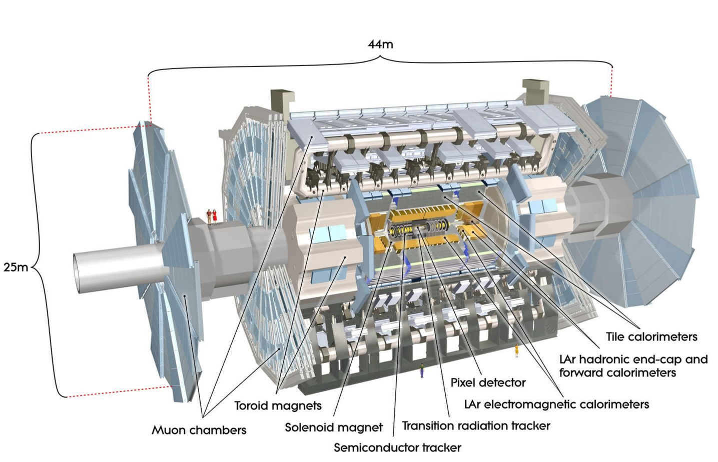
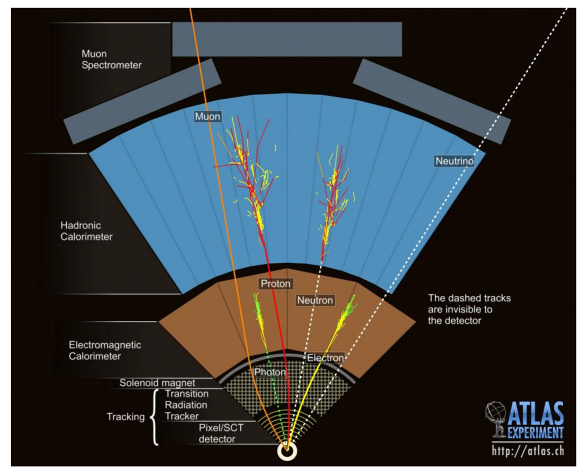
Quark/Gluon Tagging with ATLAS

By Rob Mina and Randy White
CS 229 Machine Learning

The ATLAS Experiment

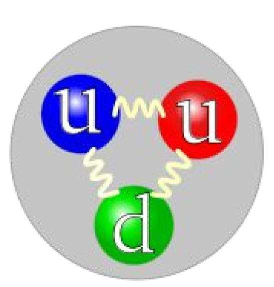
- Multi-purpose particle detector at the Large Hadron Collider (LHC) at CERN
- Works similar to a giant camera to image and measure the particles created in the collisions

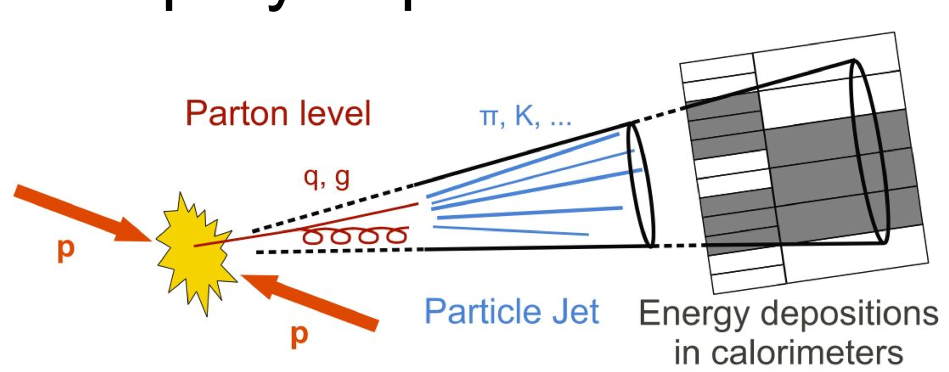




Quarks and Gluons

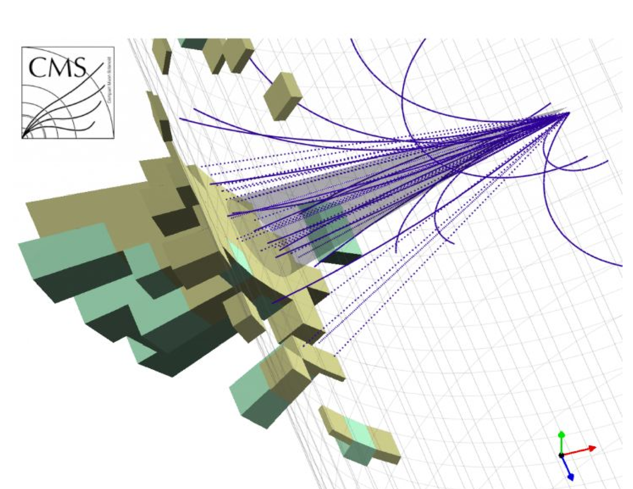
- The fundamental constituents of protons (and of most other matter) are quarks and gluons
- When produced in high energy collisions the result is a conical spray of particles

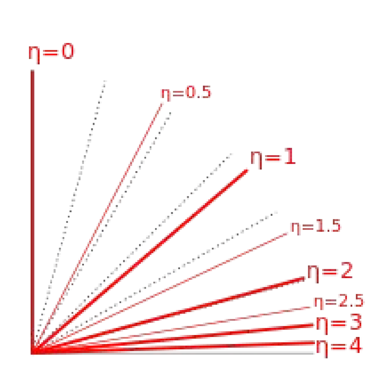




Detector Response

- These particle sprays are called “jets” and our detector records much about them
- The tracker records the paths of each charged particle as it leaves the interaction point
- Their energies are then measured in multi-layer calorimeters further from the collision region
- All of this is represented for a defined cone size around the jet axis, giving us the overall energy, momentum, shape, and constituents





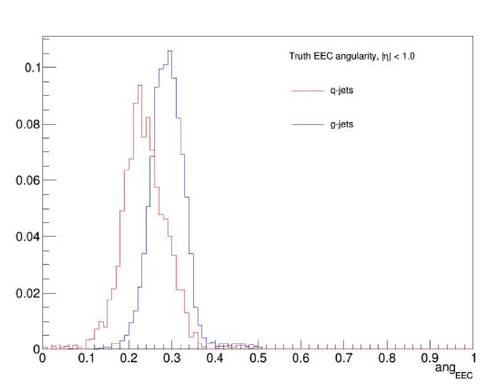
Jet Classification Problem

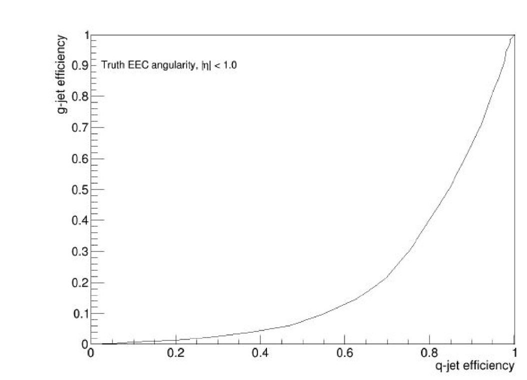
- It is crucial for certain physics results that these two types of jets are correctly identified
- We seek to perform this classification of quark- vs. gluon-initiated jets in five separate regions

Region	$\|\eta\|$ bounds	q-jet fraction	g-jet fraction
R1	$\|\eta\| < 1$	0.404	0.567
R2	$1 \leq \|\eta\| < 2$	0.296	0.302
R3	$2 \leq \|\eta\| < 2.8$	0.192	0.105
R4	$2.8 \leq \|\eta\| < 3.2$	0.0688	0.0177
R5	$3.2 \leq \|\eta\| < 4$	0.0391	0.0078

Feature Selection

- We have three overall types of variables to describe the jets, each with truth level and simulated reconstruction versions
- First: constituent count variables, including the count of all tracks and tracks above thresholds
- Second: jet width variables, related to the angular spread of constituent particles
- Third: shape information, coming from energy and momentum weighted angular distributions





Feature Performance Metrics

- Two metrics to evaluate the performance:
  - The gluon efficiency at 50% quark efficiency
  - The integral of the separation in the efficiency versus the trivial straight line

Variable	R1: g-jet eff at 50%	R1: integrated separation	R4: g-jet eff at 50%	R4: int. sep.
$N_{trk,reco1000}$	0.170	0.247	0.476	0.024
$N_{trk,reco500}$	**0.166**	**0.250**	0.476	0.023
$N_{trk,truth}$	0.126	0.285	0.138	0.267
$W_{trk,truth}$	0.162	0.204	0.137	0.240
$W_{calo,reco}$	0.202	0.208	**0.358**	**0.110**
$W_{calo,truth}$	0.114	0.239	0.099	0.251
$N_{90\%constit,reco}$	0.190	0.221	0.392	0.067
$N_{constit,truth}$	0.100	0.292	0.123	0.292
EEC_{reco}	0.190	0.216	0.367	0.101
EEC_{truth}	**0.075**	**0.297**	**0.069**	**0.298**

Model Selection

- Classifiers that works well in a high dimensional space and with our limited dataset
- Chose Logistic Regression Model, Support Vector Machine, and Neural Network

Implementation Details

- Truth and reco features treated separately
- Two and three variable versions of each model
- This is to remove redundant or highly correlated inputs and avoid overfitting
- In the N.N. we added a few non-discriminatory variables such as jet energy and charge
- Performance is tested on a separate subset of the data, using classification accuracy along with the other two efficiency measures

Region	Variables	Logistic Regression	SVM classifier	NN classifier
R1	$N_{trk,reco500} + W_{calo,reco}$	**0.81, 0.102, 0.29**	0.80, 0.122, 0.22	0.80, 0.118, 0.28
R1	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	0.81, **0.092**, 0.29	0.80, 0.122, 0.22	**0.81**, 0.093, **0.31**
R2	$N_{trk,reco1000} + W_{calo,reco}$	0.77, 0.121, 0.27	0.77, 0.137, 0.21	**0.79, 0.101, 0.29**
R2	$N_{trk,reco1000} + W_{calo,reco} + EEC_{reco}$	0.77, 0.122, 0.28	0.77, 0.130, 0.21	**0.79, 0.084, 0.29**
R3	$W_{calo,reco} + EEC_{reco}$	**0.71, 0.155, 0.23**	0.70, 0.173, 0.22	0.66, 0.253, 0.20
R3	$N_{90\%constit,reco} + W_{calo,reco} + EEC_{reco}$	**0.71, 0.155, 0.23**	0.70, 0.175, 0.22	0.66, 0.253, 0.20
R4	$N_{90\%constit,reco} + W_{calo,reco}$	0.61, **0.355, 0.11**	0.58, 0.402, 0.04	**0.64**, 0.388, 0.09
R4	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	**0.63, 0.266, 0.16**	0.58, 0.402, 0.05	0.60, 0.372, 0.08
R5	$N_{trk,reco500} + W_{calo,reco}$	0.62, 0.185, 0.12	0.71, 0.332, 0.15	**0.74, 0.148, 0.24**
R5	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	0.64, 0.216, 0.12	0.71, 0.246, 0.22	**0.74, 0.148, 0.24**

Results and Optimization

- The models yielded similar performances, with the best coming from the N.N. and the L.R.
- These results track well with previously results found in our sources for the central regions and shows how we can extend the performance into the forward regions for the coming detector upgrade
- Our optimizations came through hyper parameter searches that maximized the results versus our hidden variables: tolerances, kernel functions, and hidden layers/nodes

Sources:

[1] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. Energy Correlation Functions for Jet Substructure. JHEP, 06:108, 2013.
[2] Jason Gallicchio and Matthew D. Schwartz. Quark and Gluon Tagging at the LHC. Phys. Rev. Lett., 107:172001, 2011.
[3] Georges Aad et al. Light-quark and gluon jet discrimination in pp collisions at sqrt(s) = 7 TeV with the ATLAS detector. Eur. Phys. J., C74(8):3023, 2014.