Using Machine Learning to Improve Studies of the Higgs Field Using the ATLAS Detector at the LHC

PHYS 437A Presentations

9 September, 2020

Callum McCracken

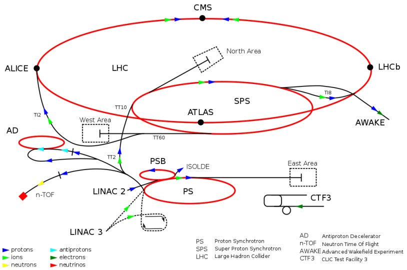
Supervisor: Maximilian Swiatlowski

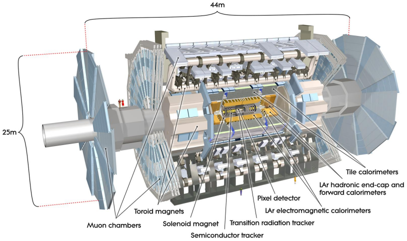
Co-Supervisor: Eduardo Martin-Martinez

Collaborators: Todd Seiss, Mel Shochet

Hello, general audience!

- Particle Physics:
  - What are the fundamental bits of the universe, and how do they interact with each other?
- LHC: Large Hadron Collider
  - Large (27km long)
  - Collides hadrons (protons)
- ATLAS:
  - General-Purpose (detects most things, not all directly)
  - Can be used to study the Higgs field

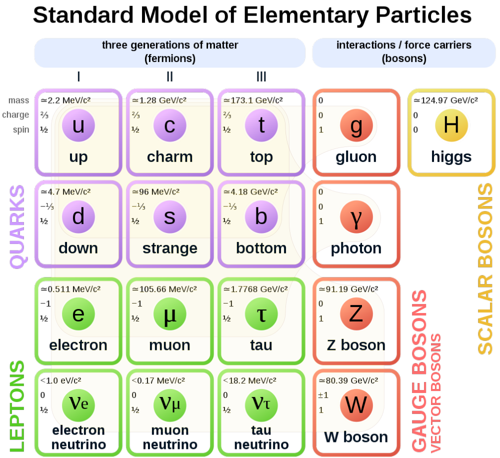




Links to image sources: CERN Accelerator Complex and ATLAS Detector

Higgs Intro

- Interaction with Higgs field responsible for particle masses
- Higgs boson: excitation of Higgs field
- Worth studying since:
  - Not very well constrained yet, good possible source of discrepancies with existing theory (new physics)
  - Understanding mass-related things might help make links with gravity (+ interaction with dark matter?)
  - New and interesting!
- Higgs potential: (scalar) field potential, characterizes interactions

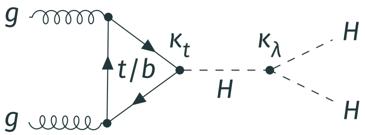


The Big Picture – Measuring the Higgs Self-Coupling

- Relevant section of SM Lagrangian for Higgs Potential:

$$V(\phi) = -\mu^2\phi^2 + \lambda\phi^4 + \ldots \text{ Taylor exp. at min } \rightarrow V_T(\phi) = -\frac{\mu^4}{4\lambda} + \frac{\sqrt{2}\mu^3}{\lambda}\phi - 4\mu^2\phi^2 + 2\sqrt{2\lambda}\mu\phi^3 + \ldots$$

- Const. and ϕ terms can be removed via change of coordinates,
  ϕ^2: mass term, ϕ^3: self-interaction or self-coupling term, not well constrained
  ($\kappa_\lambda = (2\sqrt{2\lambda}\mu)/(2\sqrt{2\lambda}\mu)_{SM}$, $\kappa_\lambda \in [-2.3, 10.3]$ at 95% confidence)

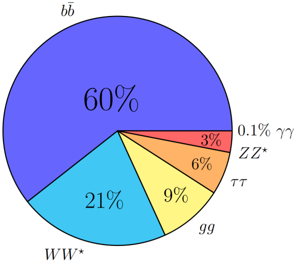


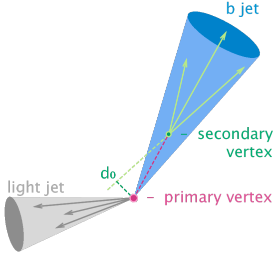
- We can measure κ_λ by considering *HH* events
- How to find these *HH* events? Using jets!

Jets, *b*-jets, And Why We Need Them

$H \rightarrow 60\% b\bar{b} \rightarrow 2 \times b$ hadrons $\rightarrow 2 \times b$-jets

- Jet: collection of particles with appx. the same direction
- ATLAS can't directly detect H or b. Instead, use b-jets, which can be directly detected.
- b-jets can be distinguished from other jets by presence of a secondary vertex (since b-hadrons have 'long' lifetimes)



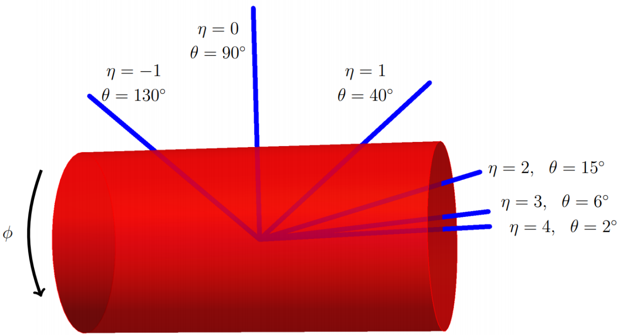


Links to image sources: branching ratios and jets

How to Describe a Jet

For every jet, ATLAS records p_T, η, ϕ, the location of the jet.

- ϕ = azimuthal angle (around beam)
- $\eta = \ln\cot\frac{\theta}{2}$, where θ = polar angle
  (η is used because it has some nice invariance properties)
- p_T = momentum transverse to beam

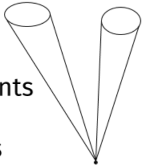


Tag, You're A *b*-jet!

Jet Reconstruction: take detector data, finds jets

Classifier: takes jets, classifies as b-jets or others

We need events with HH
⇒ 4 b-jets



Usually, b-tag ⇒ b-jet, so usually take events with 4 b-tags

To Keep in Mind

b-jets are real physical objects, *b*-tags are classifier results.

Why Focus On 3 b-tag Events?

Classifier is (for good reason) 77% efficient. ~23% of b-jets go untagged.

If we have 4 b-jets, that gives probabilities of

$$P_{4\text{ tags}} = \binom{4}{4}(0.77)^4(0.23)^0 \approx 0.35 \qquad P_{3\text{ tags}} = \binom{4}{3}(0.77)^3(0.23)^1 \approx 0.42$$

$$P_{2\text{ tags}} = \binom{4}{2}(0.77)^2(0.23)^2 \approx 0.19 \qquad P_{1\text{ tag}} = \binom{4}{1}(0.77)^2(0.23)^3 \approx 0.037$$

$$P_{0\text{ tags}} = \binom{4}{0}(0.77)^0(0.23)^4 \approx 0.003$$

So if we can use 3 b-tag events, we can more than double our number of HH events.

(and trying to work with 2- or 1-tag events may not be worth the effort)

Problem Statement

Given an event with ≥ 4 jets and 3 b-tags,
can we detect the 4th untagged b-jet if it exists,
or say there is no 4th if it does not exist?

Problem Statement

Given an event with ≥ 4 jets and 3 *b*-tags,
can we detect the 4th untagged *b*-jet if it exists,
or say there is no 4th if it does not exist?

To evaluate a solution, we can look at 5 categories of performance, (green = good, red = bad, yellow = no worse than if we do nothing):

<table>
<tr><td colspan="2"></td><td colspan="3">Truth-Matching</td></tr>
<tr><td colspan="2">(# events=304182)</td><td colspan="2">4th tag exists
(172391)</td><td>No 4th tag
(131791)</td></tr>
<tr><td rowspan="3">4th Jet Reco</td><td rowspan="2">4th jet found
(0)</td><td>Correct 4th jet</td><td>0.0%
(0)</td><td rowspan="2">0.0%
(0)</td></tr>
<tr><td>Incorrect 4th jet</td><td>0.0%
(0)</td></tr>
<tr><td>No 4th jet found
(304182)</td><td colspan="2">100.0%
(172391)</td><td>100.0%
(131791)</td></tr>
</table>

Possible Approaches

How to pick a 4th jet?

- Since the HH interaction is “hardest,” pick jet with highest p_T

Possible Approaches

How to pick a 4th jet?

- Since the HH interaction is "hardest," pick jet with highest p_T
- Since HH should have $\sum_{i=1}^{4} p_{T,i} = 0$, pick the jet that gives the smallest p_T sum.

Possible Approaches

How to pick a 4th jet?

- Since the HH interaction is "hardest," pick jet with highest p_T
- Since HH should have $\sum_{i=1}^{4} p_{T,i} = 0$, pick the jet that gives the smallest p_T sum.
- Play with classifier parameters

Possible Approaches

How to pick a 4th jet?

- Since the HH interaction is "hardest," pick jet with highest p_T
- Since HH should have $\sum_{i=1}^{4} p_{T,i} = 0$, pick the jet that gives the smallest p_T sum.
- Play with classifier parameters
- Or we could try...

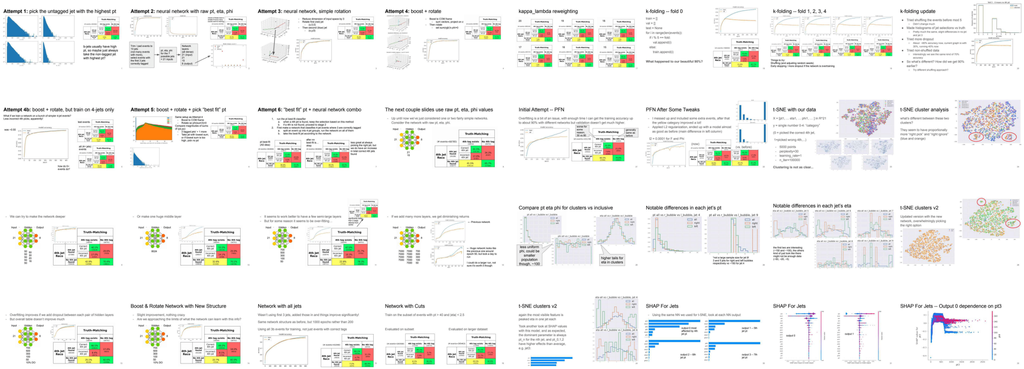
Machine Learning

Possible Approaches

How to pick a 4th jet?

- Since the HH interaction is "hardest," pick jet with highest p_T
- Since HH should have $\sum_{i=1}^{4} p_{T,i} = 0$, pick the jet that gives the smallest p_T sum.
- Play with classifier parameters
- If the data is sufficiently complex, a neural network or other form of machine learning may be the best option to use.

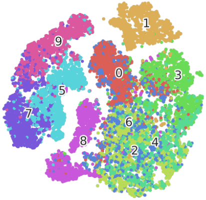
So We Tried A Bunch of Approaches…



Dataset Analysis With t-SNE

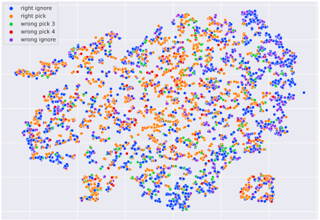
t-SNE: creates a low-dimensional embedding of a higher-dimensional dataset, can help visualize clustering and patterns in complicated data.

Example: MNIST-Fashion



Photos of types of clothing (e.g. 1=shirt). Clustering implies some nice structure behind the categories.

Our Data – 5 Table Categories



Much less clustering visible. Implies a more complex model (e.g. neural network) may be necessary.

Link to MNIST-Fashion image

The Best Network Yet

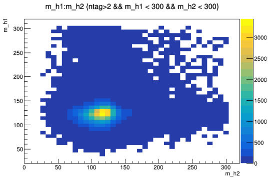
- Internal Layers: (700,500,300,100,50)
- Dropout: 10% after each internal layer
- Restrictions to match later analysis code: $p_T > 40$, $|\eta| < 2.5$

<table>
<tr><td colspan="2"></td><td colspan="3">Truth-Matching</td></tr>
<tr><td colspan="2">(# events=283508)</td><td colspan="2">4th tag exists
(113734)</td><td>No 4th tag
(169774)</td></tr>
<tr><td rowspan="3">4th Jet Reco</td><td rowspan="2">4th jet found
(114240)</td><td>Correct 4th jet</td><td>89.1%
(101296)</td><td rowspan="2">5.7%
(9667)</td></tr>
<tr><td>Incorrect 4th jet</td><td>2.9%
(3277)</td></tr>
<tr><td>No 4th jet found
(169268)</td><td colspan="2">8.1%
(9161)</td><td>94.3%
(160107)</td></tr>
</table>

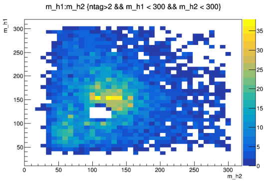
After cross-validation it looks like the accuracy may be lower than 90%, but still significantly better than previous approaches!

Integrating Back into the Main Project

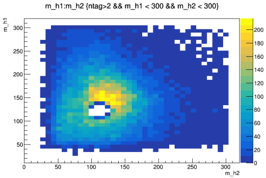
- We don’t yet know how inclusion of 3 *b*-tag events will influence the Higgs self-coupling.
- But we can look at how it affects background events.



Simulated signal



Background, 4-tag events



BG, 3+4-tag events.

- x and y axes are masses of reconstructed Hs.
- Blank spot: expected signal region, blinded for now.
- Addition of 3-tag events does not cause sculpting of background.

Images by Todd Seiss

Summary

Main Project: measuring Higgs self-coupling.

To do that, we need many *HH* events.

Using events with 3 *b*-tags may help, if we can correctly identify true *b*-jets (and thus keep backgrounds low).

Our neural network approach seems to enable this identification, to around 90% accuracy.

(# events=20488)			Truth-Matching: 4th tag exists (10396)	Truth-Matching: No 4th tag (10093)
4th Jet Reco	4th jet found (6782)	Correct 4th jet	46.2% (4804)	12.7% (1278)
		Incorrect 4th jet	6.7% (700)	
	No 4th jet found (13706)		47.1% (4891)	87.3% (8815)

⟶

(# events=283508)			Truth-Matching: 4th tag exists (113734)	Truth-Matching: No 4th tag (169774)
4th Jet Reco	4th jet found (114240)	Correct 4th jet	89.1% (101296)	5.7% (9667)
		Incorrect 4th jet	2.9% (3277)	
	No 4th jet found (169268)		8.1% (9161)	94.3% (160107)

Previous Best (≈ **65%**) vs. New Solution (≈ **90%**)

More testing with real data to follow!

Questions?

Comments?

Any other techniques we should try?

Backup slides

Link to Google Slides with more images