Undergraduate Research Opportunity Program:

SUEP X SubMIT

Christoph Paus, Chad Freer, Luca Lavezzo, Agustin Valdes

About Me

- 1st year undergraduate student from Live Oak, Texas
- Interested in physics, electrical engineering, and computer science
- Involvement in the SUEP project started in early January where I've been working under Chad Freer and Luca Lavezzo

Soft Unclustered Energy Patterns (SUEP)

- Soft unclustered energy patterns - anomalies existing in the QCD background of an event
    - Large multiplicity of soft (low transverse momentum particles)
    - SUEP candidates are found by reclustering tracks with large radius cone using **FastJet**
    - Hidden valley model with SUEP particles as connection to dark sector
- Looking at MC and 2018 data from CMS

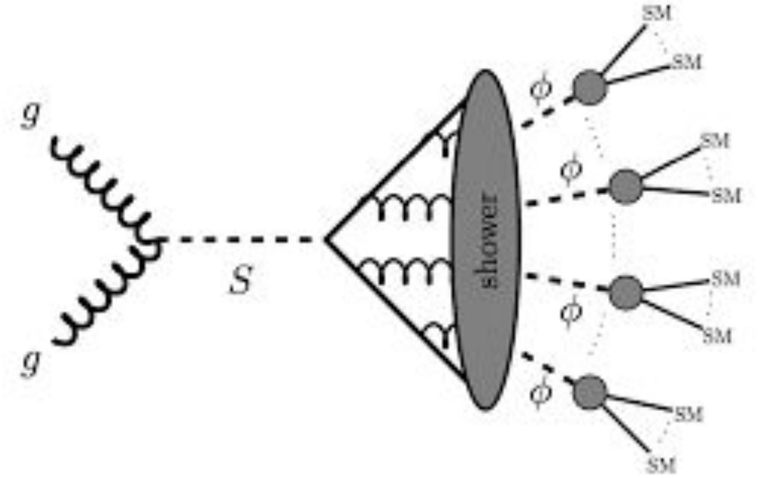
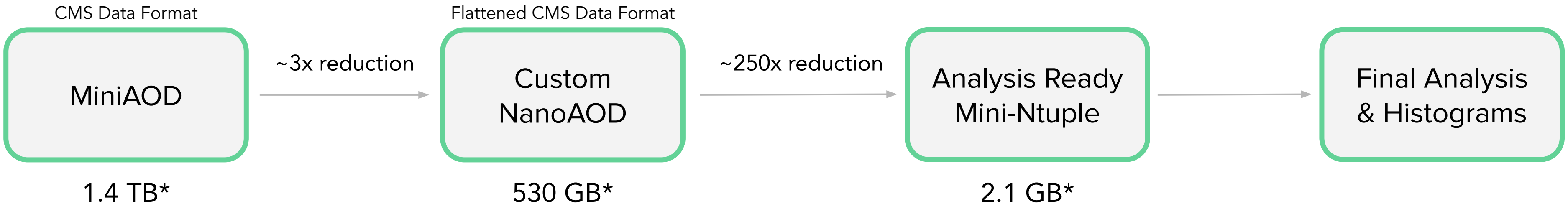


Fig. 1: Feynman Diagram of SUEP

Workflow Overview



Few Days	*Few Hours*	*Few Seconds*
Flattening via Kraken processing framework	Selection of interesting events and further slimming	Histograms & plots in Jupyter Hub
Stored in Tier 2 Hadoop	Stored in SubMIT directories /work/submit/user	Run through SubMIT machines

*Memory values from QCD_PT_170to300 dataset file

Workflow on SubMIT

- Specialized **NanoAOD** files (additional track information) are created through Kraken and stored on the Tier 2 Hadoop
  - Files are GB in size
  - **Kraken**: a processing framework used to breakdown large
  - files into smaller, often flat, files
  - **Hadoop**: storage system across computing clusters that can accessed remotely (Tier 2 vs. 3)
- Analysis-ready ntuples are created using HTCondor on the **Tier-2, Tier-3, and the CMS Global Pool** computing clusters (Fig. 2)
  - Columnar analysis framework is used (Coffea Singularity)
    - Iterative processes replaced with columnar operations
  - Tracks are clustered via **FastJet** algorithm with novel Awkward Array input (Fig. 3)

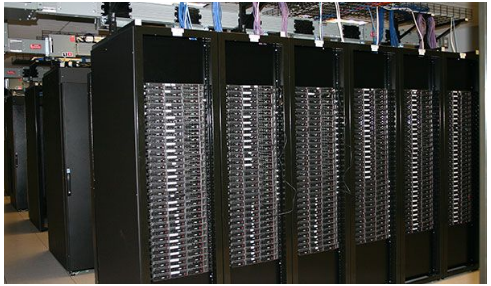


Fig. 2: Bates Lab tier 2 computing cluster

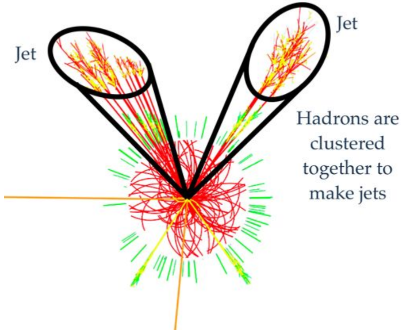


Fig. 3: FastJet anti-kt clustering, R=1.5

Workflow on SubMIT

- Histograms are created directly on SubMIT machines and can be plotted through SubMIT-hosted **JupyterHub (**Fig. 4)
  - By the time data files are accessed by the SubMIT machines, they are MB in size
- Despite decrease in file size, even small fraction of collision data from CMS is computationally-intensive to analyze
  - 20,000 - 30,000 files for 2018 analysis alone
- SUEP_coffea.py, where most of our analysis happens, is run in Singularity shell

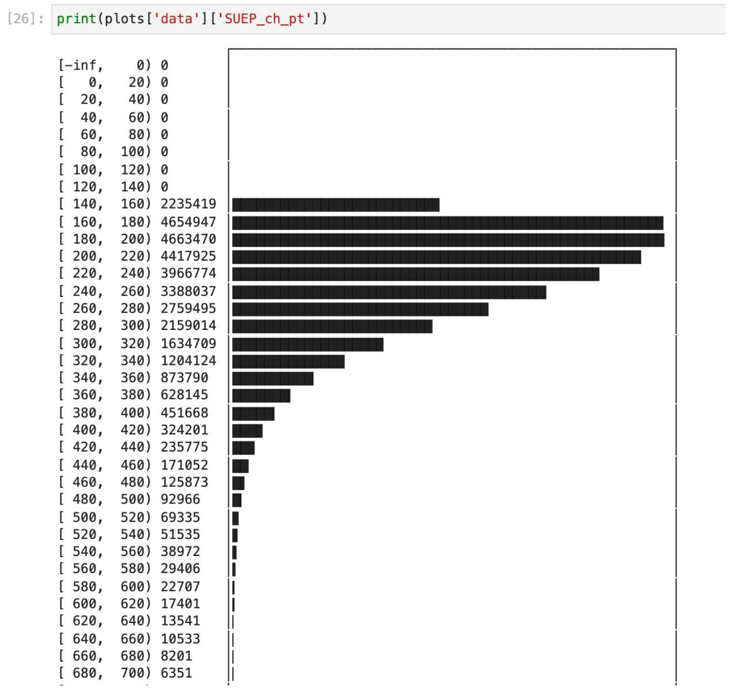


Fig. 4: Viewing produced histogram via JupyterHub

JupyterHub Interface

- A more user-friendly option is the JupyterHub interface which allows for more convenient plot generation and analysis of the data in an internet browser (Fig. 4)
- Can access all SubMIT data - home and work directories, Tier 2 Hadoop, etc.

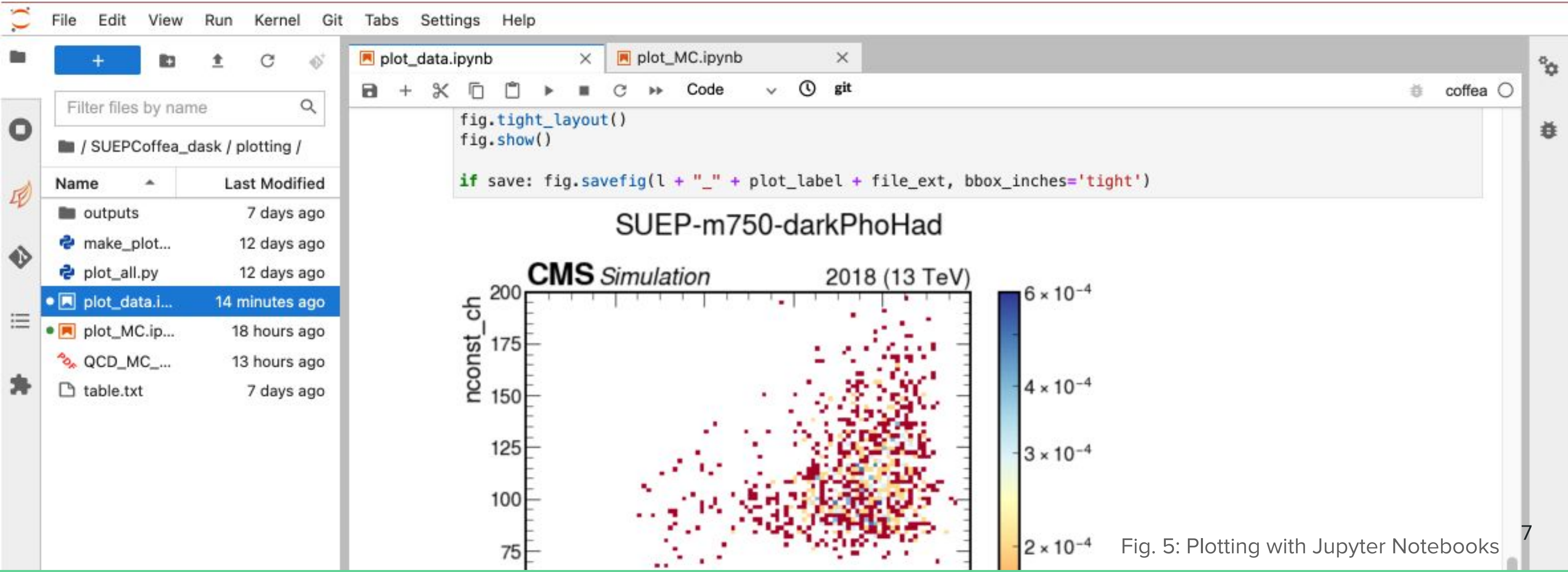


Fig. 5: Plotting with Jupyter Notebooks

Methodology - Clustering and Trigger Events

- Using **FastJet** to cluster the tracks into jets, we must sort through which jets are potential SUEP candidates
- Given sheer amount of data from proton-proton collision, triggers must be implemented to determine jets of interest (Fig. 7)
- Selection
  - QCD (background) events with HT > 1200 GeV
  - Number of tracks in jet > 1
  - At least one large radius jet with pT >150 GeV

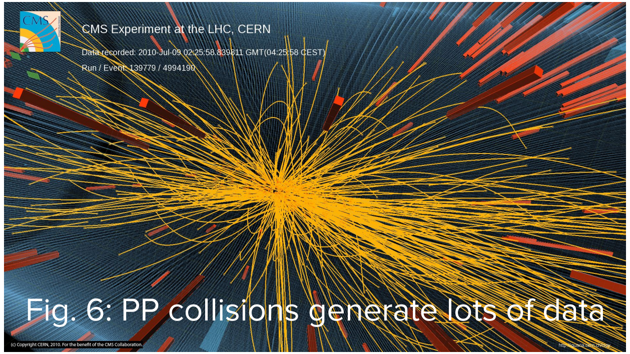


Fig. 6: PP collisions generate lots of data

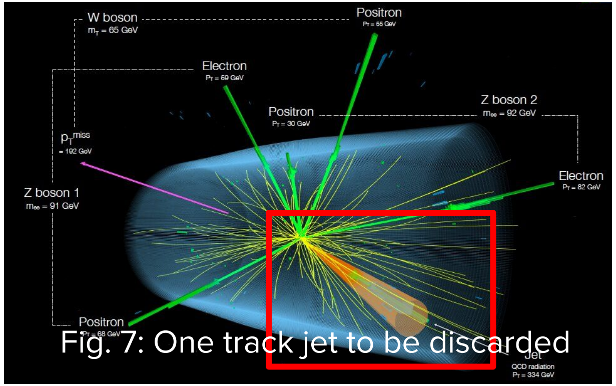


Fig. 7: One track jet to be discarded

Methodology - Variables of Interest

- Sphericity (***spher***): a measure of how uniformly distributed particles are from a point of interest (Fig. 5)
- Number of constituents (***nconst***): the total number of particles present in a SUEP candidate
- Transverse momentum (***pT***): the momentum perpendicular to the beamline (Fig. 6)
    - Conserved along this plane and gives an idea of how energetic a particle is along its track
- HT: the scalar sum of a AK4 jets' transverse momenta

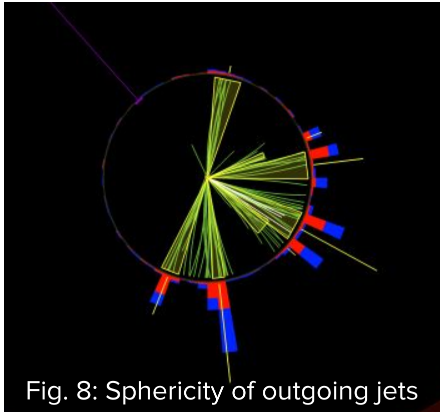


Fig. 8: Sphericity of outgoing jets

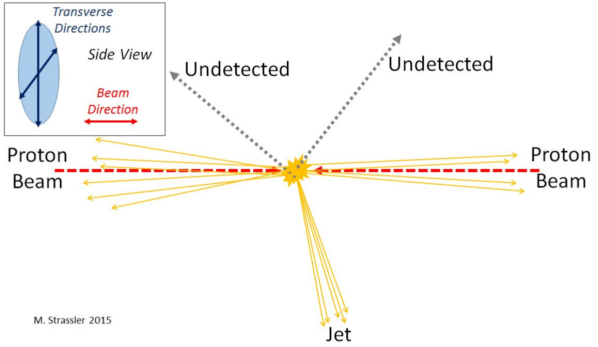
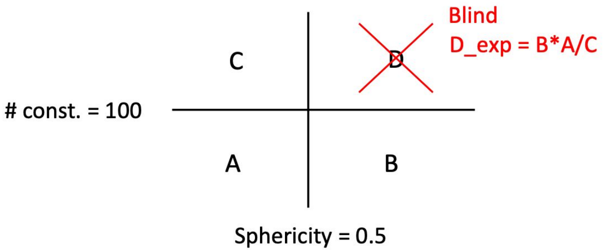


Fig. 9: Jet propelled into transverse plane

ABCD Method

- After running files through **Coffea**, we have data that is ready to be plotted and analyzed
- If a SUEP event is present, it would occur in where **nconst**. and **sphericity** are relatively greatest (D region of the Fig. 10)
- To avoid biasing the data, we predict what the D region will look like based on A,B, and C (data is blinded)



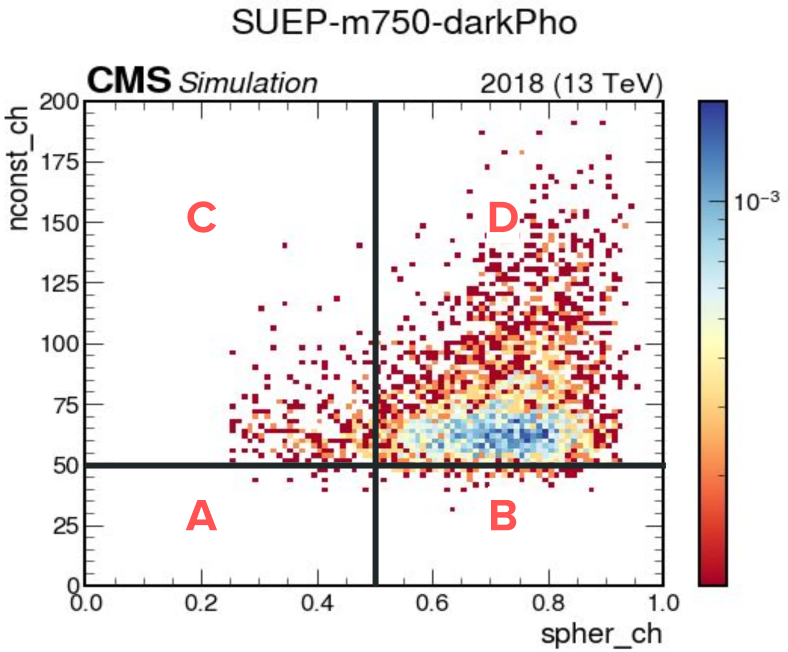


Fig. 10: SUEP sample subdivided by region for illustrating ABCD Method

QCD_MC vs. Data Discrepancy

MC QCD background prediction vs. actual SUEP sample

- Data discrepancy not visible from this 2D plot alone, must focus-in on one region at a time

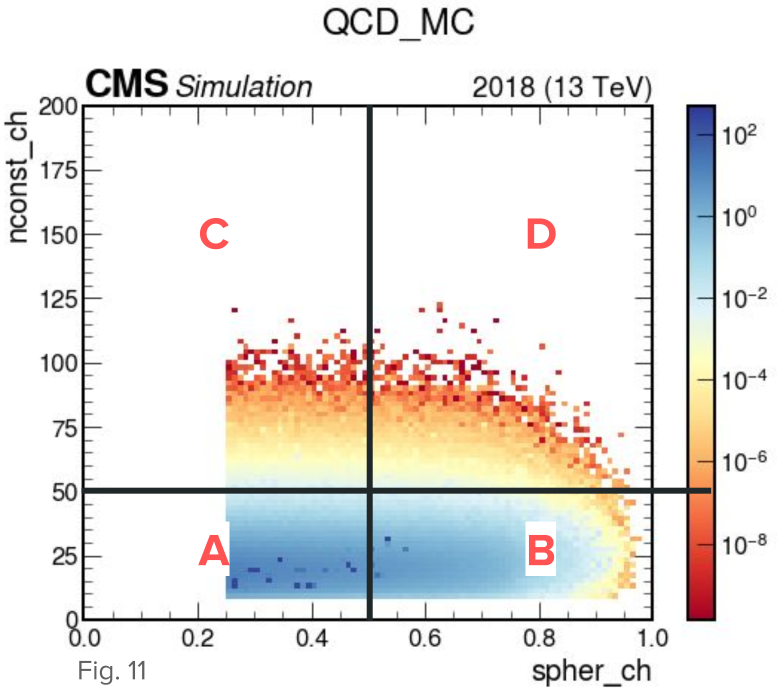


Fig. 11

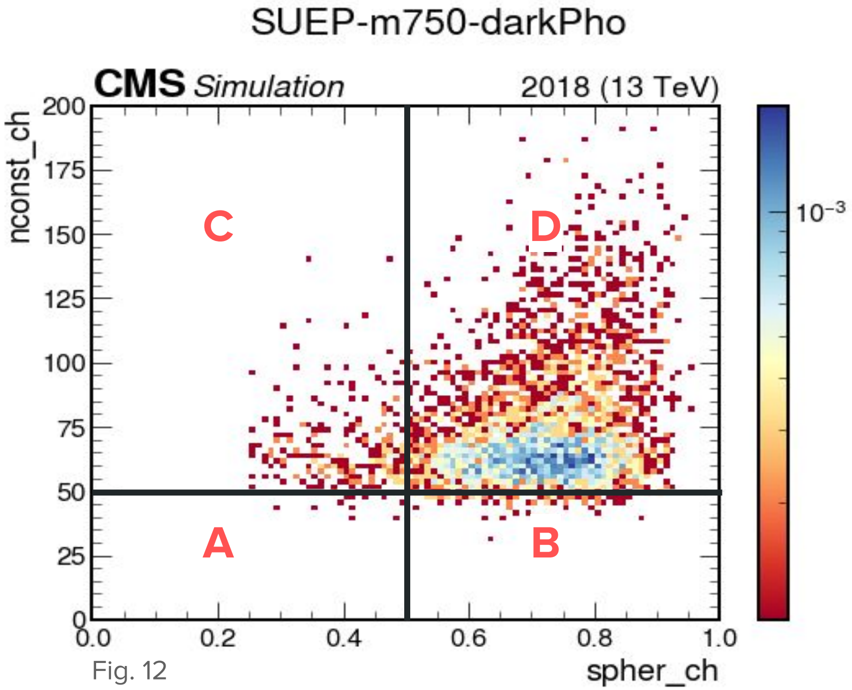


Fig. 12

QCD_MC vs. Data Discrepancy

pT plot revealed major disagreement for low pT SUEP events

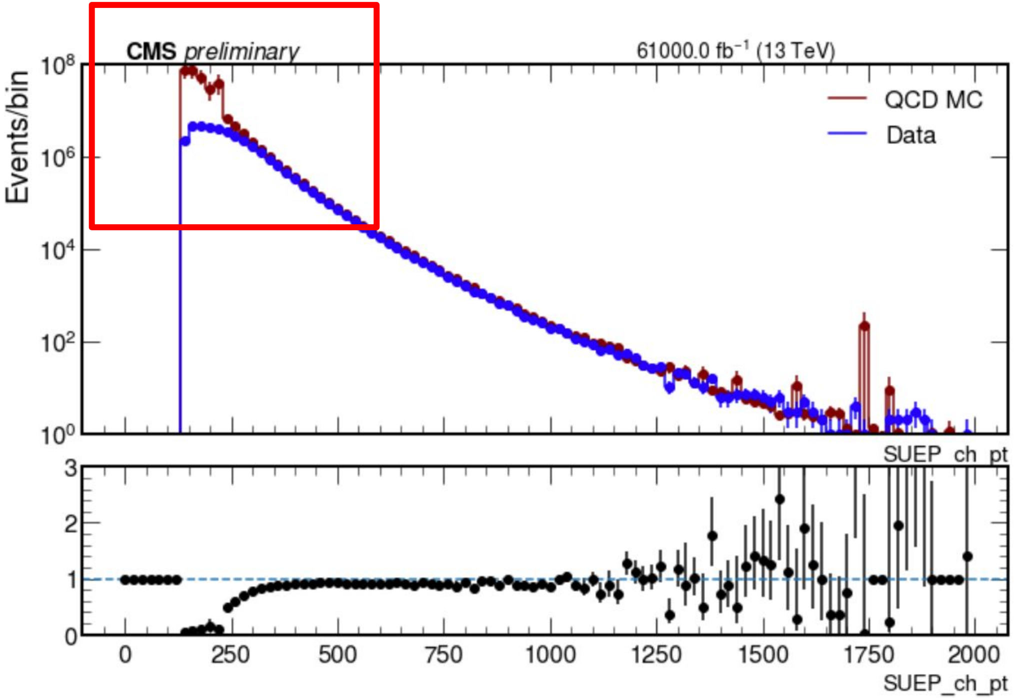


Fig. 13

Course of action: recreate plots for SUEP events with only pT>300 GeV

QCD_MC vs. Data After pT<300 Cut

Removing constituents with a pT<300 yields much better fit

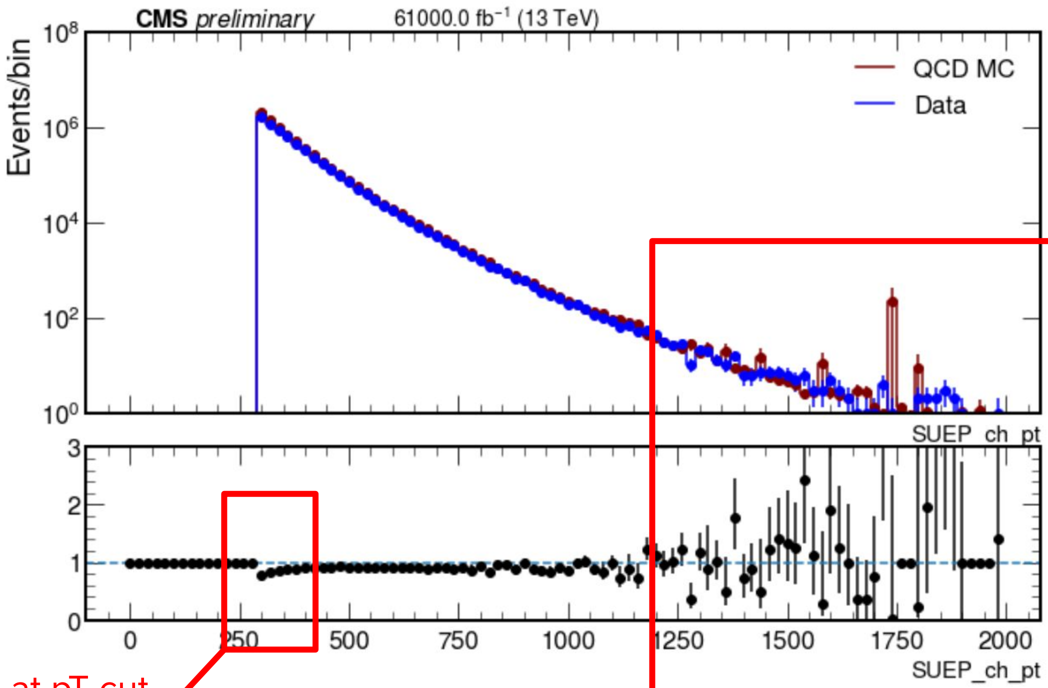


Fig. 14

Expected behavior as # of events approaches 1

Discrepancy right at pT cut

A Region (Before and After PT Cut)

This agreement becomes more clear when looking at one region at a time:

- And thanks to JupyterHub, we're able to generate these plots in a matter of seconds

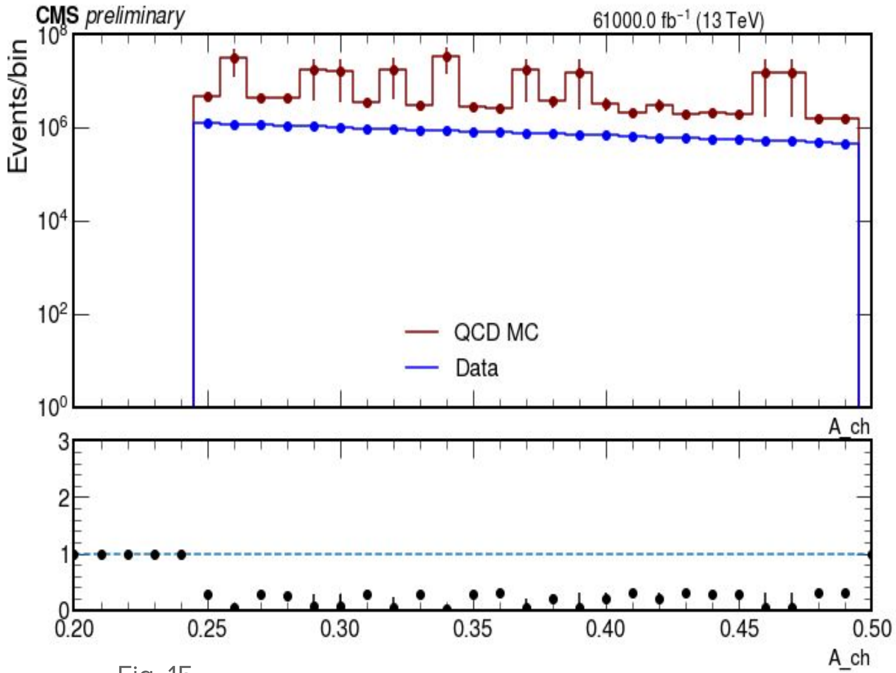


Fig. 15

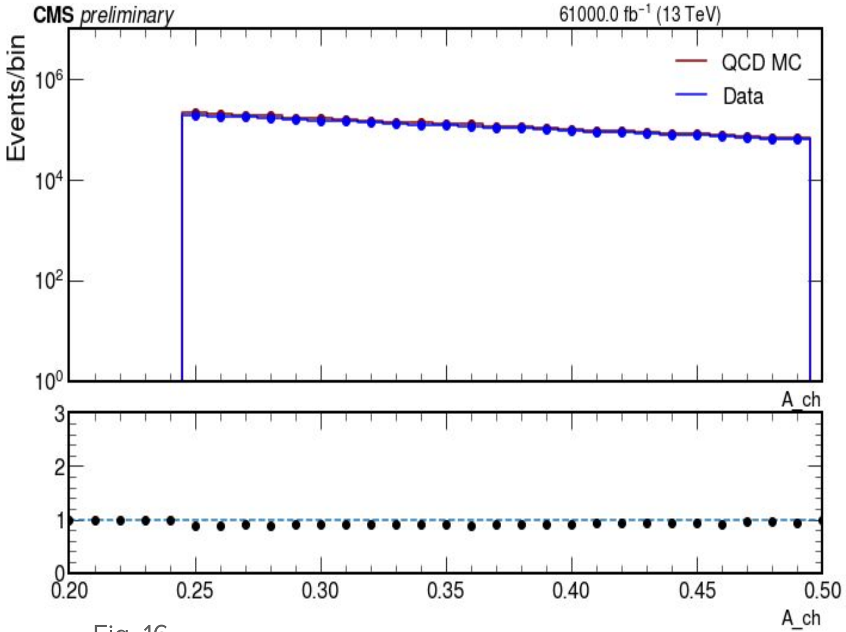


Fig. 16

B Region (Before and After PT Cut)

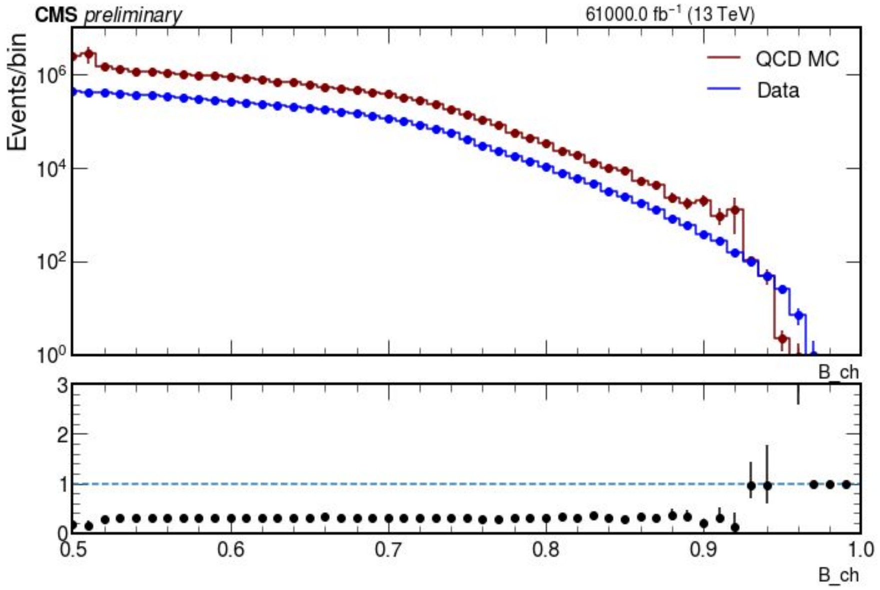


Fig. 17

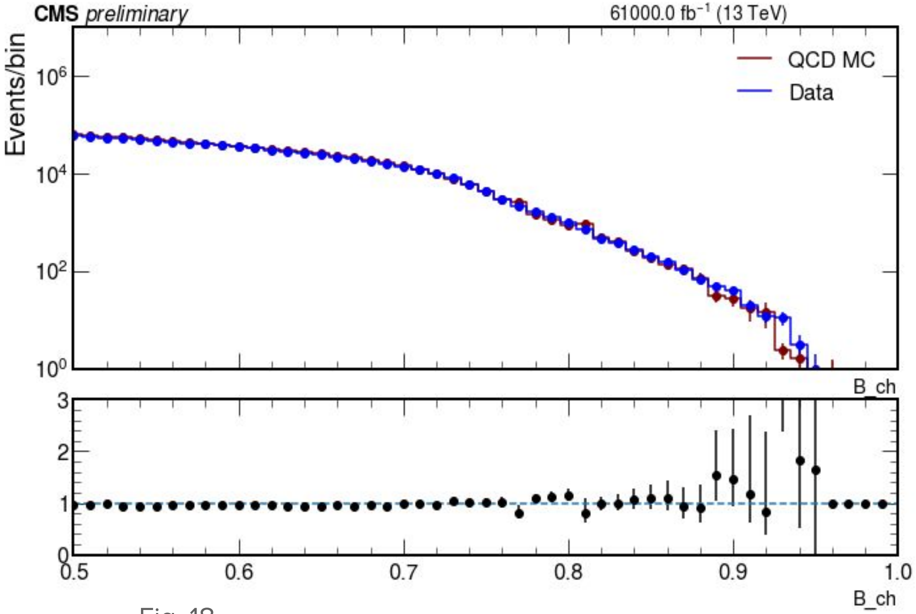


Fig. 18

C Region (Before and After PT Cut)

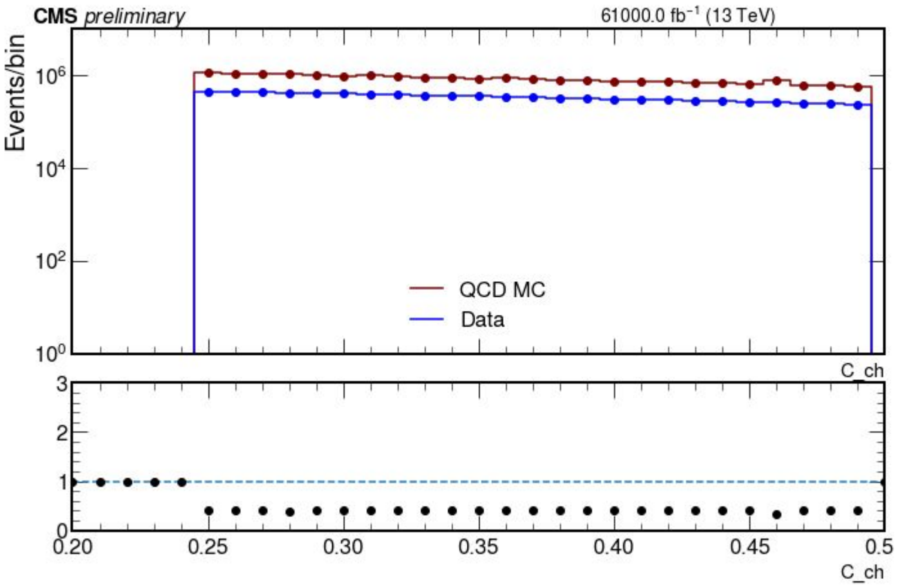


Fig. 19

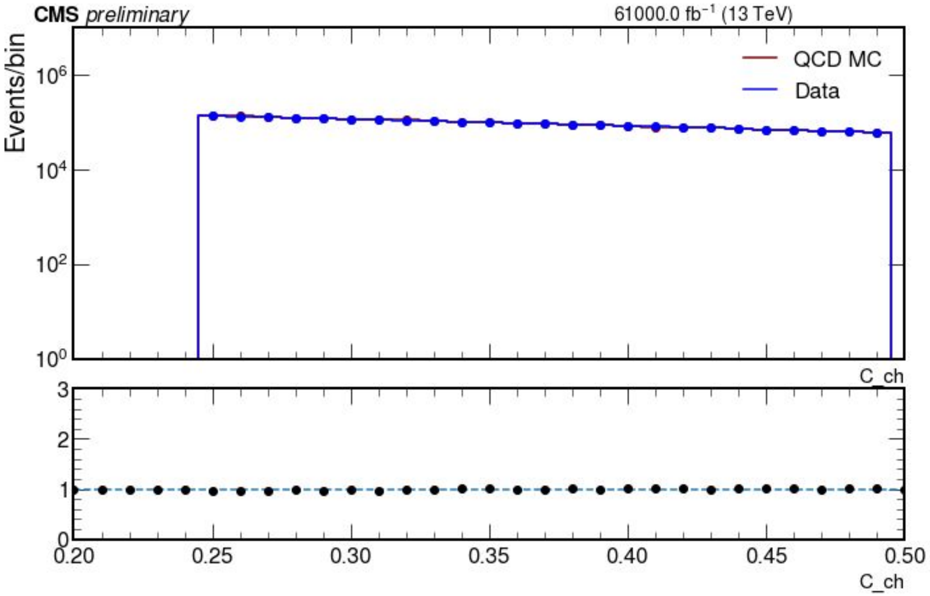


Fig. 20

D Region (Before and After PT Cut)

- Changes in other regions cumulate in the D region, with near perfect agreement save for the low bin tail

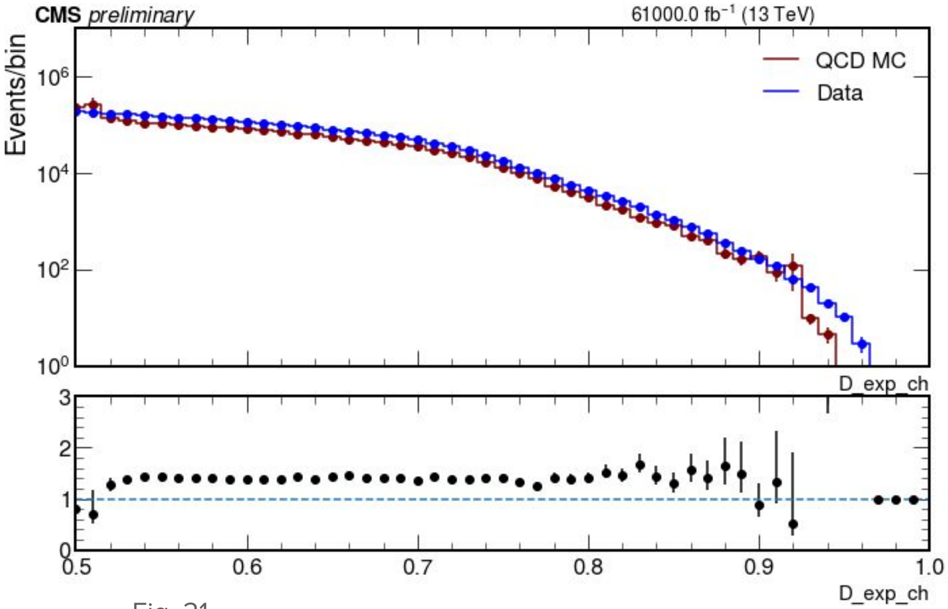


Fig. 21

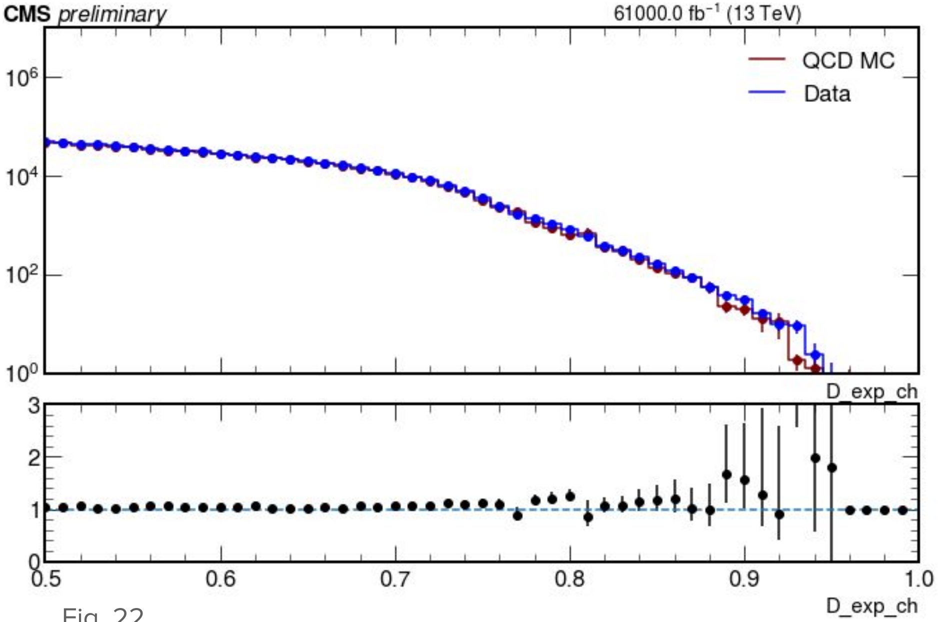


Fig. 22

Future Steps in Workflow Development

Incorporating neural network into analysis

- Train NN to distinguish between QCD background and SUEP particles
- Using SubMIT machines with GPUs to train identification algorithm
- **Triton**, an open-source platform for GPU-driven neural networks, used to do the inference
    - Already included within Coffea Singularity; scale up analysis of files

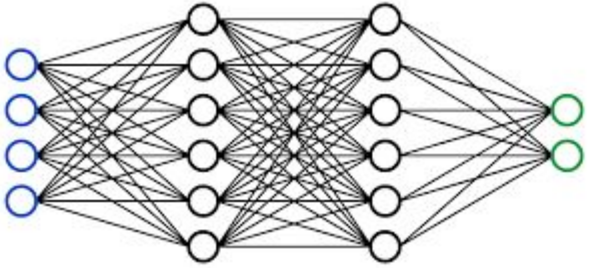


Fig. 23

Calculating and plotting limits

- How sensitive one selection is compared to another
- Run straight from JupyterHub

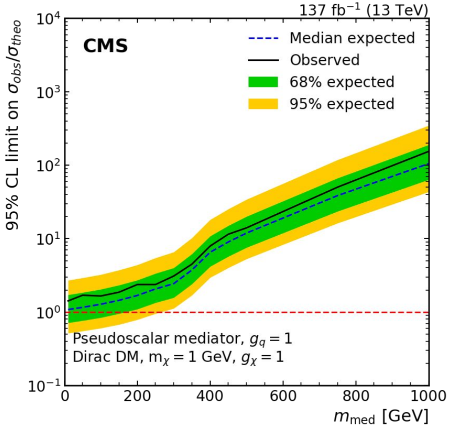


Fig. 24

Sources

Images:

Fig. 1 - http://t3serv001.mit.edu/~paus/suep/2020.08.06.kdp.SUEPsforLPCDM.pdfhttp://t3serv001.mit.edu/~paus/suep/2020.08.06.kdp.SUEPsforLPCDM.pdf

Fig. 2 - https://bateslab.mit.edu/high-performance-research-computing-facility

Fig. 3 - https://www.mdpi.com/2218-1997/5/5/114/htm

Fig. 6 -https://cms.cern/news/new-two-particle-correlations-observed-cms-detector-lhc

Fig. 7 - https://phys.org/news/2020-12-triple-threat-massive-gauge-bosons.html

Fig. 23 - https://victorzhou.com/series/neural-networks-from-scratch/

Sources

Other Resources:

- Soft Unclustered Energy Patterns (SUEP)
  - https://inspirehep.net/literature/800288
  - https://arxiv.org/pdf/2011.06599.pdf
  - https://profmattstrassler.com/articles-and-posts/relativity-space-astronomy-and-cosmology/dark-matter/searching-for-dark-matter-at-the-lhc/
- ABCD Method
  - https://twiki.cern.ch/twiki/pub/Main/ABCDMethod/ABCDGuide_draft18Oct18.pdf
- Methodology - Clustering and Trigger Events
  - https://link.springer.com/article/10.1007/s13538-014-0212-z
  - https://atlas.cern/updates/blog/what-happens-when-energy-goes-missing
- Methodology - Variables of Interest
  - https://arxiv.org/abs/1005.3299
  - https://cms-opendata-workshop.github.io/workshop-lesson-jetmet/aio/index.html
  - https://cds.cern.ch/record/1447810/files/epjc.72.2124.pdf

Questions?