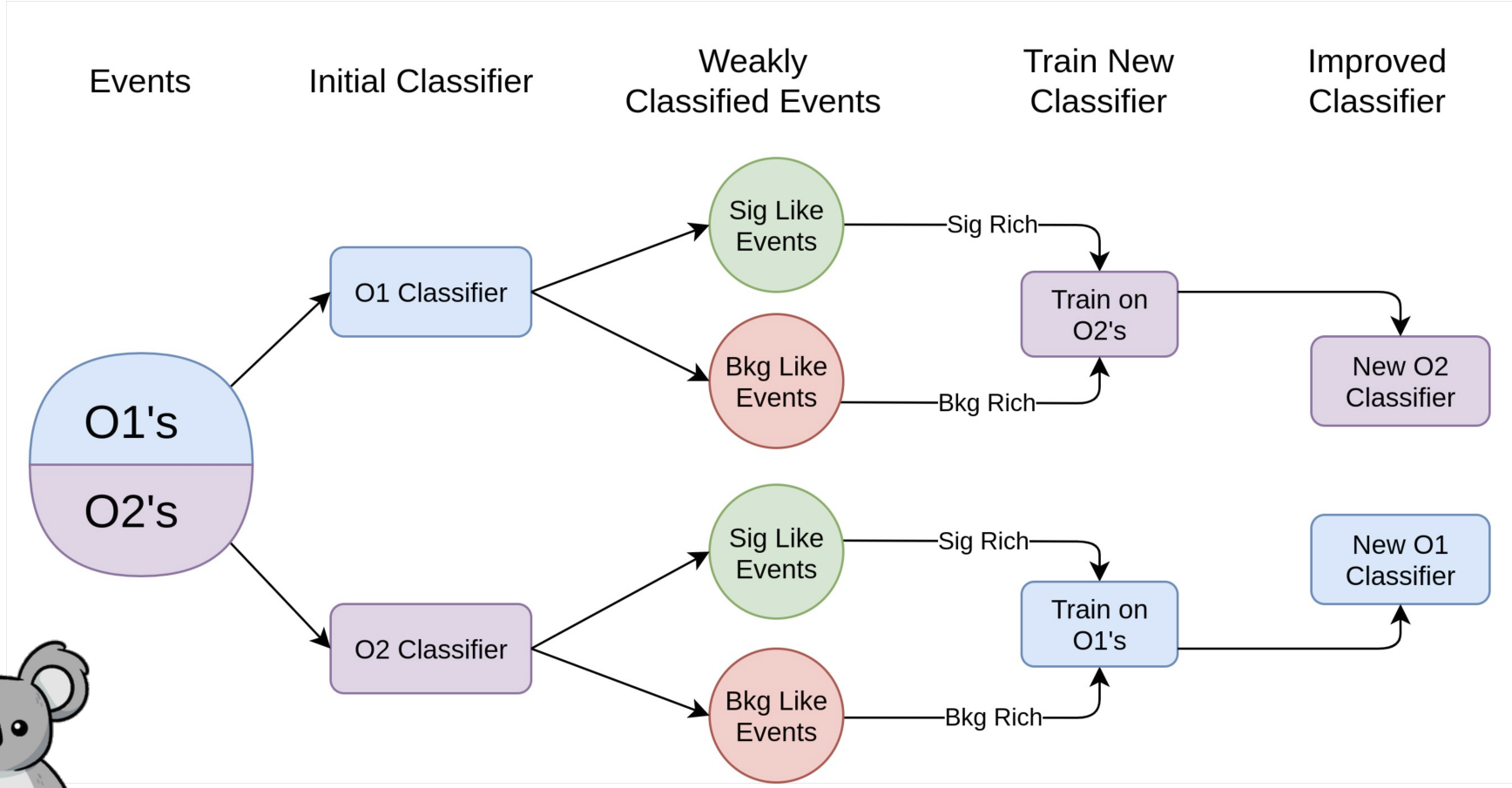

TNT

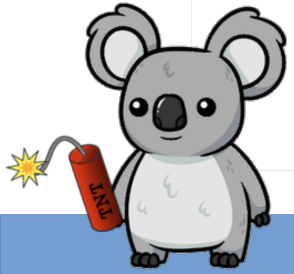
Tag N' Train


Events
Initial Classifier
Weakly Classified Events
Train New Classifier
Improved Classifier
O1's
O2's
O1 Classifier
Sig Like Events
Bkg Like Events
Sig Rich
Bkg Rich
Train on O2's
New O2 Classifier

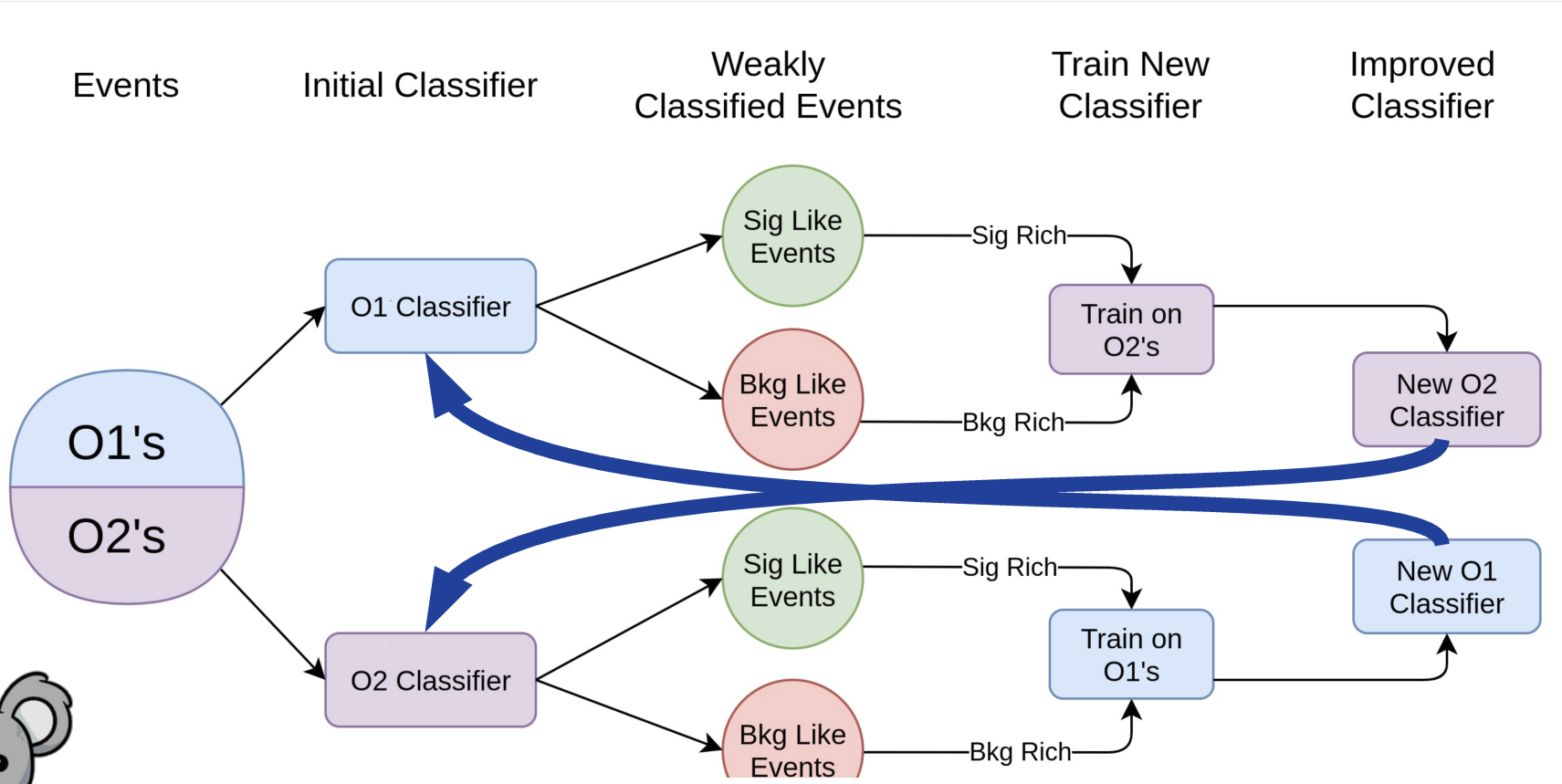

TNT

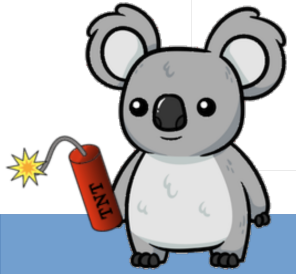
Tag N' Train


Events
Initial Classifier
Weakly Classified Events
Train New Classifier
Improved Classifier
O1's
O2's
O1 Classifier
O2 Classifier
Sig Like Events
Bkg Like Events
Sig Like Events
Bkg Like Events
Sig Rich
Bkg Rich
Sig Rich
Bkg Rich
Train on O2's
Train on O1's
New O2 Classifier
New O1 Classifier


TNT

Tag N' Train





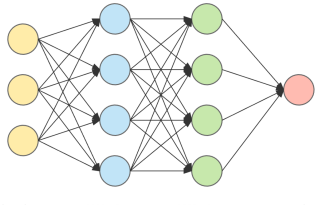
And do it again!

Dijet Anomaly Search

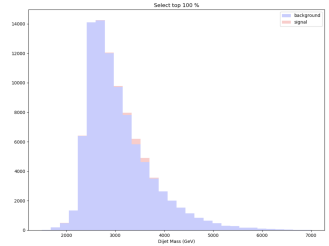
Applying TNT to a Resonance Search

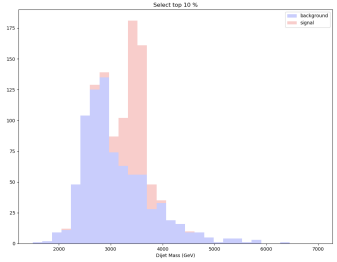
Events A

Events B

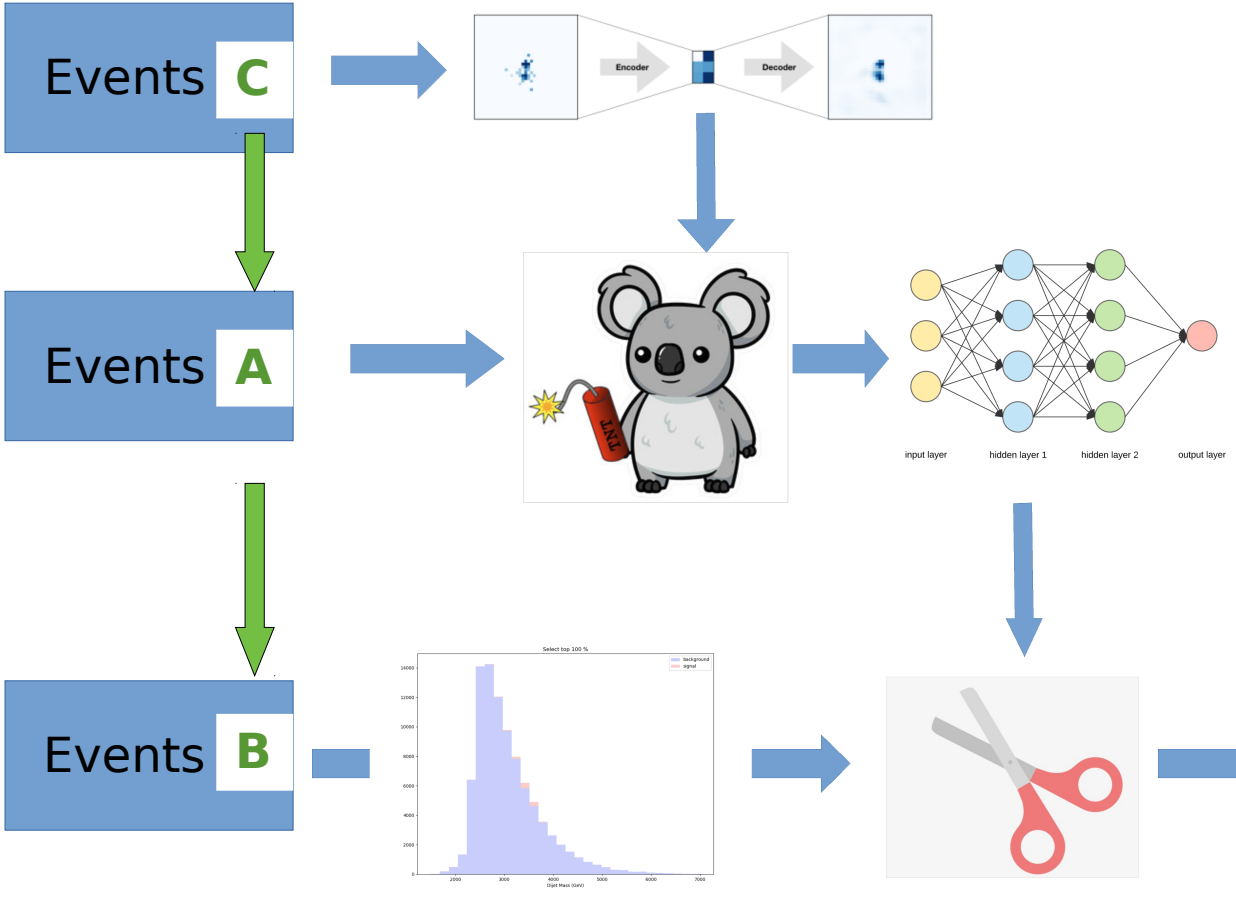


Events C





Applying TNT to a Resonance Search

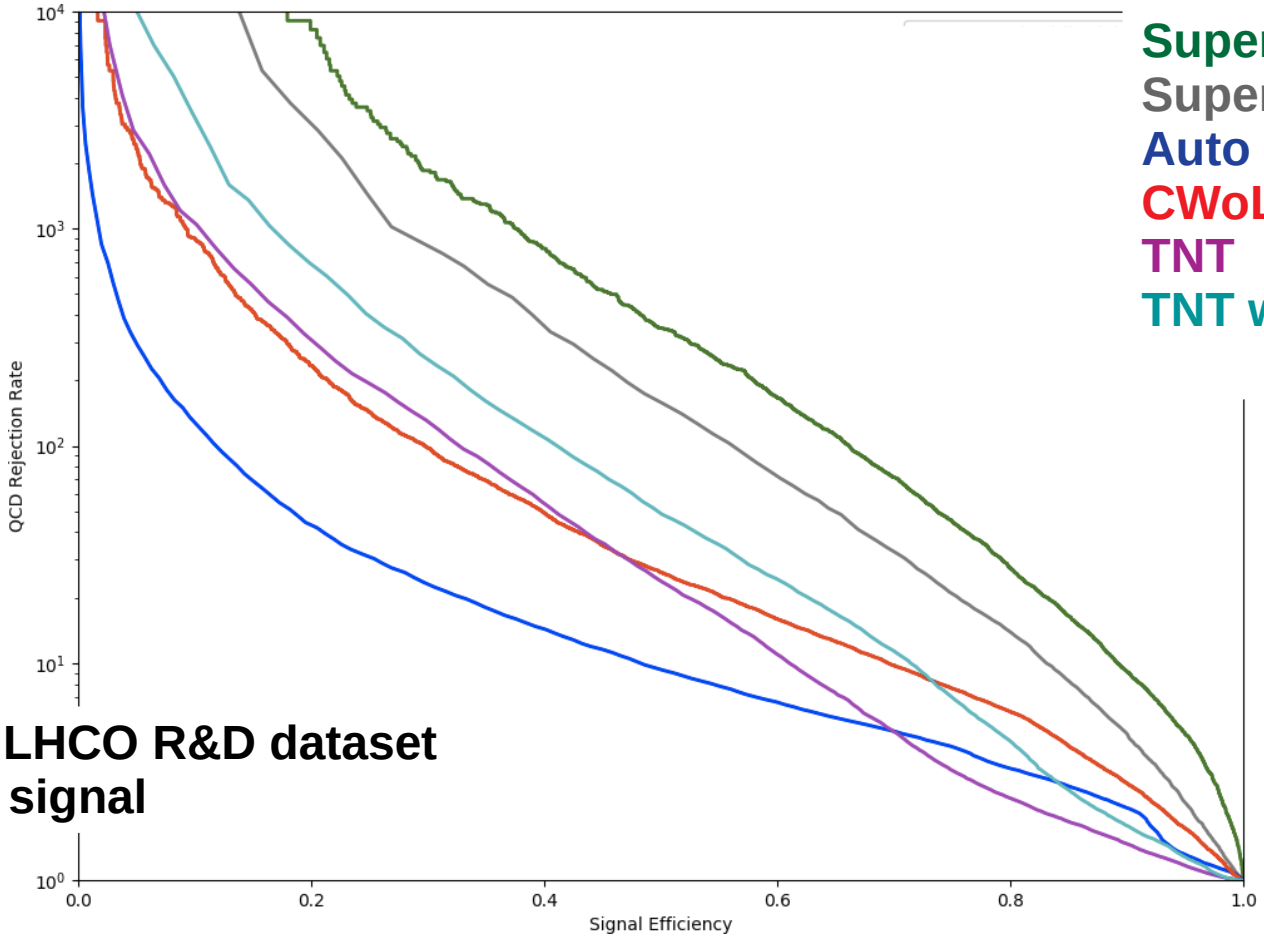


Recycle!

Technical Details

- 2 objects: heavy jet and light jet in event
- CNN Classifiers
- Top 20% ‘sig-like’, bottom 40% ‘background-like’
  - Optional dijet mass window
- Combine 2 classifiers into 1
  - Require both jet’s scores be in top X% of scores

Classification Performance

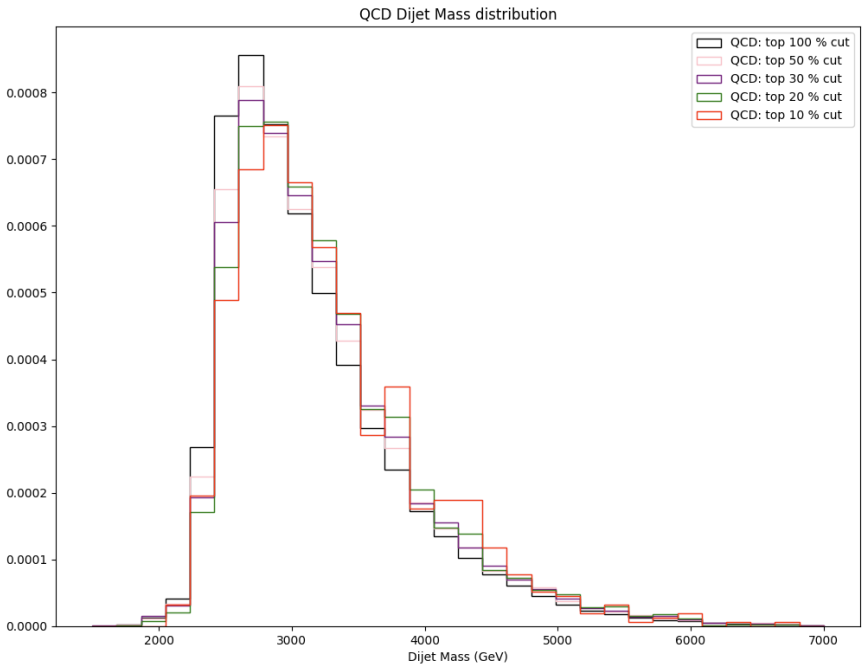


Supervised (both jets)
Supervised (sep. jets)
Auto Encoders
CWoLa Hunting
TNT
TNT with dijet mass window

Using a LHCO R&D dataset with 1% signal

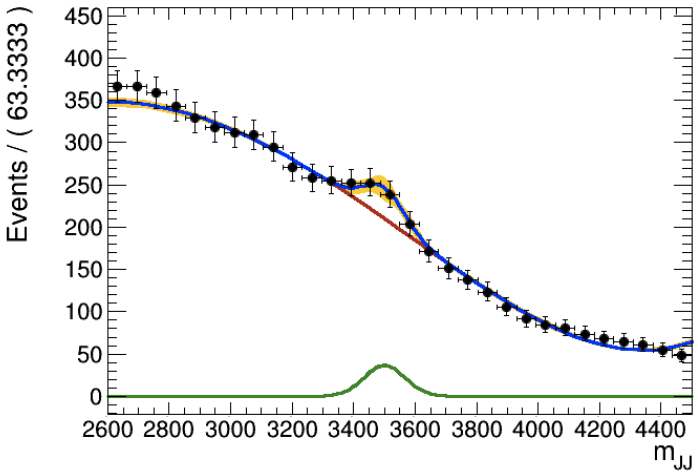
All networks using similar CNN architecture

Dijet Mass Sculpting

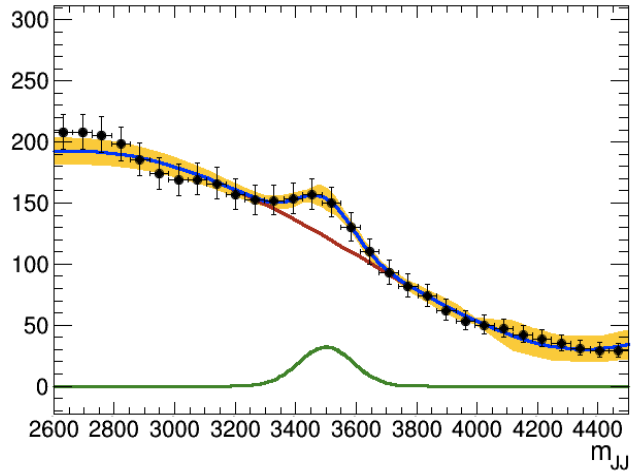


- No sculpting of dijet mass!
  - Images p_T normalized
- Re-weighting mixed samples to have same jet p_T distribution also possible

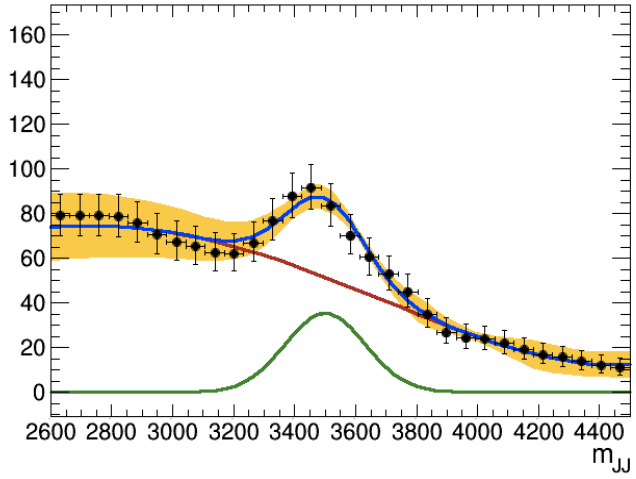
Bump Hunting



3.1σ



4.2σ



7.7σ

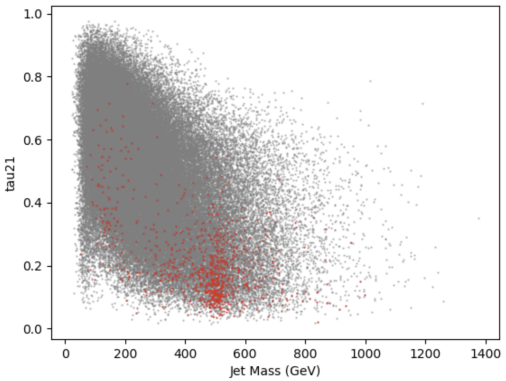
Tighter Selection

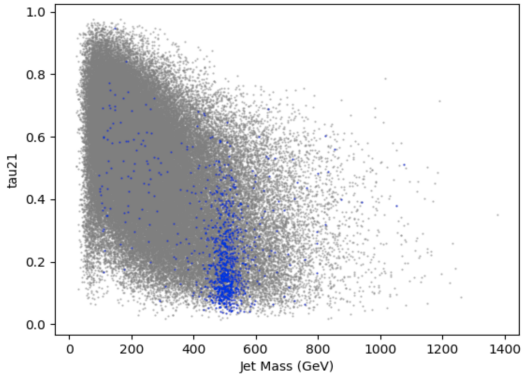
Understanding Signal

TNT Selection

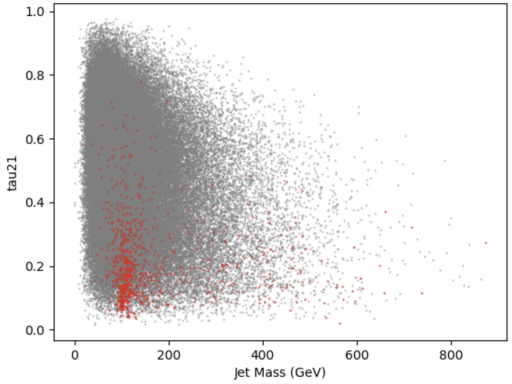
Truth

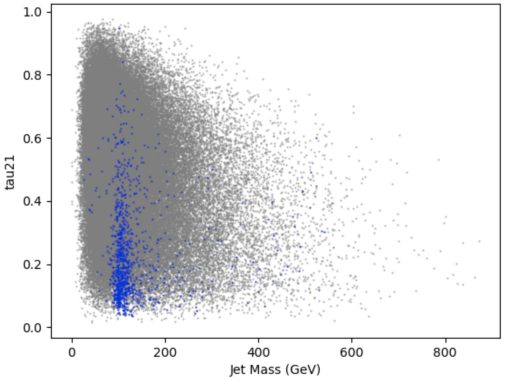
Heavy Jet



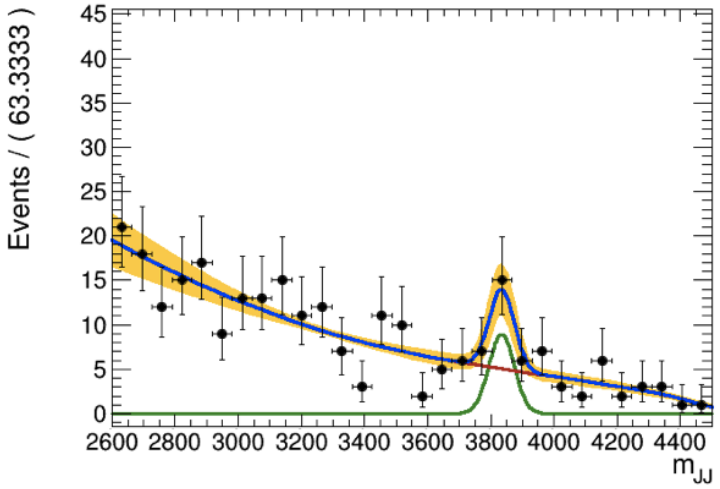


Light Jet





Black Box Results



- Resonance at ~ 3800 GeV
- 4 sigma evidence after combining samples
- Nothing seen in quick scan of black boxes 2 and 3

Future Work

Future Work

- Try other architectures (high level features?)
- Non-resonant search
  - Background estimate?
  - Sub-dominant backgrounds with ‘interesting’ jets (e.g. ttbar)?
- Can current searches be incorporated within this framework?
  - e.g. Start with a W and top classifiers trained in MC

Conclusions

- **Tag N' Train** algorithm for training on data
  - Assumes decomposition of event into 2 objects
- Natural application to dijet anomaly search
- Lot's of room to explore
- Paper in prep!

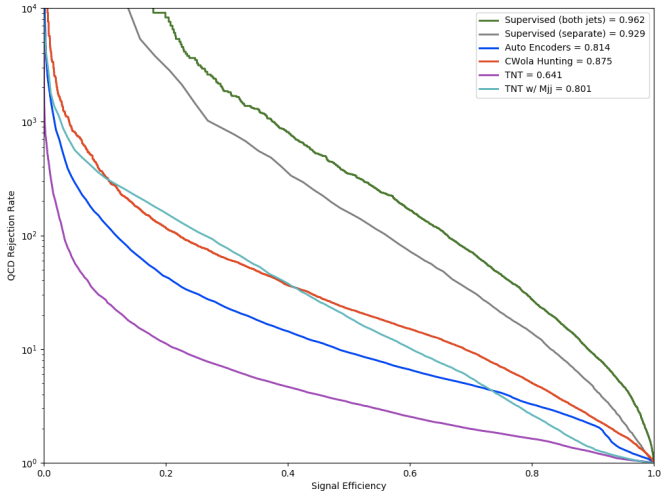
Conclusions

- **Tag N' Train** algor[illegible] for[illegible]ming on data
  - Assumes decompos[illegible]vent into 2 objects
- Natural appli[illegible]ati[illegible] [illegible]anomaly search
- Lot's of room to exp[illegible]
- P[illegible]

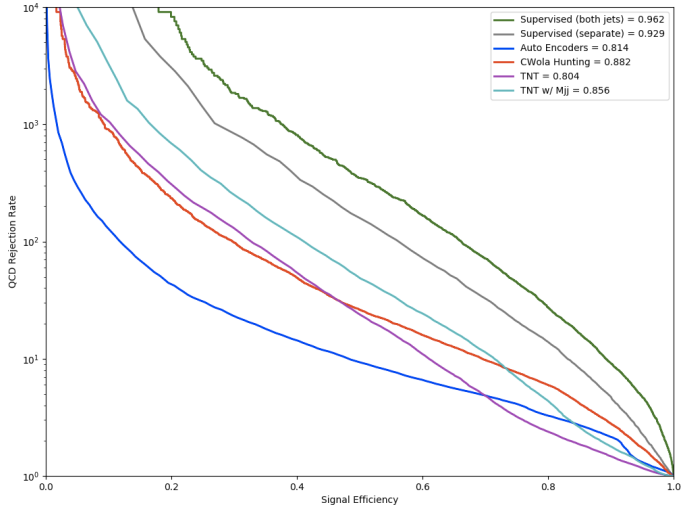
Thanks!

Backup

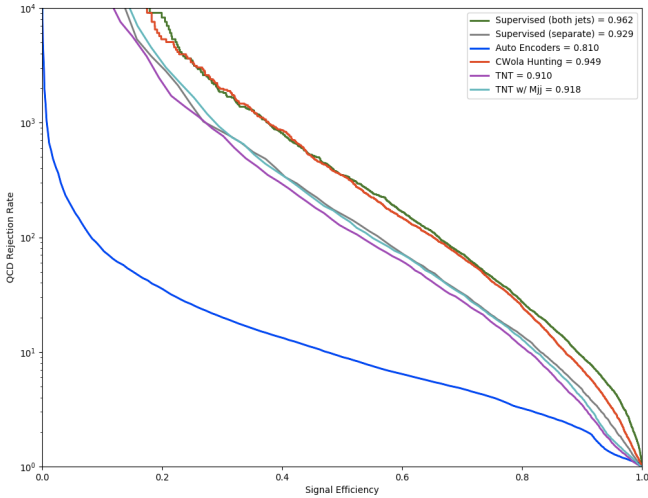
Classification Performance



0.3%



1%

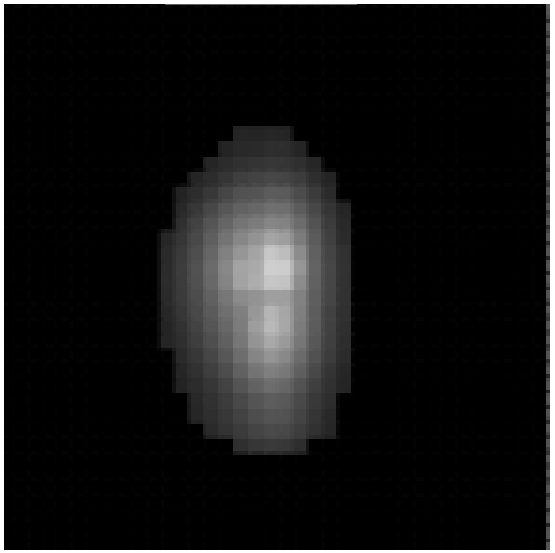


9%

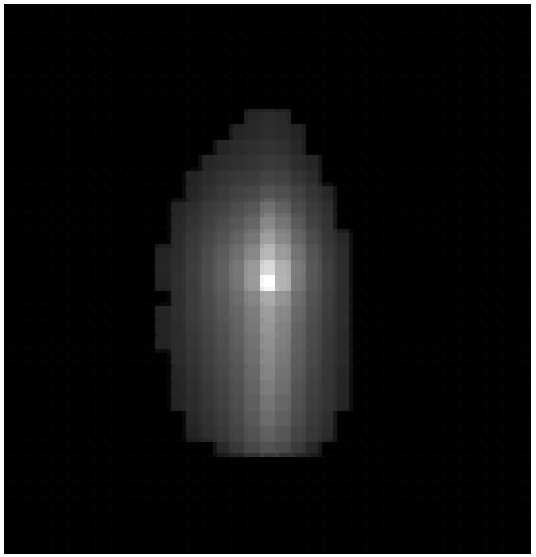
Supervised (both jets)
Supervised (sep. jets)
Auto Encoders
CWoLa Hunting
TNT
TNT with dijet mass window

Jets Images

Top jet



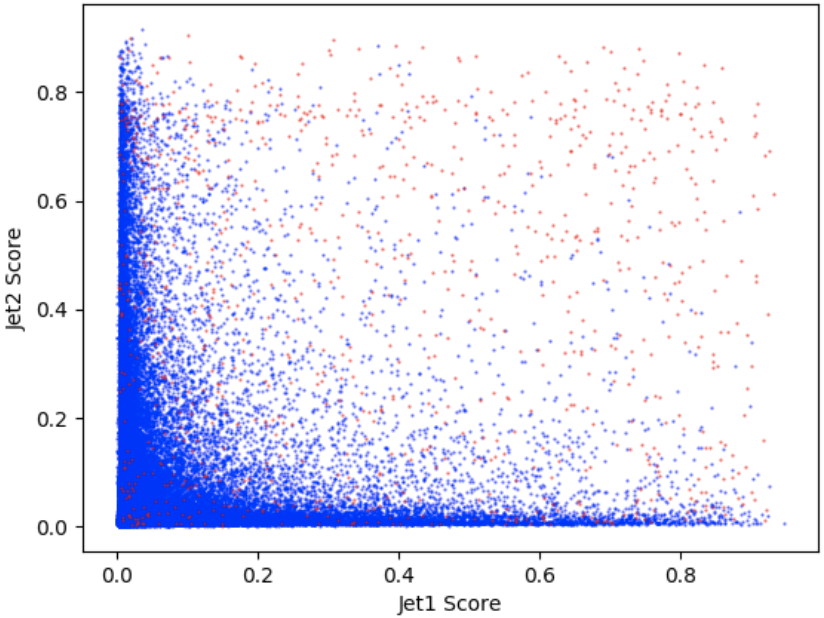
QCD jet



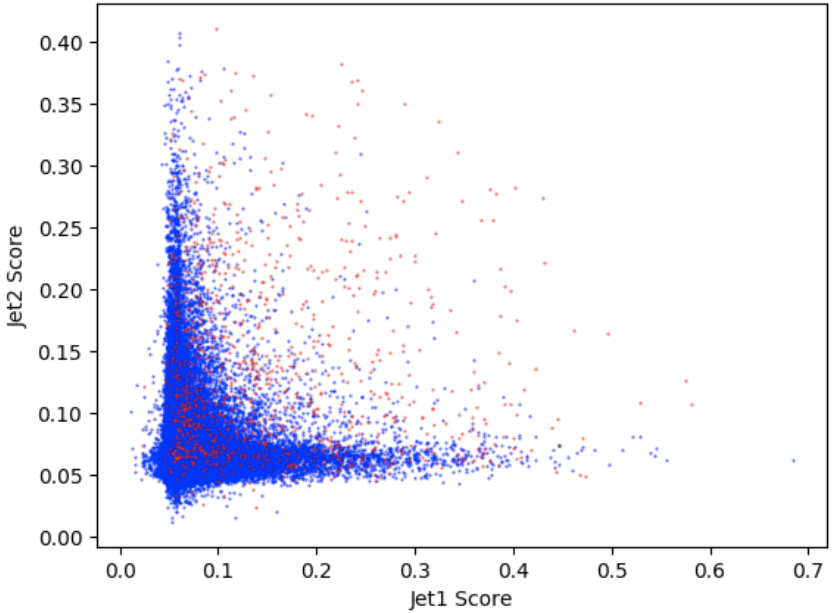
arxiv:1803.00107

- Compact, low level representation of jet
- Pre-process to center, rotate, etc.

Correlations

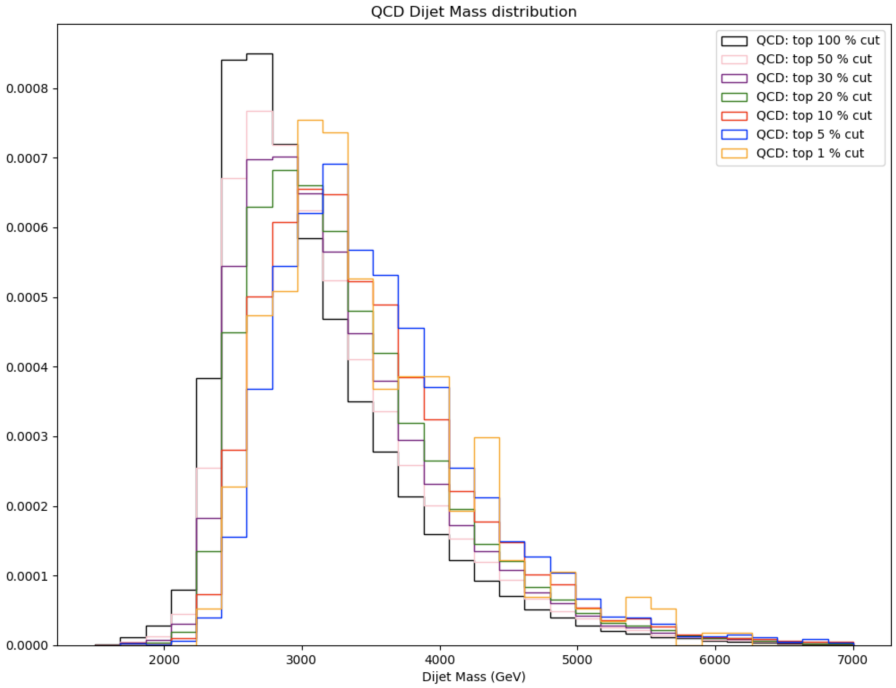


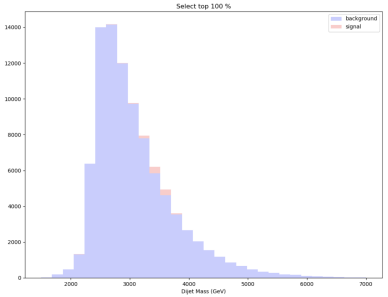
Supervised

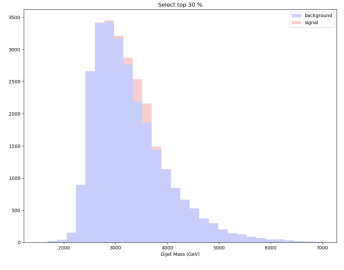


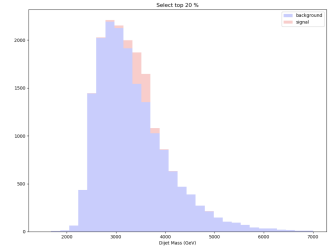
TNT (1% signal)

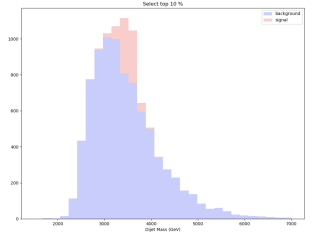
CWoLa Hunting* Dijet Mass Sculpting

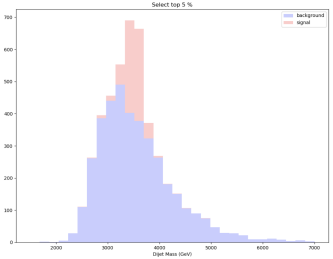


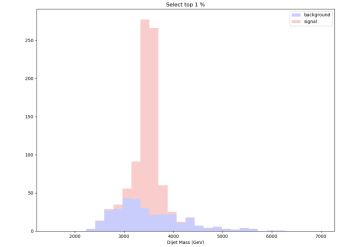












*Using same inputs as in the original paper, not CNN like prev. slide

TNT Dijet Mass Sculpting