The Impact of Topo-Cluster Splitting on Boosted Object Identification in ATLAS

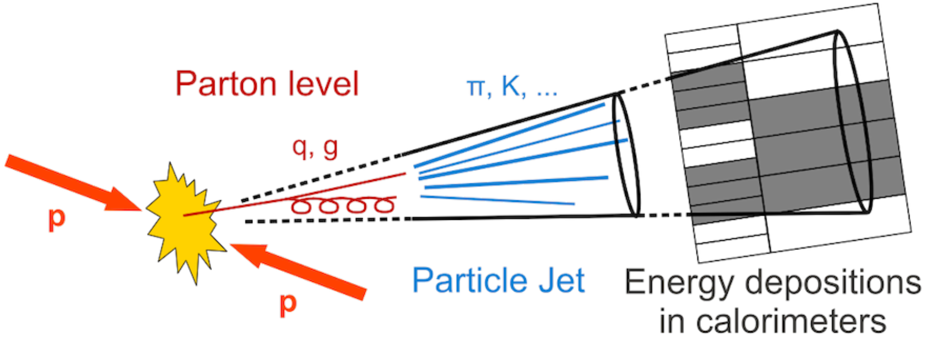
Defense Presentation

Nicolai Weitkemper

Bologna | September 29, 2025

Jets in ATLAS

- pp collisions at LHC produce quarks and gluons
- Hadronization → collimated sprays of particles called jets
- Tools for reconstruction rather than physical objects
- Built from calorimeter and/or track data

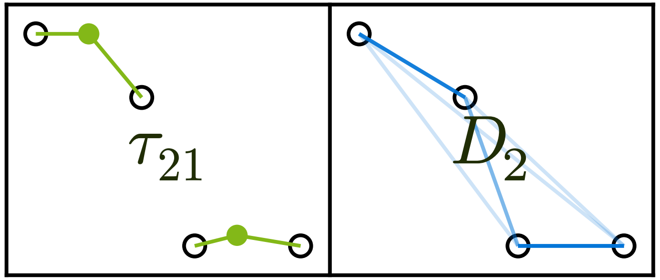


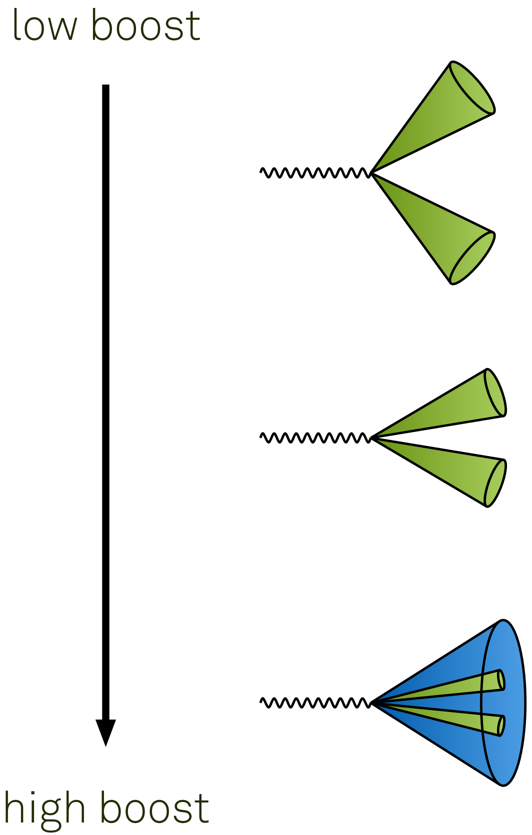
[ATLAS software documentation]

calorimeter cells → topo-clustering → constituents (clusters) → seq. recombination → jets

Substructure

- Decay products of a boosted particle tend to overlap
- Substructure variables help identify them:
  - Operate on jet constituents (here: topo-clusters)
  - τ_N: p_T-weighted summed distance to subjet axes (determined separately)
    - Ratios $\tau_{21} = \tau_2/\tau_1$ etc.
    - Low $\tau_{21} \rightarrow$ likely 2-prong
  - D_2: sensitive to 2-prong like τ_{21}, but independent of subjet axes





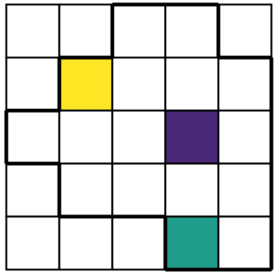
Topo-Clustering

- Intermediate step between cells and jet formation; yields constituents
- Reduces noise (pile-up, electronics)
- Operates on cell significance $\varsigma_{(\mathrm{cell})} = \frac{|E_{\mathrm{cell}}|}{\sigma_{\mathrm{cell}}}$

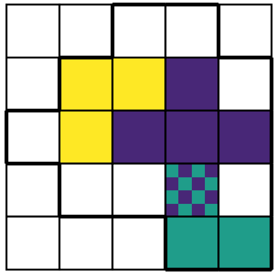
seed cells ($\varsigma \geq 4$) | add growth cells ($\varsigma \geq 2$) | repeat growth until exhausted | add border cells ($\varsigma \geq 0$)

Topo-Cluster Splitting

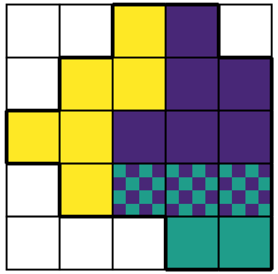
- Split overly large clusters
- Based on re-clustering from local maxima that fulfill
  - $E_{\text{cell}} \geq 500$ MeV
  - Within specific sampling layers
  - ≥ 4 neighboring cells



seed cells
($E_{\text{cell}} \geq 500$ MeV)



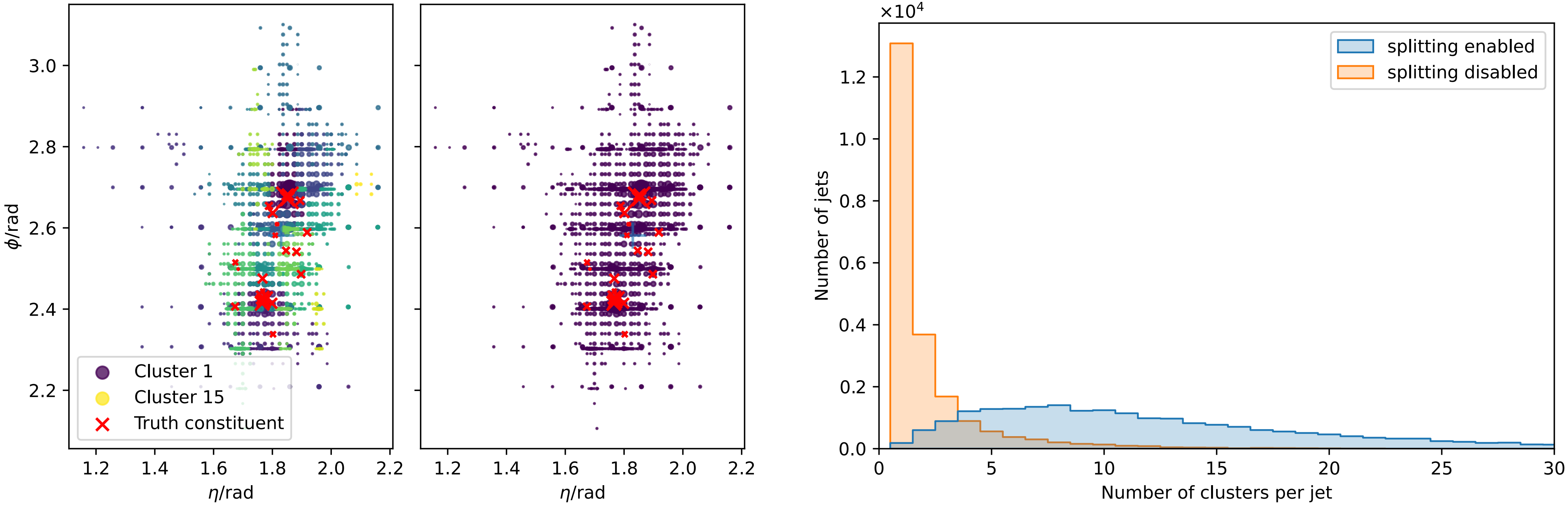
grow to cells within
original cluster



repeat growth
until exhausted

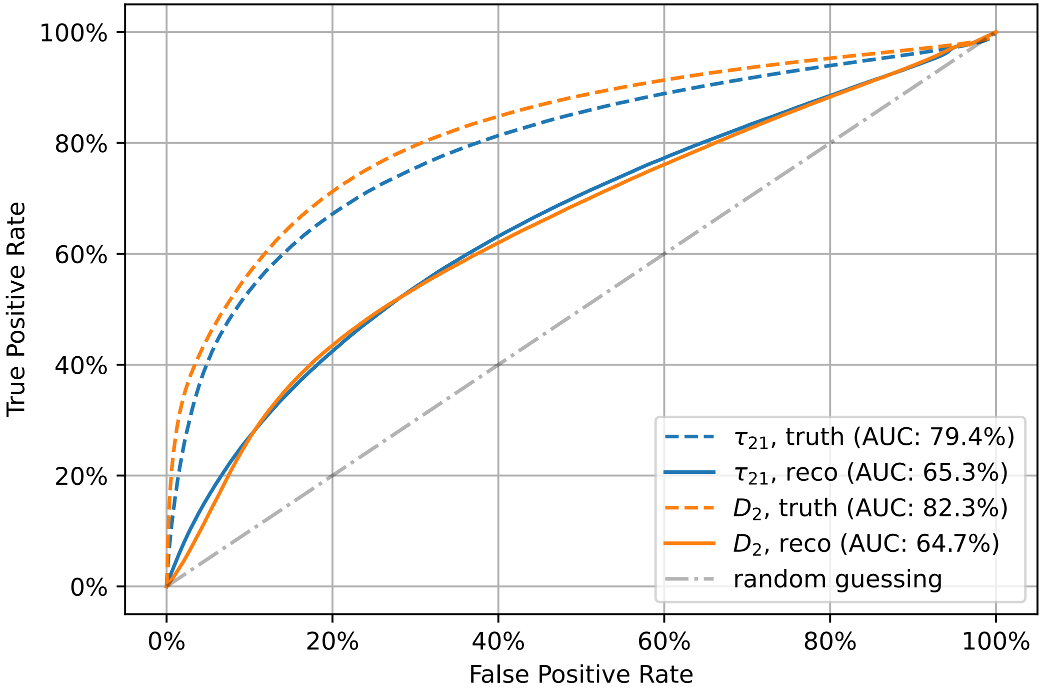
Splitting Prevents Over-Growth

- Clusters grow overly large without splitting
- But substructure variables operate on clusters (not cells)
  → Loss of information

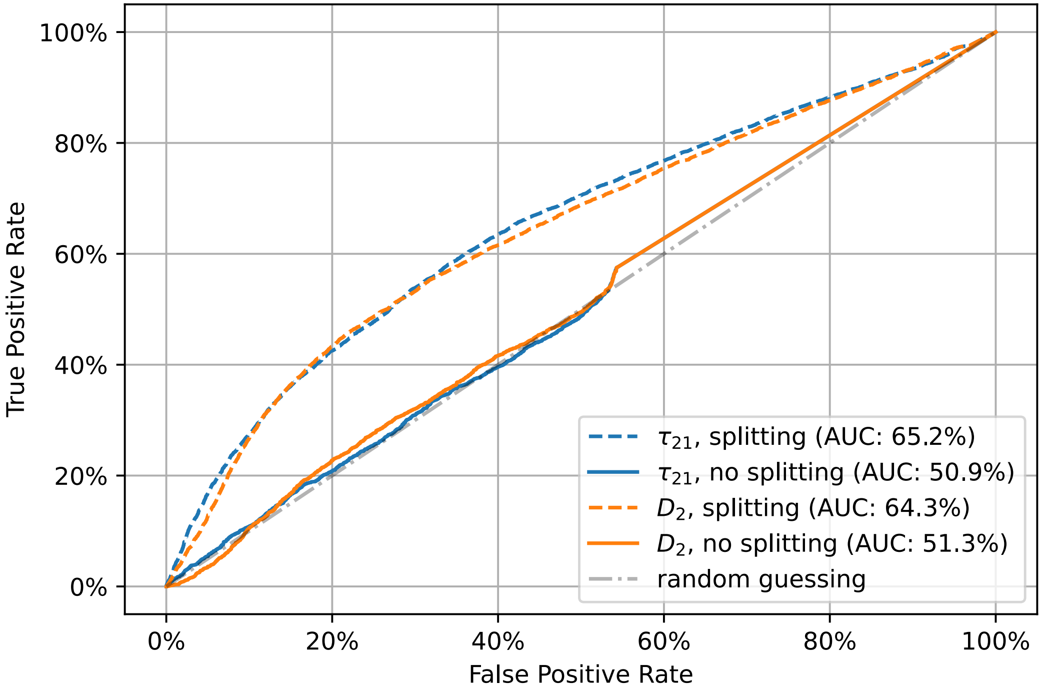


No Splitting – No Substructure

- ROC curves for W' (2-prong) vs. background (no prongs / 1-prong) classification
- Performance: truth $\gg$ reco $\gg$ reco without splitting



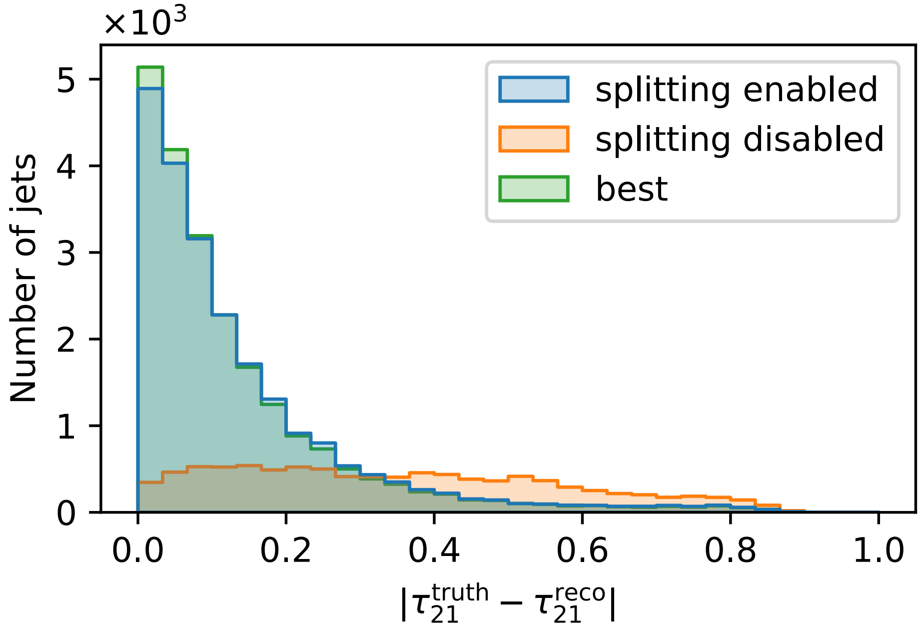
truth vs. reconstructed



splitting vs. no splitting

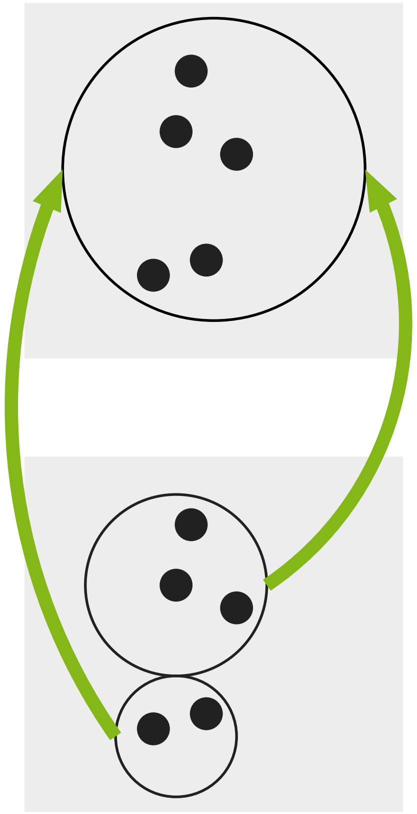
Is Splitting Always Better?

- Compare to theoretical optimum:
  - Always choose the option (splitting enabled/ disabled) closest to truth
- Splitting disabled is better $8\,\%$ of the time for τ_{21}
  - $< 5\,\%$ for D_2 and τ_{32}



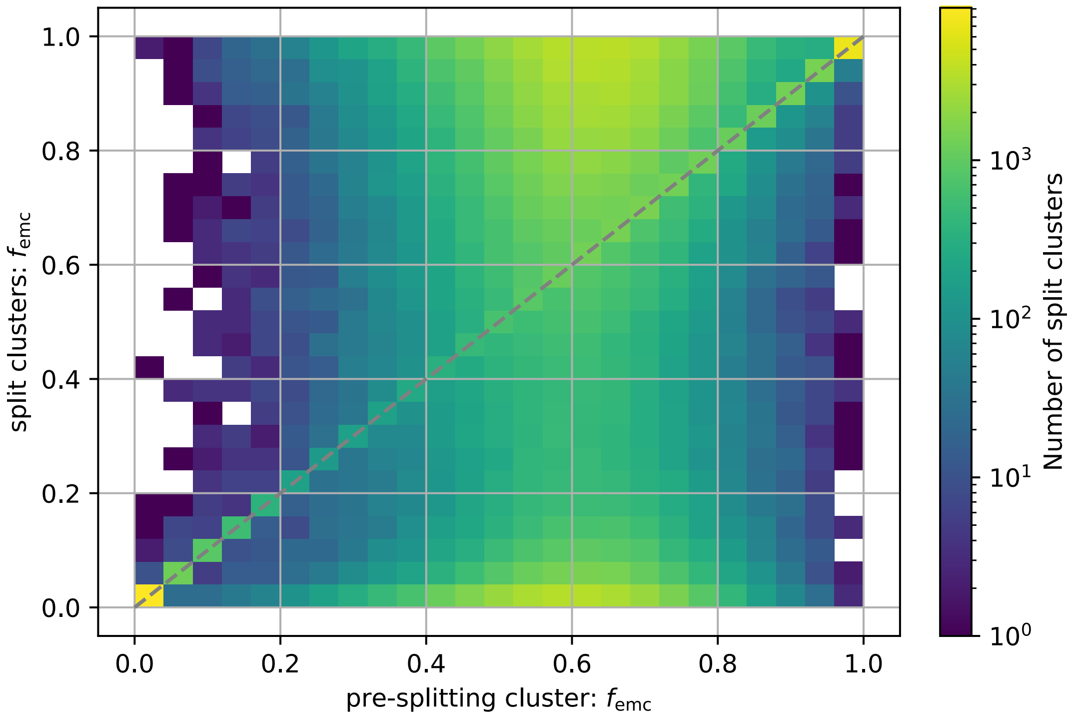
Matching (Non-)Split Clusters

- Motivation:
  - Compare individual clusters before/after splitting, not just overall trends
  - Verify expectations about splitting effects
  - Find variables that might guide future splitting optimizations
- Approach:
  - Check $\mathbf{cells}_{\text{post-splitting},j} \subseteq \mathbf{cells}_{\text{pre-splitting},i}$
- Issues:
  - Often no direct comparison to truth
  - One-to-many matching
  - Hidden correlations (e.g. cluster size)



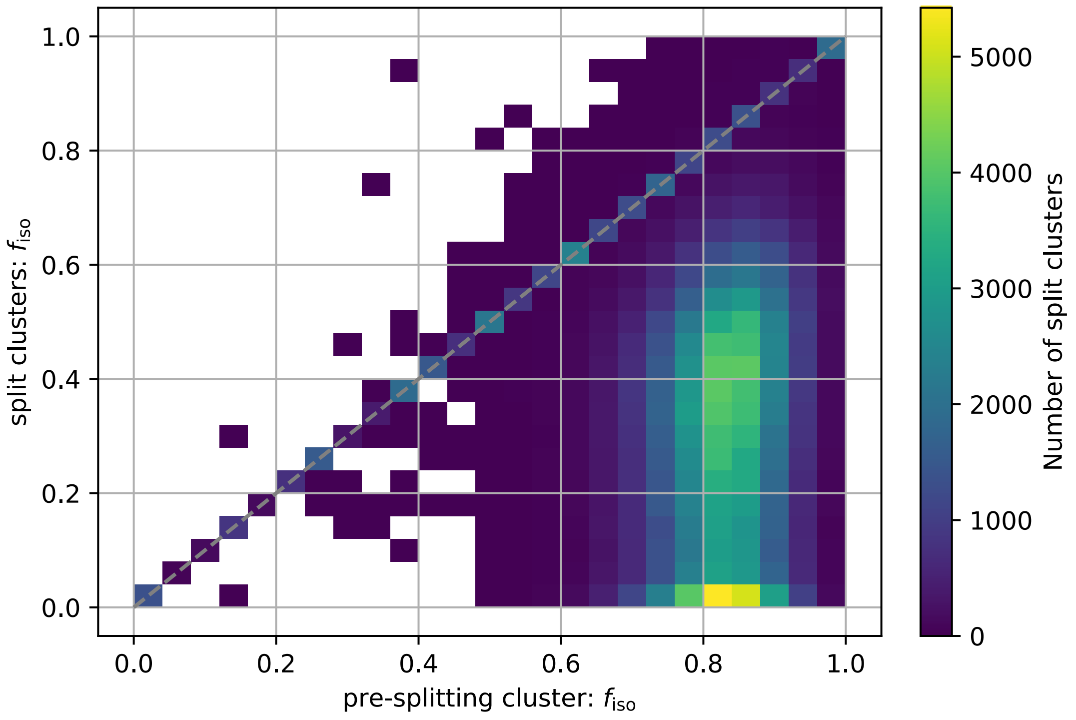
Matching Results: f_{emc}

- f_{emc}: fraction of energy in EM calorimeter
- Diagonal: not split (before = after)
- $(0,0)$ and $(1,1)$: already fully within hadronic / EM clusters
- Top/bottom high-density regions
  → Splitting makes "clearer cut" between sampling layers



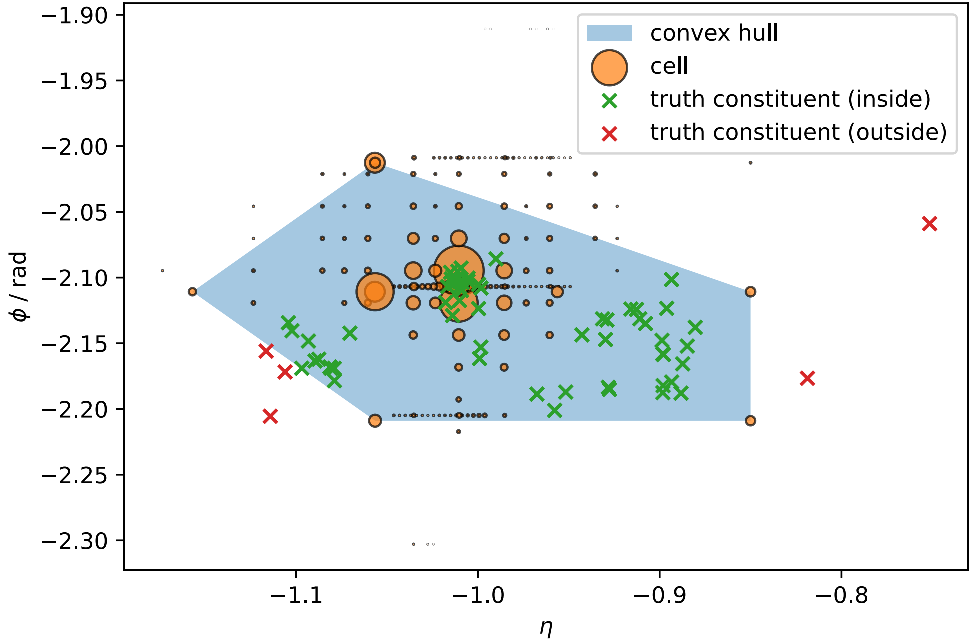
Matching Results: f_{iso}

- f_{iso}: fraction of adjacent cells that are not part of another cluster
- Relevant for calibration: Is energy lost or within another cluster?
- Pre-splitting isolation always ≈ 0.8
- Strong reduction in isolation after splitting



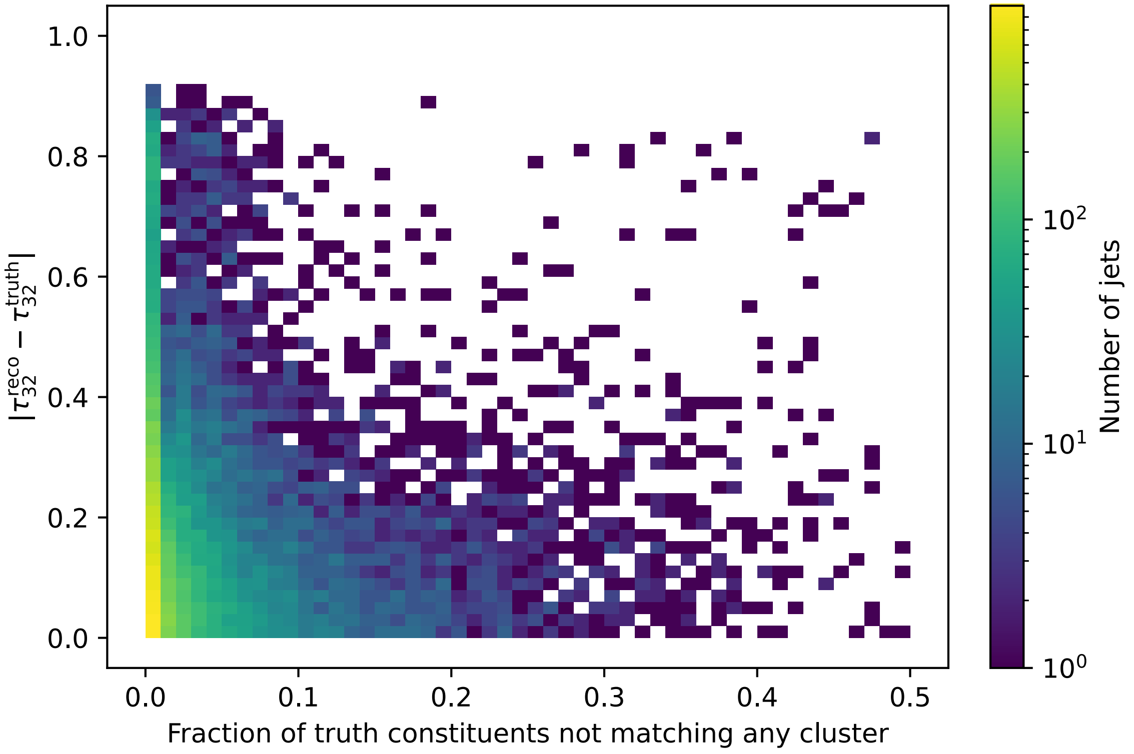
How Does Splitting Relate to Truth Constituents?

- MC gives access to truth constituents
- What is different for clusters with few/many matched constituents?
- How well do clusters capture truth constituents?
- Can cluster features help identify clusters that align well with truth constituents?
- → Match truth constituents to clusters using different methods:
  - Convex hull (shown)
  - Nearest-neighbor search per sampling layer



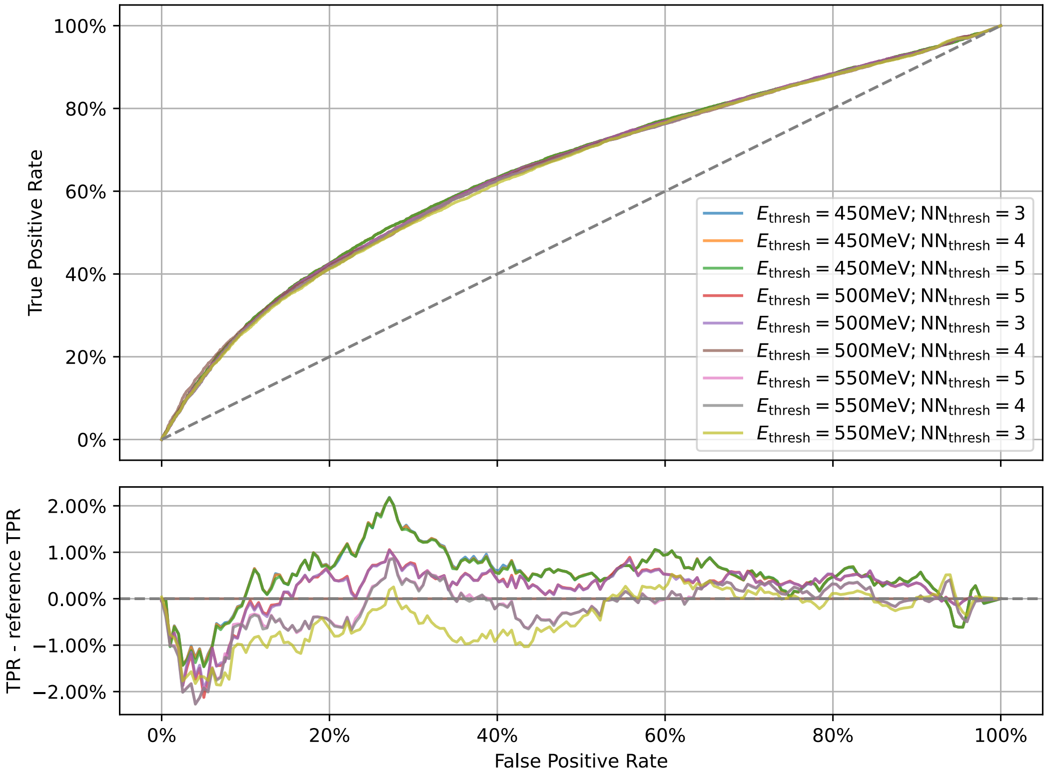
Relation of Matching to Reconstruction Error

- Most truth constituents match with at least one cluster
  - Slight overestimation due to convexity assumption
- Less matches $\not\Rightarrow$ higher error
  - Diminishing effect of low-E truth constituents
  - Hidden correlation (higher-E jet $\rightarrow$ more irrelevant truth constituents)



Splitting Parameters Are Near-Optimal

- Grid search of splitting parameters:
  - Minimum neighbors of seed cell: $(3, 4, 5)$
  - Minimum energy of seed cell: $(400, 500, 600)\ \mathrm{MeV}$
- Differences in ROC-AUC of up to $1.4\,\%$
- Inconsistent trend between substructure variables
- Future directions:
  - Larger variations
  - Other parameters (e.g. sampling layers)



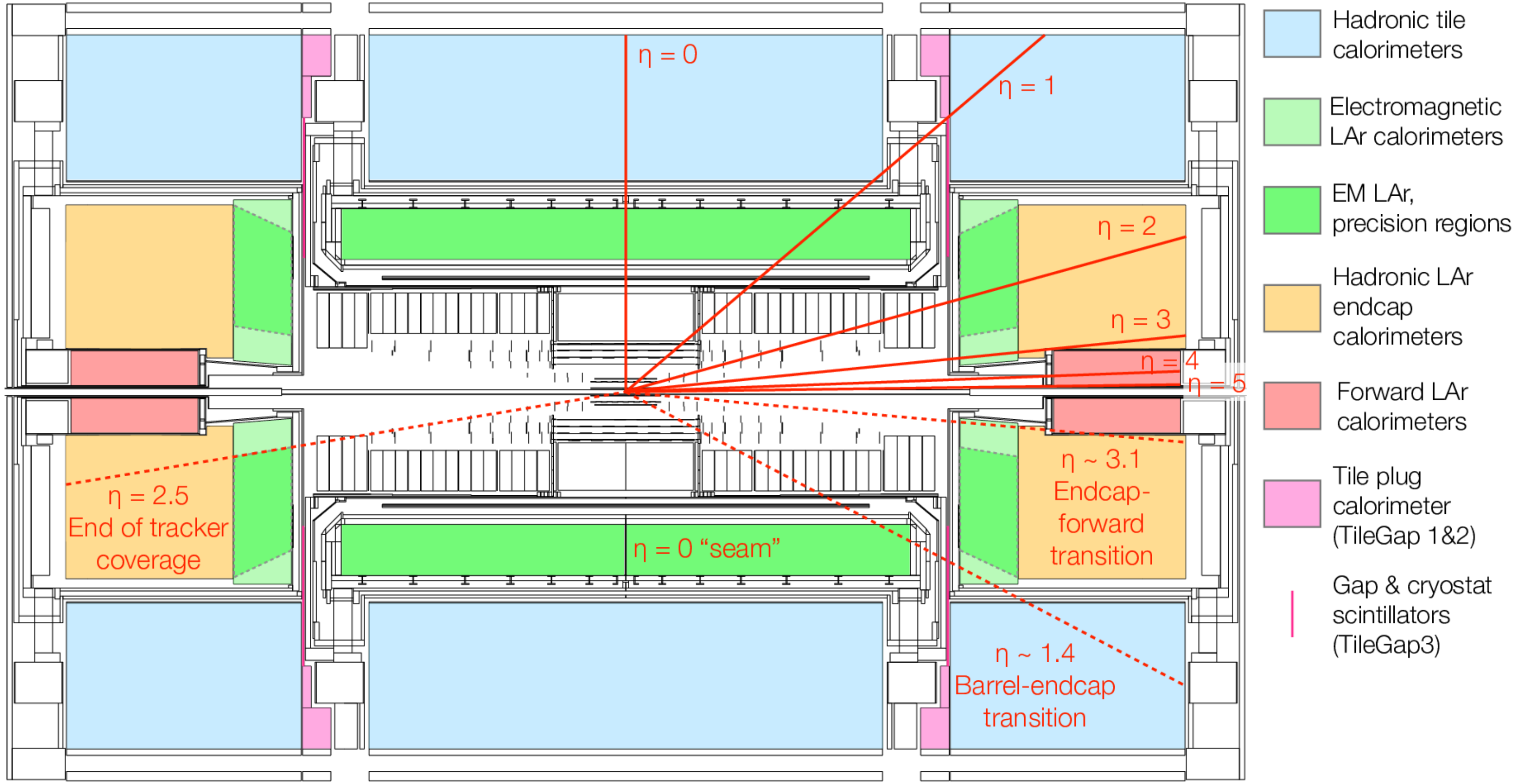
ROC curves for $\tau_{21}^{\mathrm{reco}}$

Conclusions

- Topo-cluster splitting is essential for boosted object identification
- Splitting parameters have limited impact on performance
- More fundamental changes to the algorithm needed for meaningful improvements
- Ideas for future work:
  - Tighter integration with Athena (the ATLAS software framework)
  - Clearer definition of "optimal" splitting
  - Application to more recent jet algorithms (UFOs instead of LCTopo)
  - ML methods to guide or replace splitting

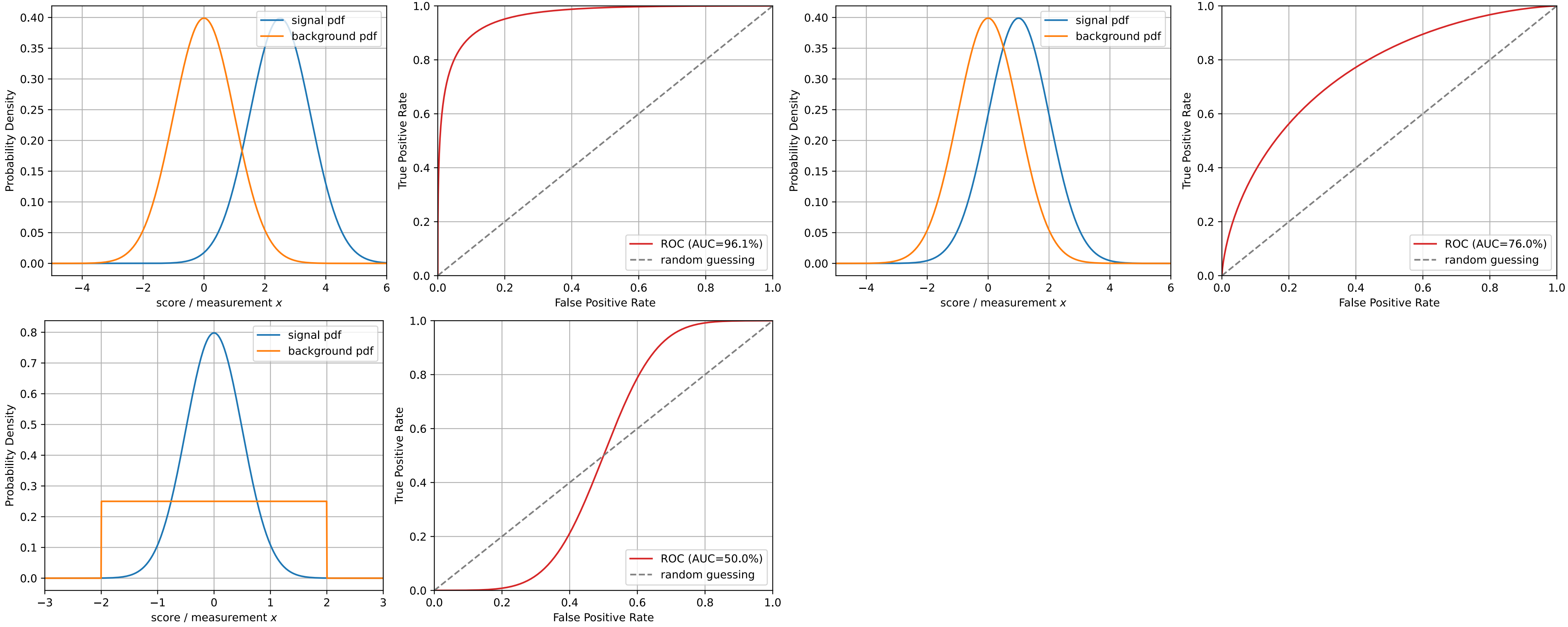
Thanks for your attention!

The ATLAS Calorimeters



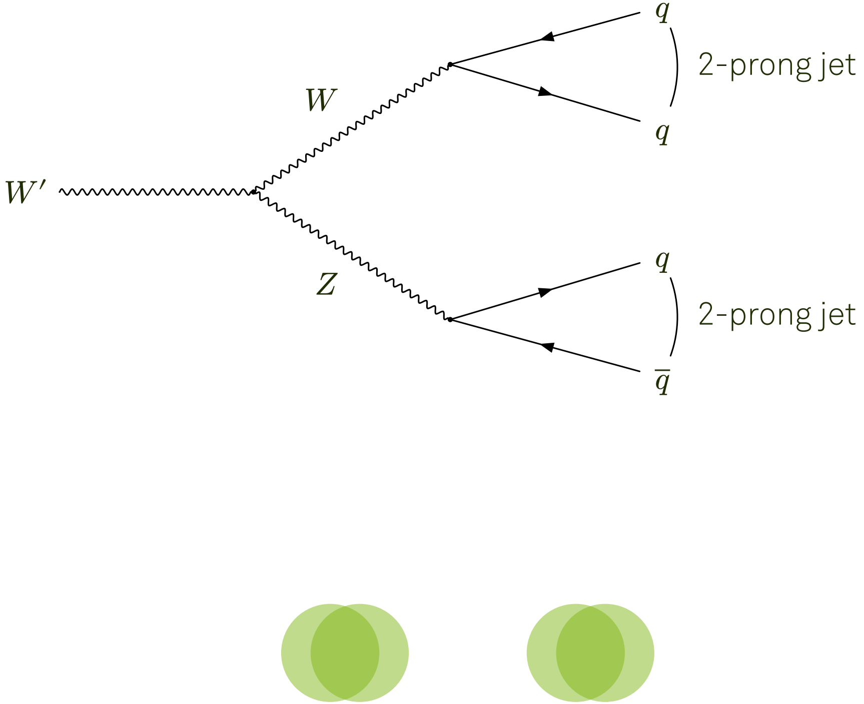
[source]

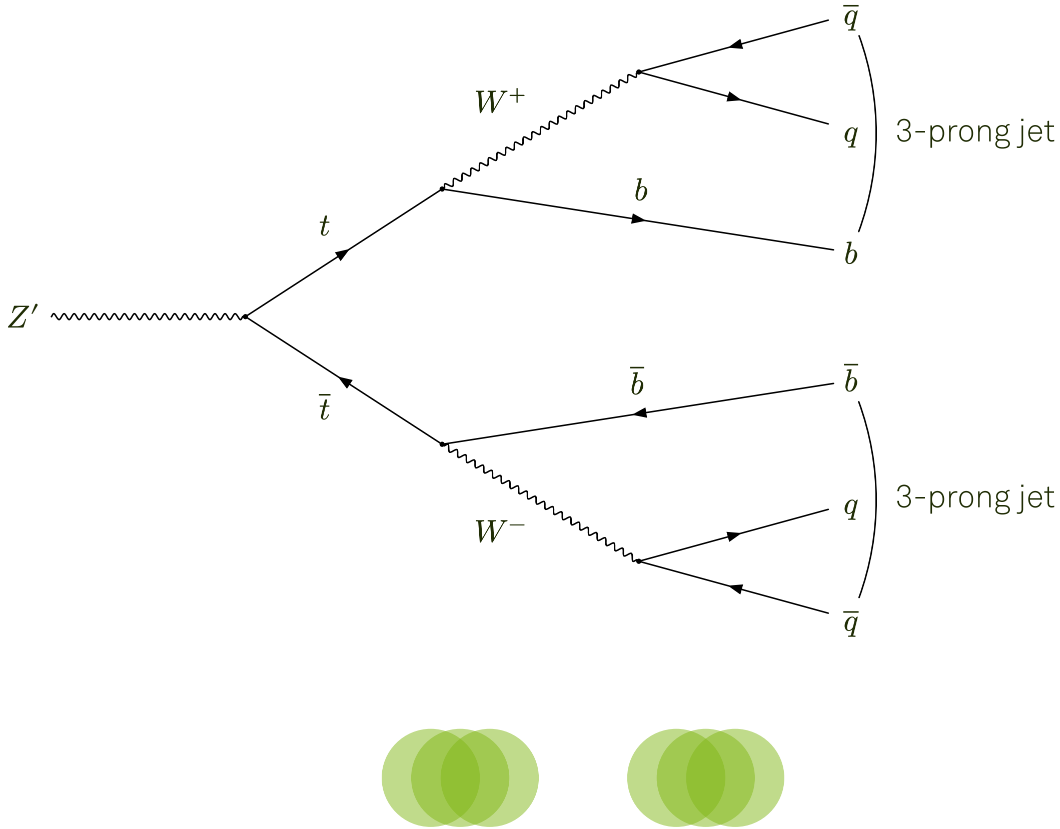
ROC Curves: Examples



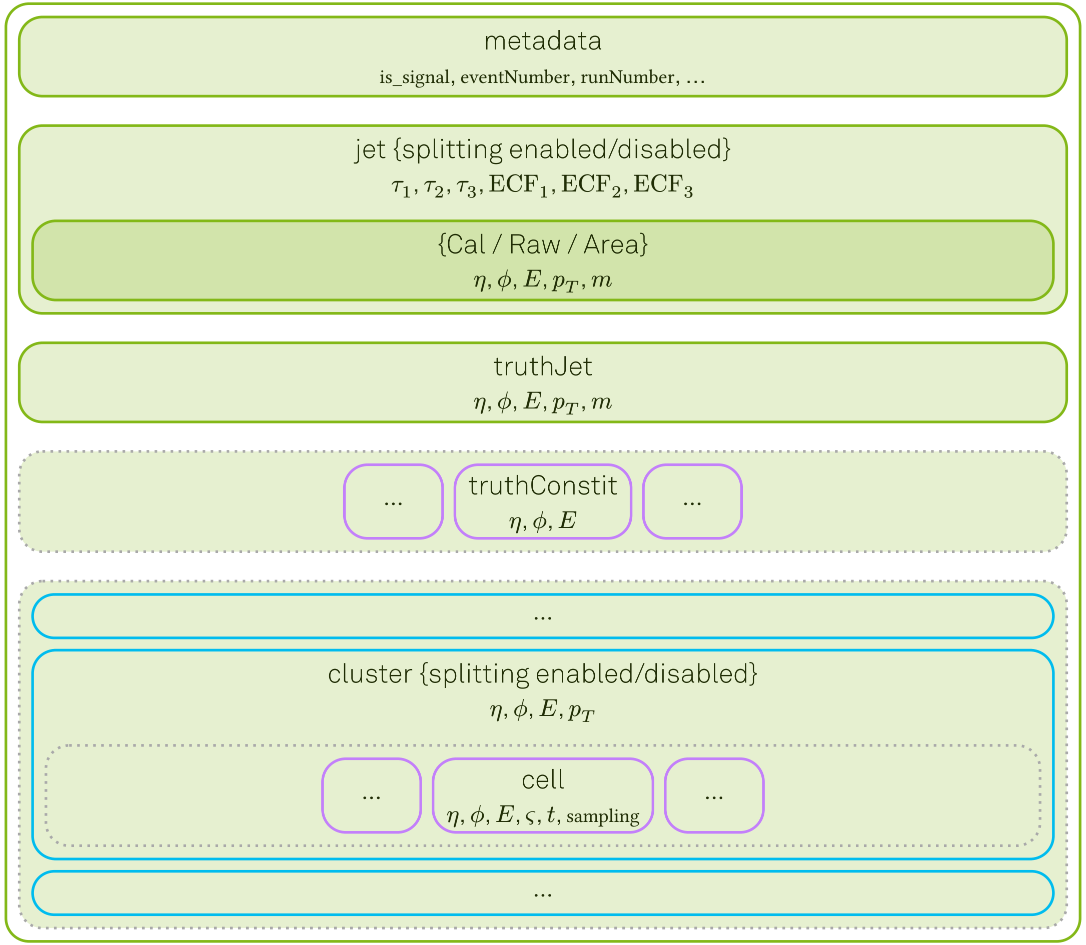
Feynman Diagrams and Prongs

- hypothetical heavy W', Z' + QCD dijets used in this analysis as a benchmark
- $m_{W'/Z'} = 2\,\mathrm{TeV}$



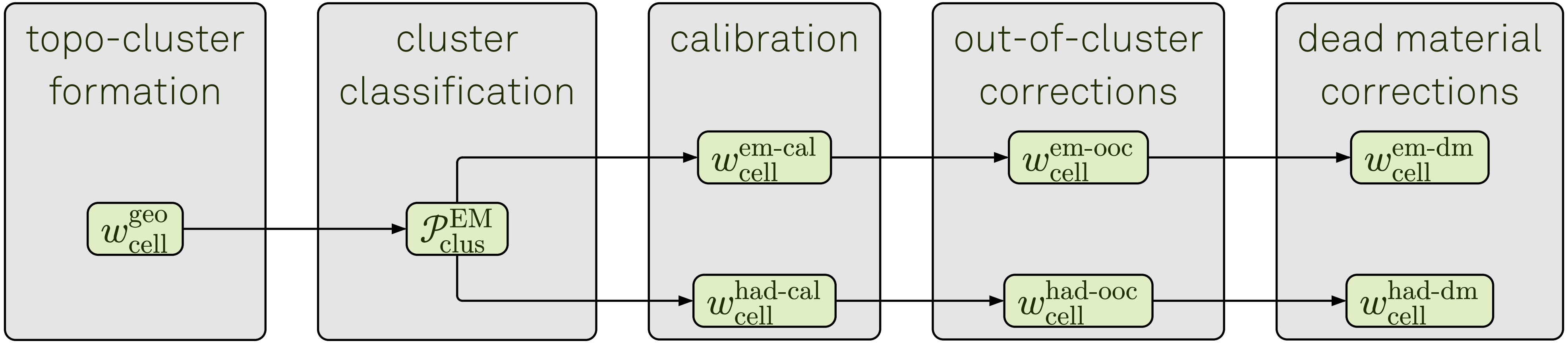


Data Structure

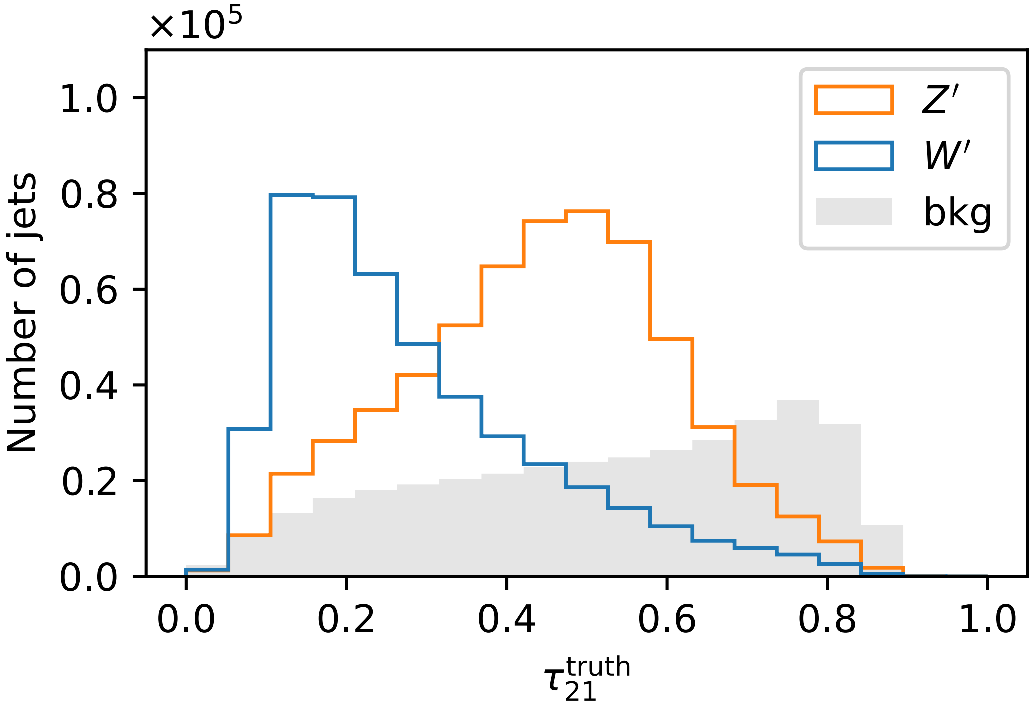


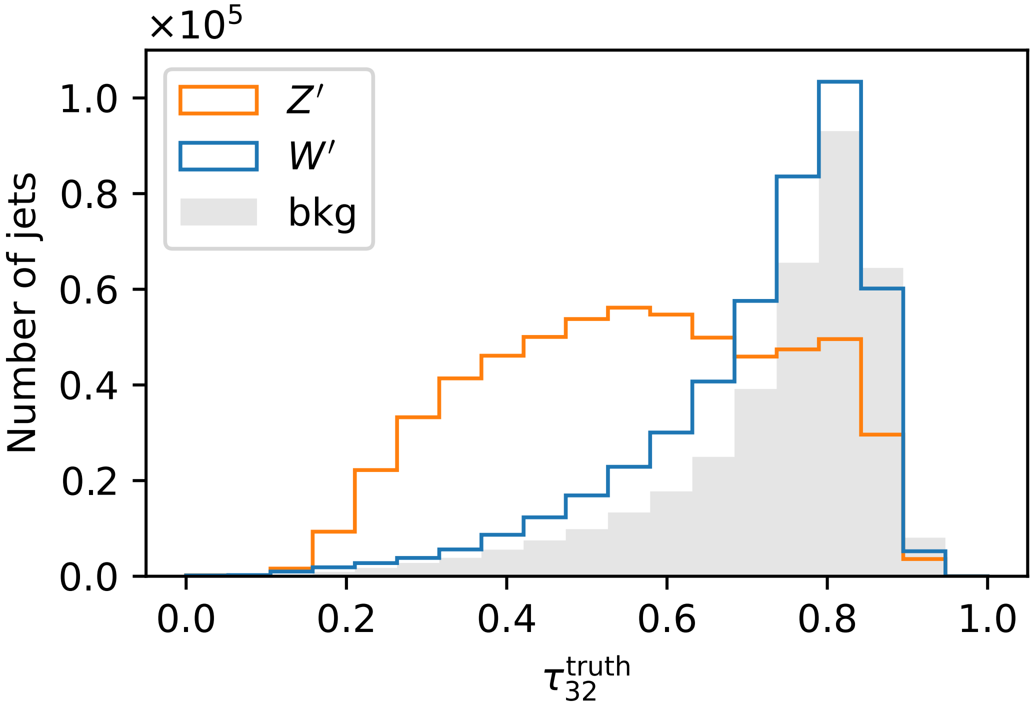
LCW

- Local Cell Weighting (LCW) calibration scheme applied to all clusters

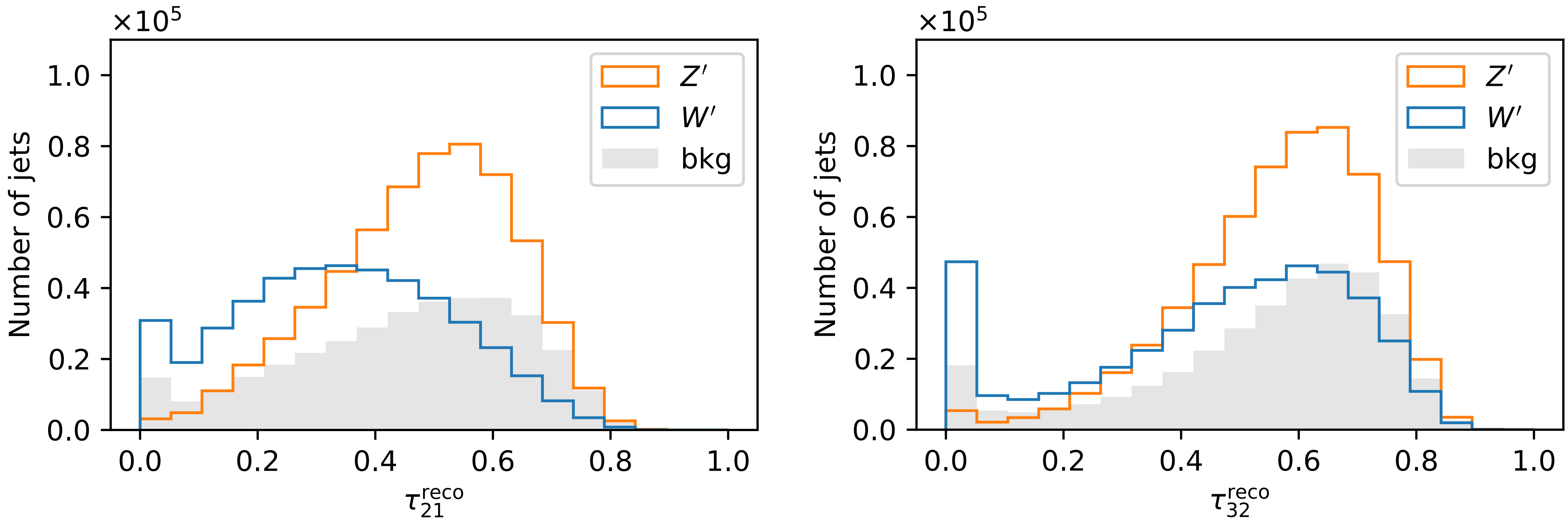


Underlying Distributions: truth

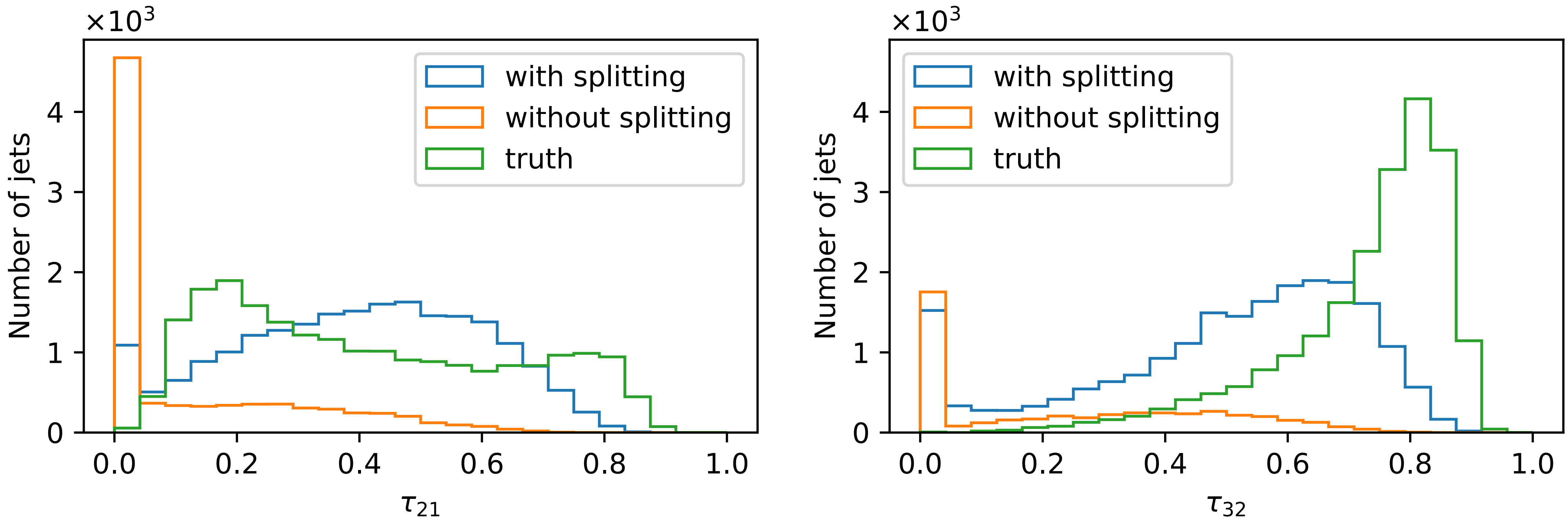




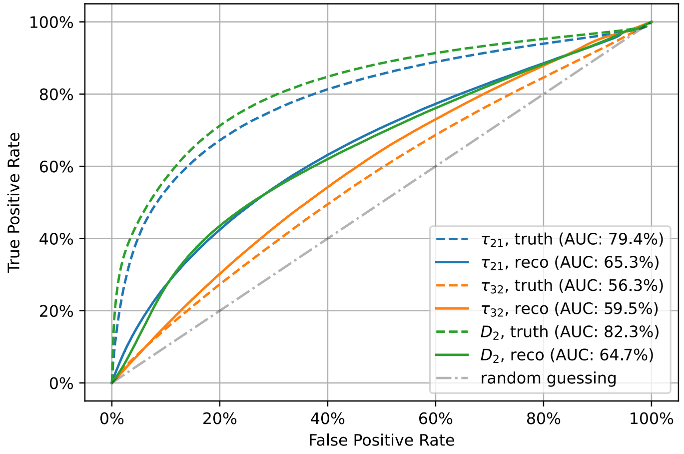
Underlying Distributions: reco



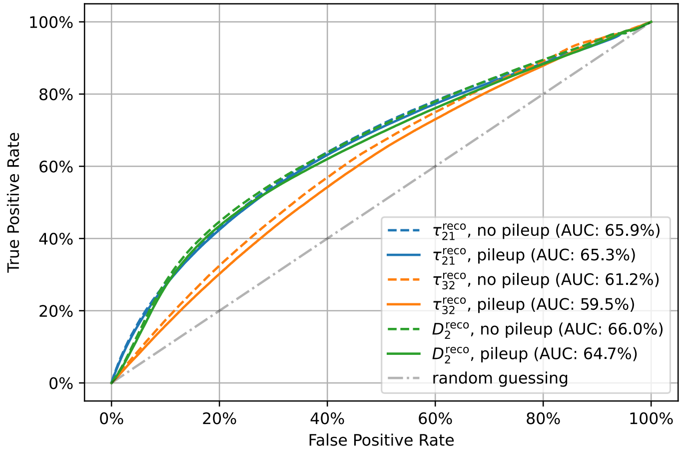
Underlying Distributions: splitting



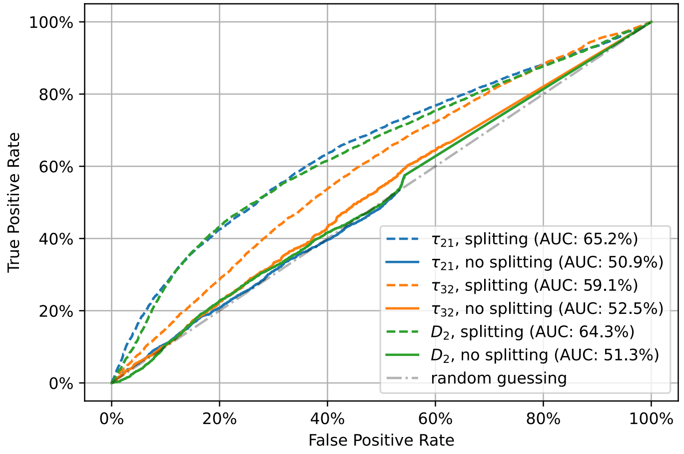
ROC Curves



truth vs. reco

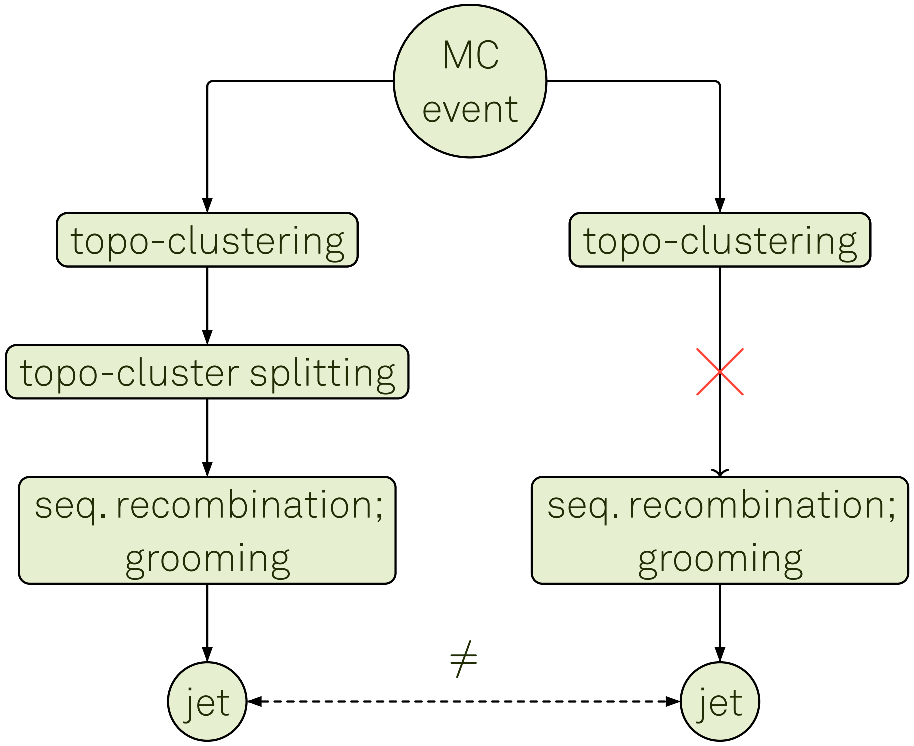


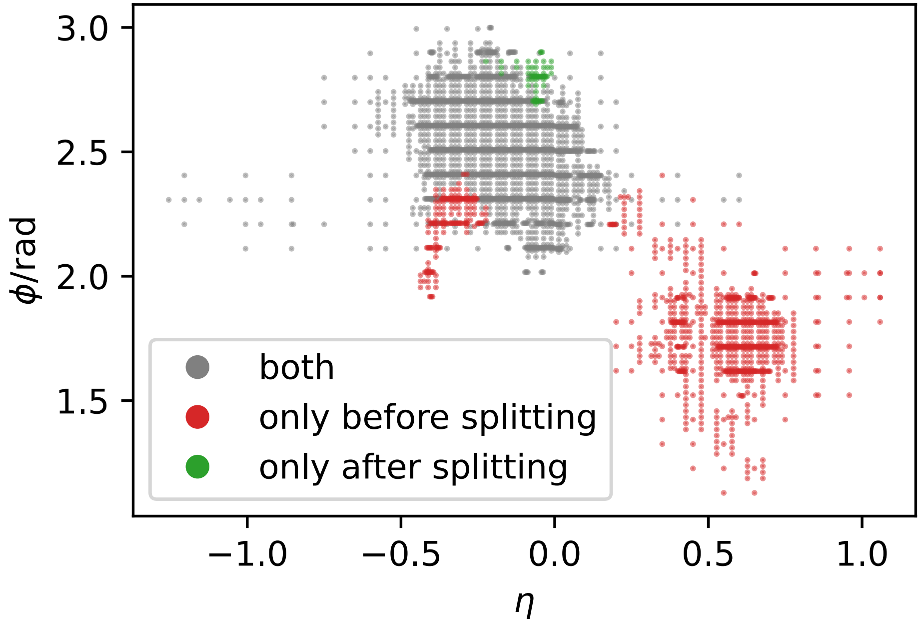
pile-up vs. no pile-up



splitting vs. no splitting

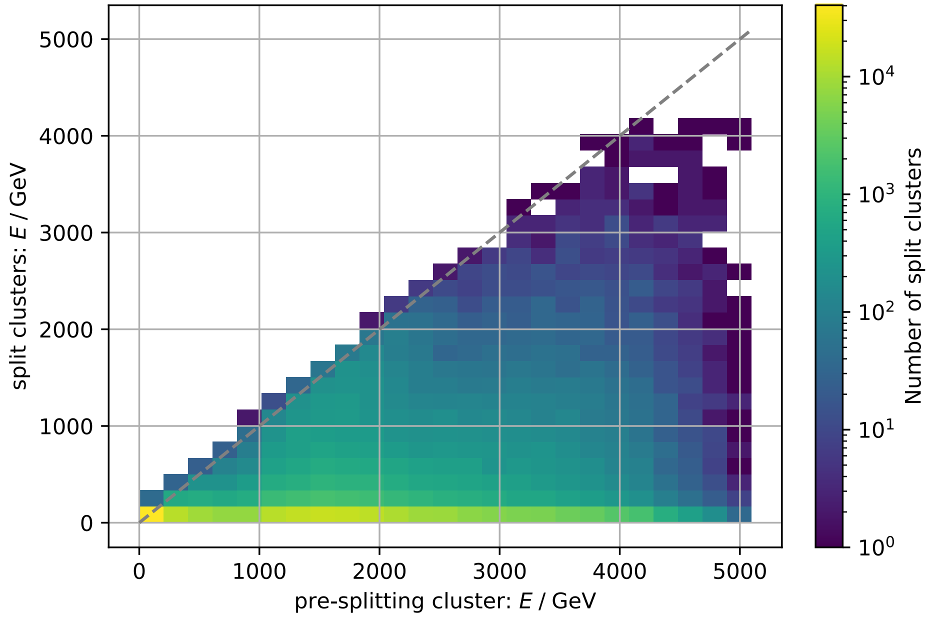
Differences due to Grooming





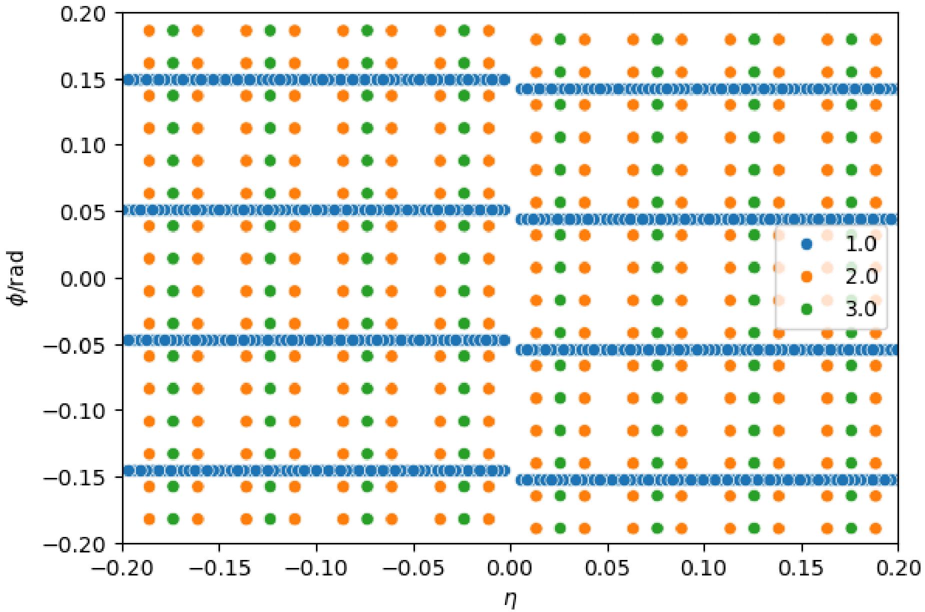
Matching Results: Cluster Energy

- concept: for each pre-splitting value, histogrammize all post-splitting values
- E: cluster energy
- always below diagonal: splitting reduces energy per cluster
- high counts at the bottom: many low-energy clusters created



Pitch of Sampling Layers

- different pitches in depending on sampling layer (color)
- different pitches in η and ϕ
- "η strips" for $\mathrm{sampling} = 1$
- "seam" at $\eta = 0$



EMDs: Split / Non-Split Clusters

variable	unit	EMD	mean ($m_i = 1$)	mean ($m_i \geq 2$)
fracE		1.885	0.013	0.924
fracE_ref		1.882	0.011	0.779
SECOND_LAMBDA	mm^2	1.552	16 065.308	452 754.603
ENG_POS	GeV	1.526	16 259.747	1 744 003.767
sumCellE	GeV	1.526	16.231	1743.671
ENG_CALIB_TOT	GeV	1.525	14.446	1798.484
E	GeV	1.524	16.032	1733.561
CELL_SIGNIFICANCE		1.523	28.925	1278.453
SIGNIFICANCE		1.521	15.880	433.215
Pt	GeV	1.473	13.230	1340.184

EMDs: Matching Truth Constituents

variable	unit	EMD	mean ($N_{\text{matching}} \leq 5$)	mean ($N_{\text{matching}} \geq 20$)
ENG_CALIB_OUT_T	GeV	2.360	0.479	1.449
ENG_CALIB_OUT_L	GeV	1.092	0.593	2.926
OOC_WEIGHT		1.079	1.347	1.041
SECOND_LAMBDA	mm²	1.070	17 920.171	175 924.693
ENG_CALIB_OUT_M	GeV	1.054	0.505	2.275
CELL_SIG_SAMPLING		1.041	1.579	6.936
CENTER_LAMBDA	mm	0.972	211.208	802.358
ISOLATION		0.934	0.552	0.304
LATERAL		0.865	0.617	0.862
AVG_TILE_Q		0.851	0.319	14.113

Grid Search Results: W'

- W' / dijet background discrimination
- τ_{21} and D_2 given for reference

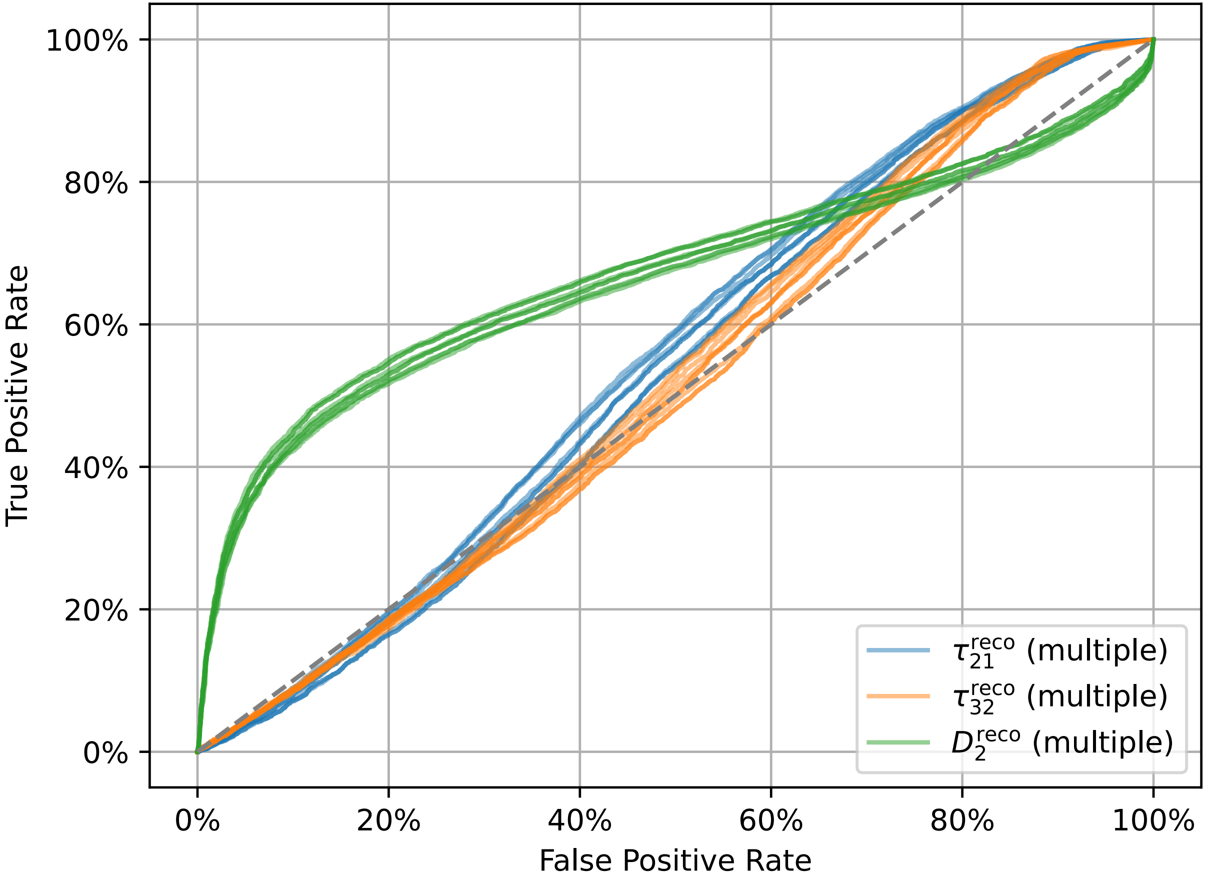
$E_{\mathrm{thresh}}/\ \mathrm{MeV}$	$\mathrm{NN}_{\mathrm{thresh}}$	$\mathrm{AUC}(\tau_{21}^{\mathrm{reco}})$	$\mathrm{AUC}(\tau_{32}^{\mathrm{reco}})$	$\mathrm{AUC}(D_2^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(\tau_{21}^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(\tau_{32}^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(D_2^{\mathrm{reco}})$	$\langle\Delta\mathrm{AUC}\rangle$
550	4	0.646	0.589	0.653	−0.001	−0.003	+0.008	0.001
550	5	0.646	0.589	0.653	−0.001	−0.004	+0.008	0.001
550	3	0.644	0.589	0.654	−0.003	−0.003	+0.009	0.000
500	4	0.647	0.593	0.644	+0.000	+0.000	+0.000	0.000
500	3	0.649	0.584	0.649	+0.002	−0.008	+0.004	−0.000
500	5	0.649	0.584	0.649	+0.002	−0.008	+0.004	−0.000
450	4	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002
450	3	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002
450	5	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002

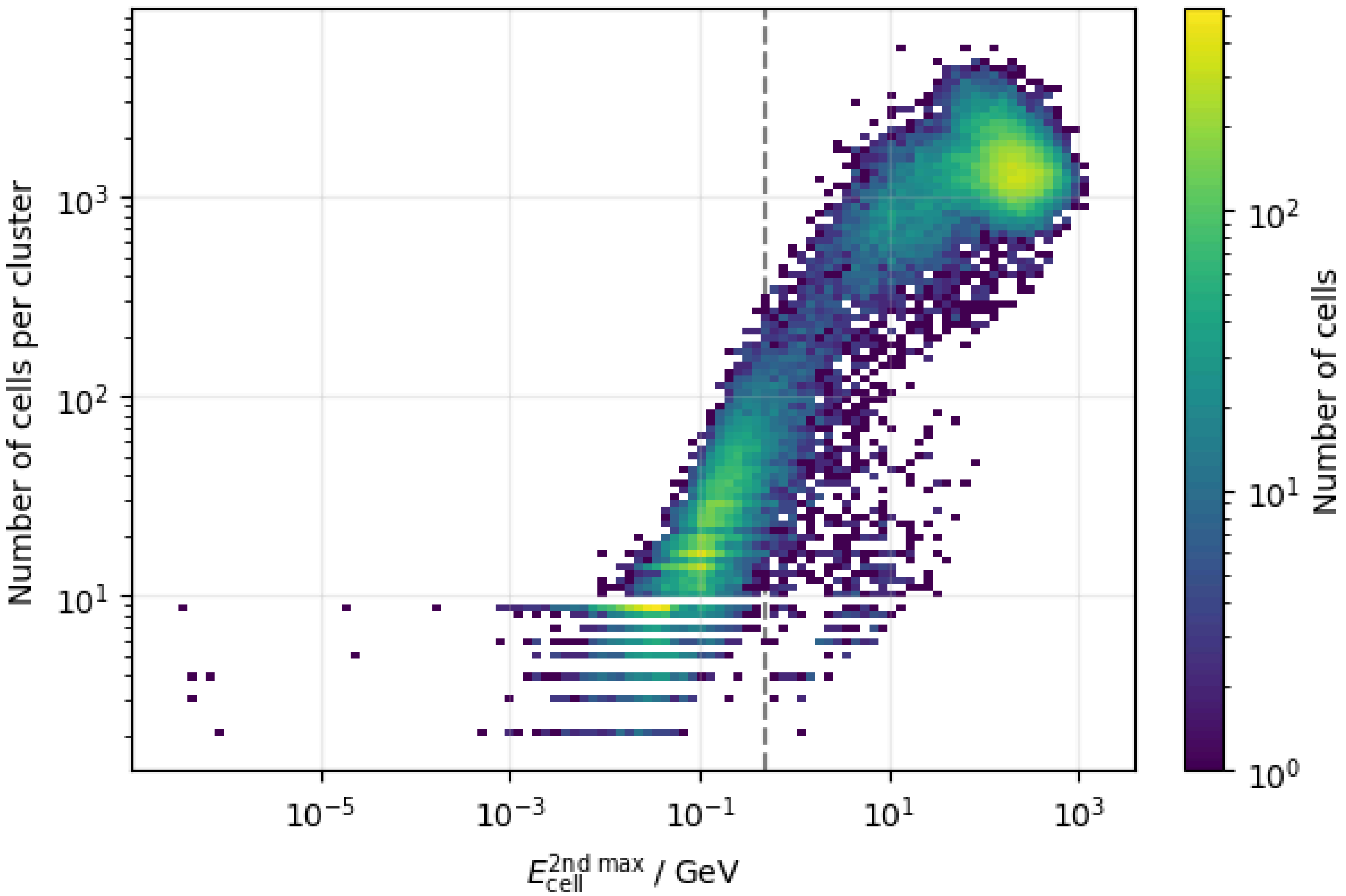
Grid Search Results: Z'

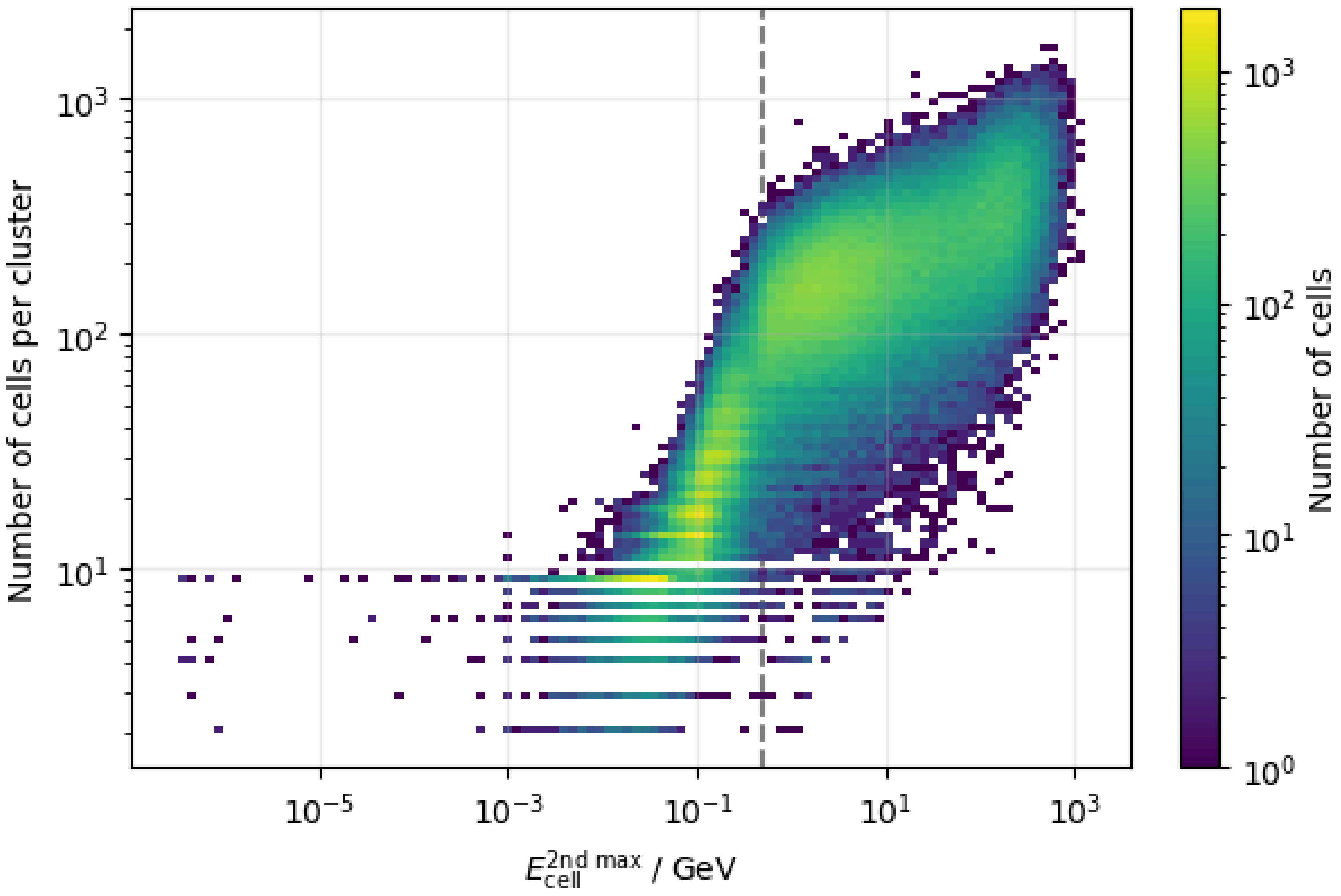
- Z' / dijet background discrimination
- τ_{21} and D_2 given for reference

$E_{\mathrm{thresh}}/$ MeV	$\mathrm{NN}_{\mathrm{thresh}}$	$\mathrm{AUC}(\tau_{21}^{\mathrm{reco}})$	$\mathrm{AUC}(\tau_{32}^{\mathrm{reco}})$	$\mathrm{AUC}(D_2^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(\tau_{21}^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(\tau_{32}^{\mathrm{reco}})$	$\Delta\mathrm{AUC}(D_2^{\mathrm{reco}})$	$\langle\Delta\mathrm{AUC}\rangle$
550	4	0.557	0.531	0.656	+0.014	+0.012	−0.010	0.005
550	5	0.552	0.528	0.657	+0.009	+0.009	−0.009	0.003
550	3	0.555	0.523	0.654	+0.012	+0.004	−0.012	0.001
500	4	0.542	0.519	0.667	+0.000	+0.000	+0.000	0.000
500	5	0.541	0.517	0.666	−0.001	−0.001	−0.000	−0.001
500	3	0.541	0.517	0.666	−0.001	−0.001	−0.001	−0.001
450	4	0.526	0.507	0.680	−0.016	−0.011	+0.013	−0.004
450	5	0.524	0.504	0.678	−0.018	−0.014	+0.011	−0.007
450	3	0.524	0.504	0.678	−0.018	−0.014	+0.011	−0.007

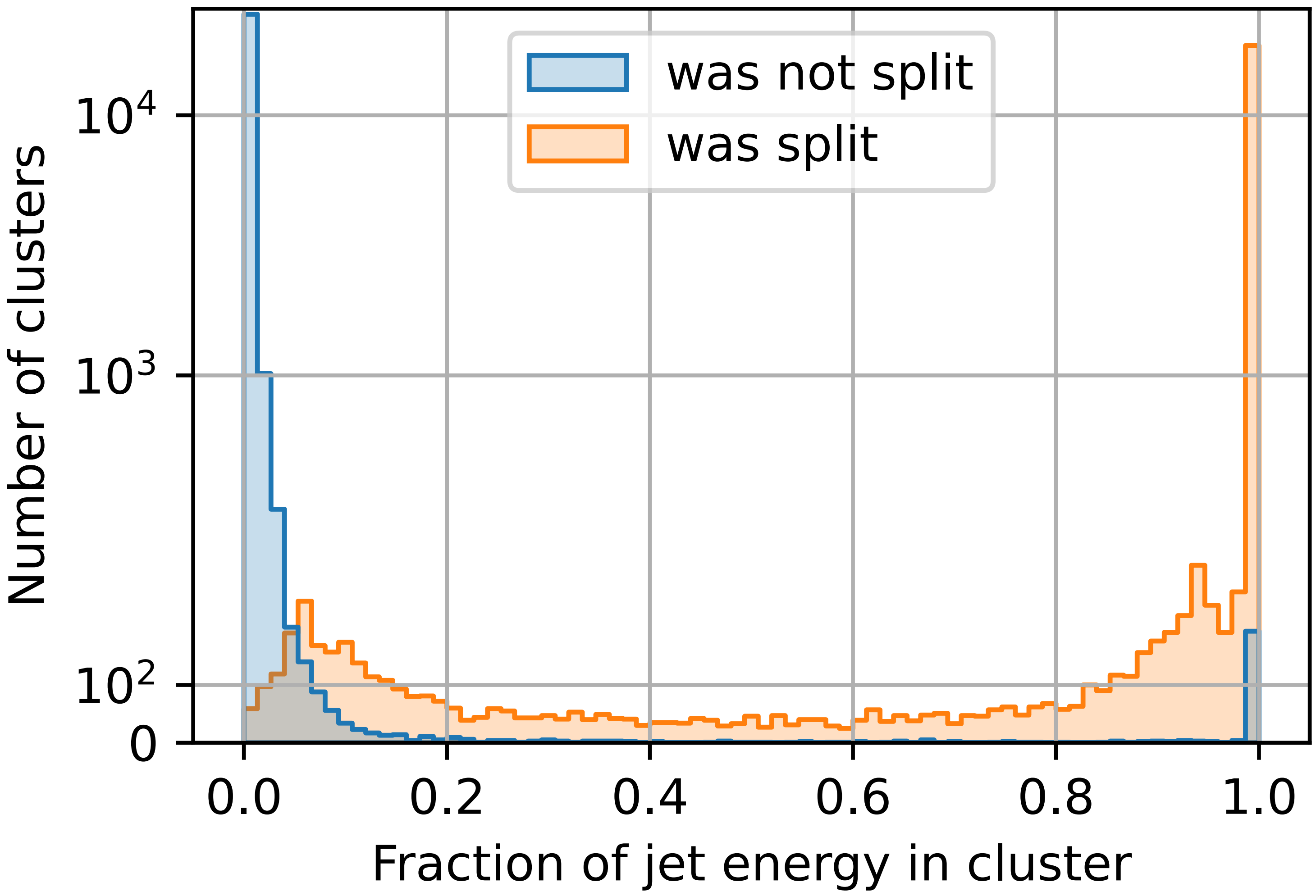
Various Other Plots

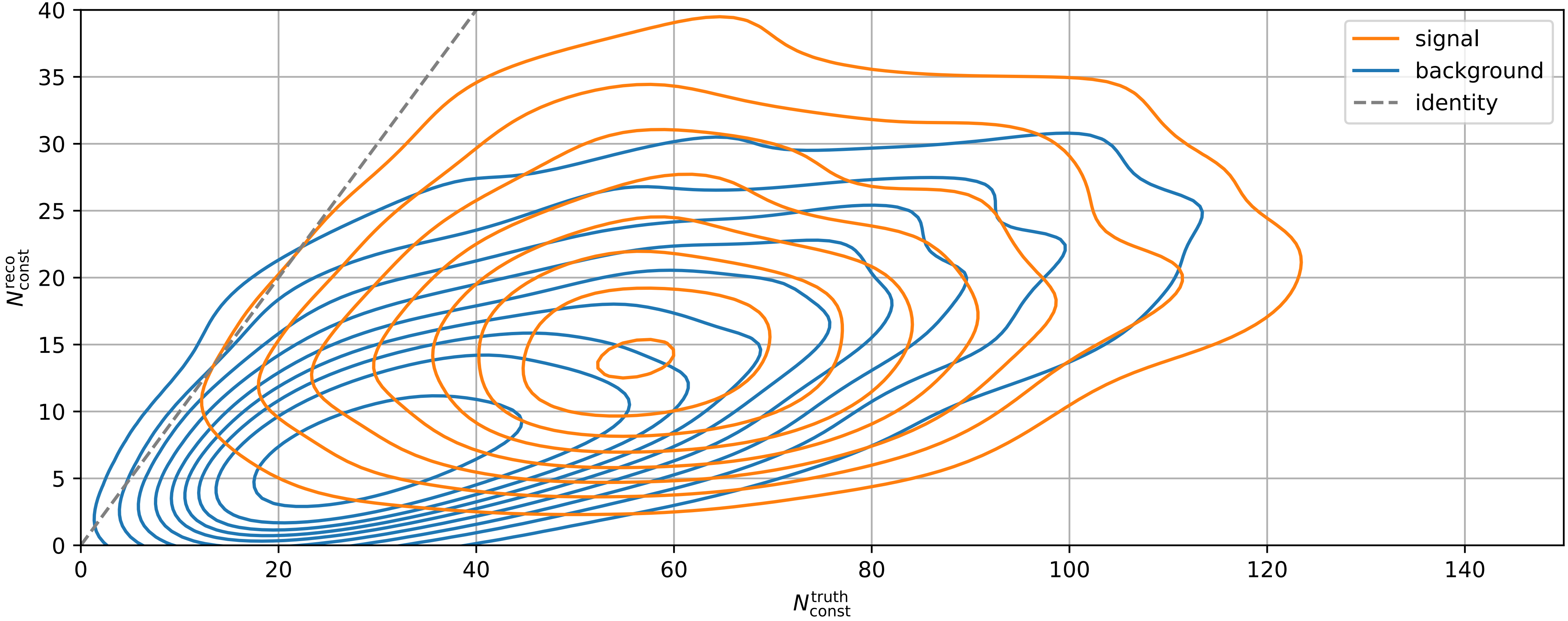



Number of cells per cluster
10^3
10^2
10^1
10^{-5}
10^{-3}
10^{-1}
10^1
10^3
$E_{\text{cell}}^{\text{2nd max}}$ / GeV
10^2
10^1
10^0
Number of cells



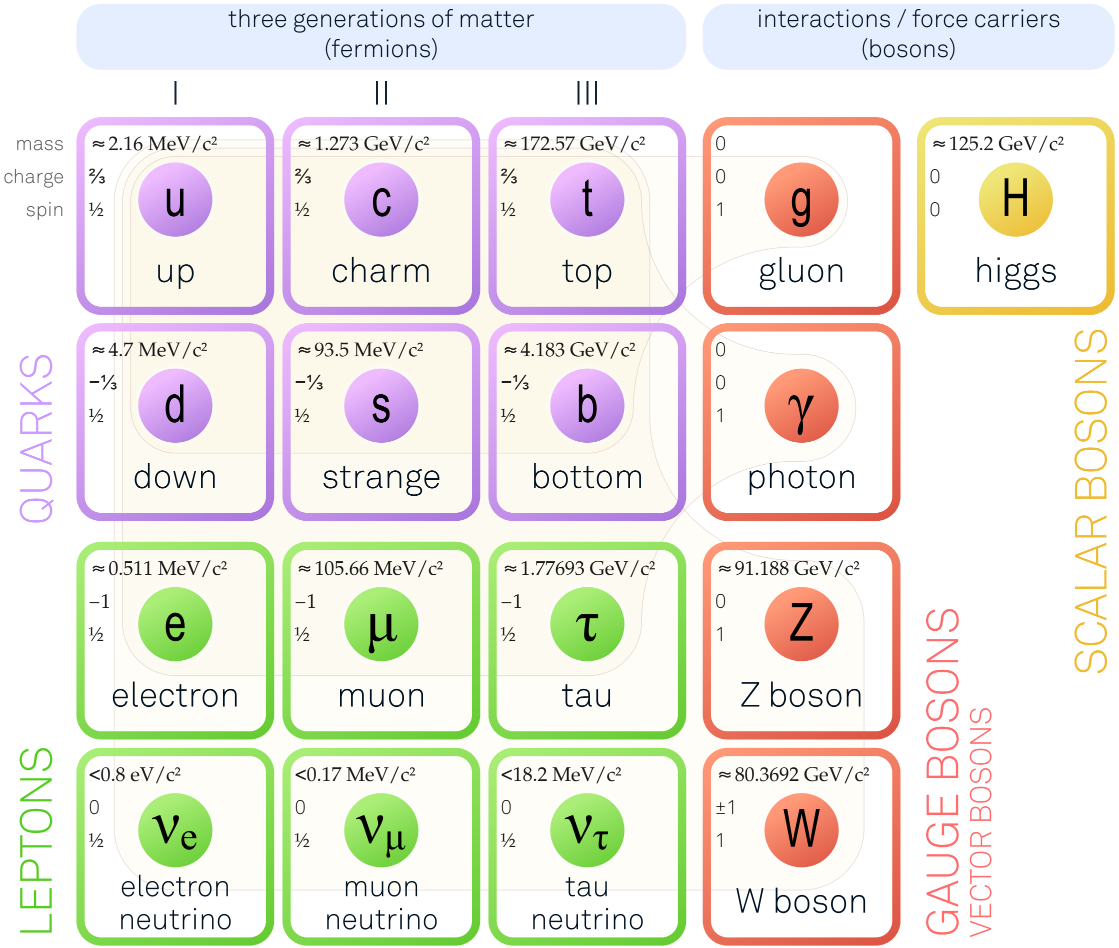
Number of cells per cluster
10^3
10^2
10^1
10^{-5}
10^{-3}
10^{-1}
10^1
10^3
$E_{\text{cell}}^{\text{2nd max}}$ / GeV
10^3
10^2
10^1
10^0
Number of cells


was not split
was split
Number of clusters
10^4
10^3
10^2
0
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of jet energy in cluster



signal
background
identity
$N_{\mathrm{const}}^{\mathrm{reco}}$
$N_{\mathrm{const}}^{\mathrm{truth}}$
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
120
140

The Standard Model



Updated Slides

Link 🔗