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, CMS, Technical Contributions

III. Introduction to Analysis

IV. Sample Generation

V. Event Selection

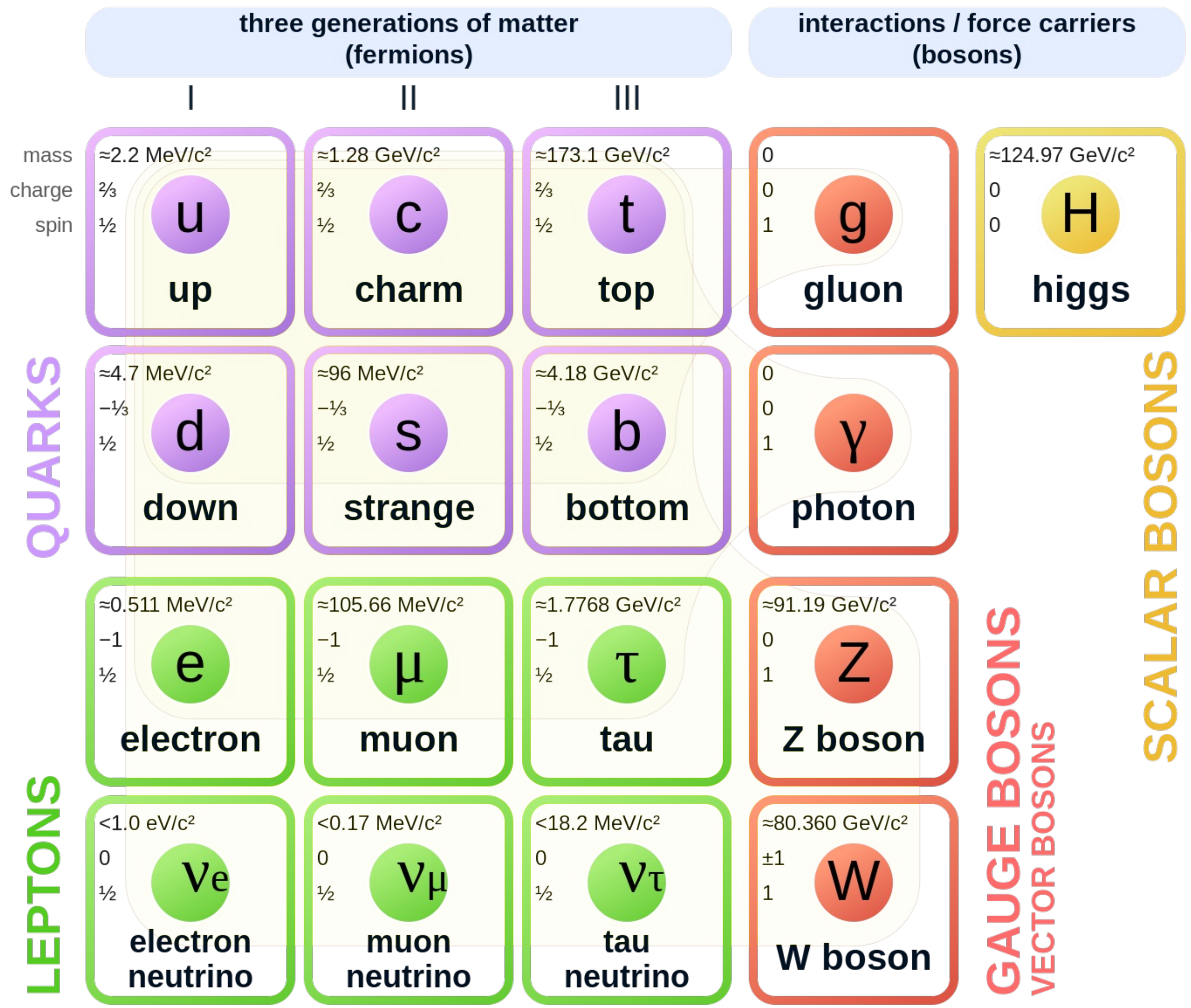
VI. Signal Modelling

VII. Advancements on Previous Studies

VIII. Summary

The Standard Model

- 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 vacuum expectation value & Yukawa couplings yield other particle masses excluding neutrinos

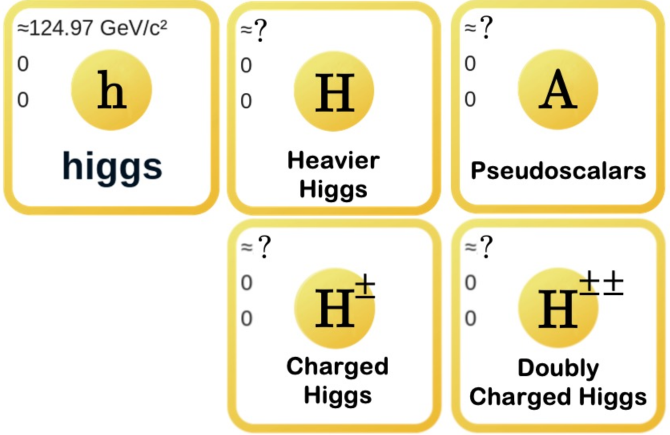


Beyond the Standard Model

- Beyond the Standard Model physics can remedy its shortcomings and predict interesting, new physics.

New phenomenologies are introduced that can be checked experimentally!

- Many models introduce an *extended Higgs sector*.
- Can remedy hierarchy problem, dark matter, and more.

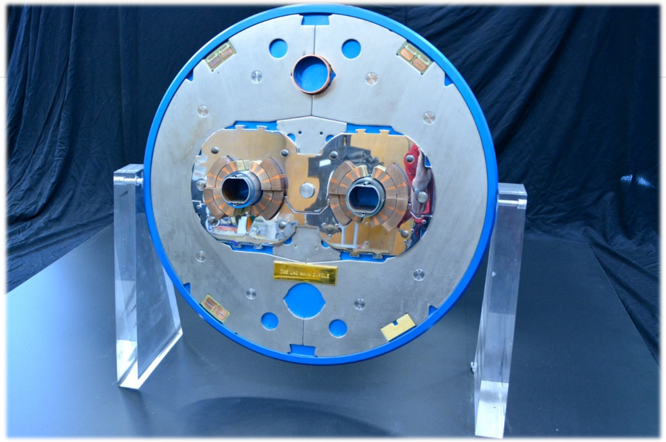


New matter content of common models:

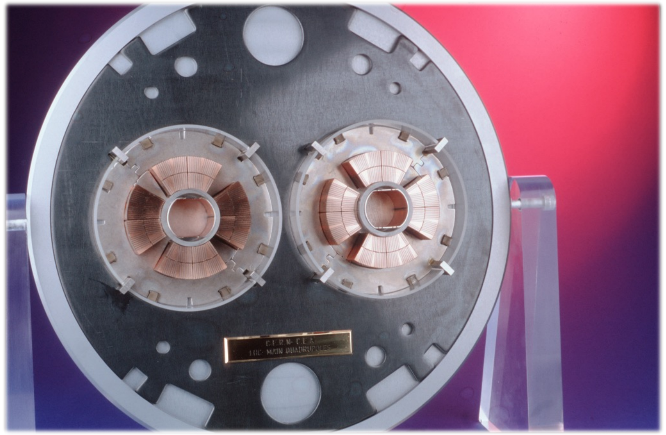
	SM	BSM
Singlet Models	h^0	S
2 Higgs Doublet Models	h^0	H^0, $H^\pm$, A^0
Higgs Triplet Models	h^0	H^0, $H^\pm$, $H^{\pm\pm}$, A^0

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

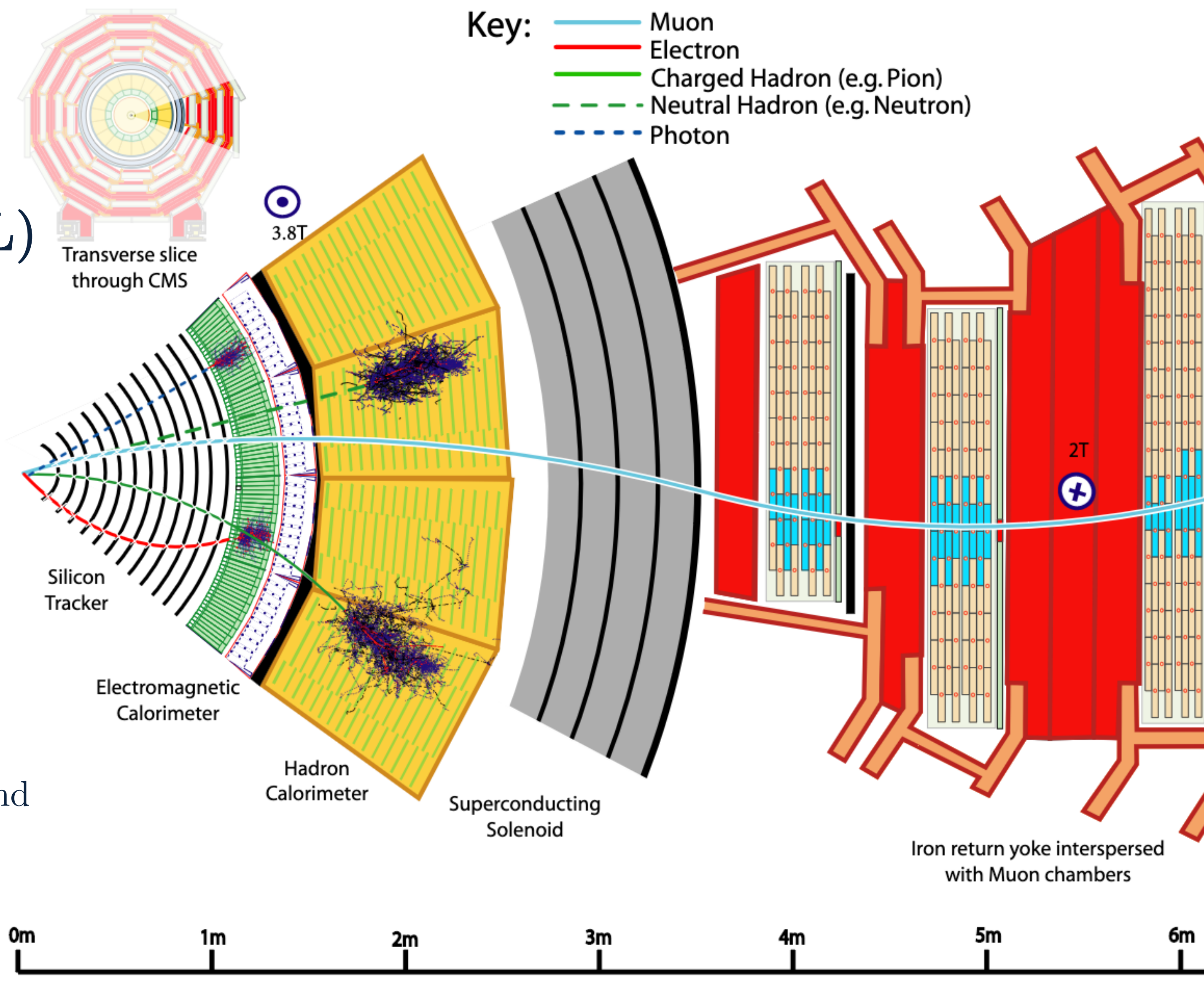


Above: Slice of LHC dipole magnets.
Below: Slice of LHC quadrupole magnets



Compact Muon Solenoid (CMS)

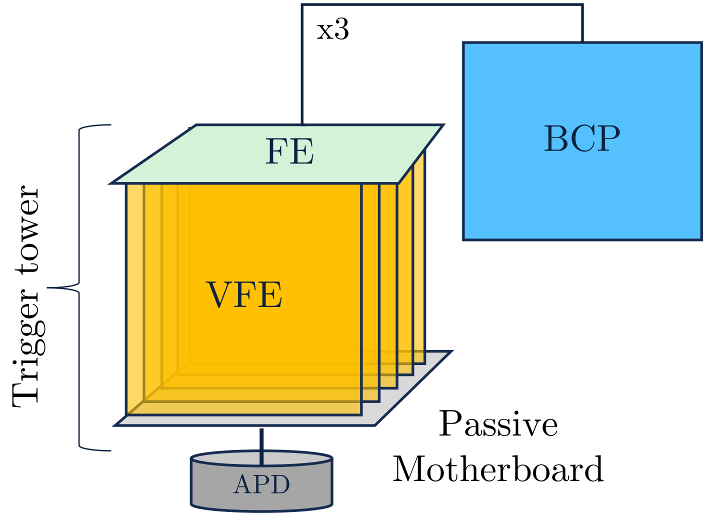
- Silicon tracker
  - Solid-state pixel and strip detectors
  - Detects charged particle tracks
- Electromagnetic calorimeter (ECAL)
  - Homogenous detector
  - Consists of 75,000 $PbWO_4$ crystals
  - Detects electromagnetically interacting particle showers
- Hadronic calorimeter (HCAL)
  - Sampling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



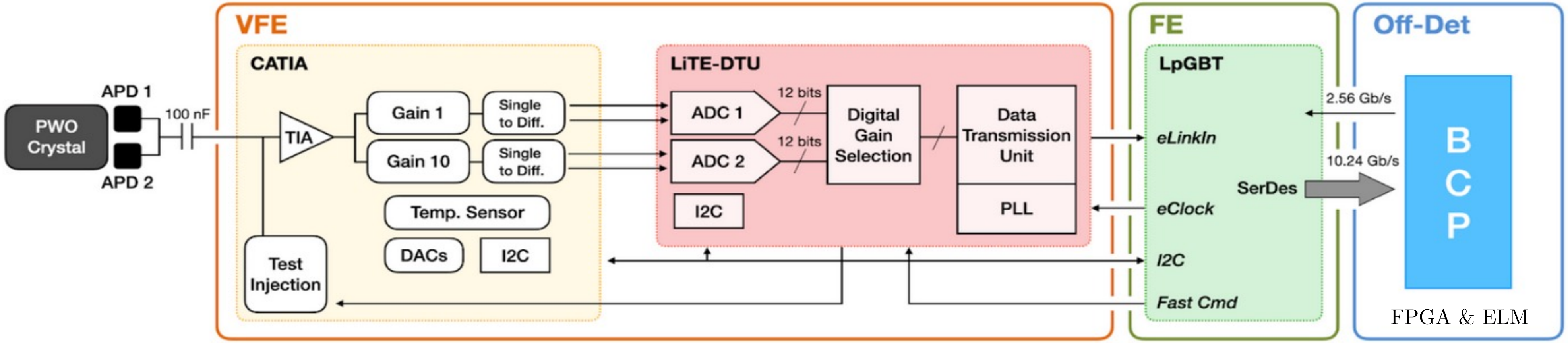
Technical Contributions (Part 1)

Working on Phase II upgrade for ECAL Barrel

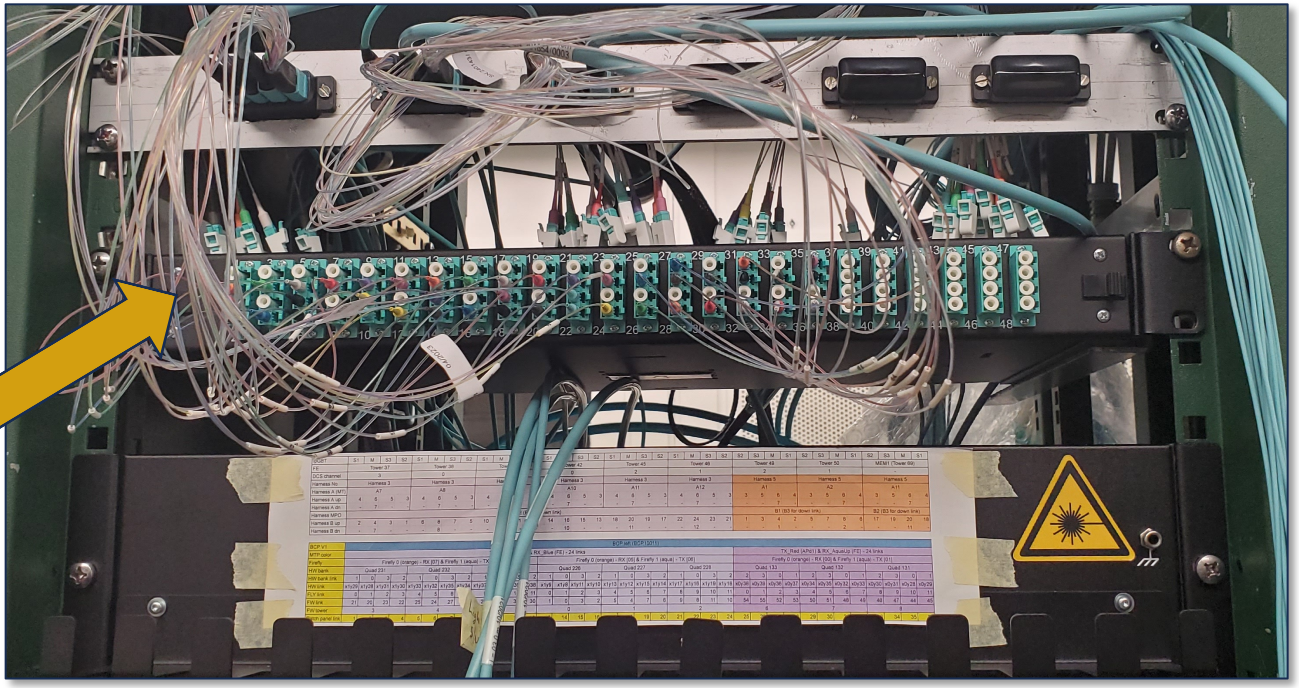
- The Phase II upgrade will entail a refurbishment of readout electronics of ECAL Barrel to accommodate the higher collision rate and latency requirements of the HL-LHC.
- Reduced crystal operating temperature 18°C ➞ 9°C.
- L1 trigger: 5x5 crystal granularity ➞ single crystal granularity



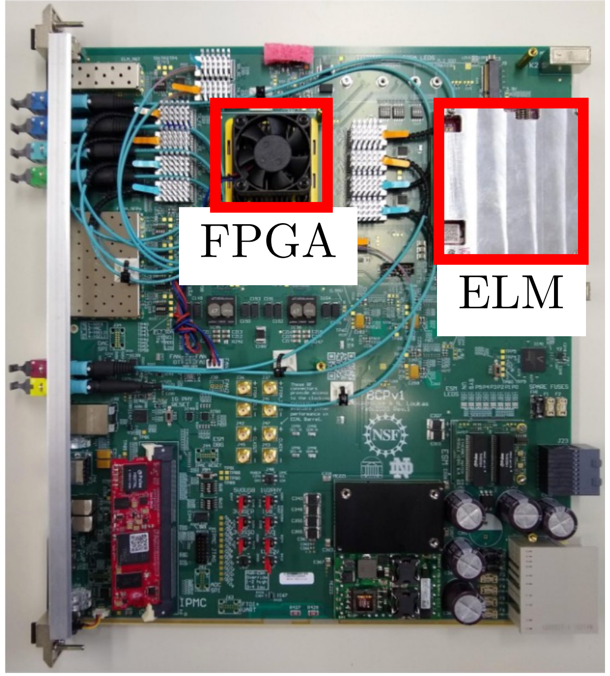
- I've been working on running systems tests for new suite of on/off-detector readout hardware.
- Developing high-level software for executing commands on hardware
- Involved in test beams and analysis

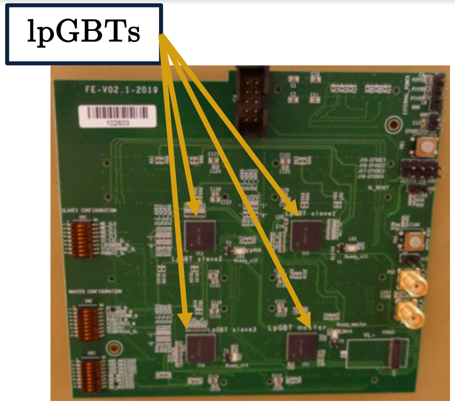


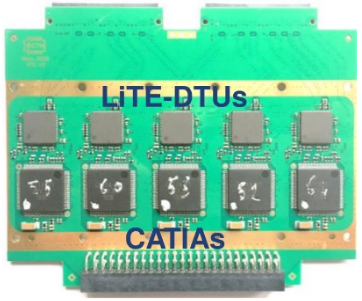
Below: Additional patch panel configuration for a second BCP before July 2023 test beam.



- Set up patch panels for additional hardware
- Write tests to check patch panel configuration
- Prepare hardware for test beams

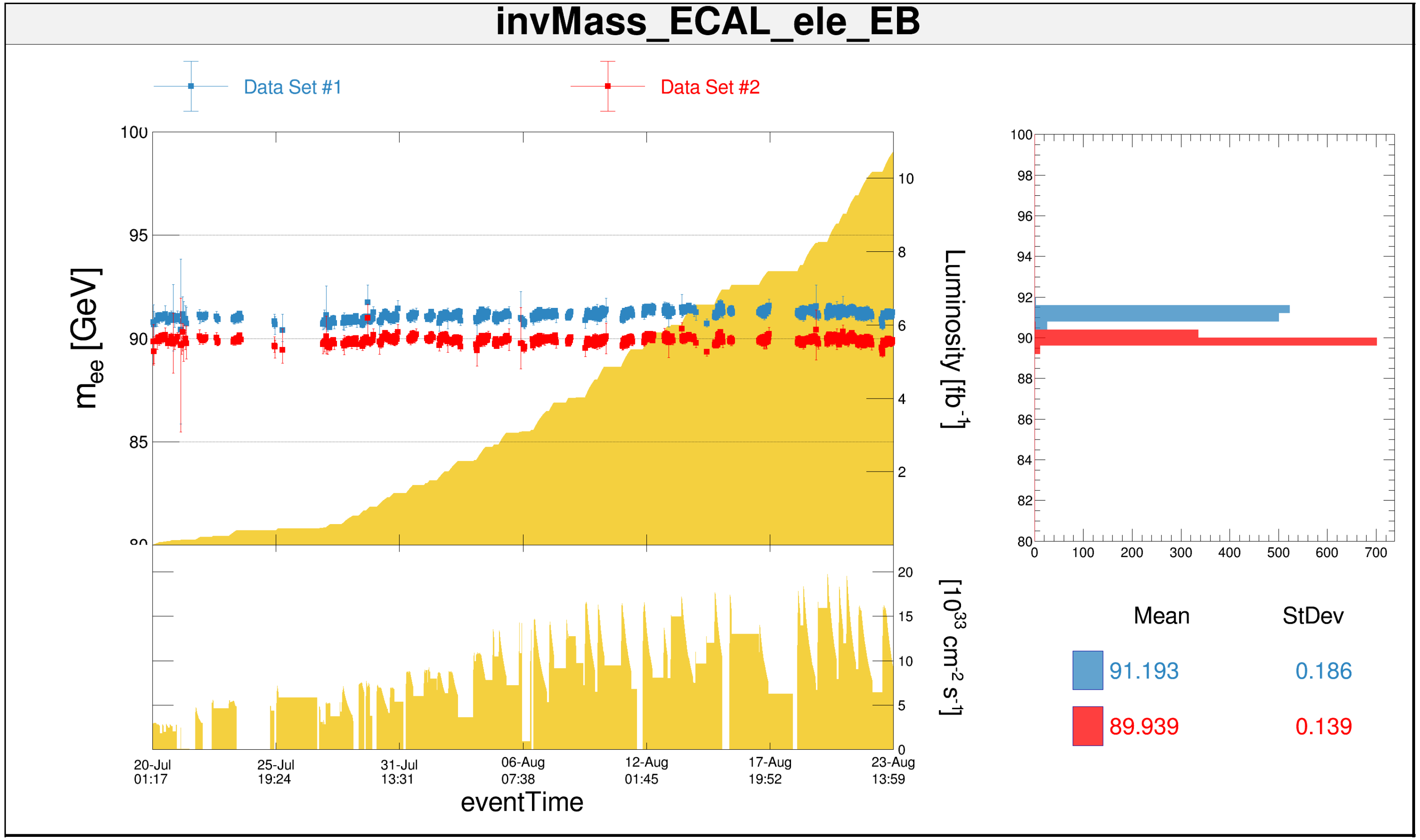






Technical Contributions (Part 2)

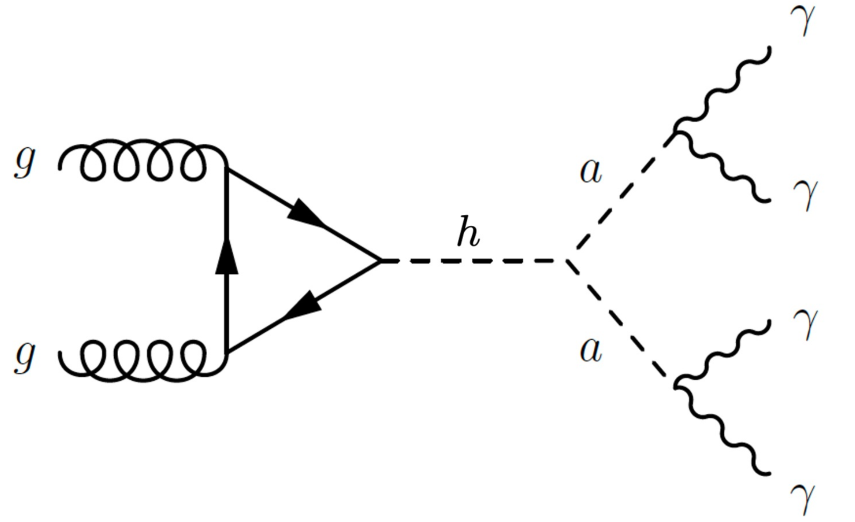
- There is a desire to streamline calibration and monitoring of ECAL.
- Developed an automatic monitoring tool alongside the new automatic ECAL calibration software. $Z \to ee$ events are one way to calibrate the detector.



Introduction to Analysis

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

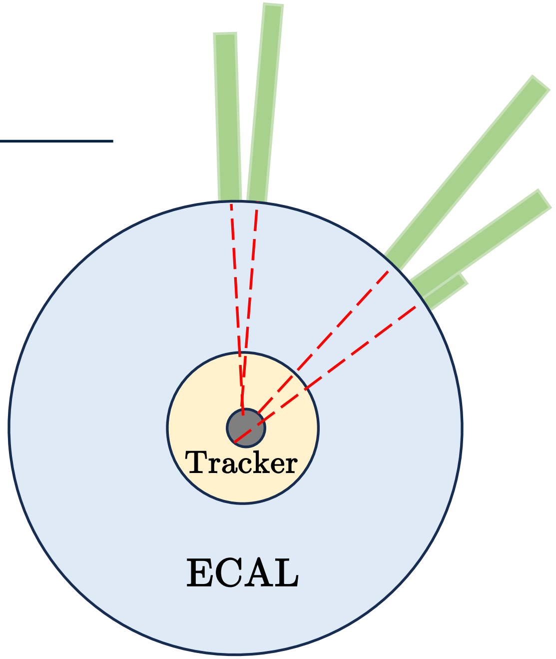
- Clean final state with low SM background
- Model agnostic search allows for greater impact of any limits or discovery
- Improved sensitivity could allow for a discovery



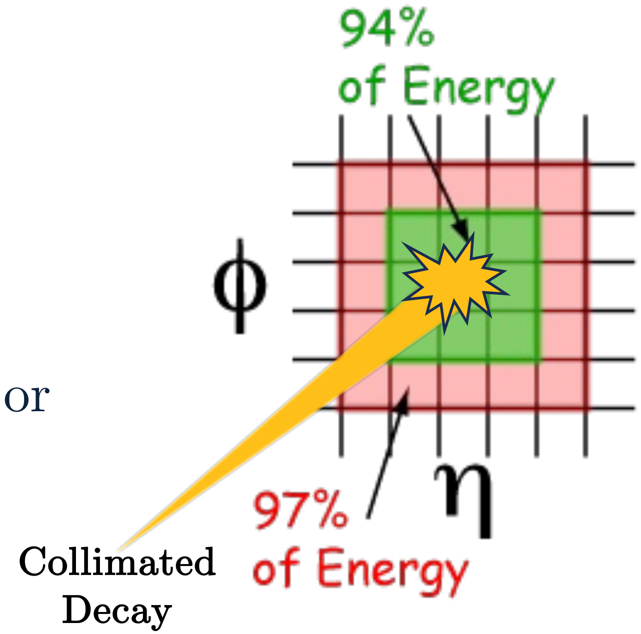
- Currently working on a proof-of-concept study with the $h \to \gamma\gamma$ group's new analysis framework that I'm helping develop.
- This proof-of-concept study consists of recreating a previous Run 2 analysis for 2018 data. The goal is to apply this work to Run 3 data and increase analysis sensitivity in hopes of producing a discovery.

Sample Generation

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

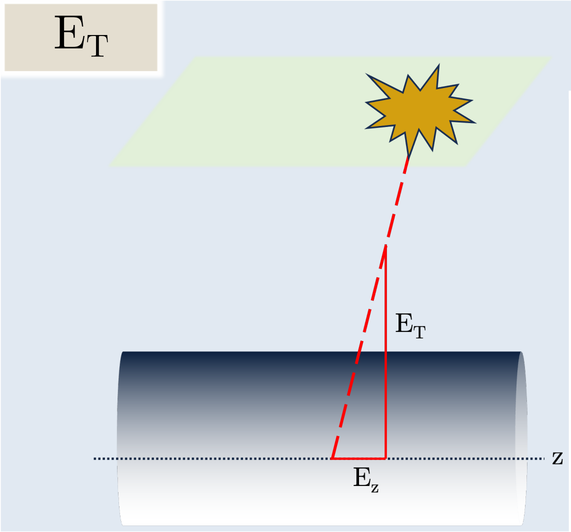


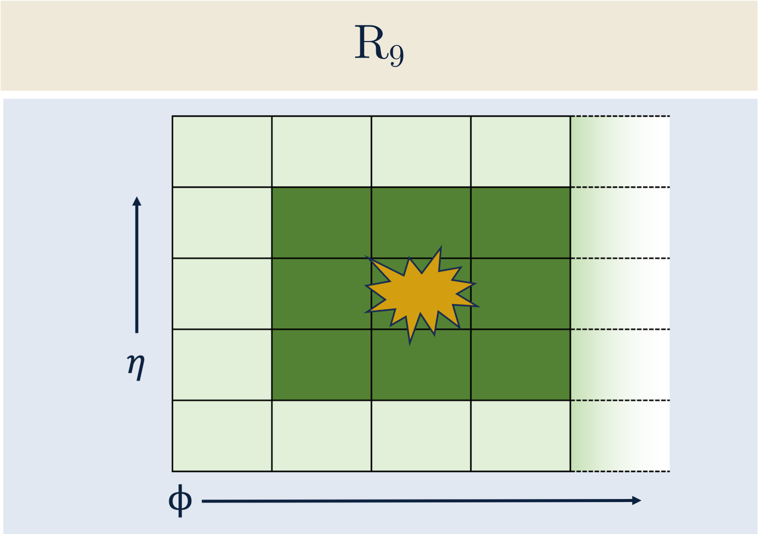
- **Background:**
  - A data-driven background is used
  - Use event mixing technique:
    - Shuffle photons in an event with photons from consecutive events
    - Removes presence of any signal while retaining background shape
  - Main backgrounds: either prompt photons from collision event or isolated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

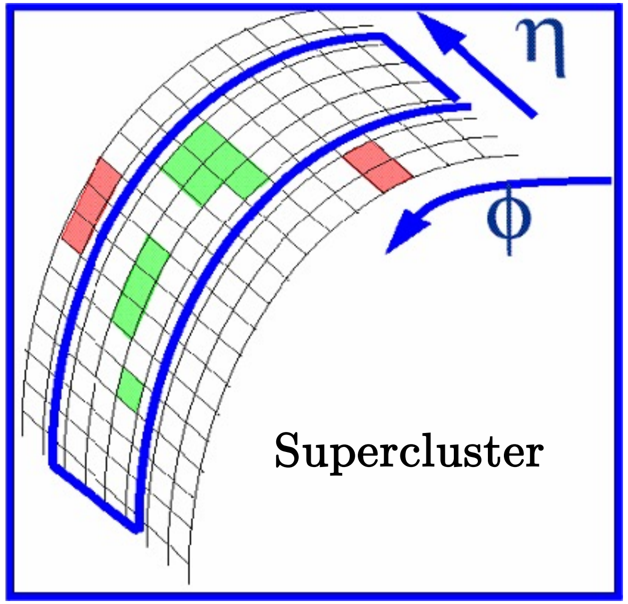


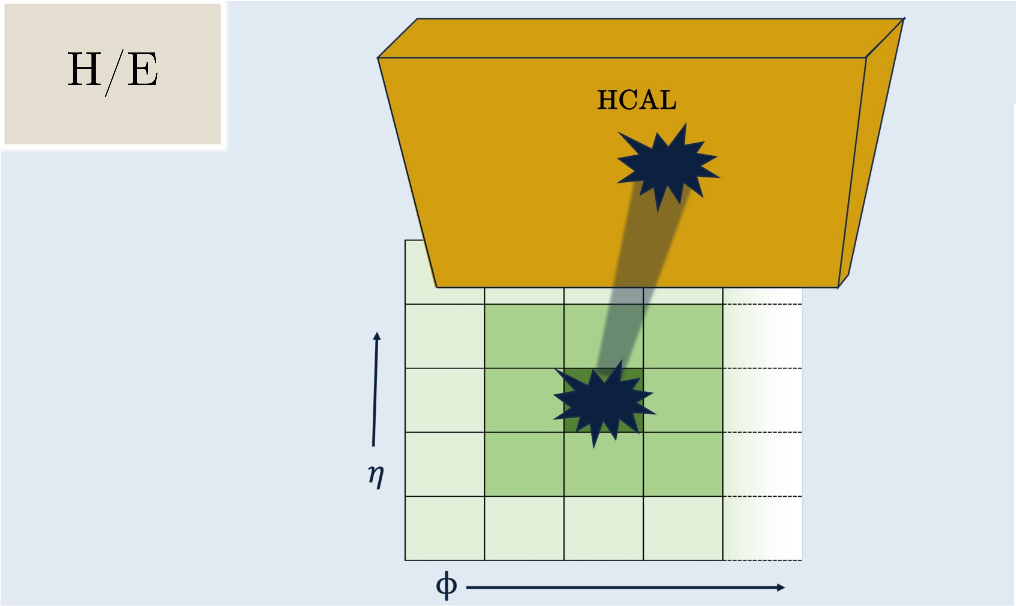
Choosing Candidate Photons

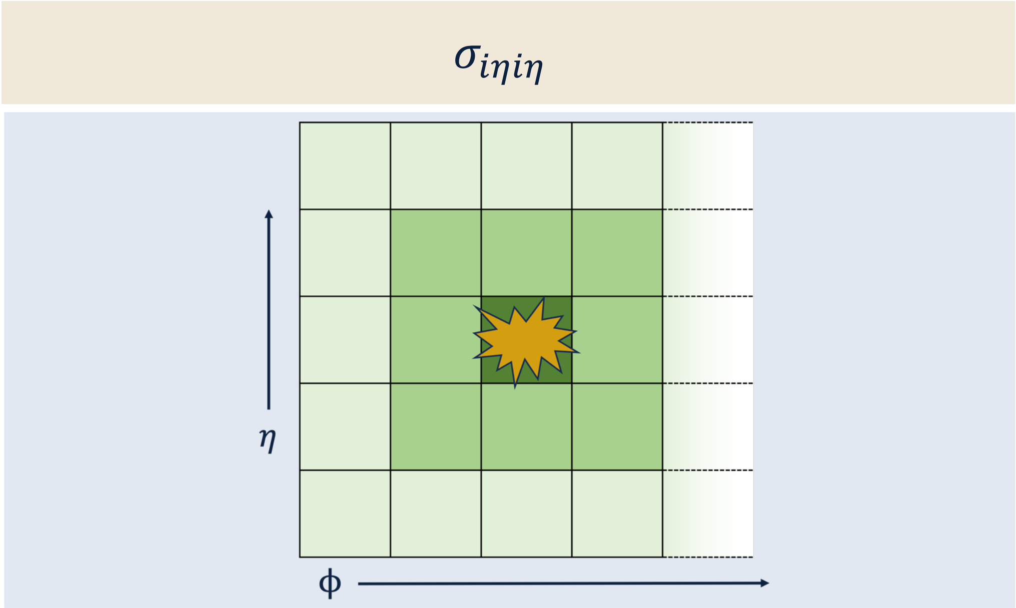
Energy deposition and shape of photon candidates:





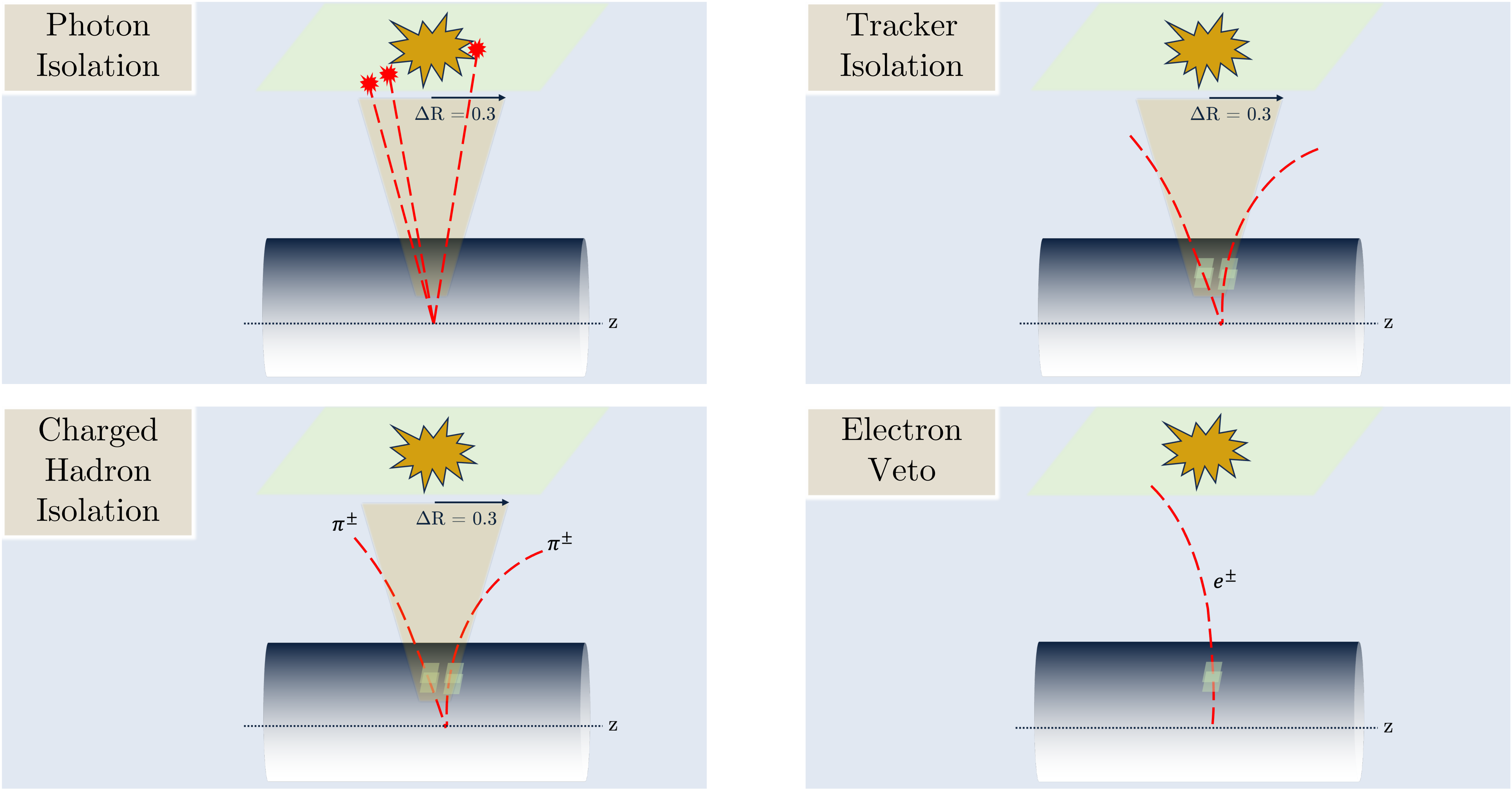






Choosing Candidate Photons

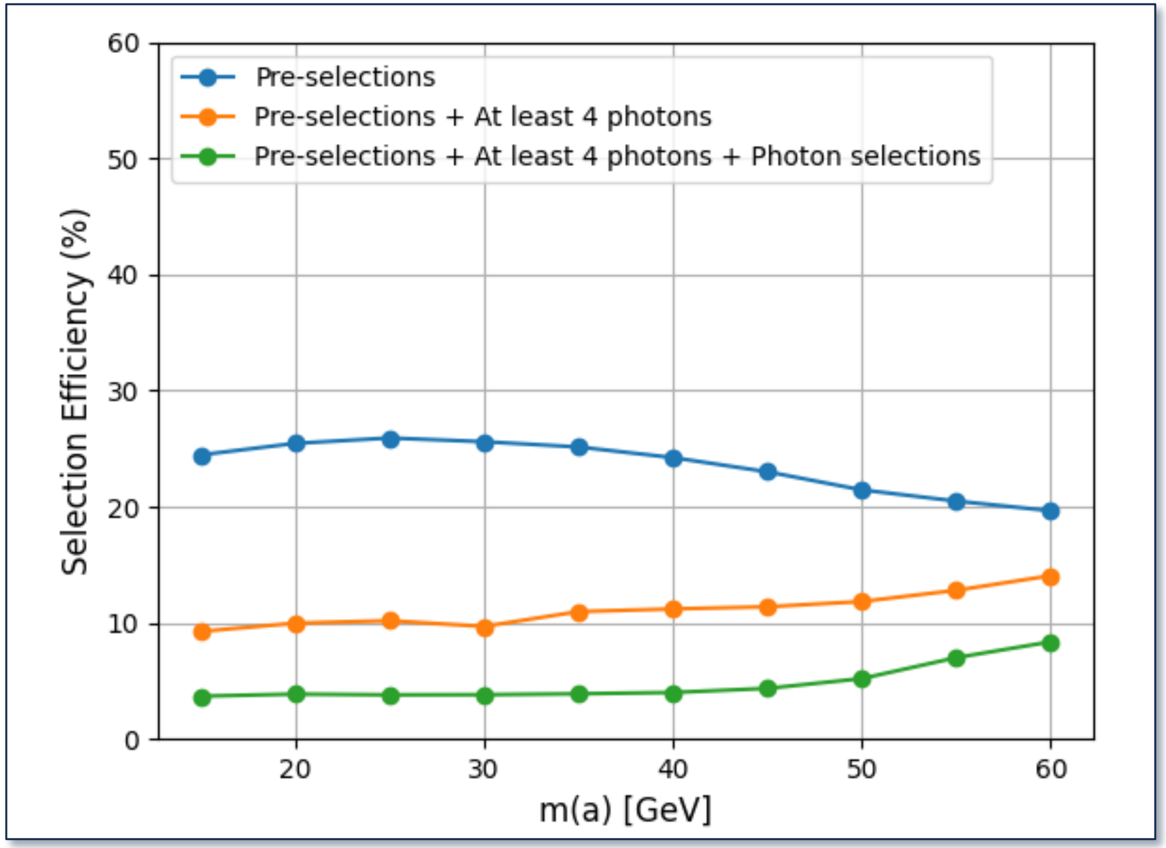
Isolation of photon candidates:



Selections

- Selections not shown are applied to photon candidates.
- Events must contain at least one diphoton and at least 4 photons.
- Additional cuts (below) are required to ensure optimal pseudoscalar reconstruction.
- $\Delta M = |m_{\gamma_a\gamma_b} - m_{\gamma_c\gamma_d}|$ is minimized to choose optimal pseudoscalar candidates

<table>
<tr><th></th><th>p_T</th><th>η</th><th>$m_{\gamma\gamma\gamma\gamma}$</th></tr>
<tr><td>γ_1</td><td>> 30.0</td><td rowspan="4">$|\eta| < 2.5$
$1.442 < |\eta| < 1.556$</td><td rowspan="4">$110 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV</td></tr>
<tr><td>γ_2</td><td>> 18.0</td></tr>
<tr><td>γ_3</td><td>> 15.0</td></tr>
<tr><td>γ_4</td><td>> 15.0</td></tr>
</table>



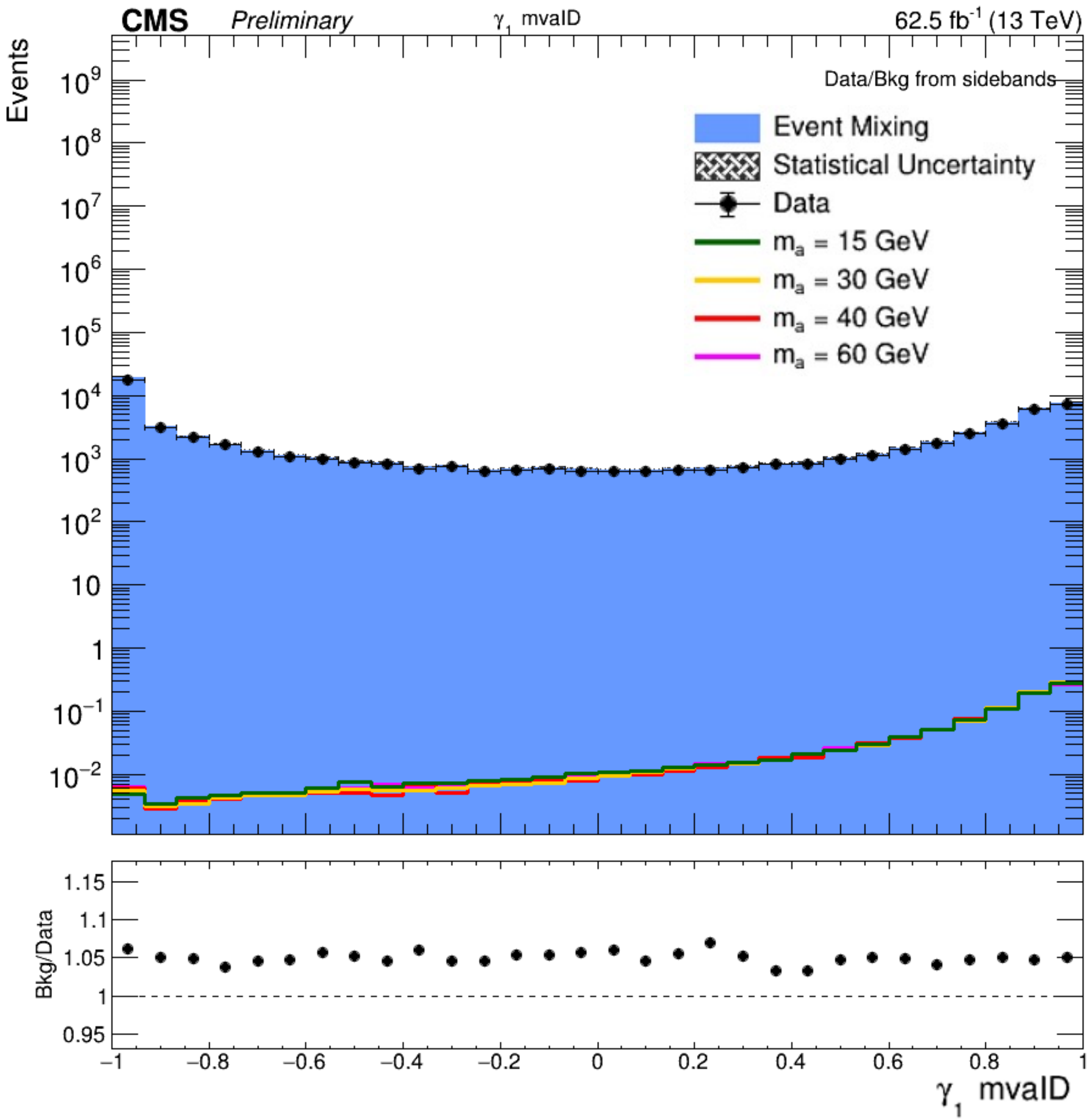
Above: Efficiencies of selections at various stages of selections for all nominal pseudoscalar mass points. This should improve at higher mass points.

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

Results of Selections

Photon Multivariate Analysis ID

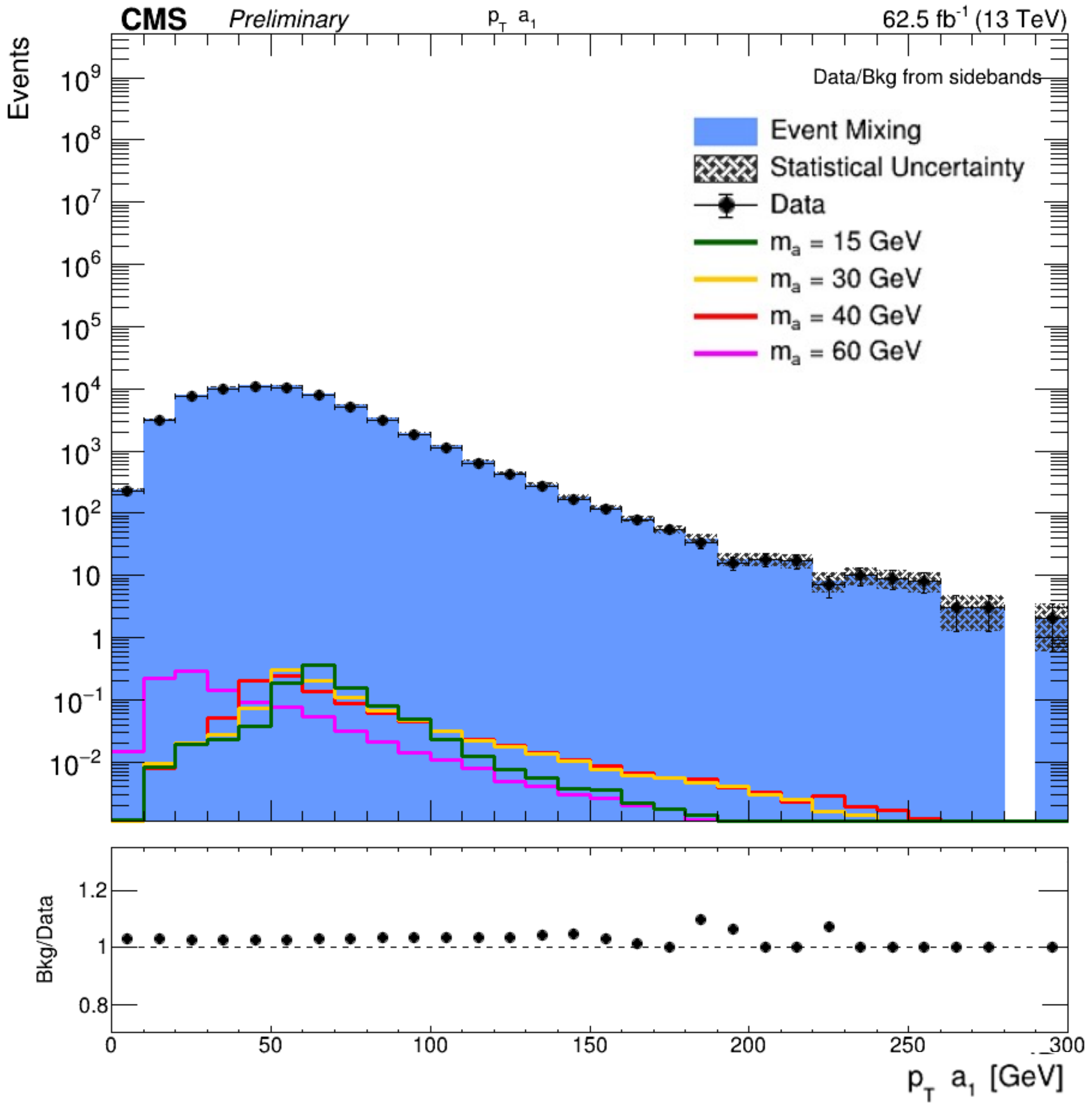
- Discriminates real photons from fake photons (collimated decays from jets)
- Likelihood score: High score (~1) corresponds to real photon-like and low score (~ -1) corresponds to fake photon-like



Results of Selections

Looking at pseudoscalar kinematics:

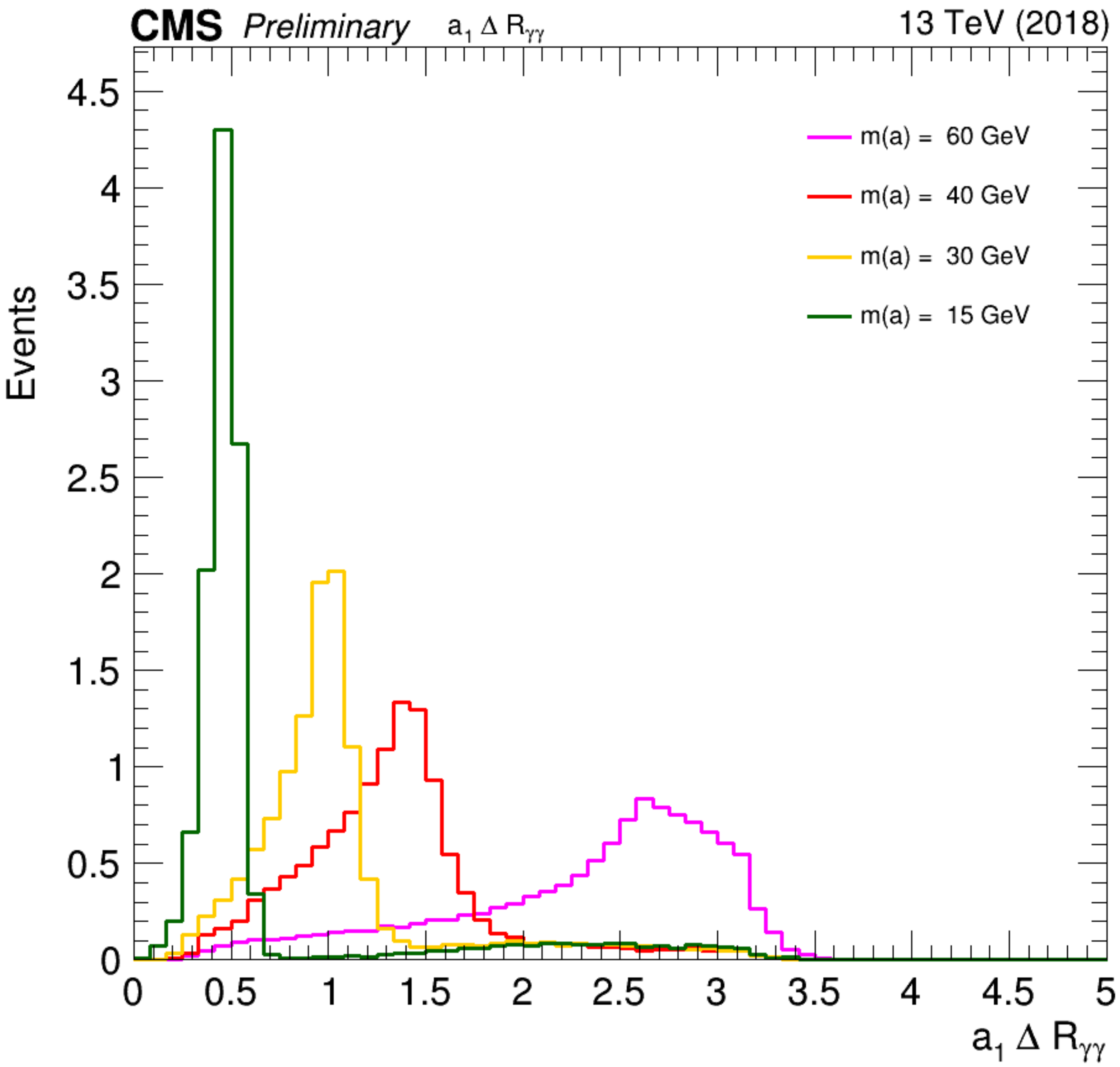
- p_T peak moves left with increasing mass
- ΔR peaks move right with increasing mass means less boosting as pseudoscalar mass increases



Results of Selections

Looking at pseudoscalar kinematics:

- 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

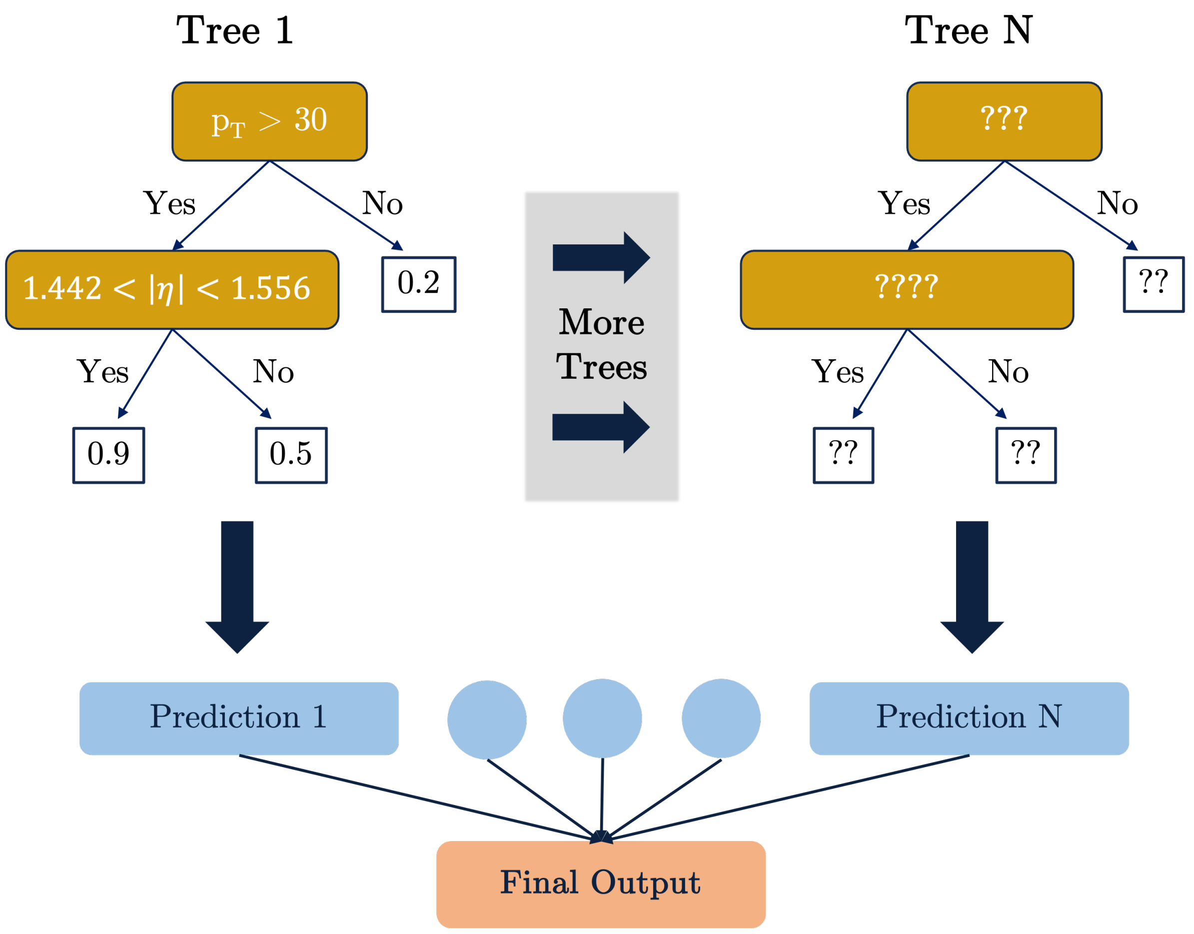
What is a BDT?

- 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.

- A BDT is used to enhance analysis sensitivity by predicting signal-like and background-like events.

- Only one BDT model is required for the entire event selection.

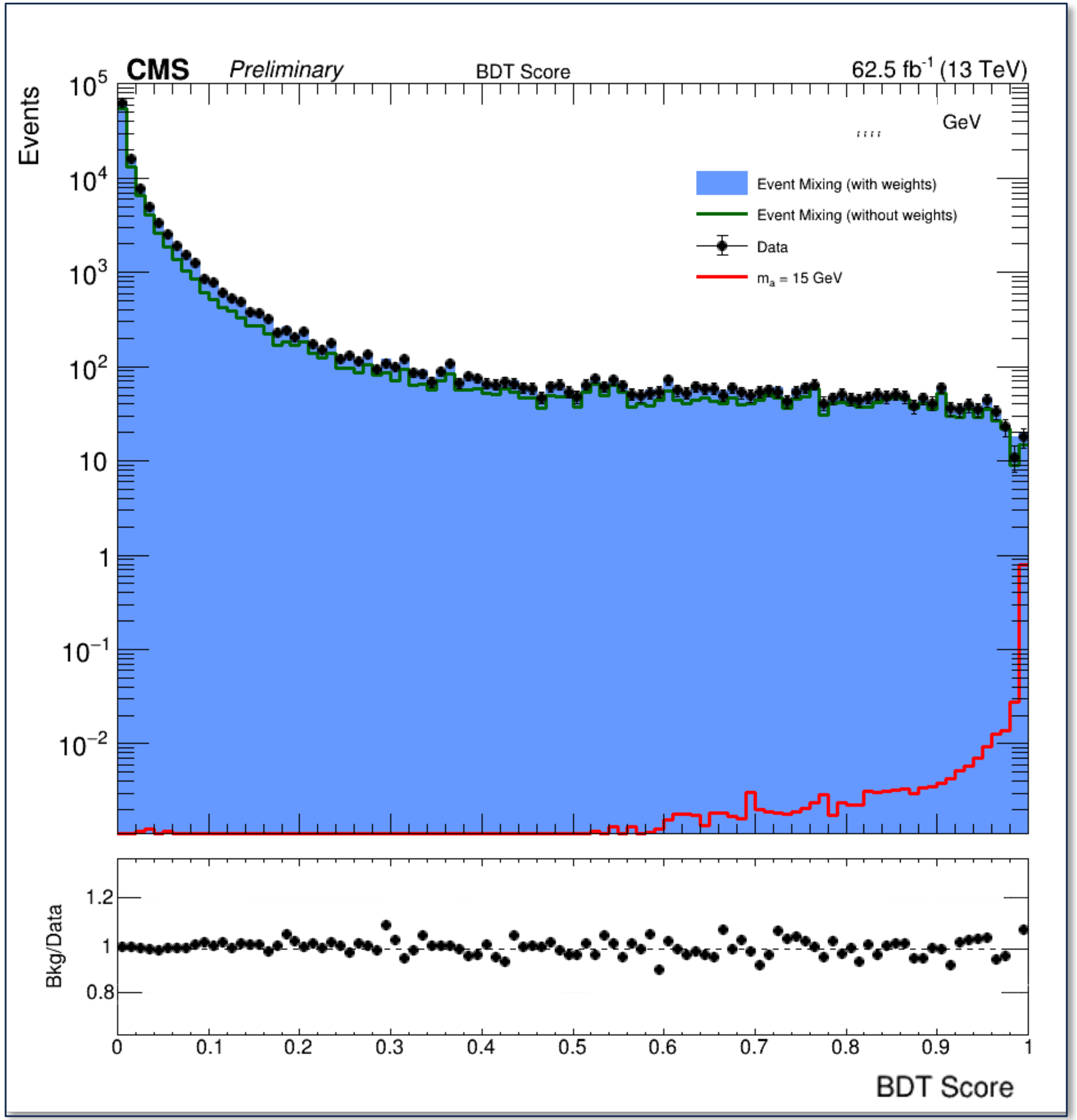


Applying the Event Selection BDT

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

For each nominal mass point:

- Signal MC, event mixed background, and data are all processed with the BDT, configured for the given mass point.
- A prediction distribution of event selection BDT score is generated for signal MC, event mixed background, and data. Shown for $m_a = 15$ GeV.



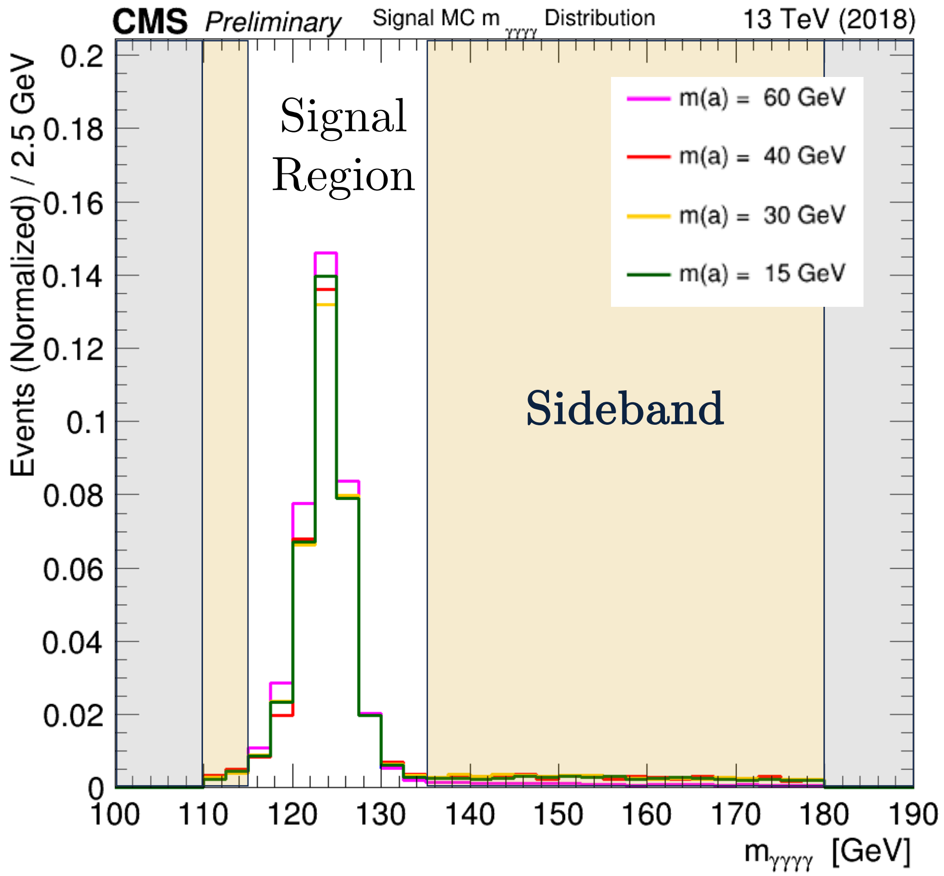
BDT Score Categorization

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

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

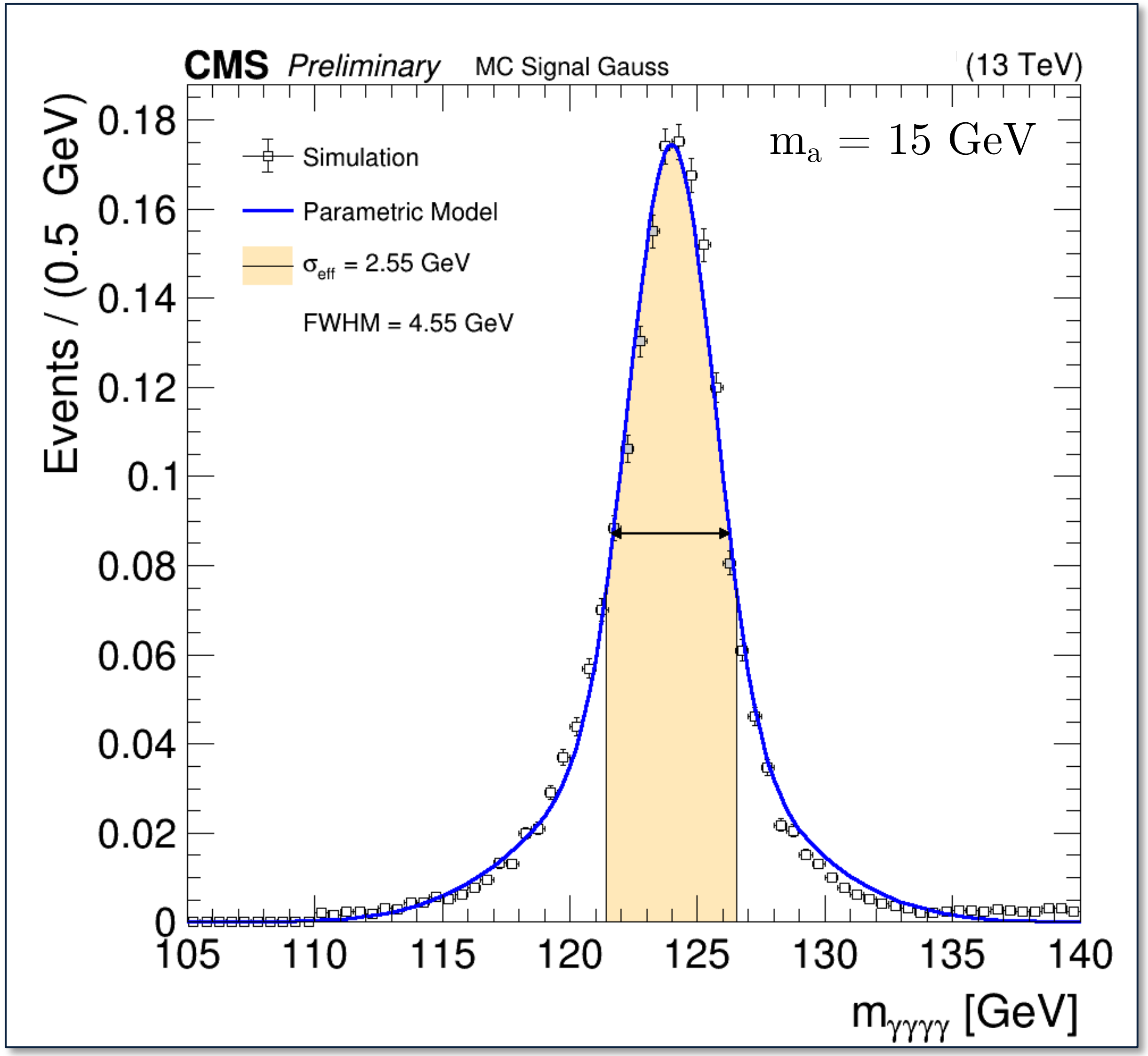
where S and B reference to the number of signal and background events.

- Need to choose what BDT score is signal-like enough. Categories are used to determine what BDT score is enough to classify an event as signal.
- Categories are constructed to maximize AMS in the signal region, where events with BDT scores inside boundaries are kept.
- 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 for optimal background estimation and data fitting.



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.



Advancements on Previous Studies

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 analysis
- 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

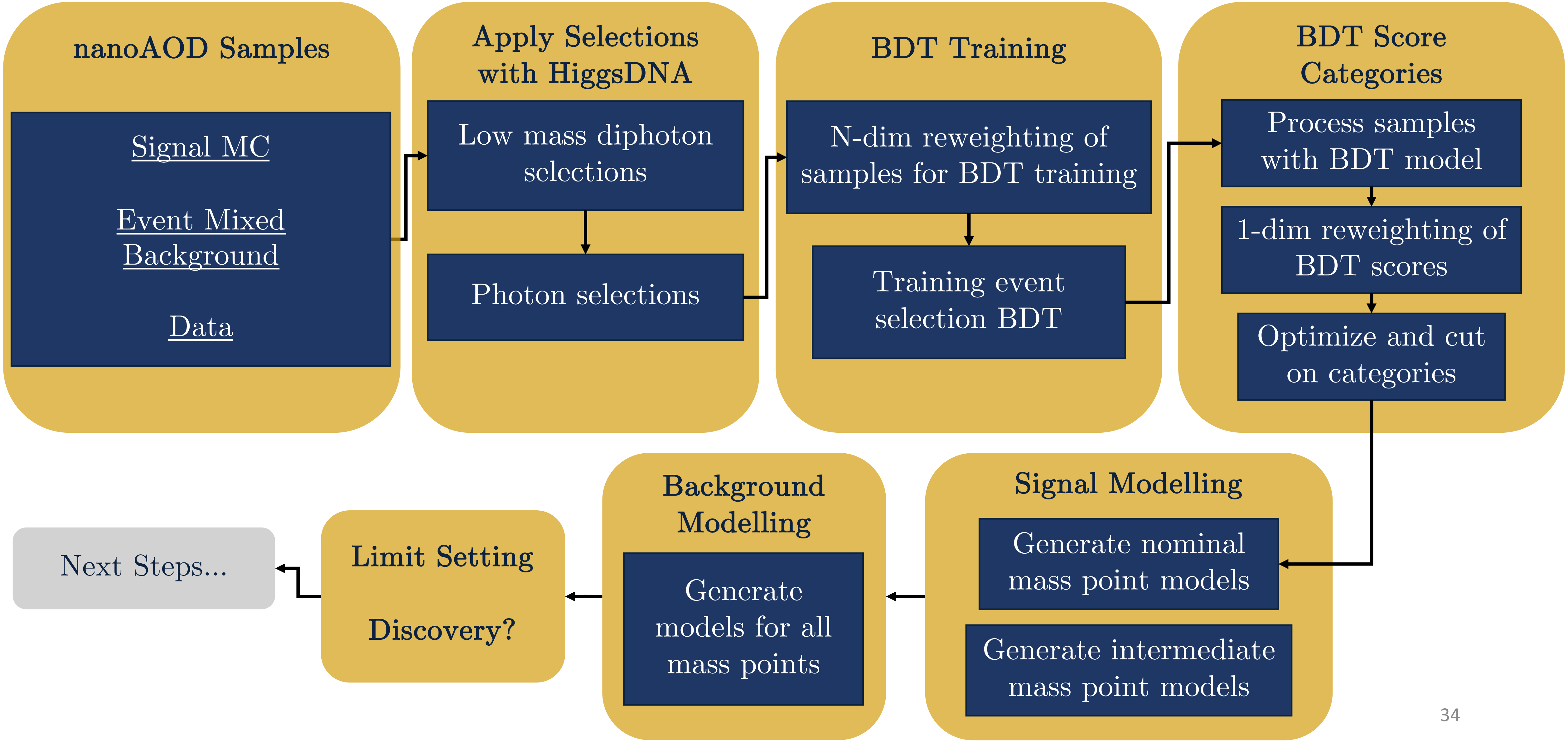
Goals:

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

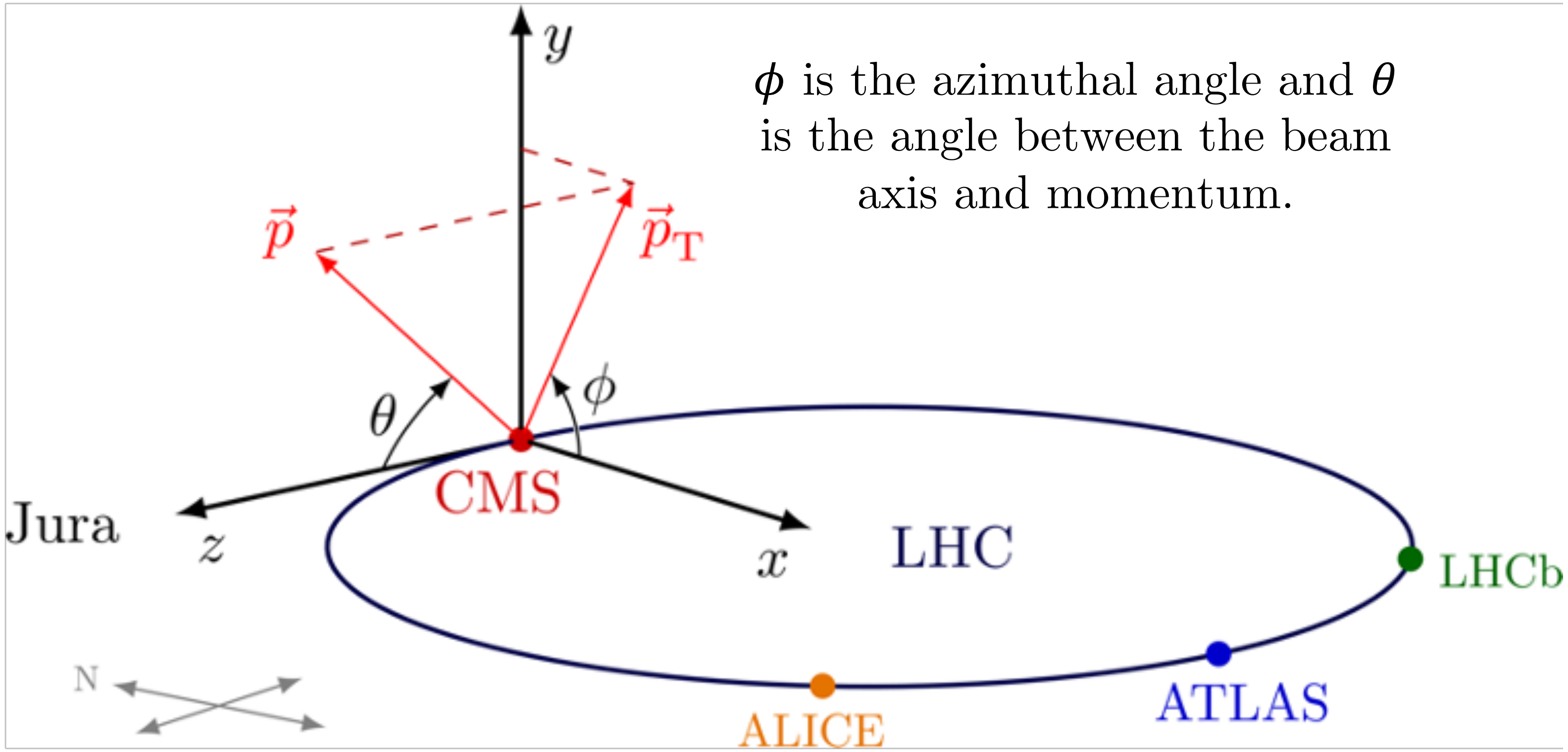
Thank You!

Backup

Analysis Overview



Quick Geometry of CMS

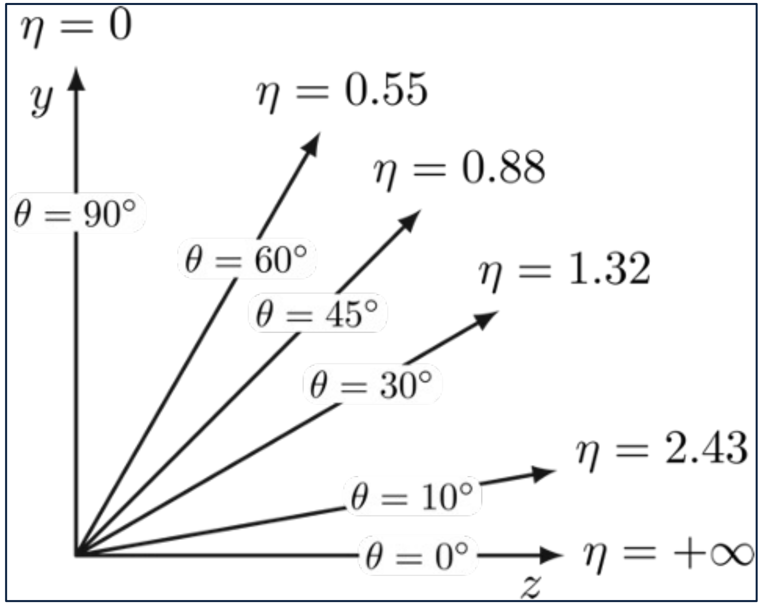


ϕ is the azimuthal angle and θ is the angle between the beam axis and momentum.

Angular Displacement:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$$

where $\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]$



Electroweak Symmetry Breaking

The Higgs get a non-zero vacuum expectation value (VEV), by convention $\langle\Phi\rangle = \frac{1}{\sqrt{2}}\binom{0}{v}$. This VEV breaks the $SU(2)_L \times U(1)_Y \longrightarrow U(1)_{EM}$.

We can now expand around the VEV and parameterize Φ as follows:

$$\Phi = \begin{pmatrix} G^+ \\ v + \frac{1}{\sqrt{2}}(H^0 + iG^0) \end{pmatrix} = \mathrm{e}^{i\vec{\chi}\cdot\vec{\tau}} \begin{pmatrix} 0 \\ v + \frac{1}{\sqrt{2}}h \end{pmatrix} \xrightarrow{\substack{\text{Unitary} \\ \text{Gauge}}} \begin{pmatrix} 0 \\ v + \frac{1}{\sqrt{2}}h \end{pmatrix} \tag{8}$$

where $\vec{\chi}$ are the eaten Goldstone bosons (GSBs) corresponding to G^+ and G^0, and $\vec{\tau}$ are the associated broken generators.

From here it is relatively straightforward to generate a mass term for every elementary particle and interactions with the SM Higgs:

$$\mathcal{L} \supset |D_\mu\Phi|^2 - y_d\bar{Q}_L\Phi d_R - y_d\bar{Q}_L\tilde{\Phi}u_R - y_e\bar{L}_L\tilde{\Phi}e_R + h.c. \tag{9}$$

Higgs Mechanism

Expanding $|D_\mu\Phi|^2 = \left| \frac{1}{\sqrt{2}} \left(\partial_\mu g T^i W_\mu^i + g' B_\mu\right) \begin{pmatrix} 0 \\ v \end{pmatrix} \right|^2$, we get

$$= \frac{v^2}{8} \left[g^2 \left(\left(W_\mu^1\right)^2 + \left(W_\mu^2\right)^2 \right) + \left(g W_\mu^3 - g' B_\mu\right)^2 \right].$$

This yields masses for W/Z and the photon:

$$W_\mu^\pm \equiv \frac{1}{\sqrt{2}} \left(W_\mu^1 \mp i W_\mu^2\right) \quad \text{with mass} \quad m_W = \frac{g\,v}{2}$$

$$Z_\mu \equiv \frac{1}{\sqrt{g^2 + g'^2}} \left(g W_\mu^3 - g' B_\mu\right) \quad \text{with mass} \quad m_Z = \frac{v}{2}\sqrt{g^2 + g'^2}\,,$$

$$A_\mu \equiv \frac{1}{\sqrt{g^2 + g'^2}} \left(g' W_\mu^3 + g B_\mu\right) \quad \text{with mass} \quad m_A = 0\,.$$

Various Extended Higgs Sector Models

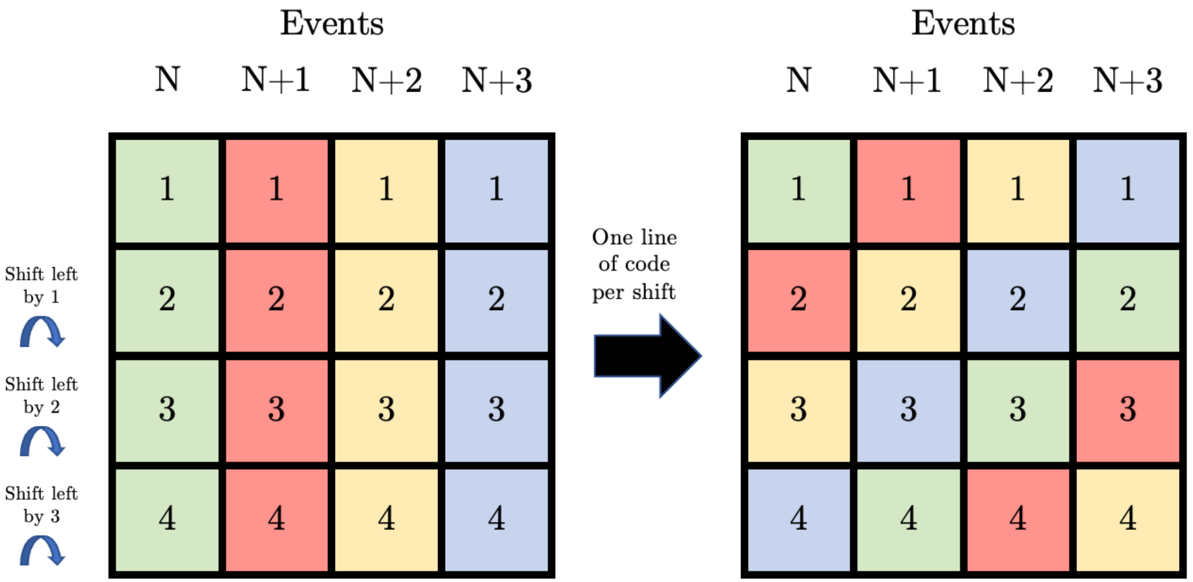
- Possibly the only decay (a->gg) of pseudoscalars depending on model (theory)
- Difference in phenomenologies
- No FCNC -> Z2 symmetries
- Rho parameter + maintaining = 1 & Restrictions in (T_3, Y) for rho = 1
- E.g. triplets (etc) need contrived VEV for rho ~= 1

Background Estimation Summary

- Why not use MC for background?
- Note about background modelling as nuisance parameter when extracting limits?
- Main background processes + background estimation summary
- Check notes for more!

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…

Higgs to Diphoton NanoAOD Framework (HiggsDNA)

- Coffea processors
- How cuts are implemented
- What this thing is for (hgg group's framework).

Pre-selections

<table>
<tr><th>γ_1 must be:</th><th>E_T</th><th>R_9</th><th>H/E</th><th>$\sigma_{i\eta i\eta}$</th><th>PF Pho Iso</th><th>Tracker Iso</th></tr>
<tr><td>EB; $R_9 > 0.85$</td><td rowspan="3">15.0</td><td rowspan="3">> 0.5</td><td rowspan="3">< 0.08</td><td rowspan="2">< 0.015</td><td rowspan="3">< 4.0</td><td rowspan="3">< 6.0</td></tr>
<tr><td>EB; $R_9 \leq 0.85$</td></tr>
<tr><td>EE; $R_9 > 0.9$</td><td>< 0.035</td></tr>
</table>

- 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

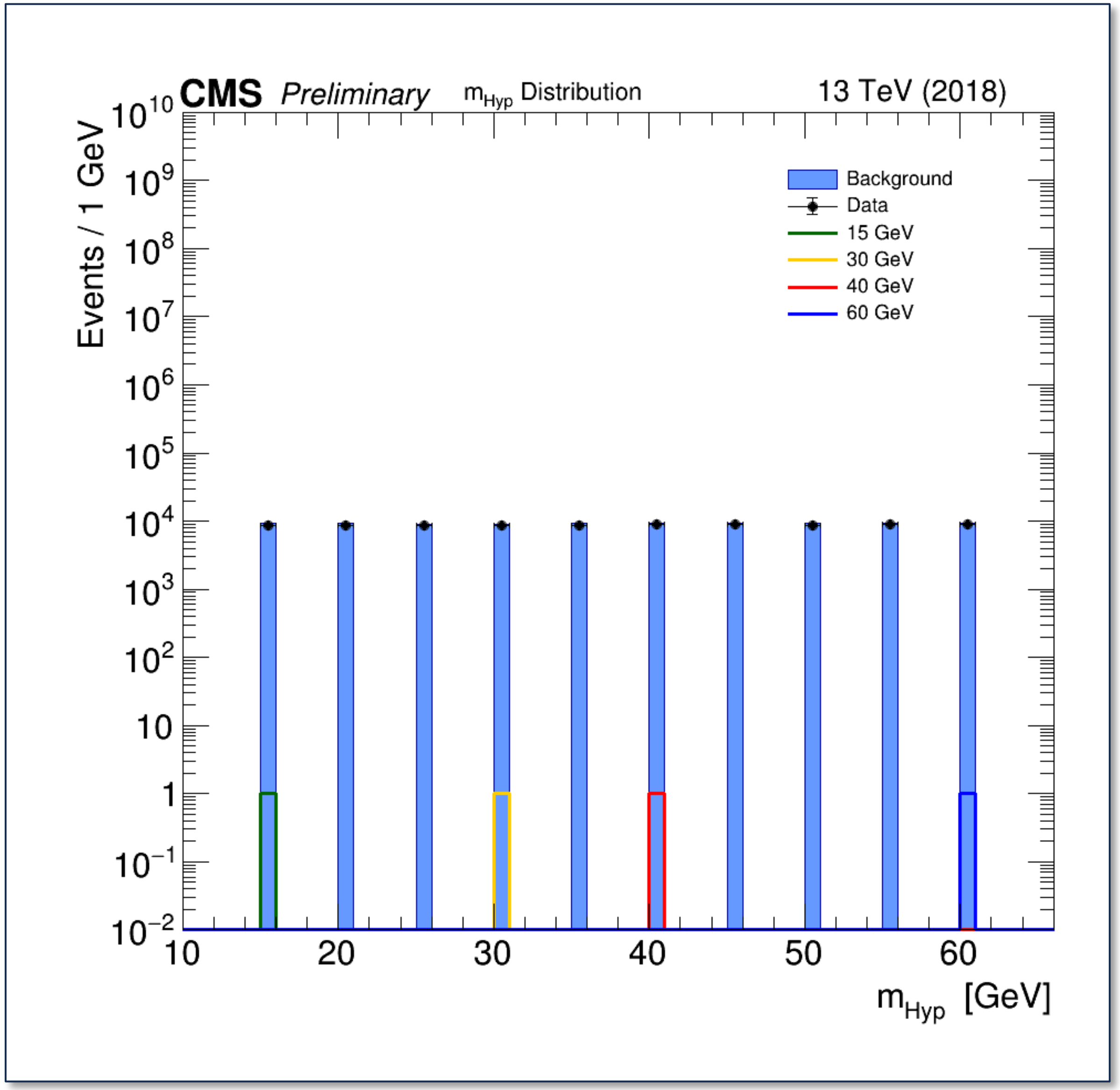
<table>
<tr><th>Option 1</th><th>Option 2</th><th>Option 3</th></tr>
<tr><td rowspan="3">$R_9 > 0.8$</td><td rowspan="3">Charged Hadron Iso < 20.0</td><td>Charged Hadron Iso/$p_T < 0.3$</td></tr>
<tr><td>IF $p_T > 14.0$</td></tr>
<tr><td>AND $H/E < 0.15$</td></tr>
</table>

Deep learning vs BDT

- Deep learning model (Keras network vs tree) (BDT: many weak learners)
  - Better with lower-level variables. Still might be possible given the work done to add extra branches to the samples anyway
- More BDT explanation: https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf

Event Selection BDT

- 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.

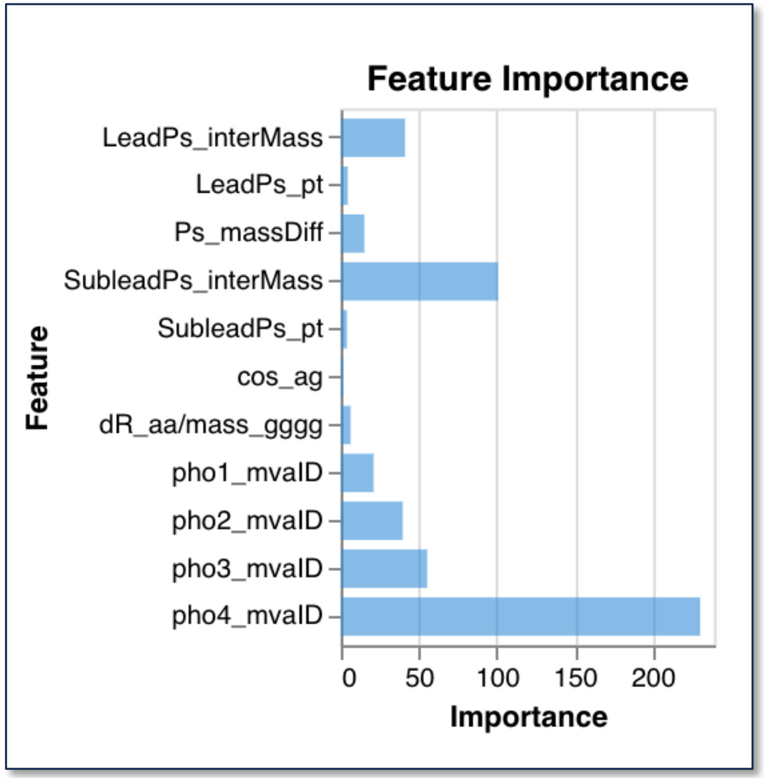


Event Selection BDT Inputs

- 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.

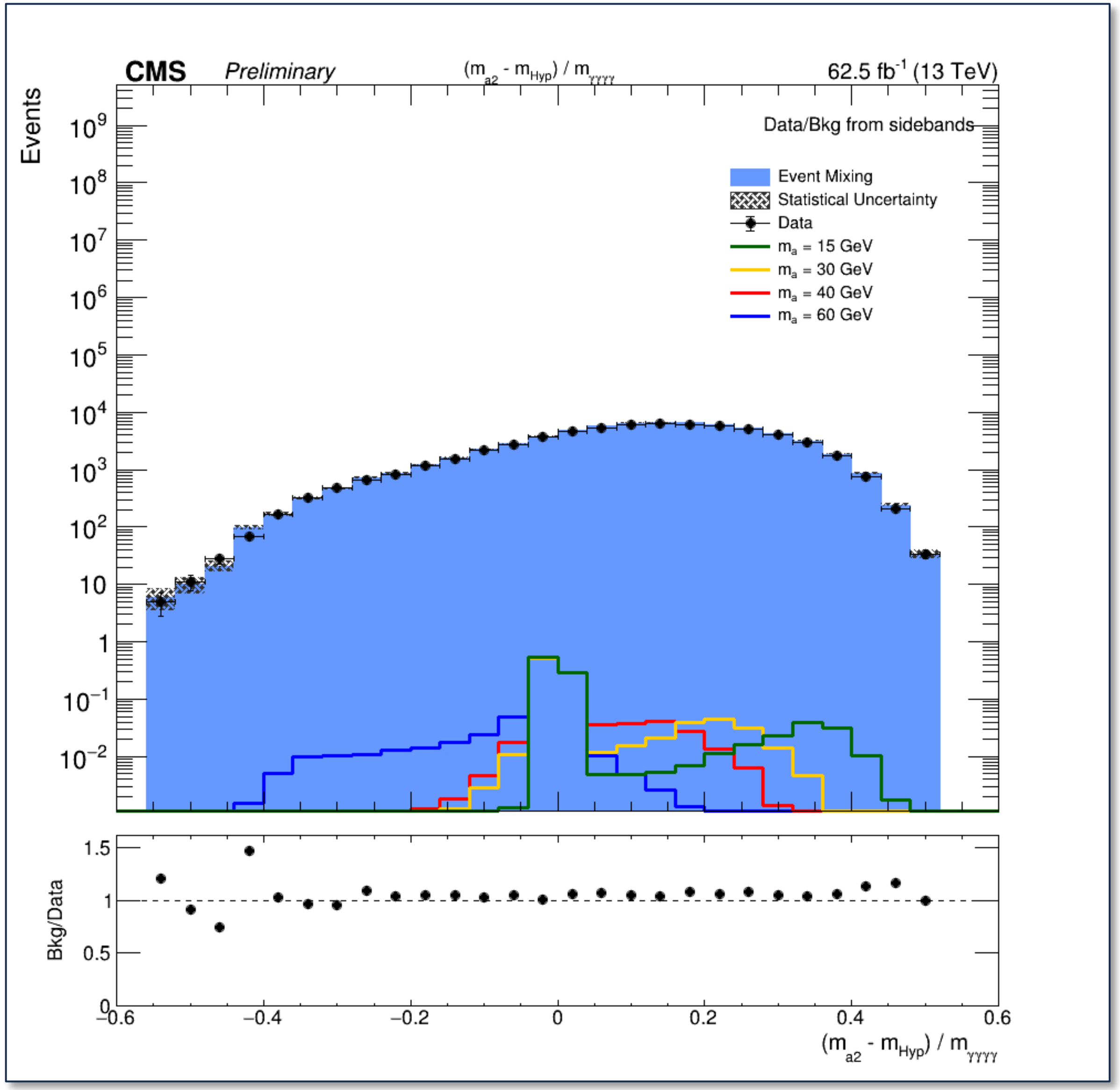
Input variables:

- $\gamma_{1,2,3,4}$ 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 $a \rightarrow \gamma\gamma$
- $\frac{m_{a1\,RECO} - m_{a\,Hyp}}{m_{\gamma\gamma\gamma\gamma}}$ $\frac{m_{a2\,RECO} - m_{a\,Hyp}}{m_{\gamma\gamma\gamma\gamma}}$

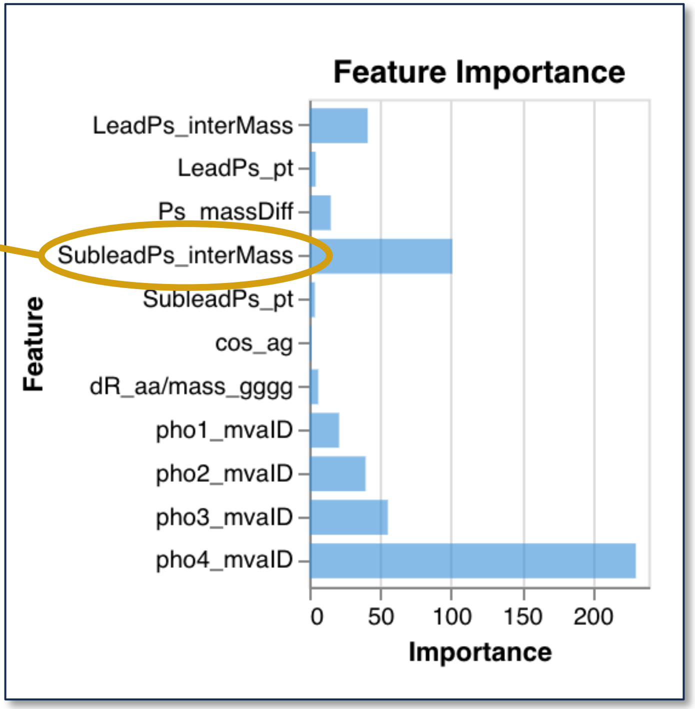


Above: Chart of feature importance for the event selection BDT input variables from one of the best performing models.

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.

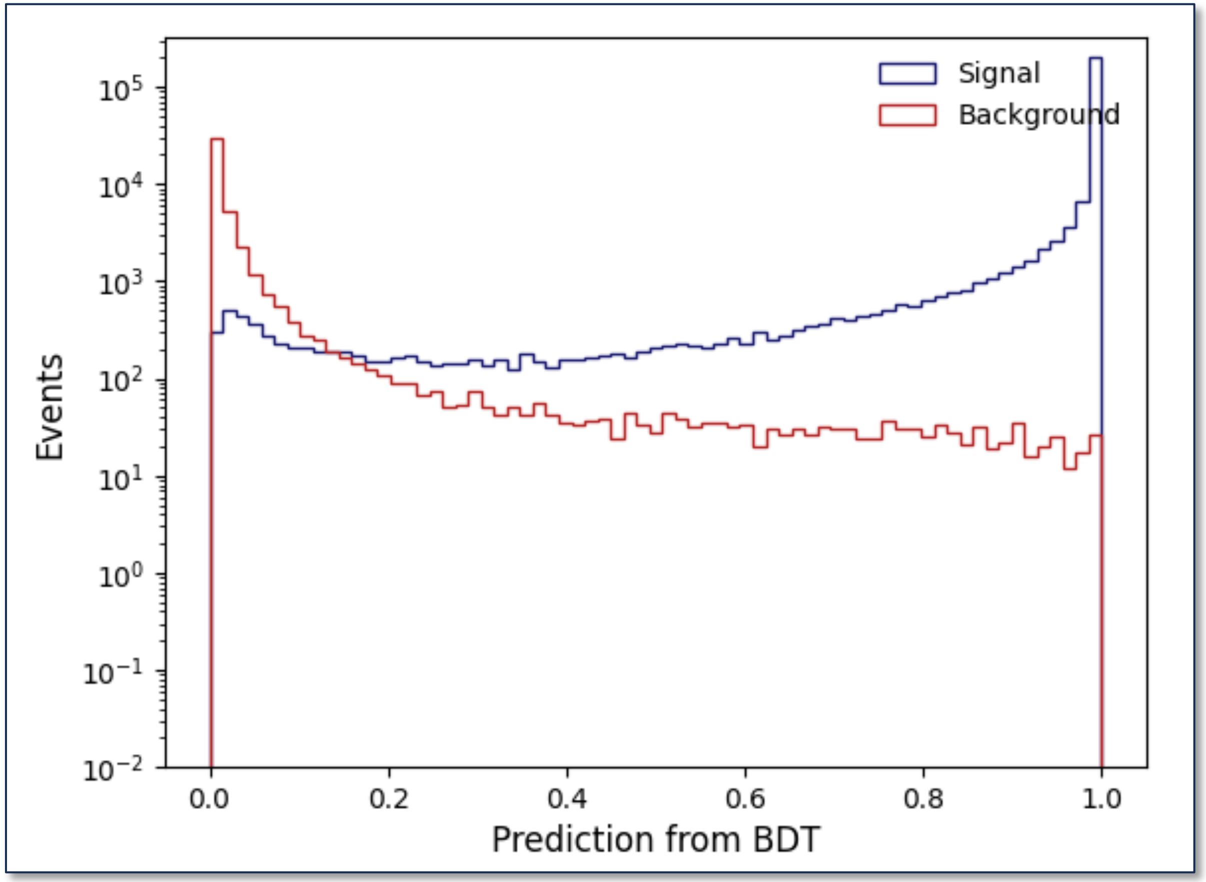


Above: Chart of feature importance for the event selection BDT input variables from one of the best performing models.

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.

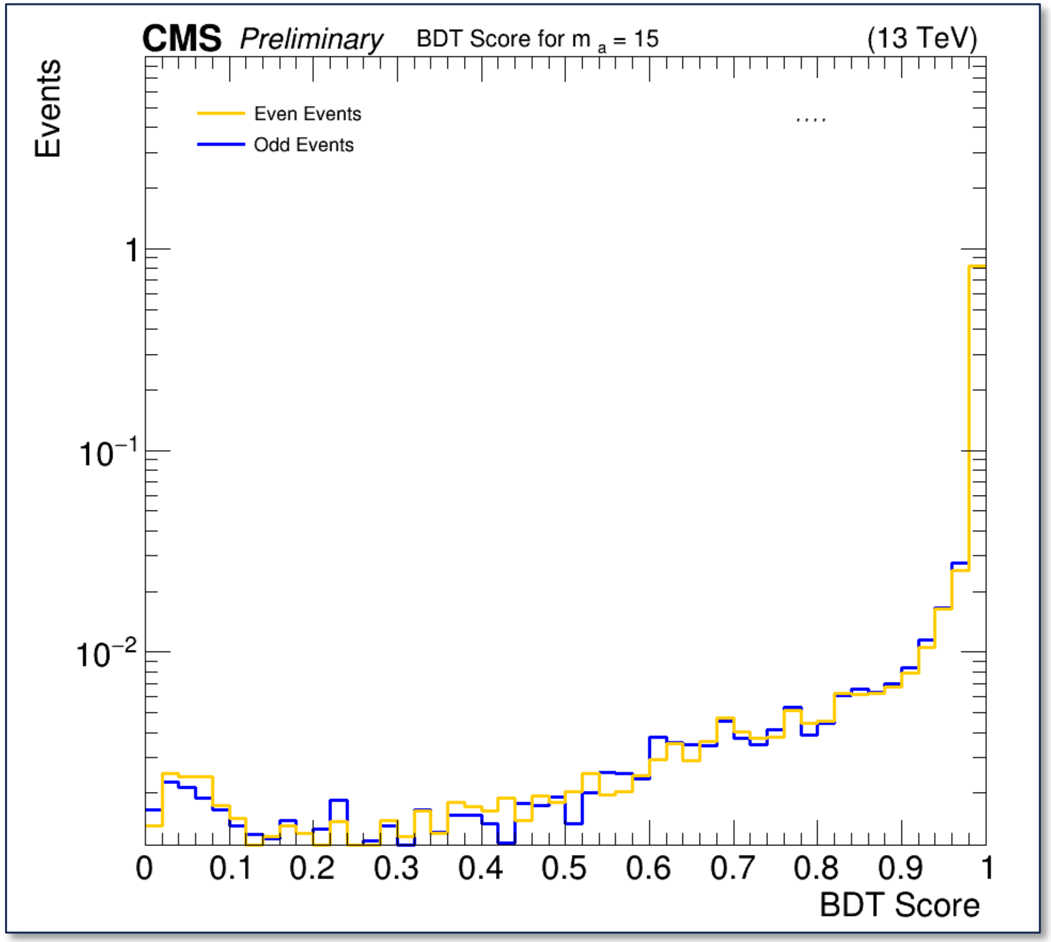
BDT Evaluation Metrics

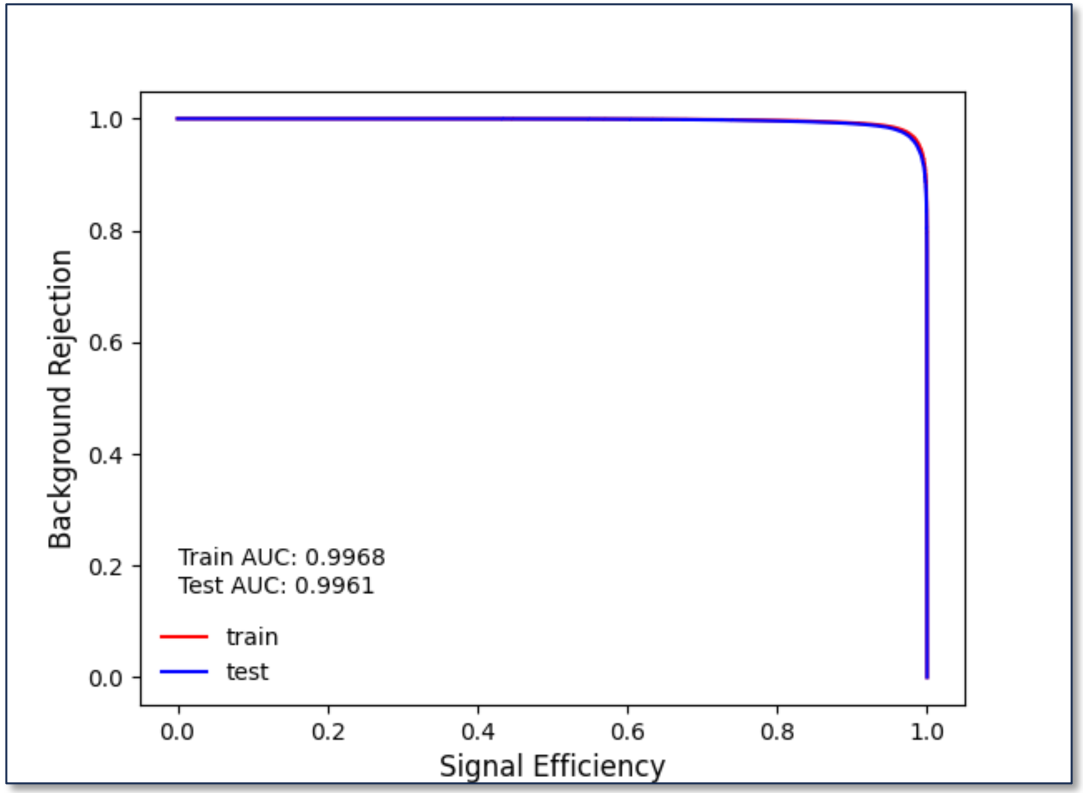
- Metrics for evaluating BDT performance:
  - AUC
  - Log loss minimization
  - Error rate
  - Average truth deviation
  - F1 score
  - ROC curve

- Mention cross-validation to prevent overtraining

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

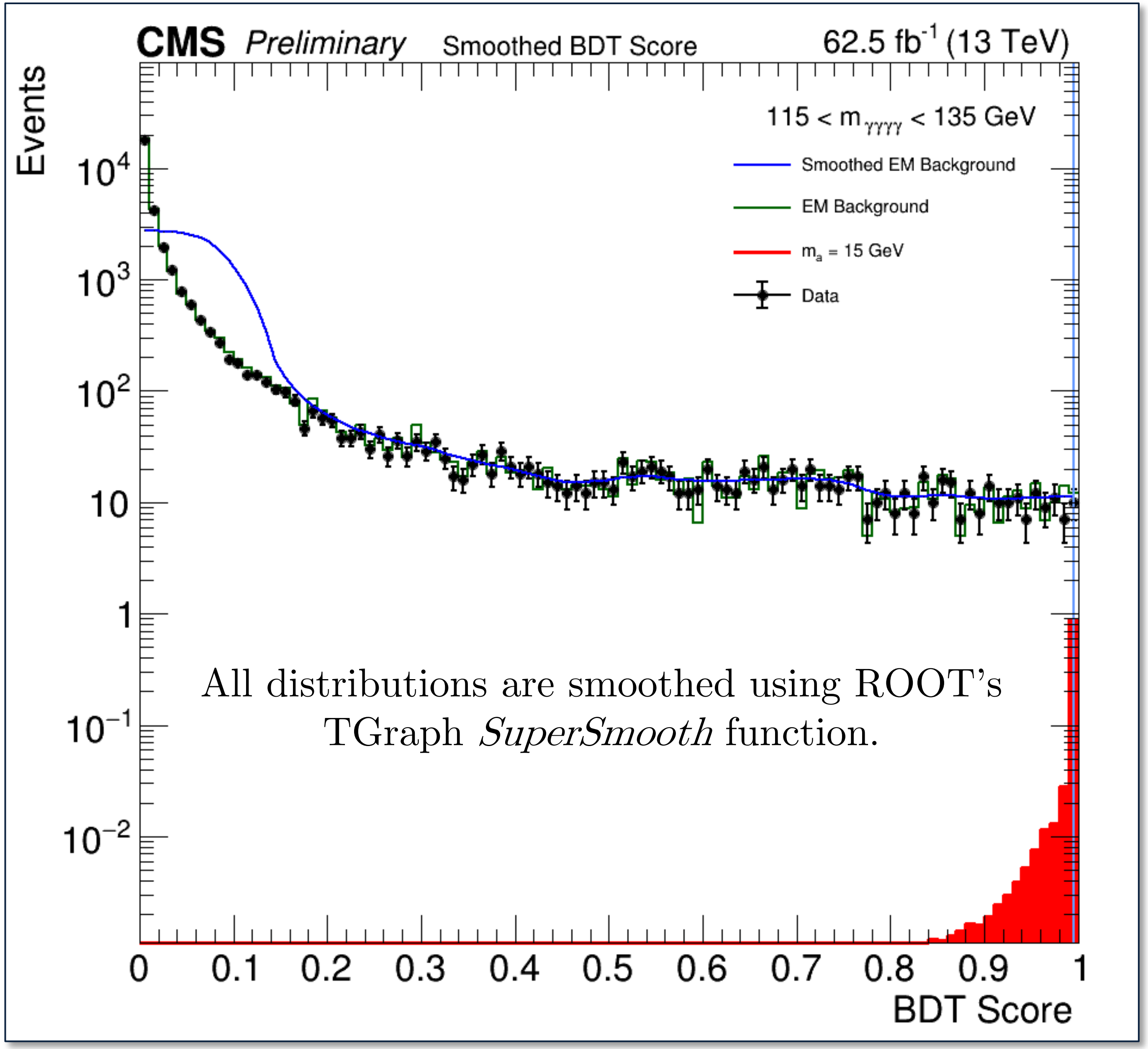
- Reminder: Approximate Mean Significance (AMS)

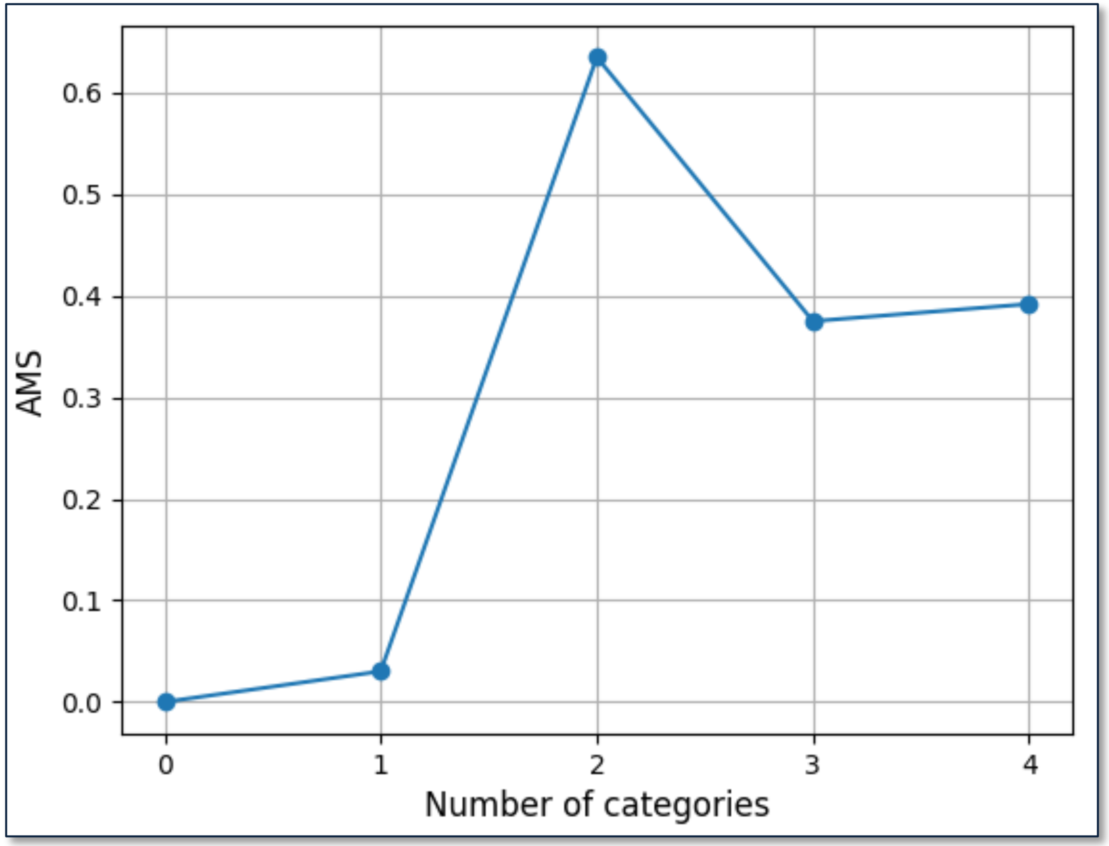
$$\text{AMS} = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right) - S\right]}$$

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 exclude background-like events such that AMS is maximized.

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.

Fitting Signal Shape

- Signal modelling: Why 2 Gaussians? Why not more? (2NLL/Chi2 discussion)

Analysis Significance

- Why choose AMS over other significance definitions?
- AMS -> S/sqrt(B) for B >> S (confirm this)

Miscellaneous

- E_T vs p_T. Involves cosh(eta) (or theta?) – measure momentum but can break up into longitudinal and transverse components. For massless particles, $E_T = p_T$.
- Higgs production cross section for all channels combined is 52 pb (https://arxiv.org/abs/1610.07922)
- Vacuum of LHC
- LHC magnets