Search for an exotic Higgs boson decay to two pseudoscalar bosons with a four photon final state

Oral Candidacy Proposal

December 13, 2023

Sergi Castells

Advisors: Colin Jessop and Nancy Marinelli

UNIVERSITY OF NOTRE DAME

I. The Standard Model and Beyond

II. LHC and CMS

III. Introduction to Analysis

IV. Sample Generation

V. Event Selection

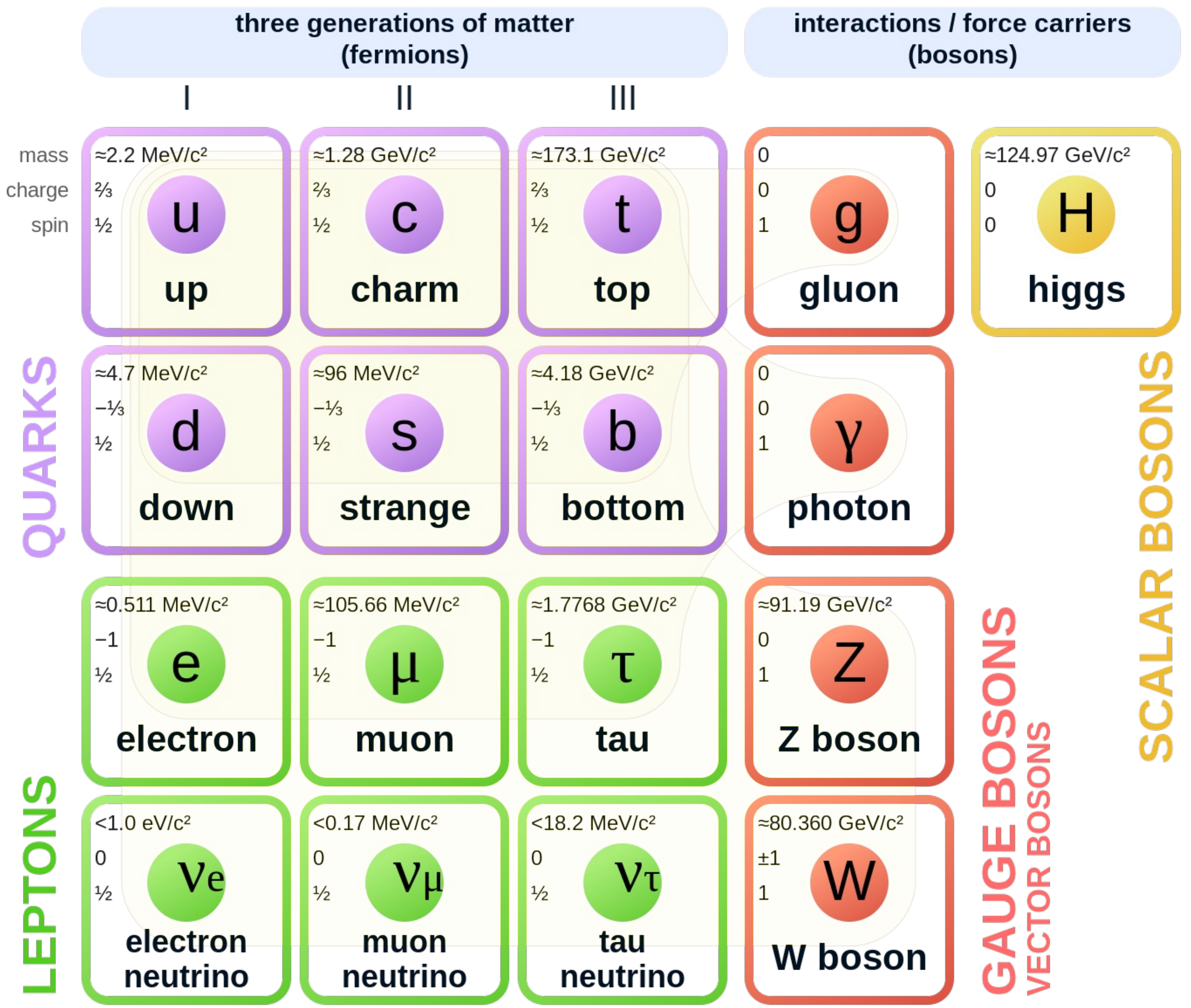
VI. Signal Modelling

VII. Advancements on Previous Studies

VIII. Summary

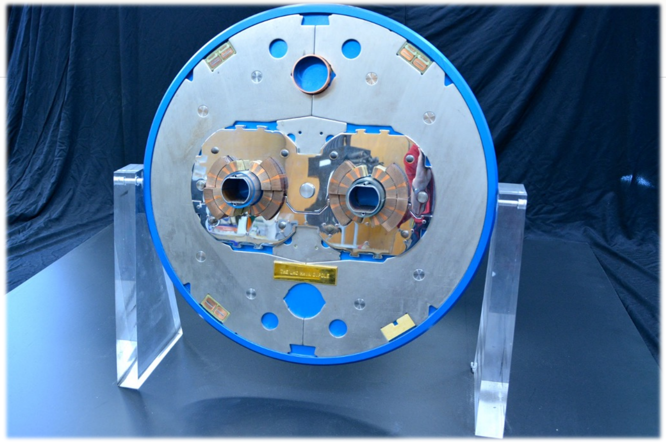
The Standard Model and Beyond

- Most complete model of particle physics that describes all known elementary particles.
- Describes three fundamental interactions:
  - Weak
  - Strong
  - Electromagnetic
- Rich Higgs sector phenomenology:
  - Higgs mechanism yields gauge boson masses
  - Higgs VEV & Yukawa couplings yield other particle masses excluding neutrinos
- Beyond the Standard Model physics can remedy its shortcomings and predict interesting, new physics.

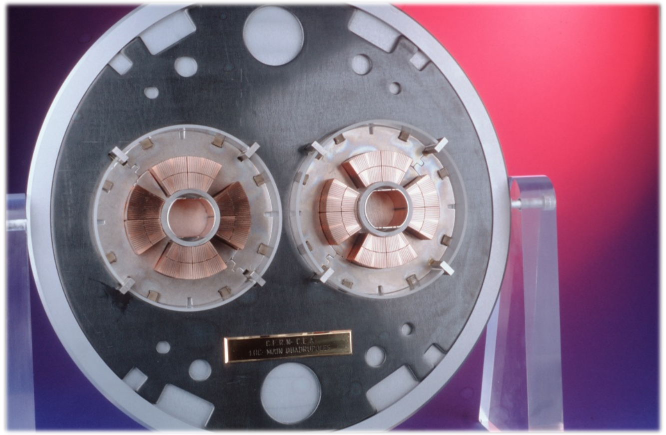


The Large Hadron Collider (LHC)

- The LHC is the world's largest particle accelerator, located on the border of France and Switzerland.
- It is 27 km in circumference and operates at a center-of-mass energy of 13.6 TeV.
- The LHC can achieve a nominal instantaneous luminosity of 10^{33} cm^{-2} s^{-1}, colliding proton beams at a frequency of 40 MHz.
- Superconducting dipole and quadrupole magnets bend and squeeze the proton beams, respectively.
- Four detectors along beamline:
  - General purpose: CMS and ATLAS
  - Specialized: ALICE and LHCb



Above: Slice of LHC dipole magnets.
Below: Slice of LHC quadrupole magnets



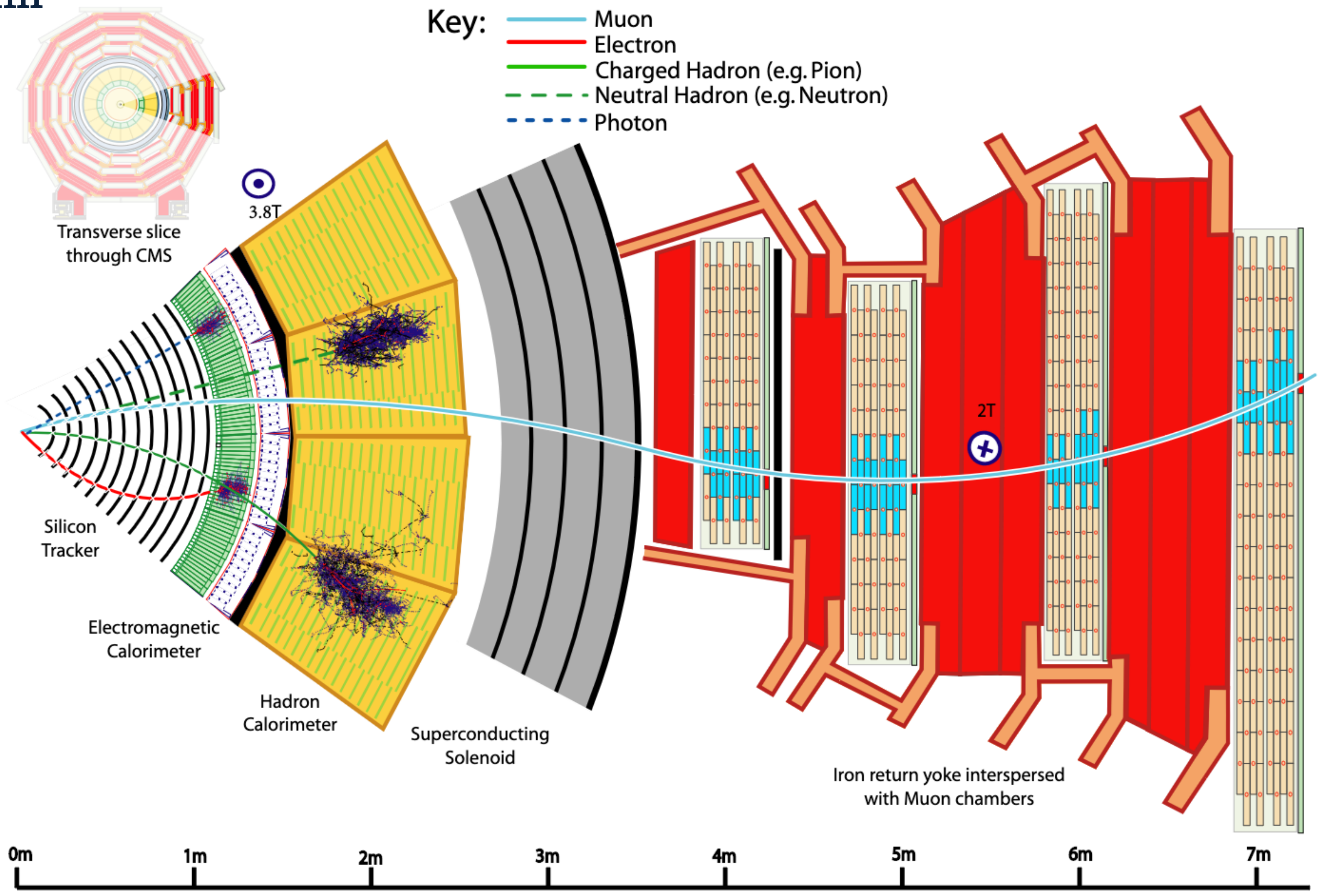
Compact Muon Solenoid (CMS)

- **Silicon tracker**
  - Solid-state pixel and strip detectors
  - Detects charged particle tracks
- **Electromagnetic calorimeter (ECAL)**
  - Homogenous detector
  - Consists of 75,000 $PbWO_4$ crystals
  - Detects electromagnetically interacting particle showers
- **Hadronic calorimeter (HCAL)**
  - Sampling calorimeter
  - Layers of plastic scintillator and brass absorbers
  - Detects neutral and charged hadrons
- **Solenoid Magnet**
  - Bends charged particles to measure momentum
- **Muon chambers**
  - Detects muons

Object Reconstruction

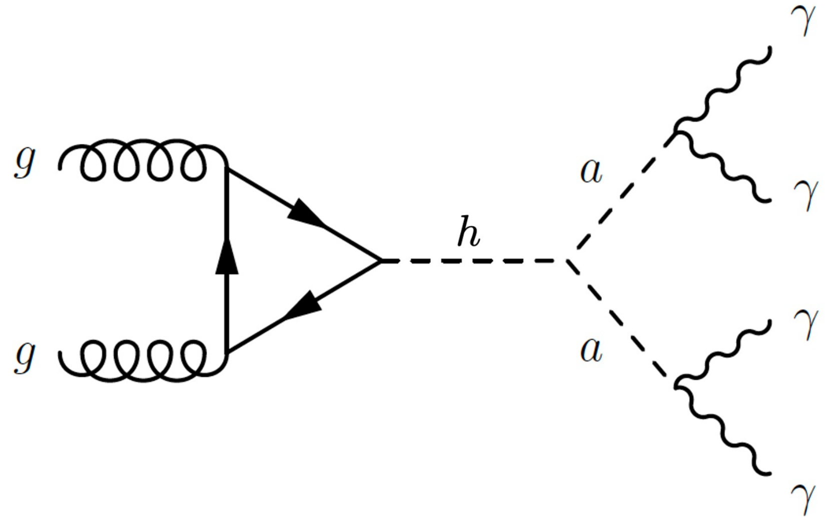
Particle Flow (PF) Algorithm

- Takes raw hits in the detector and creates PF elements
- Links PF elements between sub-detectors to create particle candidates
  - Reconstruct muons
  - Electrons and isolated photons
  - Charged/neutral hadrons and non-isolated photons
- Combines particle candidates to build a picture of a collision event



Introduction to Analysis

Signal: $h \to aa \to \gamma\gamma\gamma\gamma$



Why do this analysis?

- Clean final state with low SM background
- Model agnostic search allows for greater impact of any limits or discovery
- Improved sensitivity could allow for a discovery

- Currently working on a proof-of-concept study with the $\mathbf{h} \to \gamma\gamma$ group's new framework: Higgs to Diphoton NanoAOD Framework (HiggsDNA).

- This proof-of-concept study consists of recreating a previous Run 2 analysis for 2018 data. The goal is to apply this work to Run 3 data and increase analysis sensitivity in hopes of producing a discovery.

Sample Generation

- **Signal Monte Carlo (MC):**
  - Simulated $h \rightarrow aa \rightarrow \gamma\gamma\gamma\gamma$ samples
  - Pseudoscalar mass range of $15 < \mathrm{m_a} < 60$ GeV in steps of 5 GeV
  - Decays from a SM Higgs boson with $\mathrm{m_h} = 125$ GeV
  - Only gluon fusion production mode of the Higgs boson considered

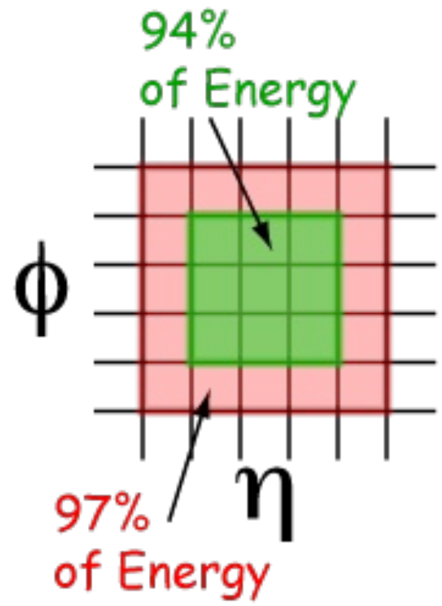
- **Background:**
  - A data-driven background is used
  - Use event mixing technique:
    - Shuffle photons in an event with photons from consecutive events
    - Removes presence of any signal while retaining background shape
  - Main backgrounds: either prompt photons from collision event or isolated photons reconstructed due to very collimated decays fragmented from jets (fake photons)

Choosing Candidate Photons

Energy deposition and shape of photon candidates:

E_T	R_9
Transverse energy of a photon candidate measured in ECAL.	Ratio of energy deposited in ECAL inside 3x3 crystals centered on the most energetic crystal in the supercluster and the supercluster energy.

Supercluster:
5×5 ECAL Crystals



H/E	$\sigma_{i\eta i\eta}$
Ratio of energy deposited in nearest HCAL tower within a cone of radius $\Delta R = 0.15$ centered on photon candidate direction and energy deposited in its supercluster.	Energy weighted standard deviation of a single crystal's η within 5x5 crystals centered at the crystal with maximum energy.

Choosing Candidate Photons

Isolation of photon candidates:

Particle Flow Photon Isolation	Tracker Isolation
Sum of E_T of all Particle Flow photon candidates within a cone of radius $\Delta R = 0.3$ excluding candidate photon.	Sum of track p_T within a cone of $\Delta R = 0.3$.

Charged Hadron Isolation	Electron Veto
Sum of p_T of all Particle Flow charged hadron candidates within a cone of radius $\Delta R = 0.3$.	Checks that no electron candidate track matches photon candidate supercluster.

Pre-selections

γ_1 must be:	E_T	R_9	H/E	$\sigma_{i\eta i\eta}$	PF Pho Iso	Tracker Iso
EB; $R_9 > 0.85$	15.0	> 0.5	< 0.08	< 0.015	< 4.0	< 6.0
EB; $R_9 \leq 0.85$						
EE; $R_9 > 0.9$				< 0.035		

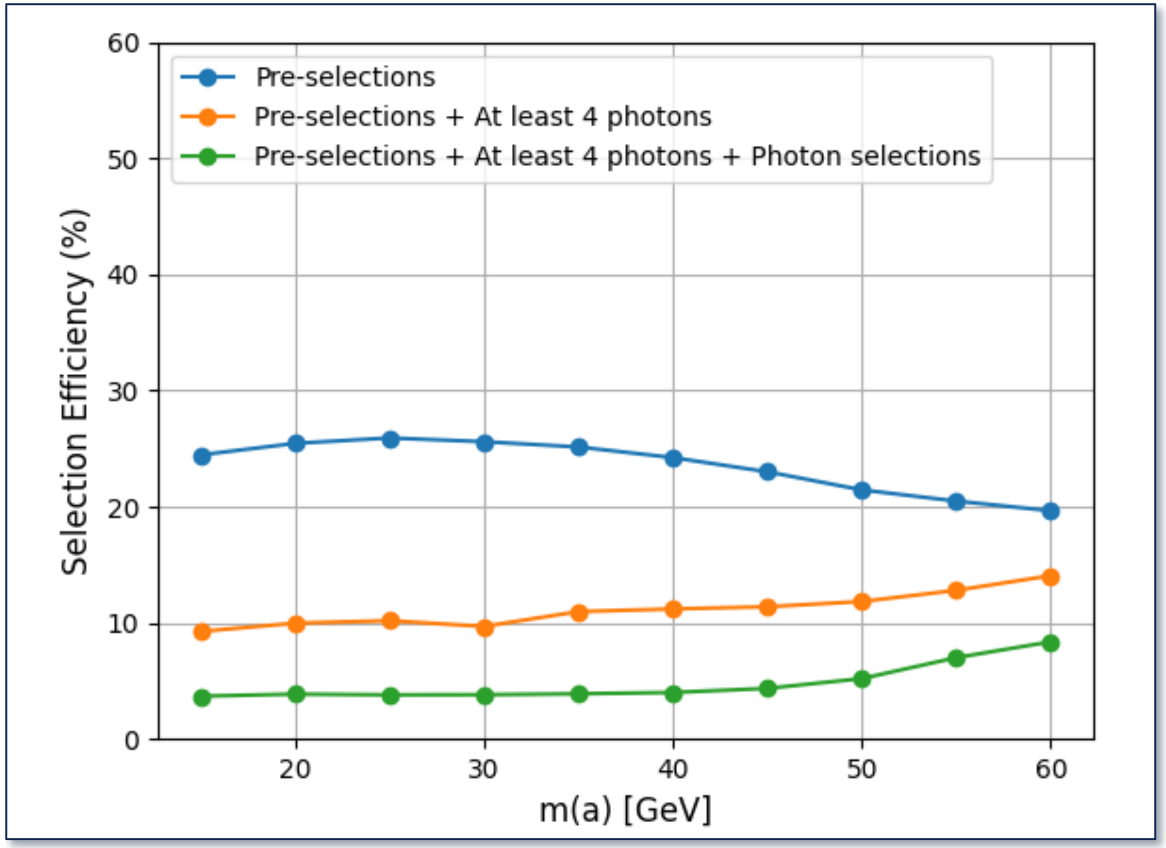
- Selections slightly more strict than High Level Trigger (HLT) are applied
- Events must pass at least one of Options 1-3 (motivated by miniAOD cuts)
- Photons must also pass an electron veto
- Events must have at least two photon candidates
- Diphoton candidates require $p_T > 30$ GeV and $p_T > 18$ GeV for leading and subleading daughter photons

Option 1	Option 2	Option 3
$R_9 > 0.8$	Charged Hadron Iso < 20.0	Charged Hadron Iso/$p_T < 0.3$
		IF $p_T > 14.0$
		AND H/E < 0.15

Photon Selections

- Require events have at least one diphoton candidate before applying photon selections
- Selections are applied to events passing pre-selections with at least 4 photon candidates
- $\Delta\mathrm{M} = |m_{\gamma_a\gamma_b} - m_{\gamma_c\gamma_d}|$ is minimized to choose optimal pseudoscalar candidates

	$\mathrm{p_T}$	η	$m_{\gamma\gamma\gamma\gamma}$
γ_1	> 30.0	$\lvert\eta\rvert < 2.5$	$110 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV
γ_2	> 18.0		
γ_3	> 15.0	$1.442 < \lvert\eta\rvert < 1.556$	
γ_4	> 15.0		



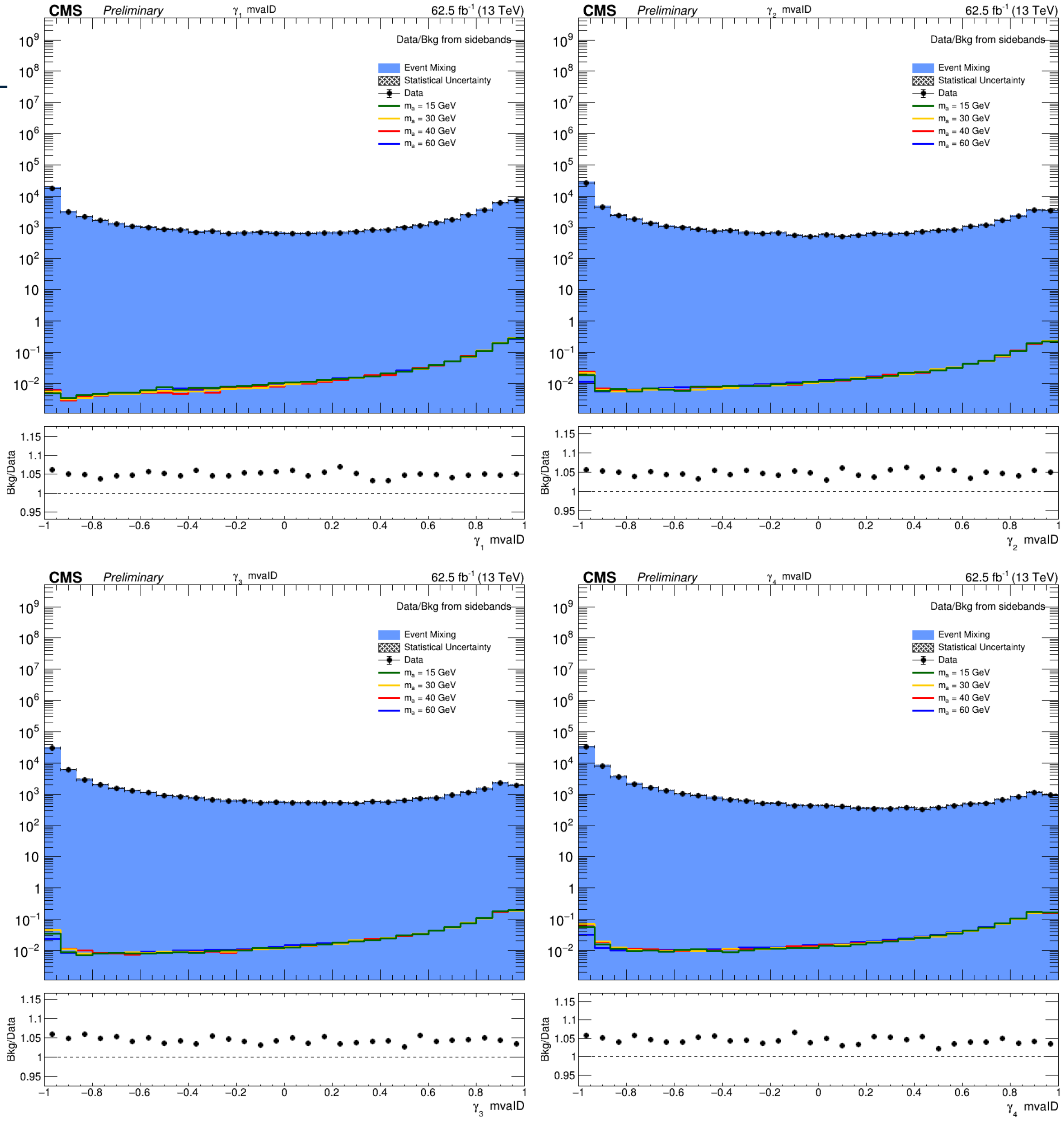
Above: Efficiencies of selections at various stages of selections for all nominal pseudoscalar mass points. This should improve at higher mass points.

Left: Additional selections applied to photons passing pre-selections.

Results of Selections

Photon MVA ID

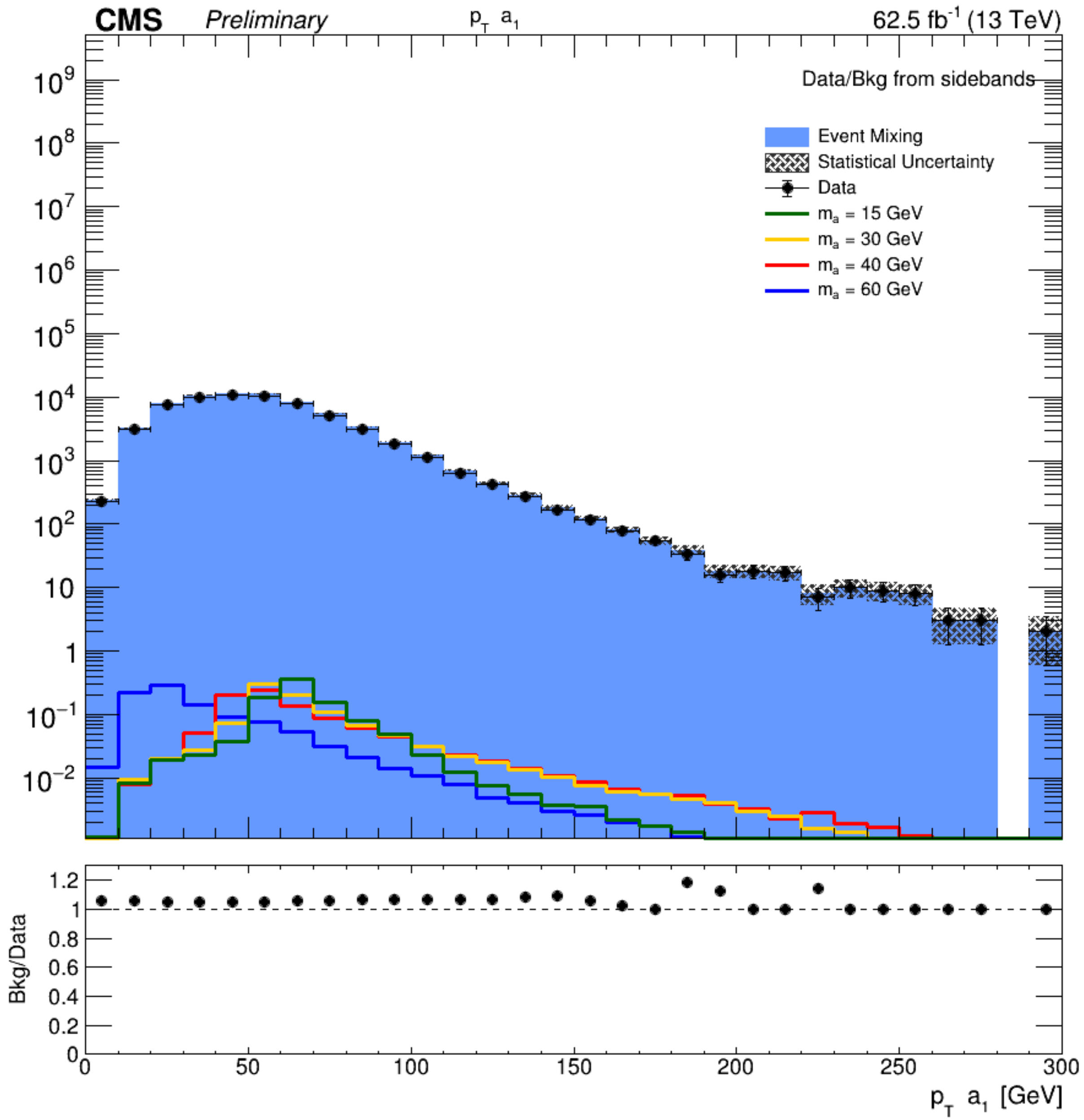
- Multivariate Analysis technique employing a boosted decision tree
- Discriminates real photons from fake photons (collimated jets, prompt photons)
- Likelihood score: High score (~1) corresponds to signal-like (real photon) and low score (~ -1) corresponds to background-like (fake photon)
- No cuts are applied to this variable to maximize training data for event selection BDT



Results of Selections

Are pseudoscalar kinematics sensible?

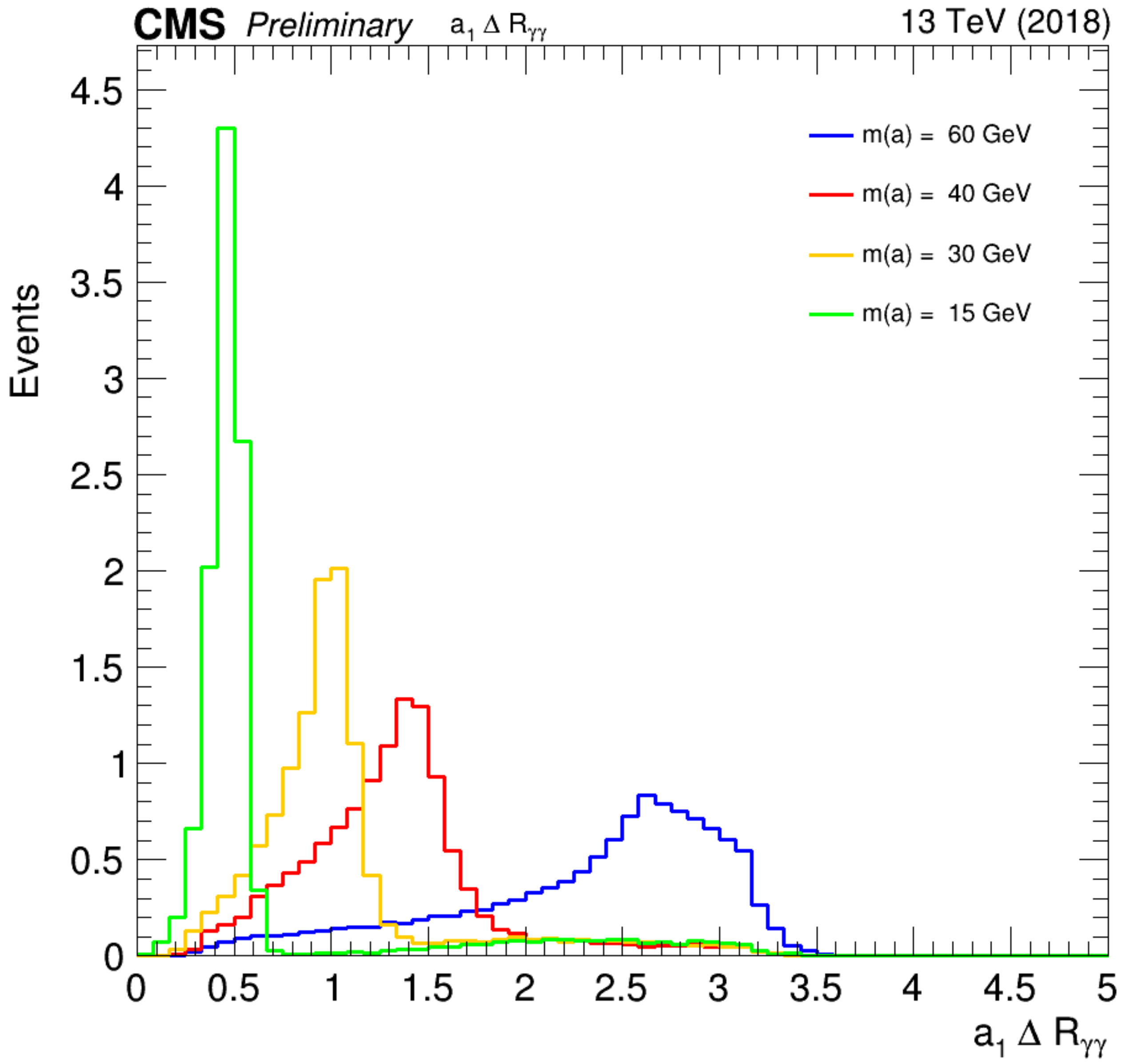
- p_T peak moves left with increasing mass
- ΔR peaks move right with increasing mass means less boosting as pseudoscalar mass increases



Results of Selections

Are pseudoscalar kinematics sensible?

- p_T peak moves left with increasing mass
- ΔR peaks move right with increasing mass means less boosting as pseudoscalar mass increases

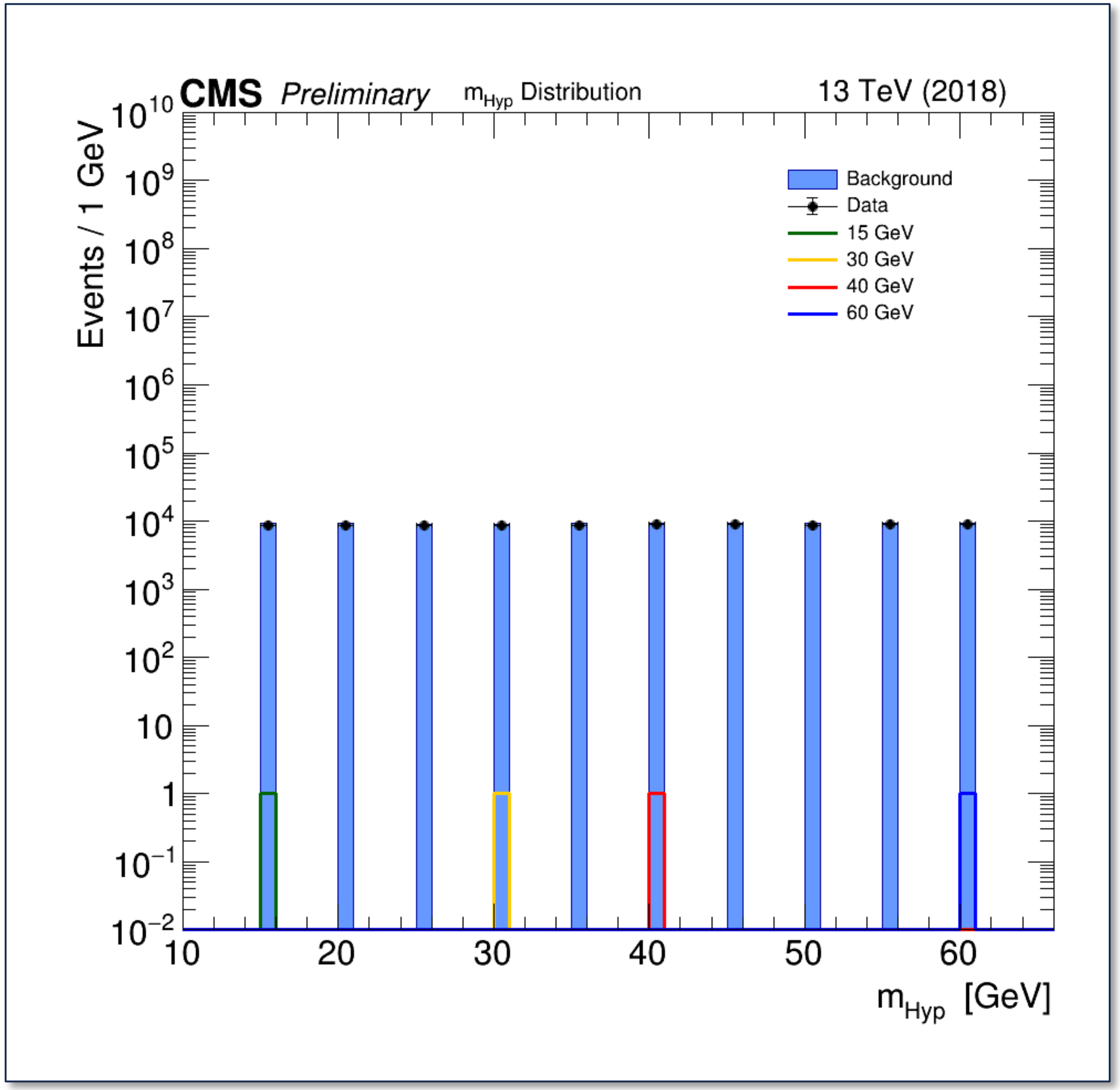


Event Selection BDT

What is a BDT?

- Boosted Decision Tree
- Combination of many weak learners (trees) into a strong classifier
- Need a BDT to better distinguish signal-like events from background-like events

- A BDT can significantly improve analysis sensitivity compared to a purely cut-based analysis.
- One event selection BDT model is produced for all mass points, made possible by using a constructed variable: hypothesis mass or m_{Hyp}
- m_{Hyp} has the associated value for a given pseudoscalar mass point for signal MC but a flat distribution for event mixed background and data, by construction.

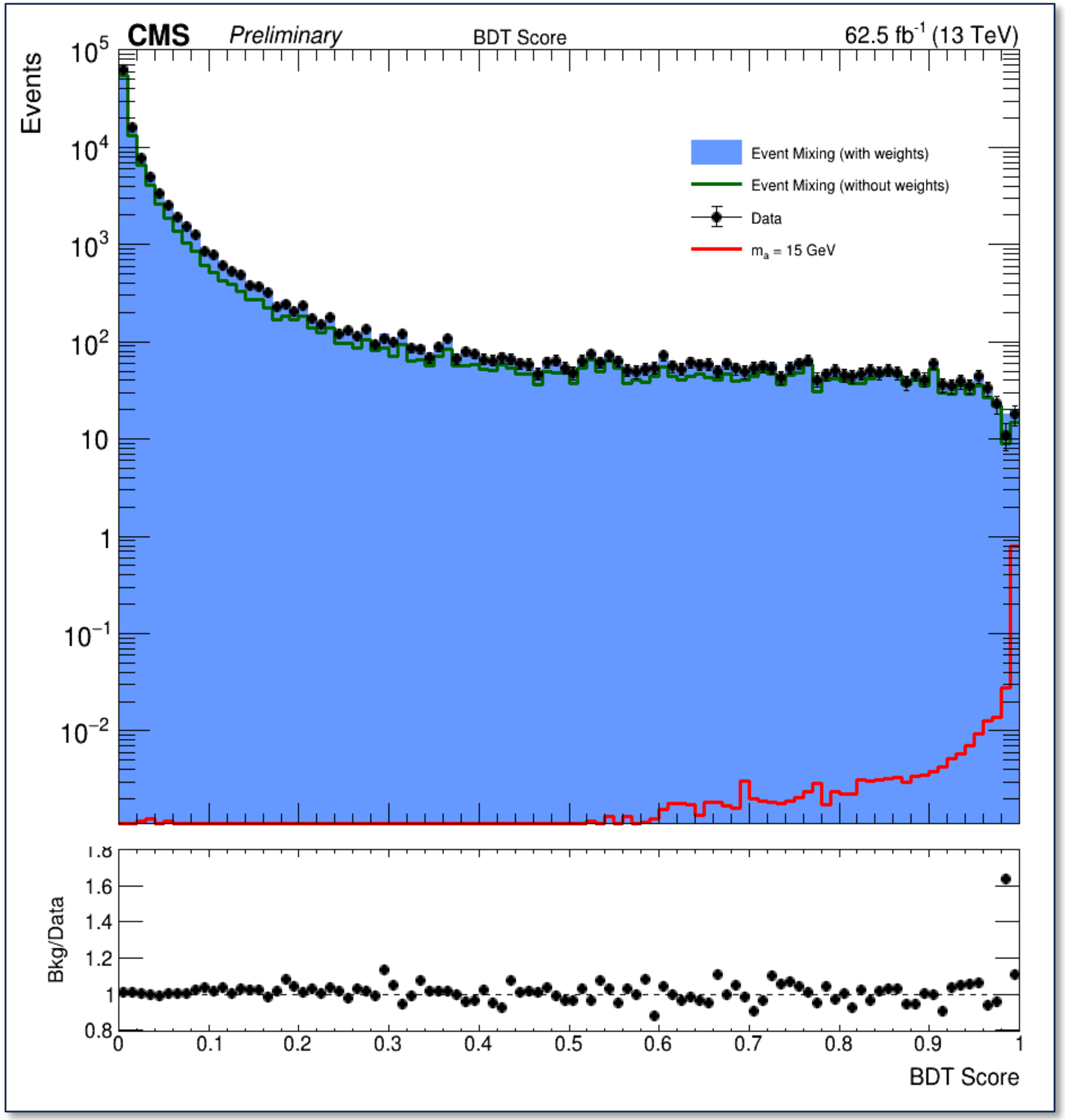


Applying the Event Selection BDT

- The event selection BDT model is applied to all samples for each nominal mass point.

For each nominal mass point:

- m_{Hyp} is set to the nominal mass point of interest for appropriate signal MC sample, event mixed background, and data.
- Input variables dependent on m_{Hyp} are recalculated.
- A prediction distribution of event selection BDT score is generated for signal MC, event mixed background, and data. Shown for $m_a = 15$ GeV.



BDT Score Categorization

- The goal of the event selection procedure is to maximize significance of the analysis. Approximate Mean Significance (AMS) is defined as:

$$\mathrm{AMS} = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right)-S\right]}$$

where S and B reference to the number of signal and background events.

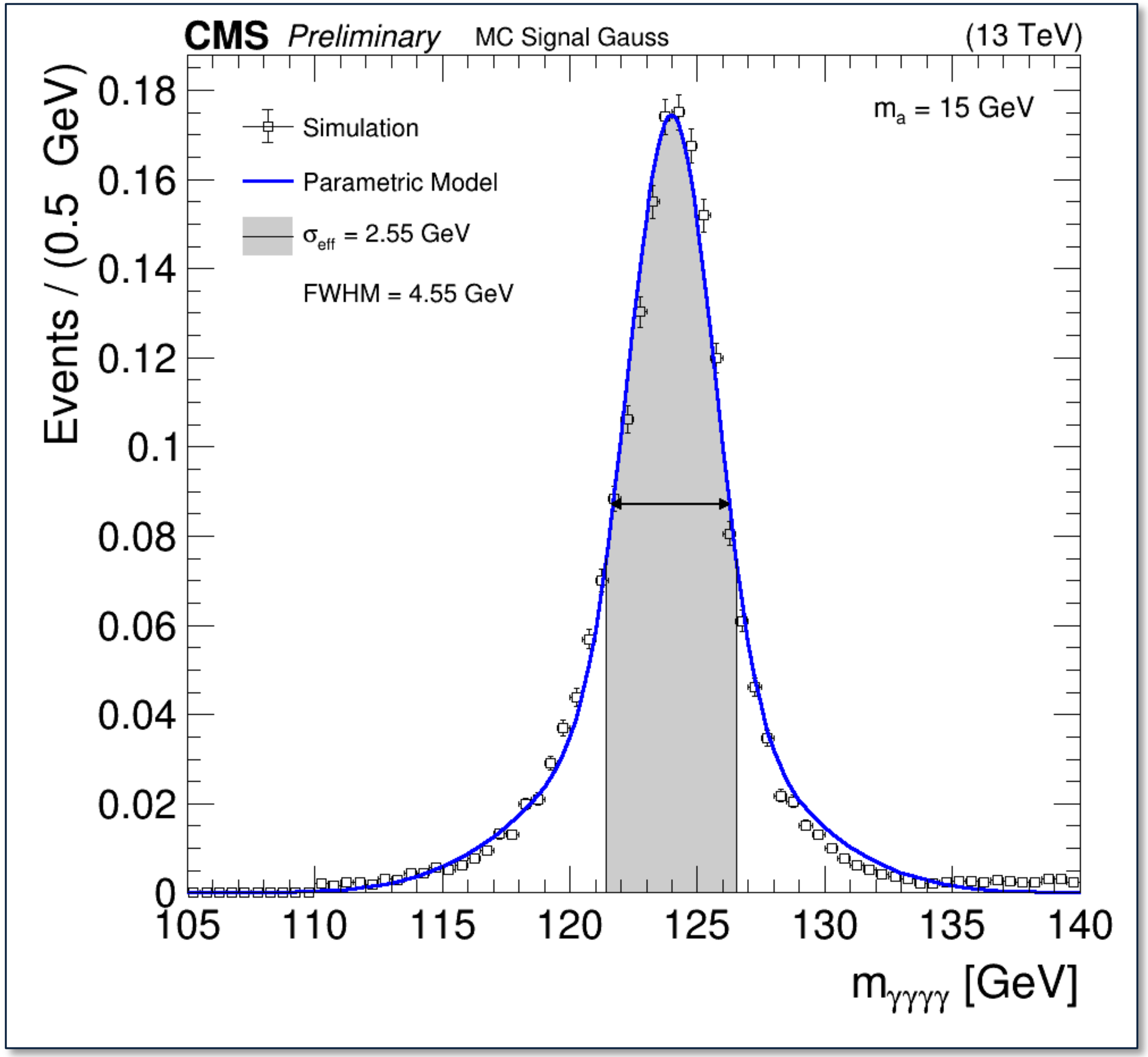
Signal Region: $115 < m_{\gamma\gamma\gamma\gamma} < 135$ GeV

Sideband Region: $(110 < m_{\gamma\gamma\gamma\gamma} < 115) \cup (135 < m_{\gamma\gamma\gamma\gamma} < 180)$ **GeV**

- Categories are constructed to maximize significance in quadrature in the signal region.
- Cuts on category boundaries are applied on top of all analysis selections for all samples. A minimum of 8 data events in sideband region are required.

Signal Modelling

- Signal models of $m_{\gamma\gamma\gamma\gamma}$ are constructed for all nominal pseudoscalar mass points.
- The parametric model is constructed from the sum of 2 Gaussians.
- Intermediate masses are modelled by the nearest nominal mass model with normalization interpolated from detector efficiency × analysis acceptance.
- σ_{eff} refers to the region in which 68% of signal events are contained and FWHM to the Full Width Half Maximum of the parametric model.



Advancements on Previous Studies

Facts about previous analysis:

- Run 2 analysis looked at 2016–2018 data
- 132 fb^{-1} of raw statistics
- Set limits on $\sigma(pp \rightarrow h) \times Br(h \rightarrow aa \rightarrow \gamma\gamma\gamma\gamma)$ of 0.80 fb for $m_a = 15$ GeV and 0.26 fb for $m_a = 62$ GeV

How to increase analysis sensitivity:

- Statistics
- Better estimation of background
- Improved event selection procedure

Ways to improve on Run 2 analysis:

- Expect Run 3 statistics to match Run 2 statistics by end-2024. Current Run 3 statistics are at ~60 fb^{-1}.
- Even more statistics if doing combined Run 2 + partial Run 3
- Improvement of $\sqrt{2}$ in precision from Run 2 to Run 3
- New background MC samples
- Deep learning model instead of a BDT for event selection

Summary

A search for the SM Higgs boson decaying to two pseudoscalars with a fully resolved, four photon final state was proposed.

Many of the tools needed to perform the full analysis are constructed:

- Event selections processor
- Event selection BDT
- BDT score category optimization tools
- Signal modelling tools for nominal mass points

Future Work:

- Fix any problems with current suite of analysis tools
- Background modelling
- $m_{\gamma\gamma\gamma\gamma}$ fitting for limit setting in Run 2 proof-of-concept study
- Applying work to Run 3 data

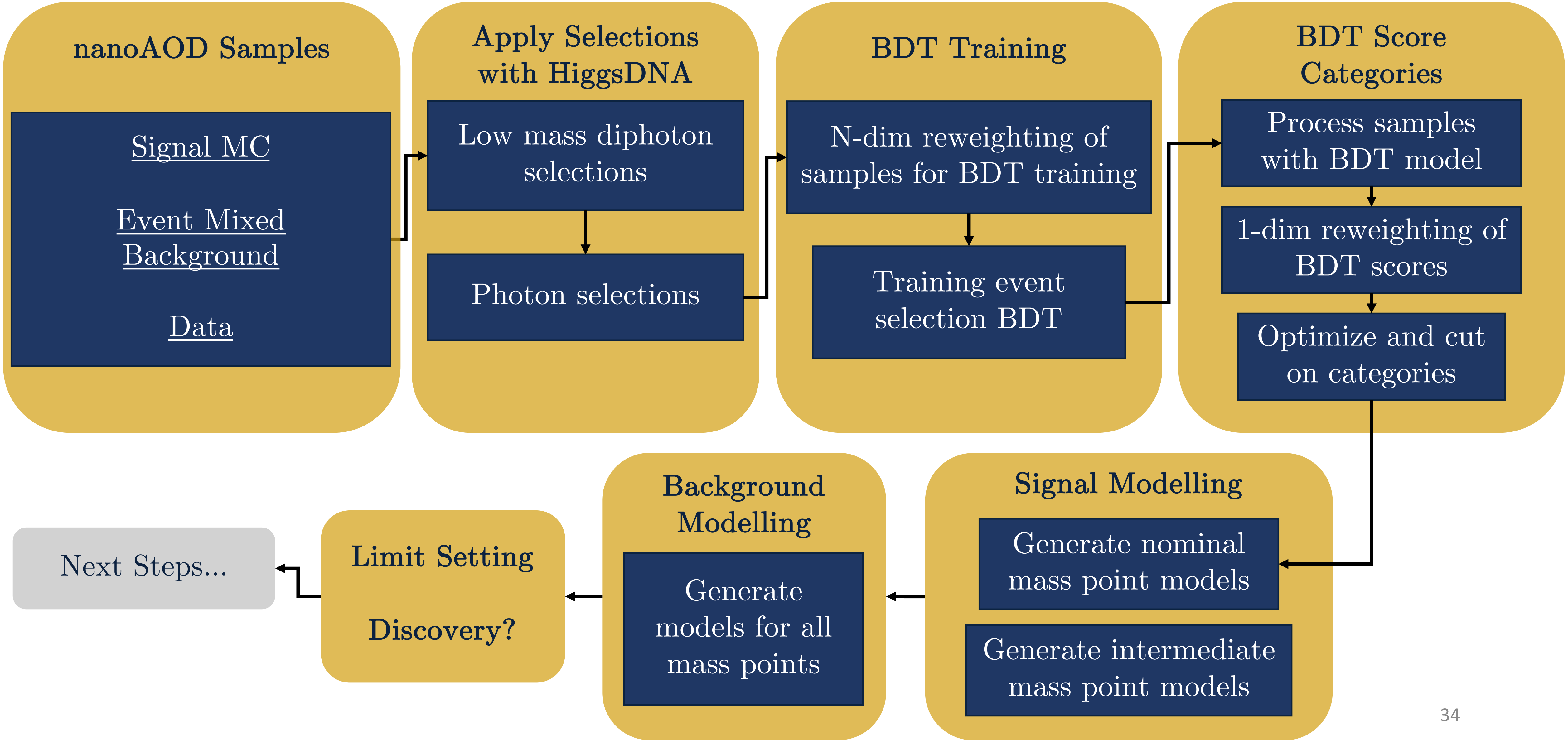
Goals:

- Complete reproduction of Run 2 analysis for 2018 subset with new HiggsDNA framework
- Increase analysis sensitivity for Run 3 data and produce a discovery

Thank You!

Backup

Analysis Overview

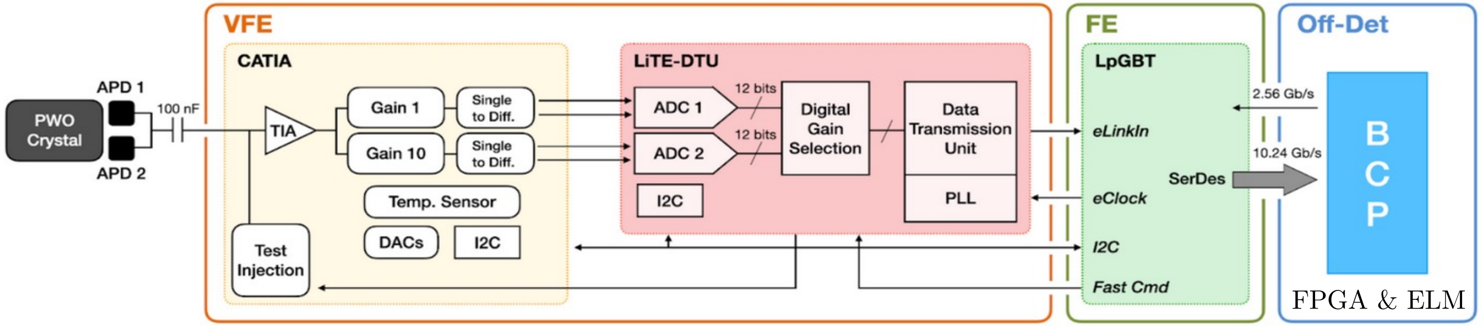


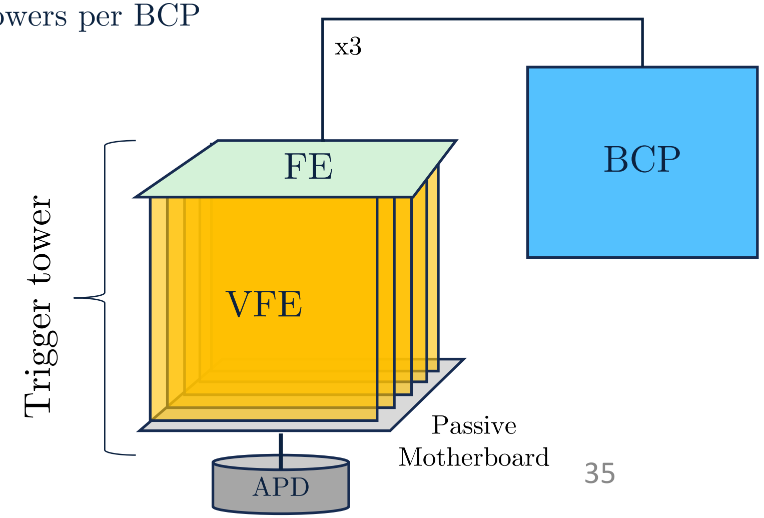
Technical Work

- Working on Phase II upgrade for ECAL Barrel
  - Phase II upgrade will entail a refurbishment of the readout electronics of the ECAL barrel to accommodate the higher rate and latency requirements of the HL-LHC
  - Running systems tests for new suite of on/off-detector readout hardware

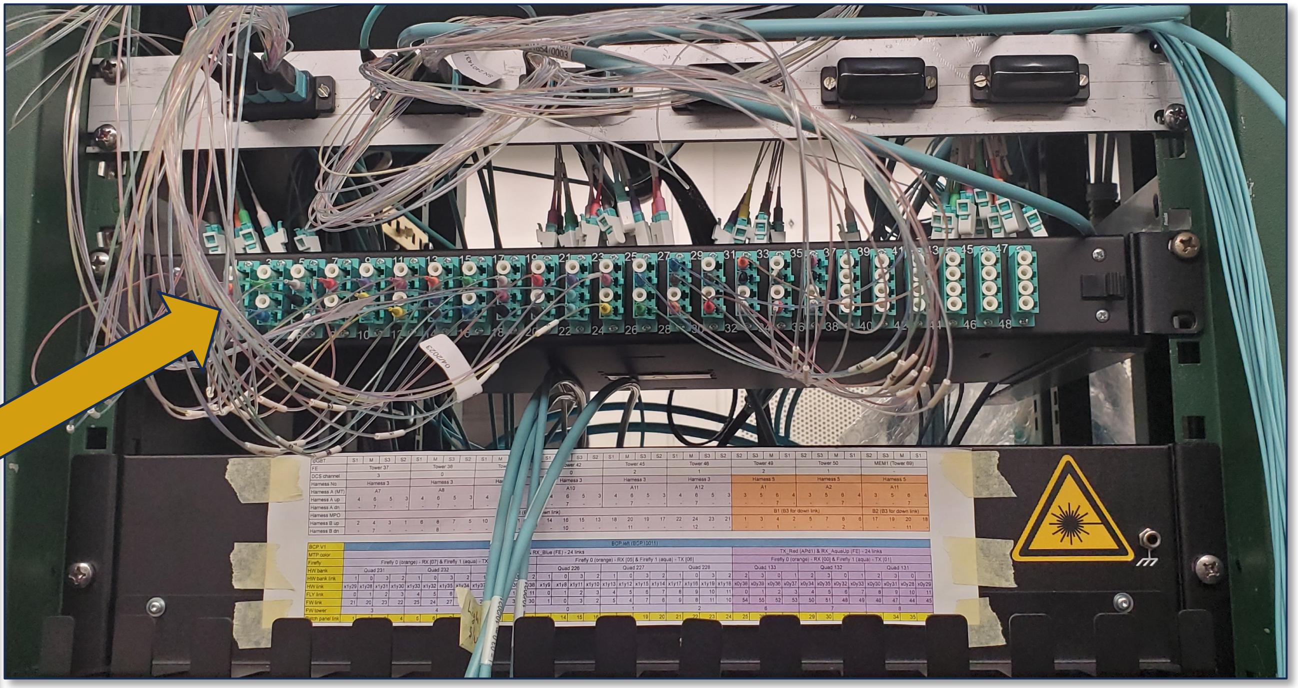
Software

- Verify firmware mapping of registers on BCP
- Each ECAL corresponds to a readout channel. Wrote tests to verify long-term channel stability (errors, alignment of messages)
- Develop high-level software for writing/reading BCP registers and executing commands on hardware (configure, align BCP, optimize pedestal ADCs from APD voltage)
  - Originally run on lpGBT (on FE) vs now running on ELM (on BCP) for increased speed
  - Updated high-level software to accommodate firmware change allowing for multiple trigger towers per BCP
- Write tests to check patch panel configuration
- Convert test software procedures for test beams





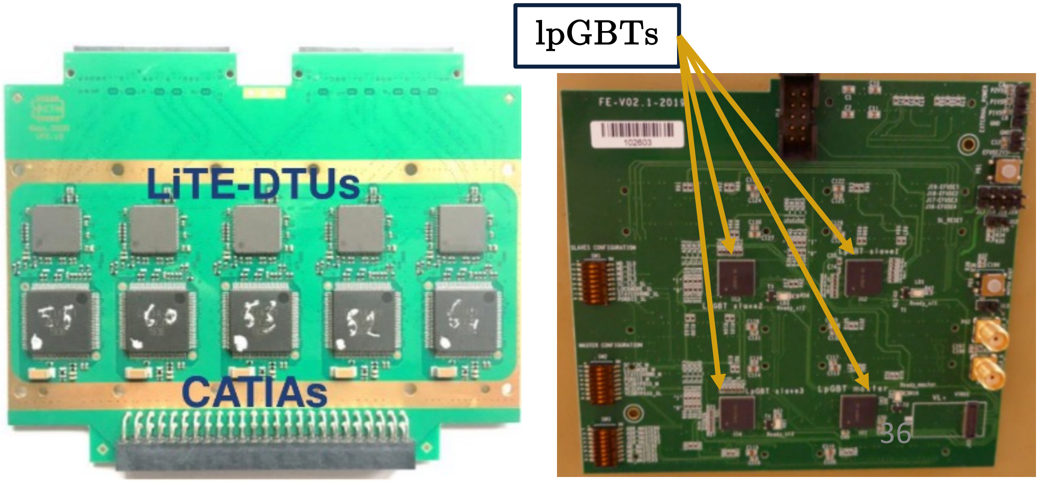
Technical Work (cont.)



Above: Patch panel reconfiguration for a new BCP firmware version.

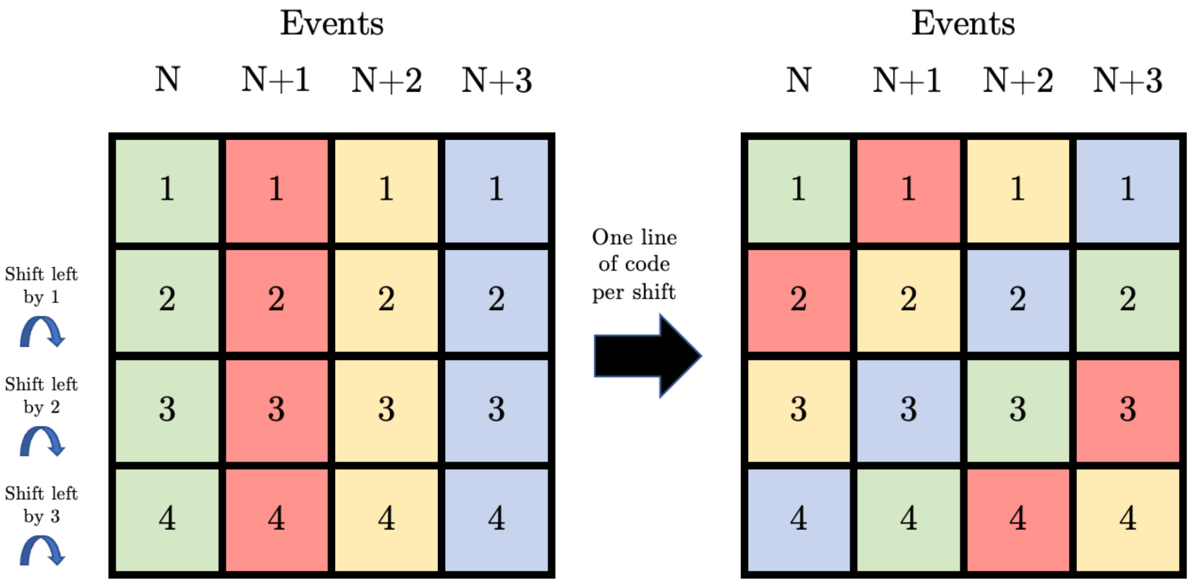
Hardware

- Set up patch panels for additional hardware (BCPs)
- Prepare hardware for test beams



Event Mixing

- Expected shuffling:



- Events are color coded and then shuffled.
- <u>How to read</u>: Event N has photon N → event N, photon 2 → event N+1, etc.

- Observed shuffling:
  - Photon 1 (Evt 1) → Photon 1 (Evt 1)
  - Photon 2 (Evt 1) → Photon 2 (Evt 2)
  - Photon 3 (Evt 1) → Photon 3 (Evt 3)
  - Photon 4 (Evt 1) → Photon 4 (Evt 4)

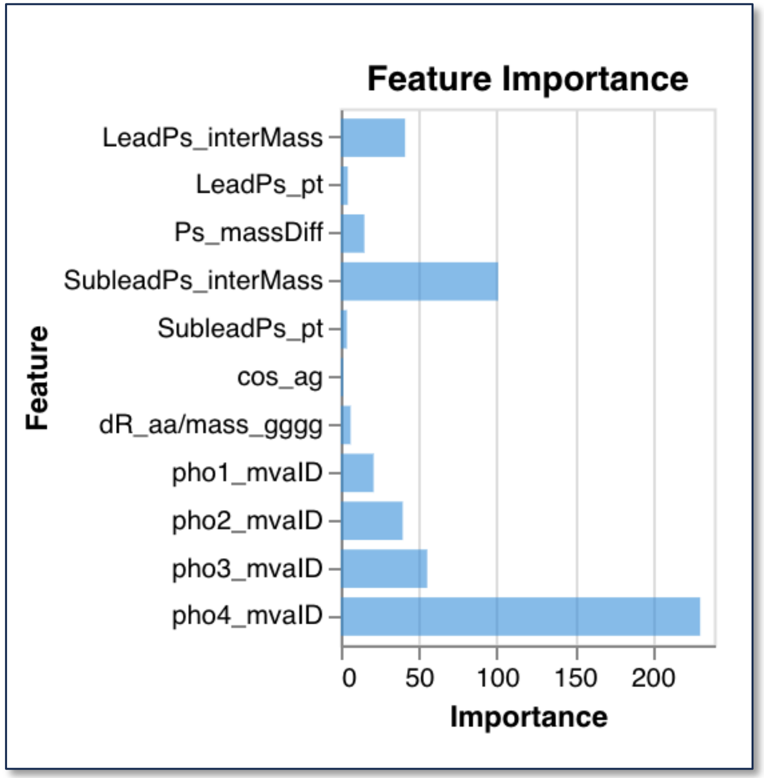
And so on for the other events...

Event Selection BDT Inputs

- Processed signal MC samples are combined into one signal-like sample. Processed event mixed background is used as the background-like sample.
- An N-dim reweighting is performed to correct discrepancies between event mixed background and data before training.

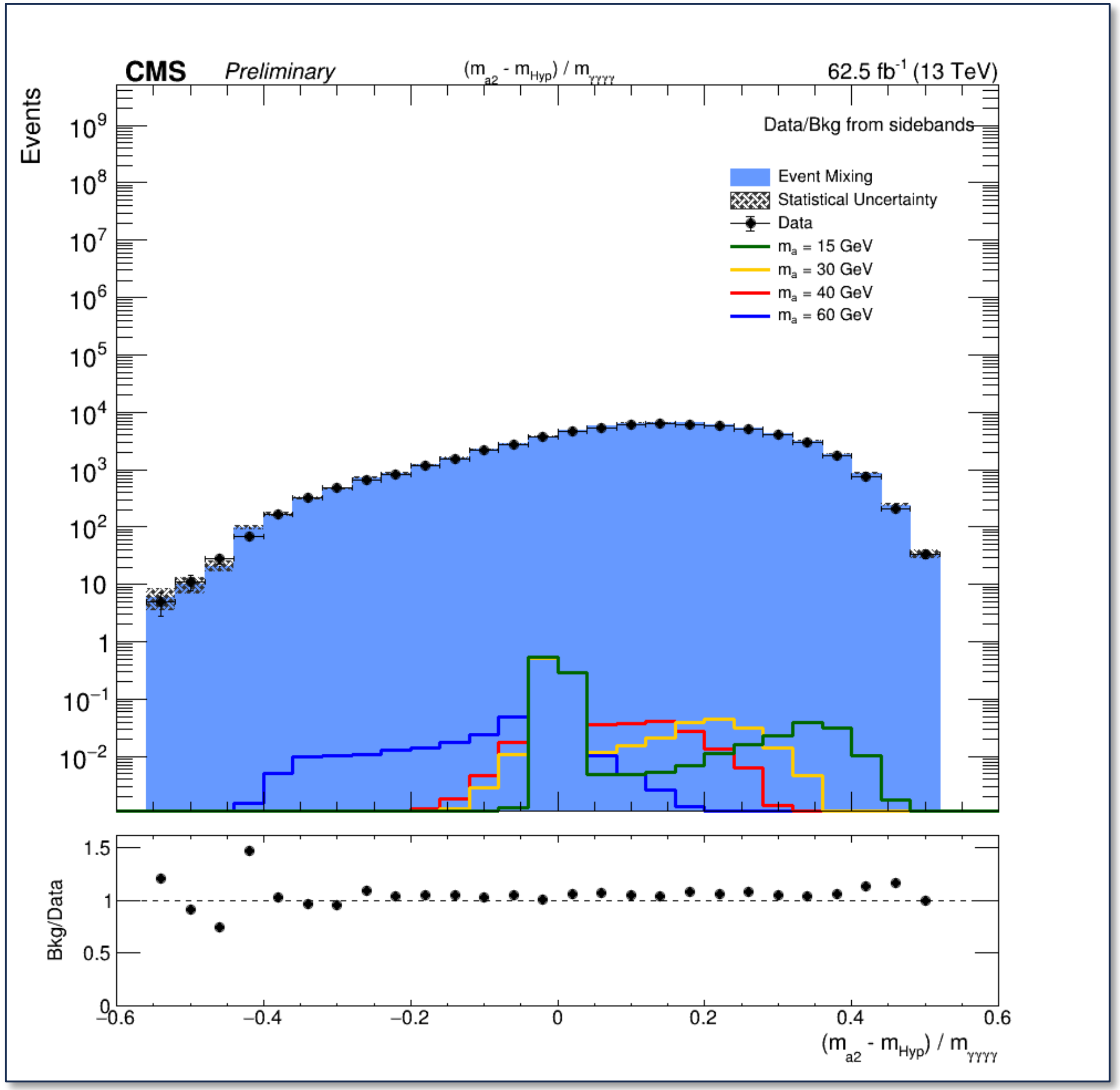
Input variables:

- $\gamma_{1,2,3,4}$ MVA ID
- a_1 p_T a_2 p_T
- $\Delta R(a_1,a_2)/m_{\gamma\gamma\gamma\gamma}$
- $m_{a1} - m_{a2}$
- $\cos(\theta_{a\gamma})$, where $\theta_{a\gamma}$ is the angle between the leading photon coming from the leading pseudoscalar and the direction of $a \to \gamma\gamma$
- $\frac{m_{a1\,RECO} - m_{a\,Hyp}}{m_{\gamma\gamma\gamma\gamma}}$ $\frac{m_{a2\,RECO} - m_{a\,Hyp}}{m_{\gamma\gamma\gamma\gamma}}$

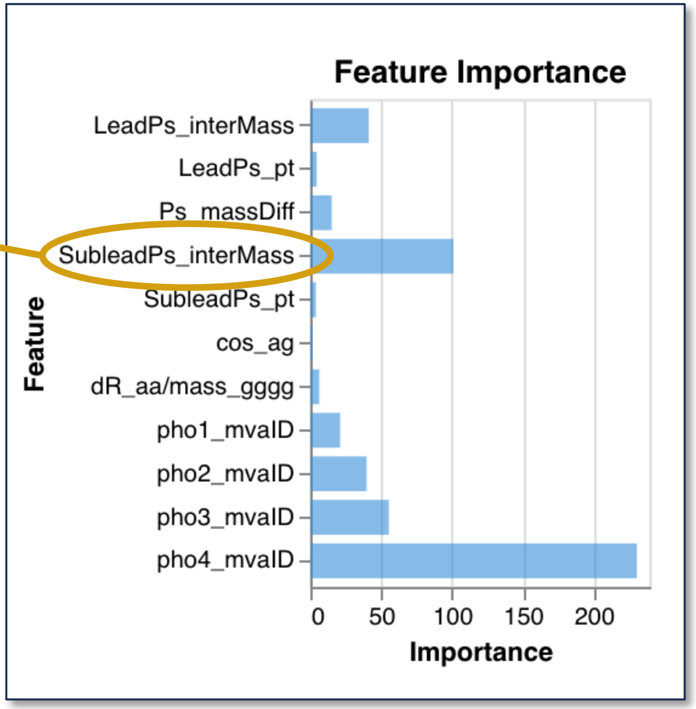


Above: Chart of feature importance for the event selection BDT input variables from one of the best performing models.

Important BDT Input Variables



Left: Distribution of $(m_{a2\ RECO} - m_{a\ Hyp})/m_{\gamma\gamma\gamma\gamma}$ for subleading pseudoscalar candidate. Shown for signal MC at 15, 30, 40, 60 GeV pseudoscalar mass points, event mixed background, and data. All selections have been applied.

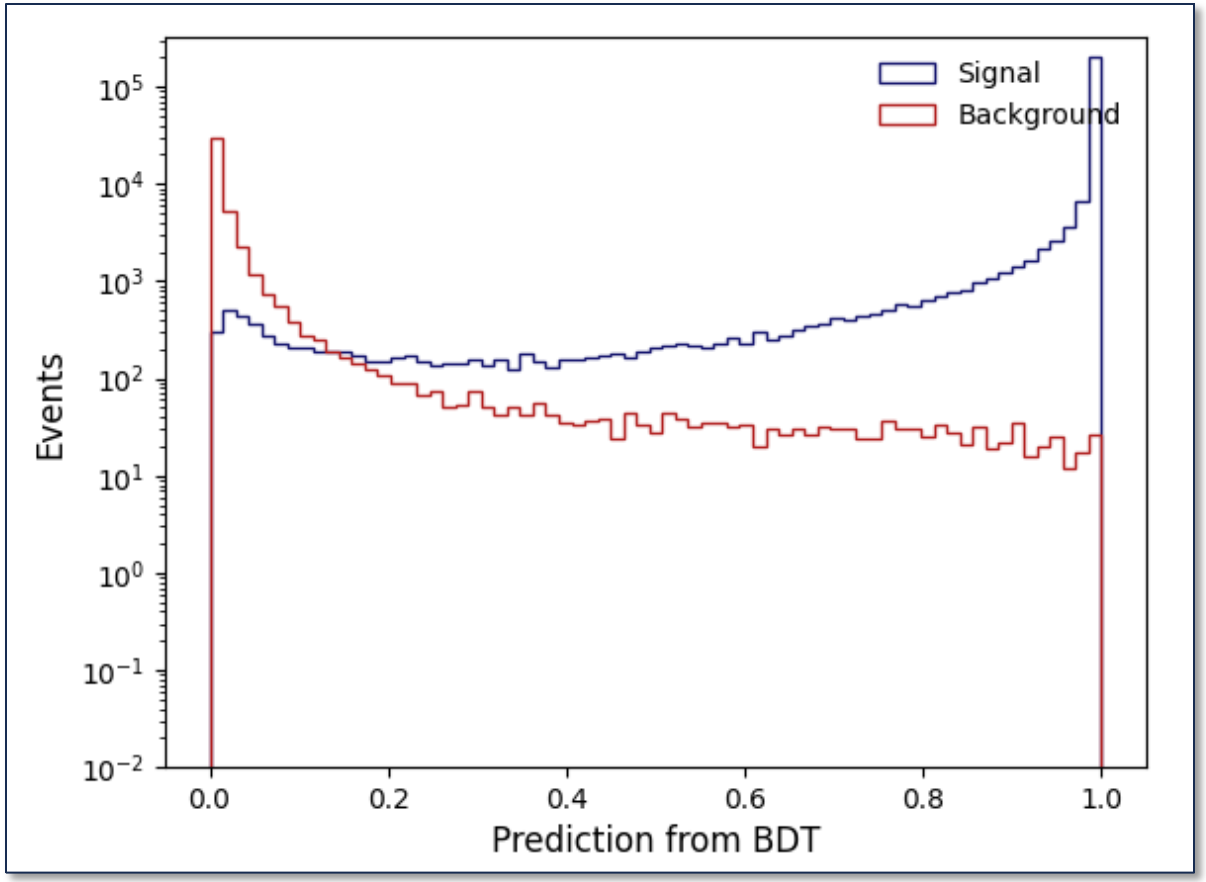


Above: Chart of feature importance for the event selection BDT input variables from one of the best performing models.

Training the Event Selection BDT

- A 1-dim reweighting of BDT score is required for each pseudoscalar mass point after training the event selection model.
- Training/testing is split amongst odd/even events and is performed over the full $m_{\gamma\gamma\gamma\gamma}$ mass range.
- Samples are weighted by the appropriate luminosity and the per-event N-dim reweights before training.

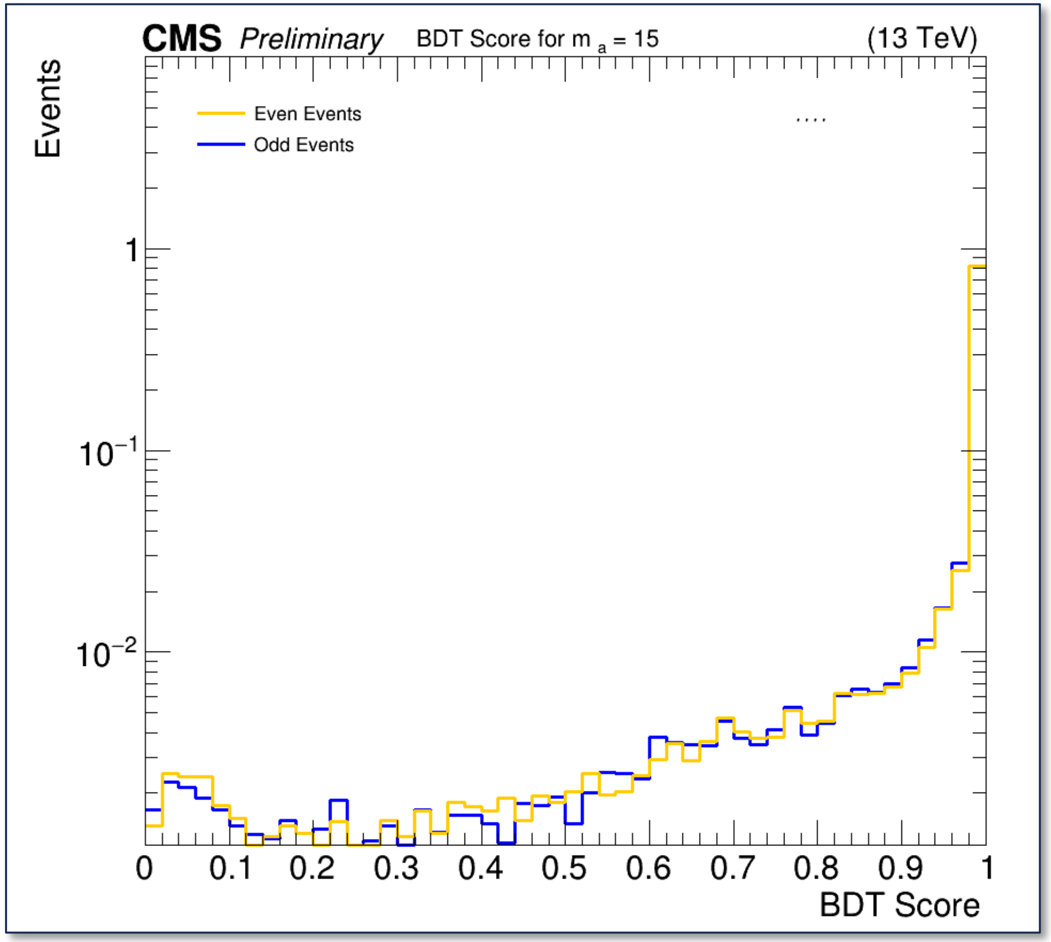
- Metrics for evaluating BDT performance:
  - AUC
  - Log loss minimization
  - Error rate
  - Average truth deviation
  - F1 score
  - ROC curve

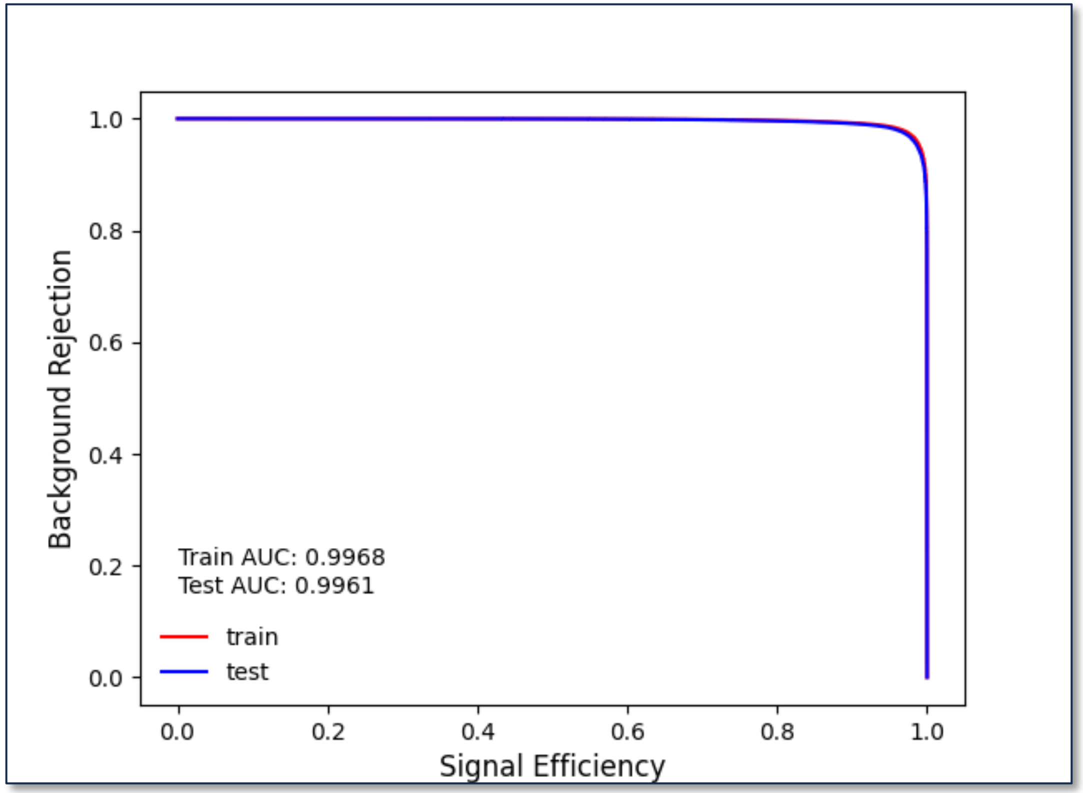


Above: Event selection BDT predictions for training set using processed signal MC and event mixed background.

Checking the Event Selection BDT

- Checks for overtraining are done to ensure an optimal event selection model is produced.
- ROC checks can be done for the entire model at once and even/odd checks must be done by nominal mass point.





Above: ROC curve for one of the best event selection BDT models. A match between train/test curves and their associated area under curve (AUC) indicates a lack of overtraining.

Left: Overtraining check by comparing prediction distribution of trained event selection BDT model for $m_a = 15$ GeV. Odd/even (train/test) samples should have a similar distribution which would indicate a lack of overtraining.

BDT Score Categorization Details

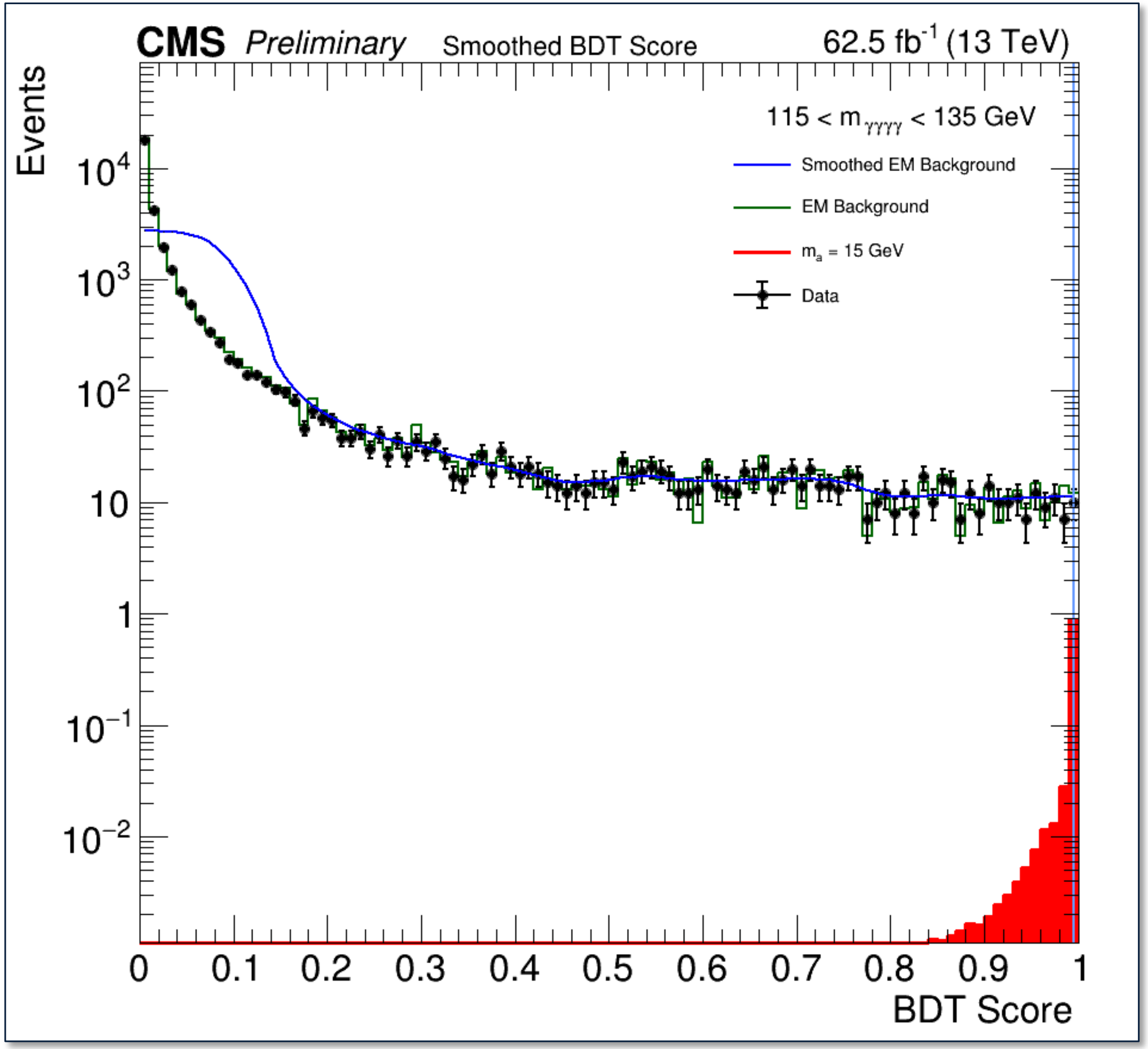
- Reminder: Approximate Mean Significance (AMS)

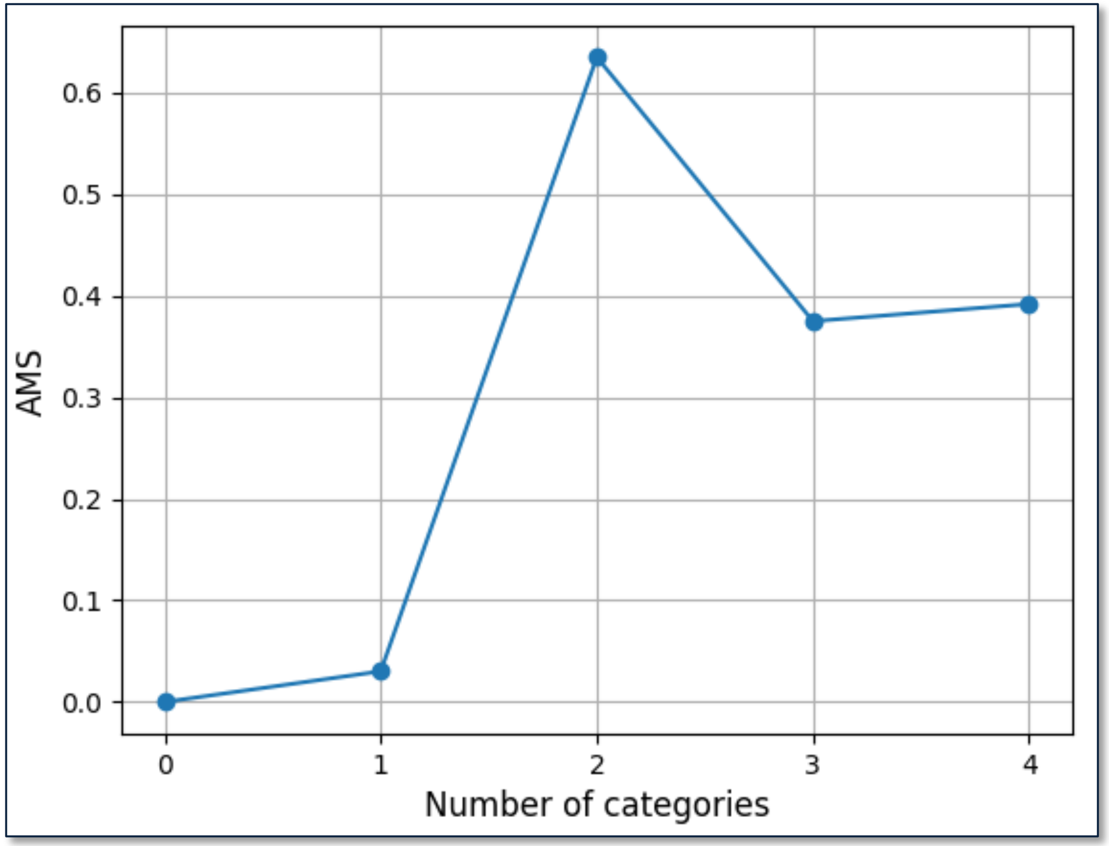
$$\mathrm{AMS} = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right)-S\right]}$$

where S and B reference to the number of signal and background events, respectively.

- The BDT score distributions for signal MC, event mixed background, and data are smoothed for each nominal mass point to account for statistical fluctuations when using very fine binning.
- Category boundaries are optimized to maximize total AMS in the signal region while requiring at least 8 data events in the sideband region.
- In the case of multiple categories, AMS is added in quadrature.
- Cuts at category boundaries will be applied to all samples.

BDT Score Category Optimization





Top: Total AMS for up to 4 categories. Ideally, total AMS would always increase and then plateau.

Left: Event selection BDT predictions for signal MC, event mixed background, and data for $m_a = 15$ GeV in the signal region. Smoothed background is shown. To properly smooth the background distribution, a cut at ~0.1 is needed. A category boundary line is shown in light blue near 1.