WMLQ 2024 Conference
07.06.2024

NATIONAL CENTRE FOR NUCLEAR RESEARCH ŚWIERK

ASTROCENT
Particle Astrophysics Science and Technology Centre
International Research Agenda

CEAI
Center of Excellence in Artificial Intelligence

Reconstruction of muon bundles in KM3NeT detectors using machine learning methods

Piotr Kalaczyński on behalf of the KM3NeT Collaboration

Work supported by:

NATIONAL SCIENCE CENTRE POLAND
grant numbers: 2021/41/N/ST2/01177
2015/18/E/ST2/00758

We gratefully acknowledge the funding support by program “Excellence initiative—research university” for the AGH University in Krakow as well as the ARTIQ project: UMO-2021/01/2/ST6/00004 and ARTIQ/0004/2021.

European Funds Smart Growth

Republic of Poland

FNP Foundation for Polish Science

European Union
European Regional Development Fund

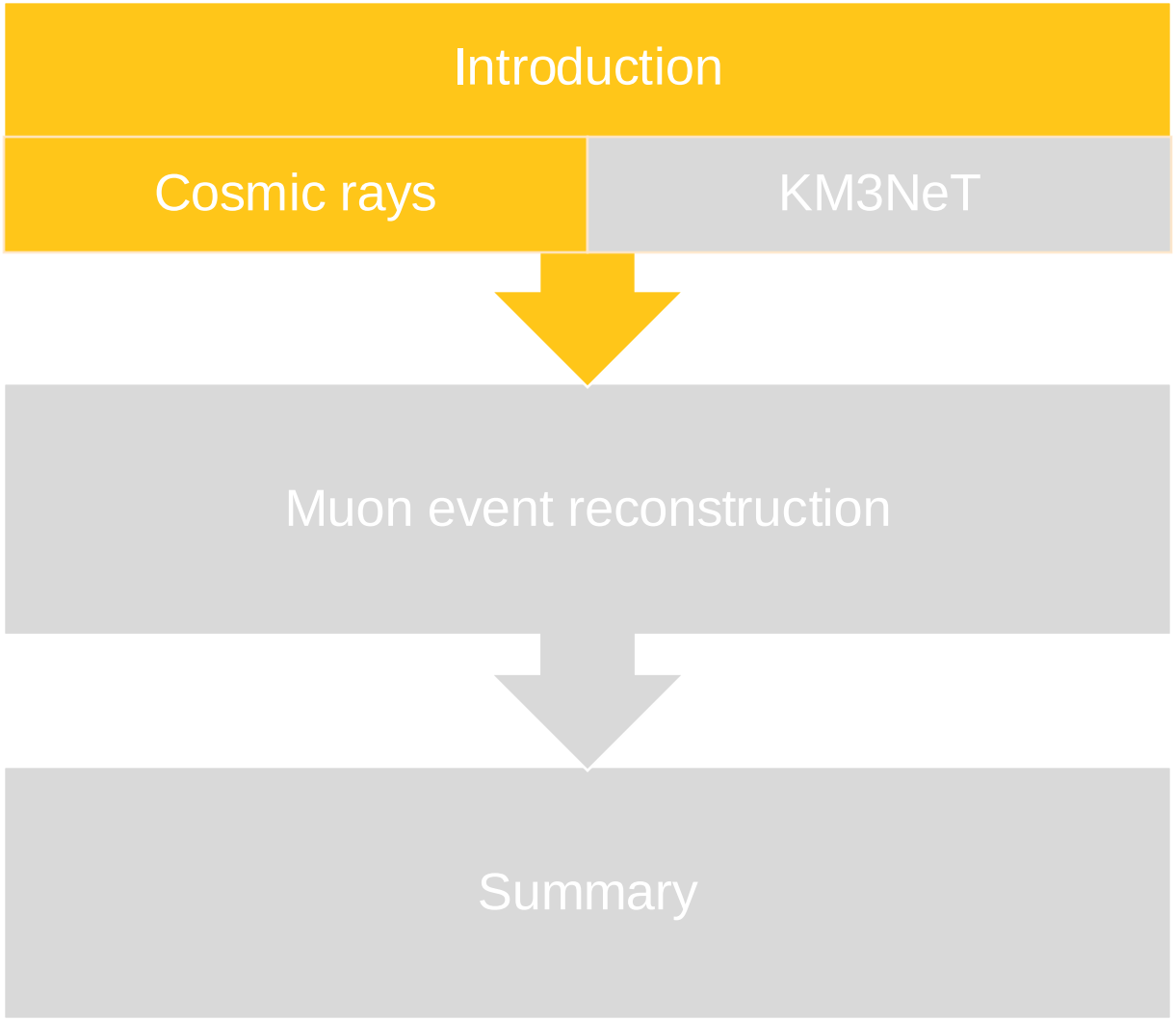
Outline

Introduction

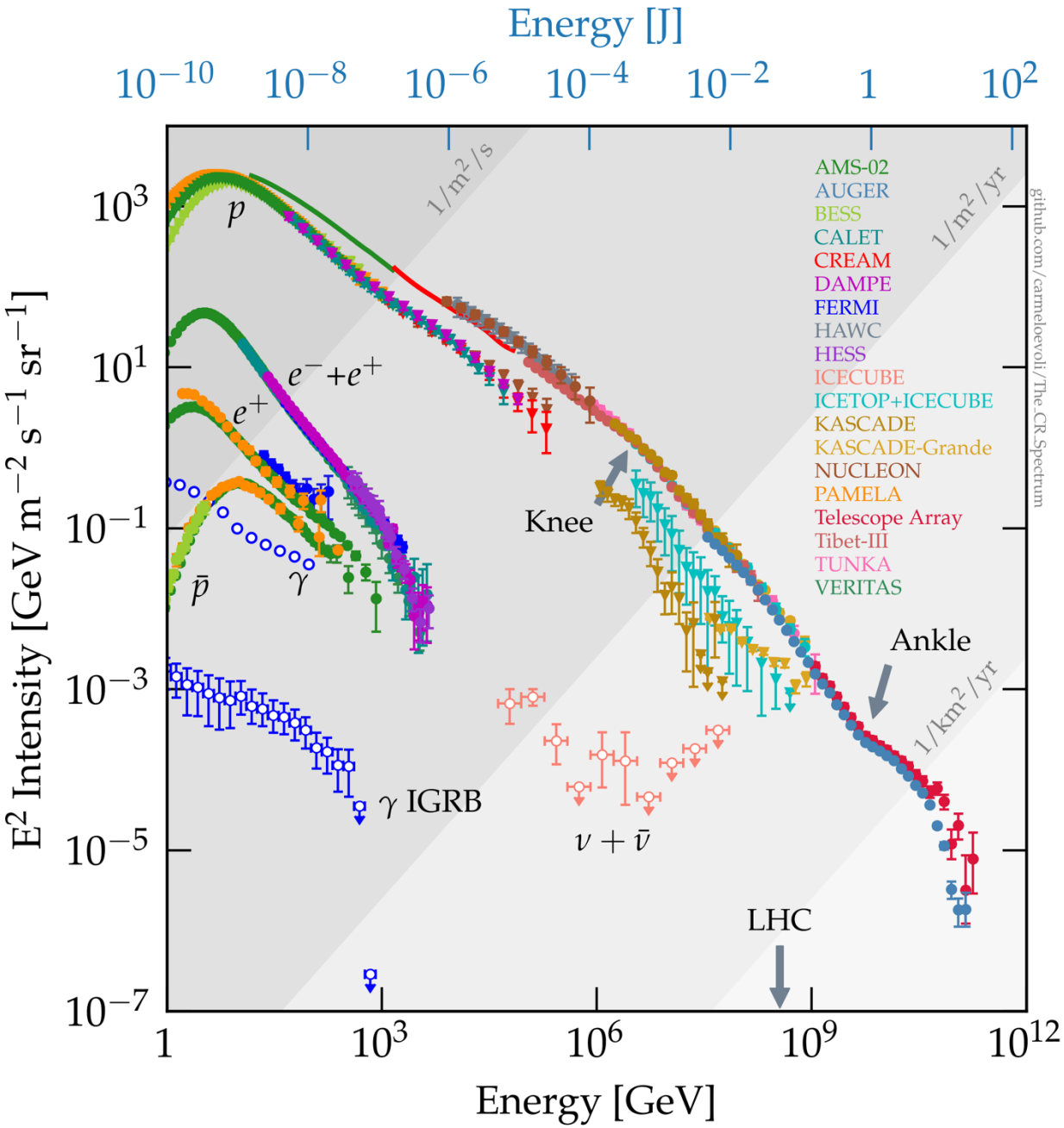
Muon event reconstruction

Summary

Outline


Introduction
Cosmic rays
KM3NeT
Muon event reconstruction
Summary

Cosmic rays



Cosmic rays (CR):

High-Energy particles and atomic nuclei from outer space that reach the Earth

Discovered by Victor Hess in 1912
(Nobel Prize in 1936)

Quite a few measurements since then …

But wait, muons are not in the figure … !?

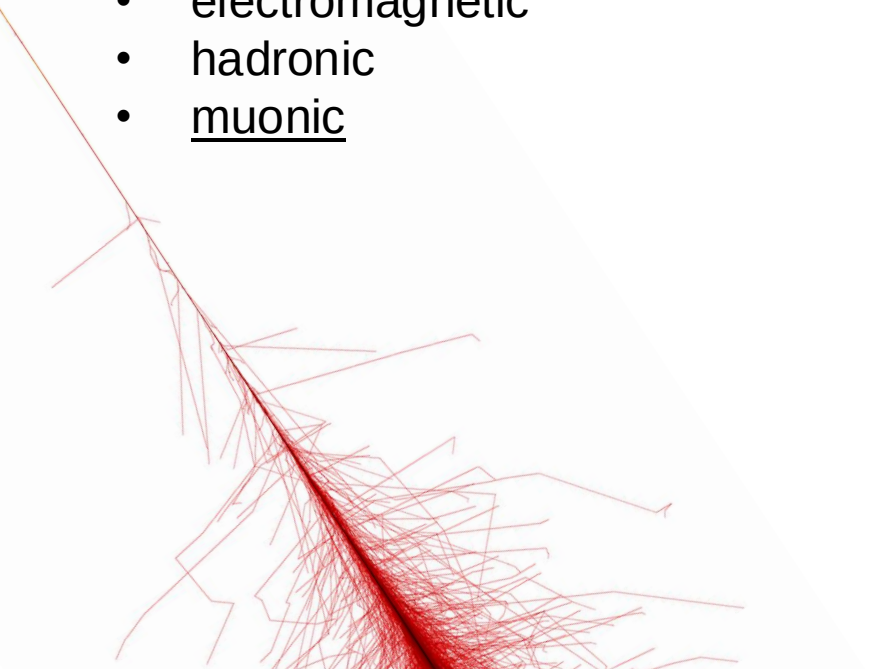
Extensive air showers (EAS)

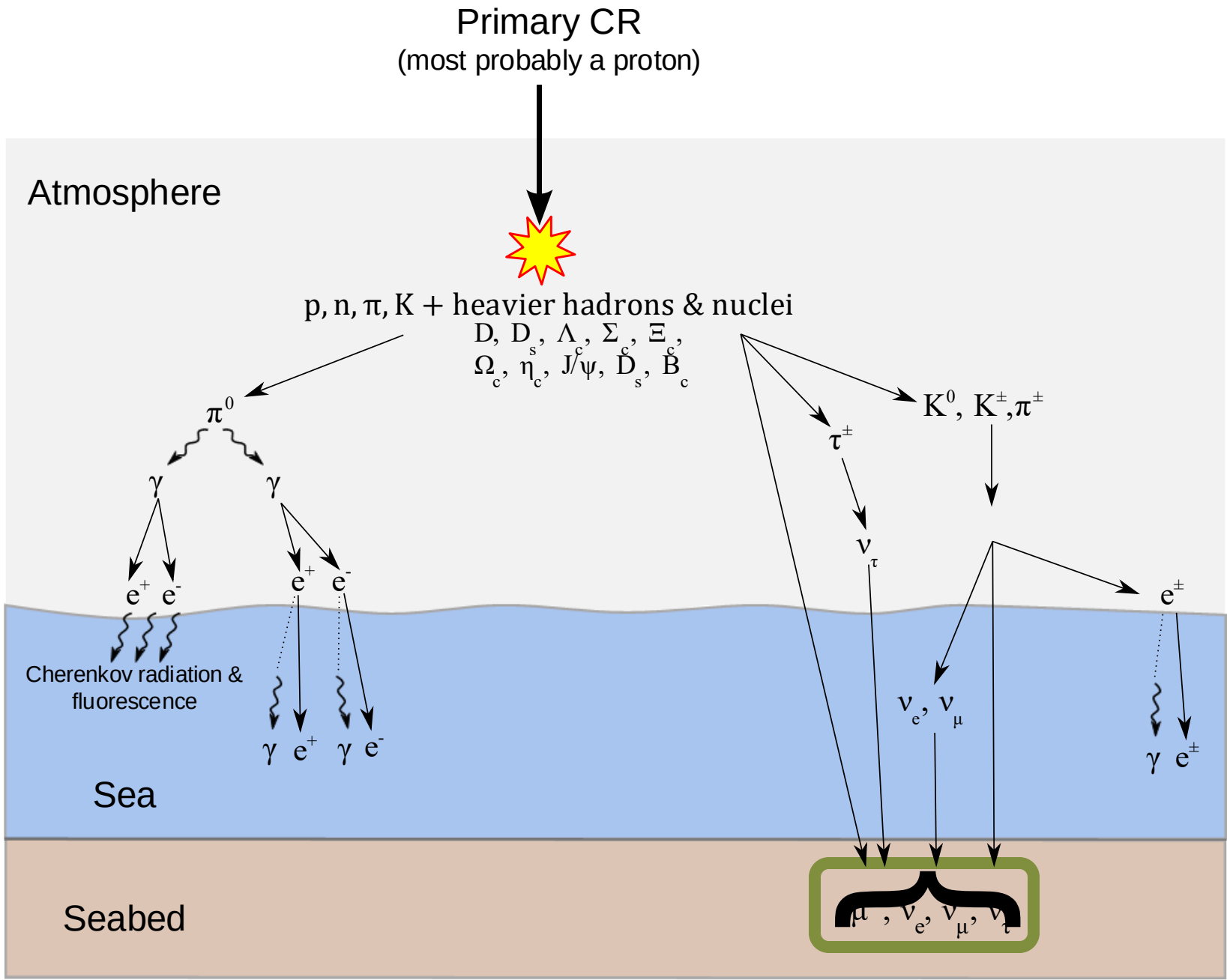
Muons are <u>not</u> primary CRs!

They are secondaries produced in …

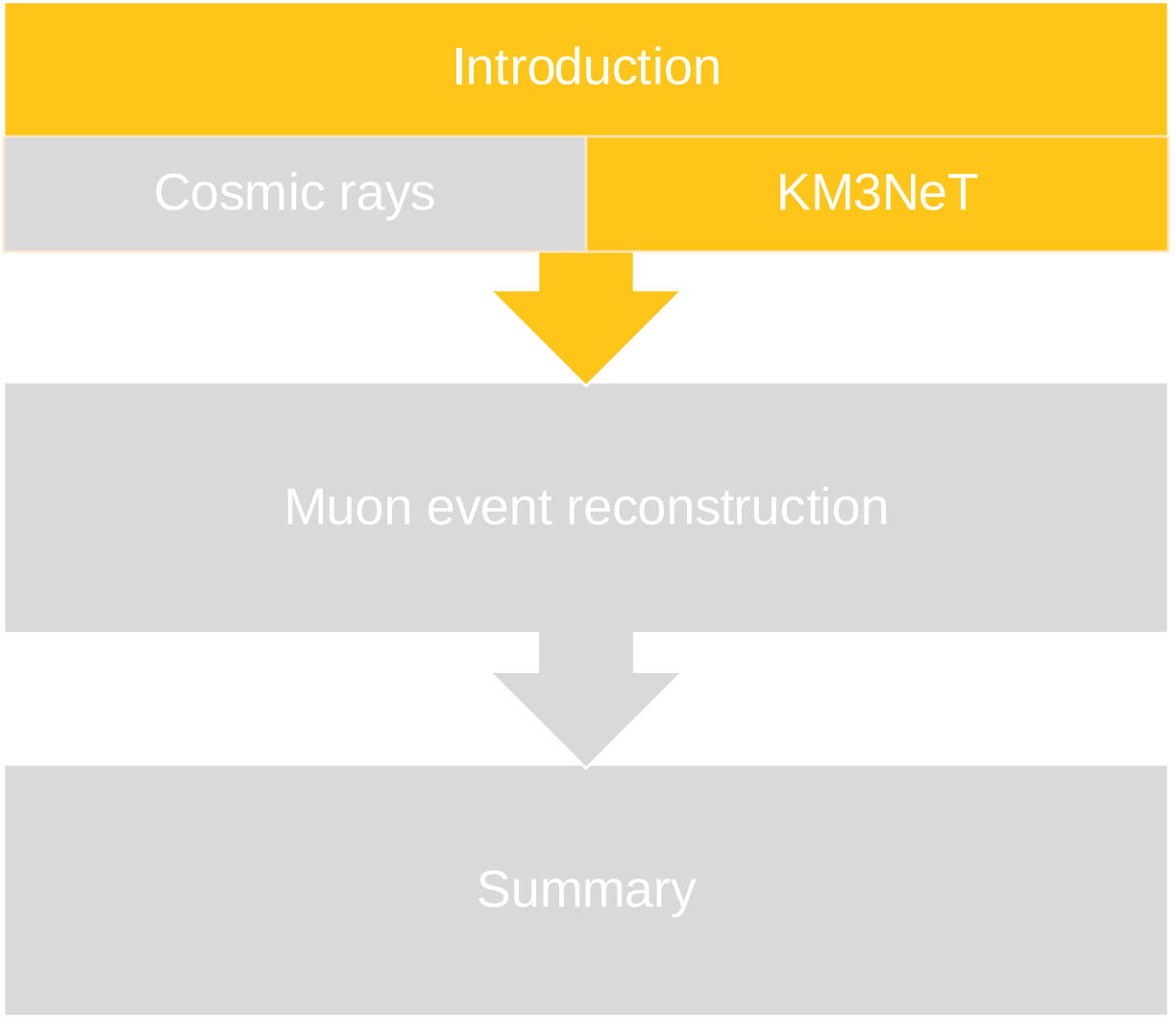
Extensive air showers (EAS):

- Caused by primary CR
- Most start at $h \sim 30 - 40\,\mathrm{km}$
- 3 components:
  - electromagnetic
  - hadronic
  - <u>muonic</u>





Outline

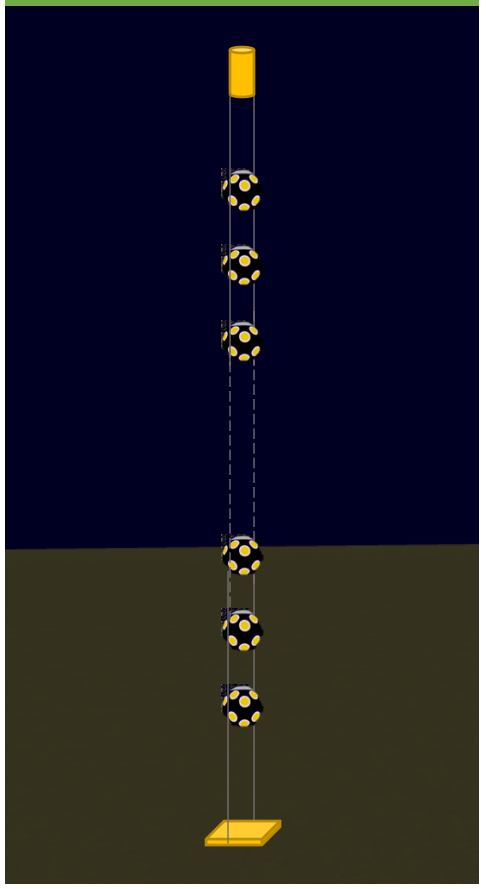

Introduction
Cosmic rays
KM3NeT
Muon event reconstruction
Summary

Detector design summary

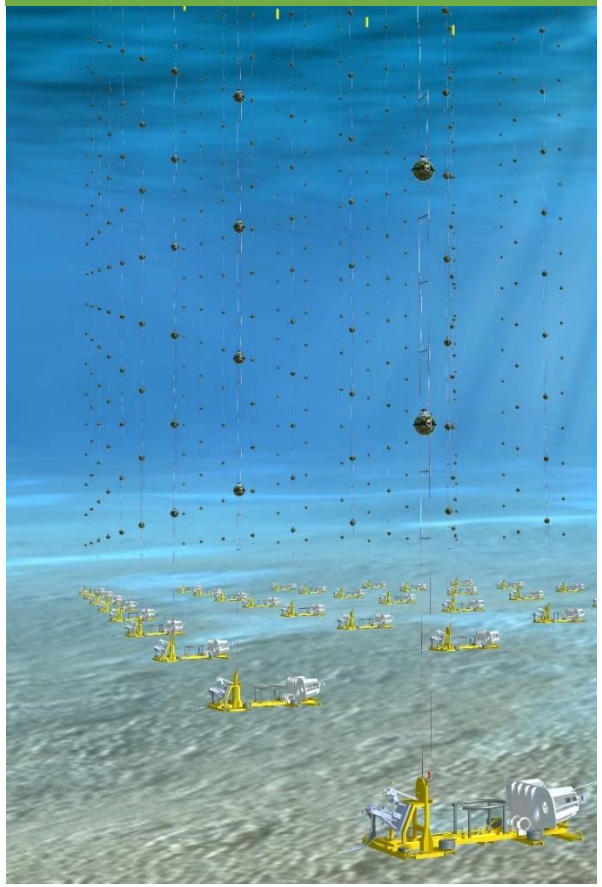
DOM:
71 unique components

1 DOM:
31 PMTs

1 string (DU):
18 DOMs



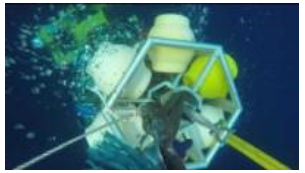
1 building block:
115 DUs



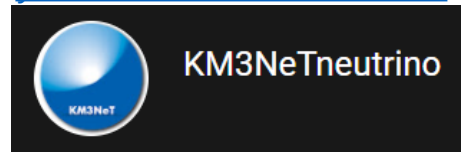
[DOM production:(@Nikhef)]

[Preparation for deployment:]

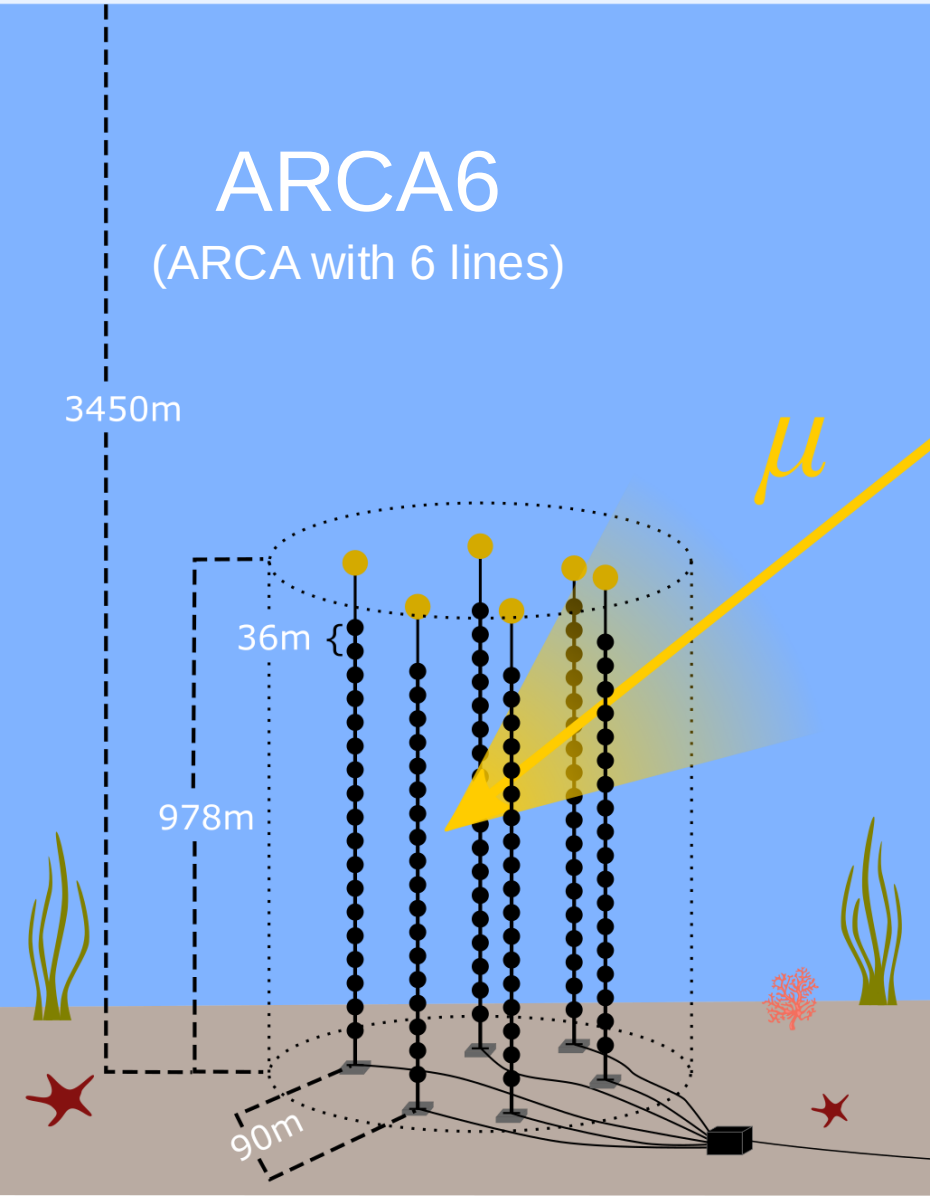
[String deployment:]



More at:
youtube.com/KM3NeTneutrino

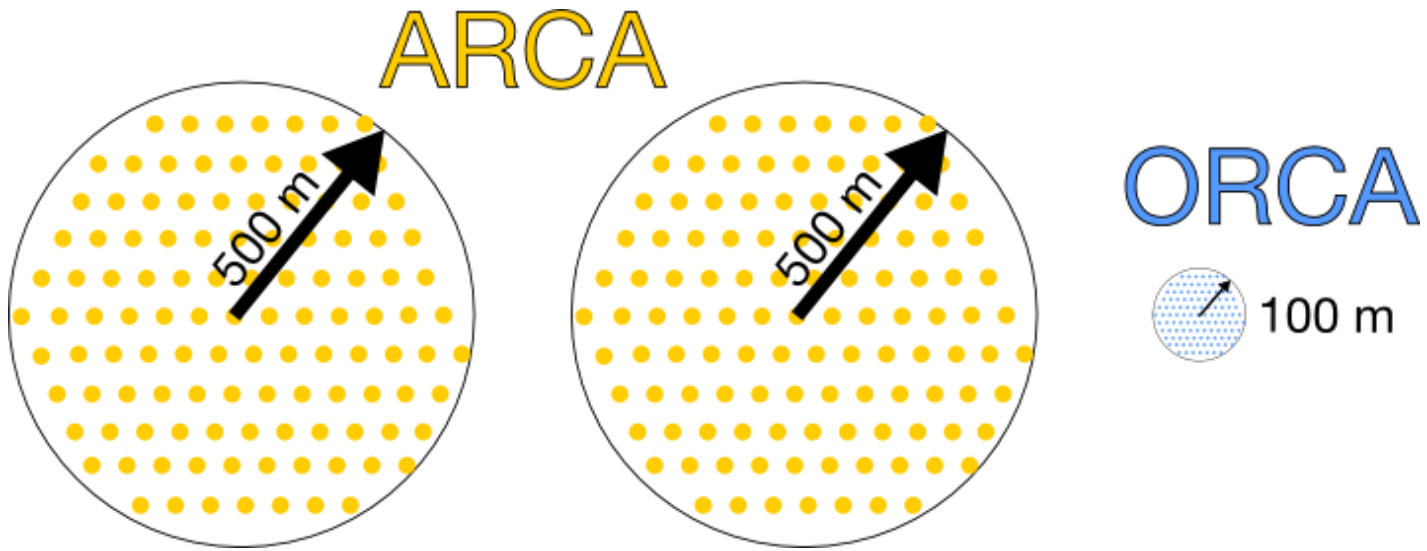


KM3NeT detectors: brief summary



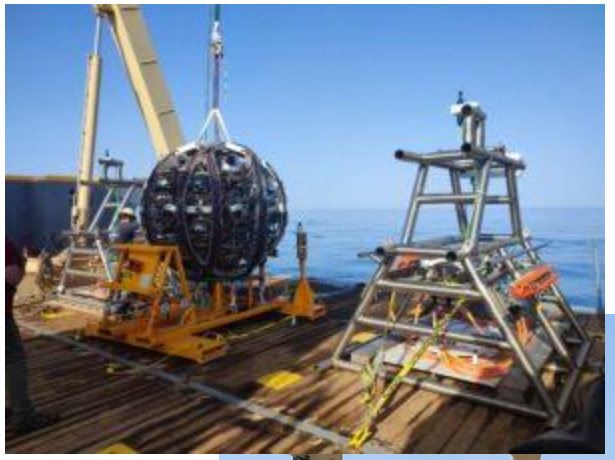
Detector	ARCA	ORCA
Depth	3.5 km	2.45 km
Volume	$1\ \mathrm{km}^3$ (1Gton)	$0.007\ \mathrm{km}^3$ (7Mton)
# lines	28 / 2x115	18 / 115
Topic	Astroparticle RCA*	Oscillation RCA*
Goal	ν_{astro}	m_ν hierarchy

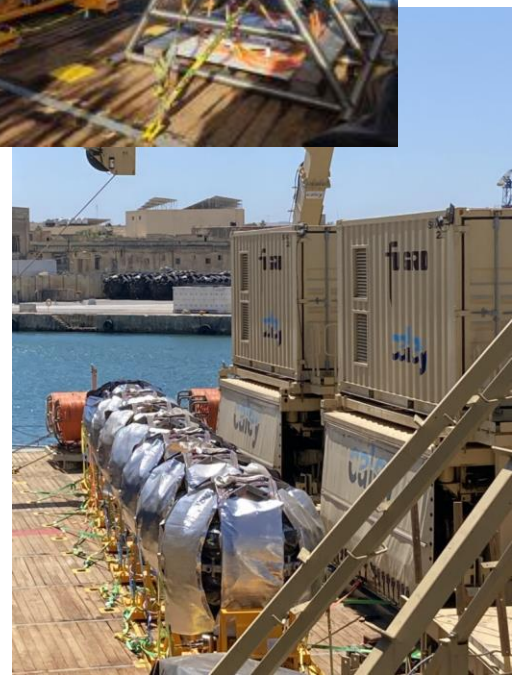
*RCA : Research with Cosmics in the Abyss



ARCA timeline

- Dec 2015: first strings
- … more details
- Sep 2023: +9 **strings**
- Currently: **ARCA28!**
- …
- Sep 2024: next deployment
- …
- Nov 2031(?): ARCA2x115





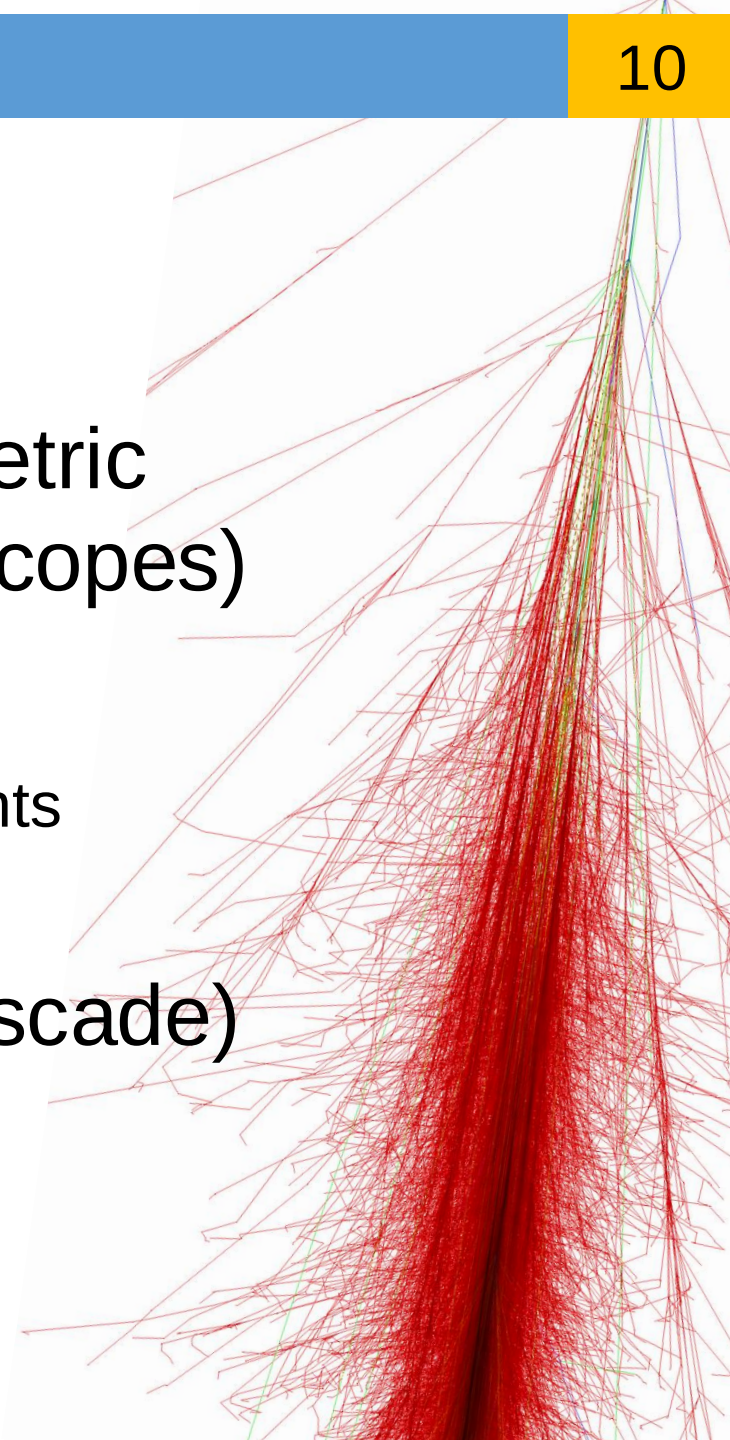
For more follow us at:
https://www.km3net.org

ORCA timeline

- Dec 2017: first strings
- … more details
- Jun 2024: +4 **strings**
- Currently: **ORCA22!**
- …
- Summer 2024: next deployment
- …
- Jan 2031(?): ORCA115

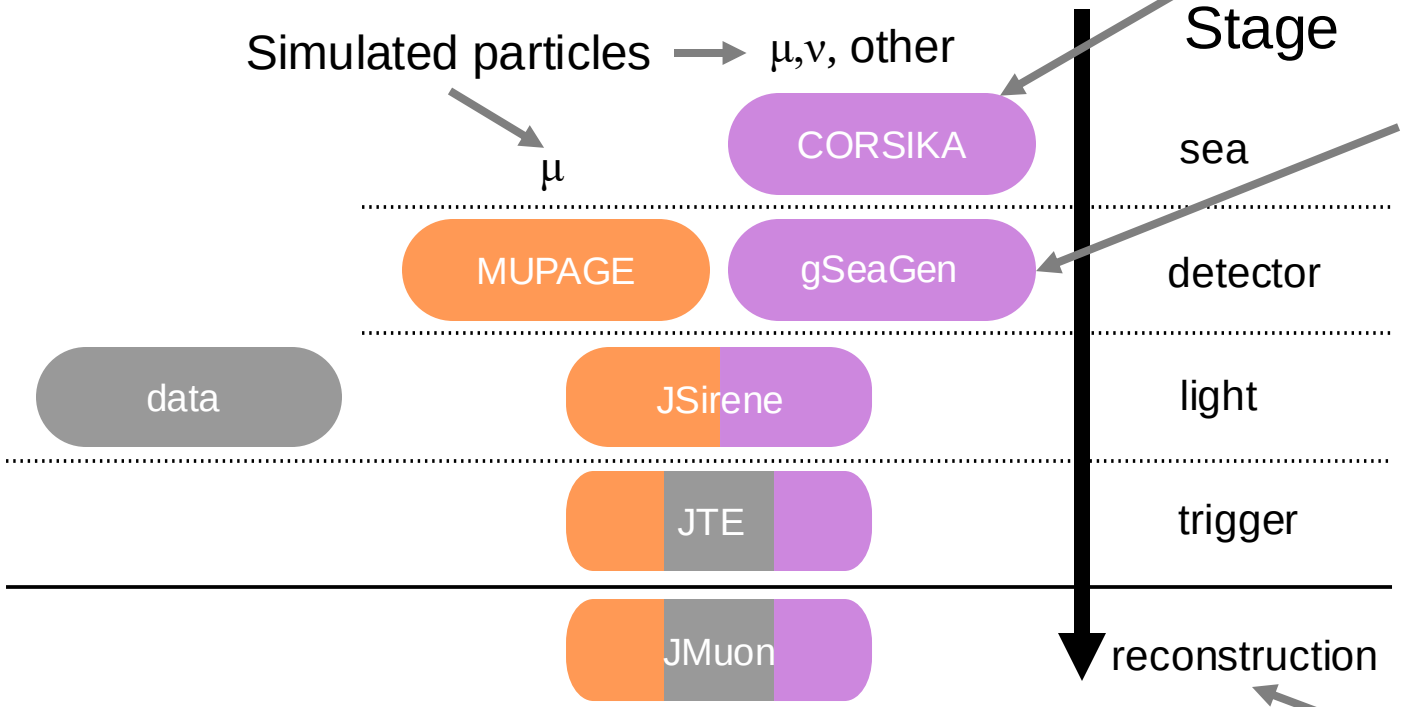
We have 2 options:

1. MUPAGE (atmospheric **MU**ons from **PA**rametric formulas: a fast **GE**nerator for neutrino telescopes)
   - developed for ANTARES
   - fast muon MC generator
   - based on parametric formulas and MACRO measurements
   - parameters can be freely tuned
2. CORSIKA (**CO**smic **R**ay **SI**mulations for **KA**scade)
   - developed for KASCADE
   - full simulation of air showers
   - customizable (models, primaries, etc.)



KM3NeT simulation chain

Standard KM3NeT muon simulation scheme



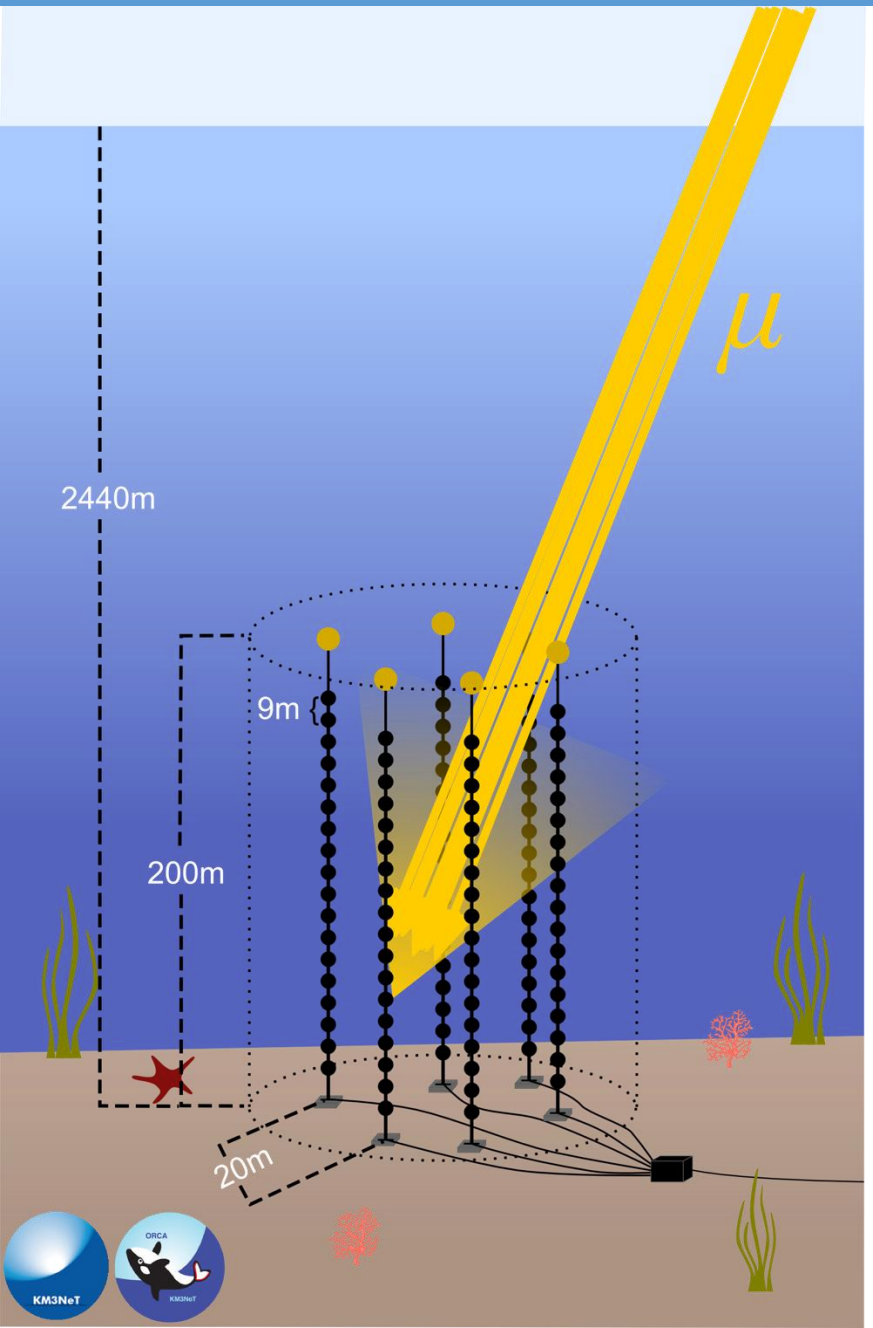
My involvement:

- ❖ Production of CORSIKA simulation(s):
  - Test prods
  - Mini-prods (systematic uncertainties)
  - First ever mass (>6PB) production
- ❖ Development of gSeaGen:
  - Implementation of CORSIKA support
  - Rework of the propagation routine:
    - Speed-up: for EeV 1w -> 40min
    - Included Earth's curvature
    - Statistics optimization: muon range tolerance margins, possibility to retry propagation etc.

  More in [in progress]:
  "gSeaGen code by KM3NeT: an efficient tool to propagate muons simulated with CORSIKA"
  (planned journal: Comp. Phys. Comm.)
- ❖ Reconstruction of new observables
  (see next Section)

Legend:
sea – sea surface above km3net detectors
detector – cyllindrical volume around the detector
light – simulation of photon emission and detection near the detector (including environmental background)
trigger – selection of potentially interesting events e.g. 3DMuon selects muon tracks
reconstruction – reconstruction of various observables, e.g. energy, direction

Muon bundles



Muon bundle: muons originating from a single EAS

Some important properties:

Observable	Description	Standard reco?
$\cos\theta_{\text{zenith}} = \frac{\sum \cos\theta_{\text{zenith}}}{N_\mu}$	Direction (zenith)	good
$E_{\text{bundle}} = \sum E_\mu$	Bundle energy	rather bad (focused on bundles with a single muon)
E_{primary}	Energy of the primary CR	no
N_μ	Muon multiplicity	no

That's what I focused on

Outline

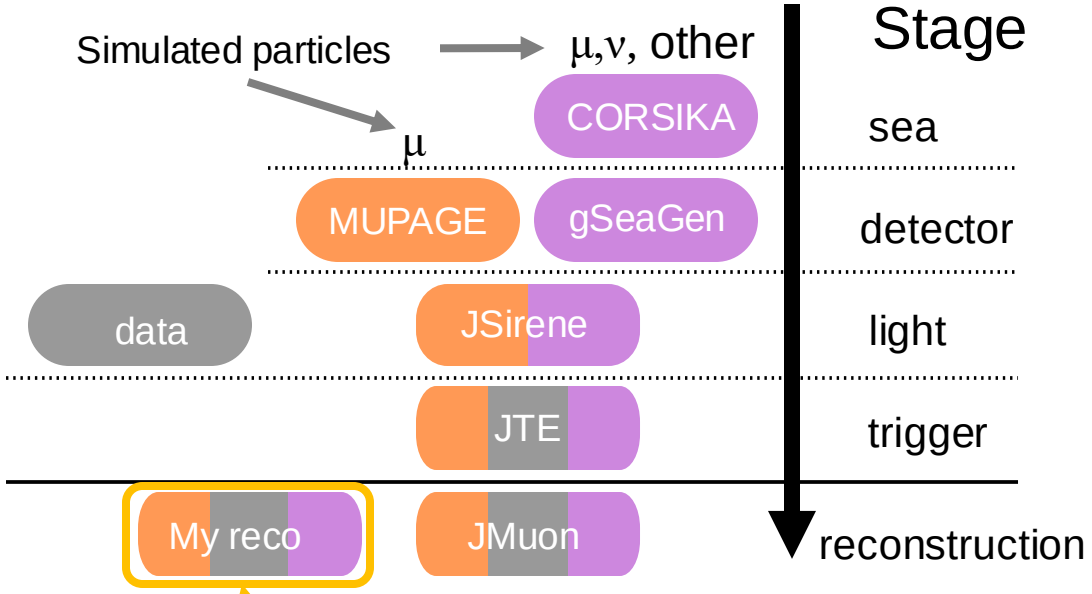
Introduction

Muon event reconstruction

Summary

My reconstruction of muon bundle properties

KM3NeT muon simulation scheme



(some)
Tools:

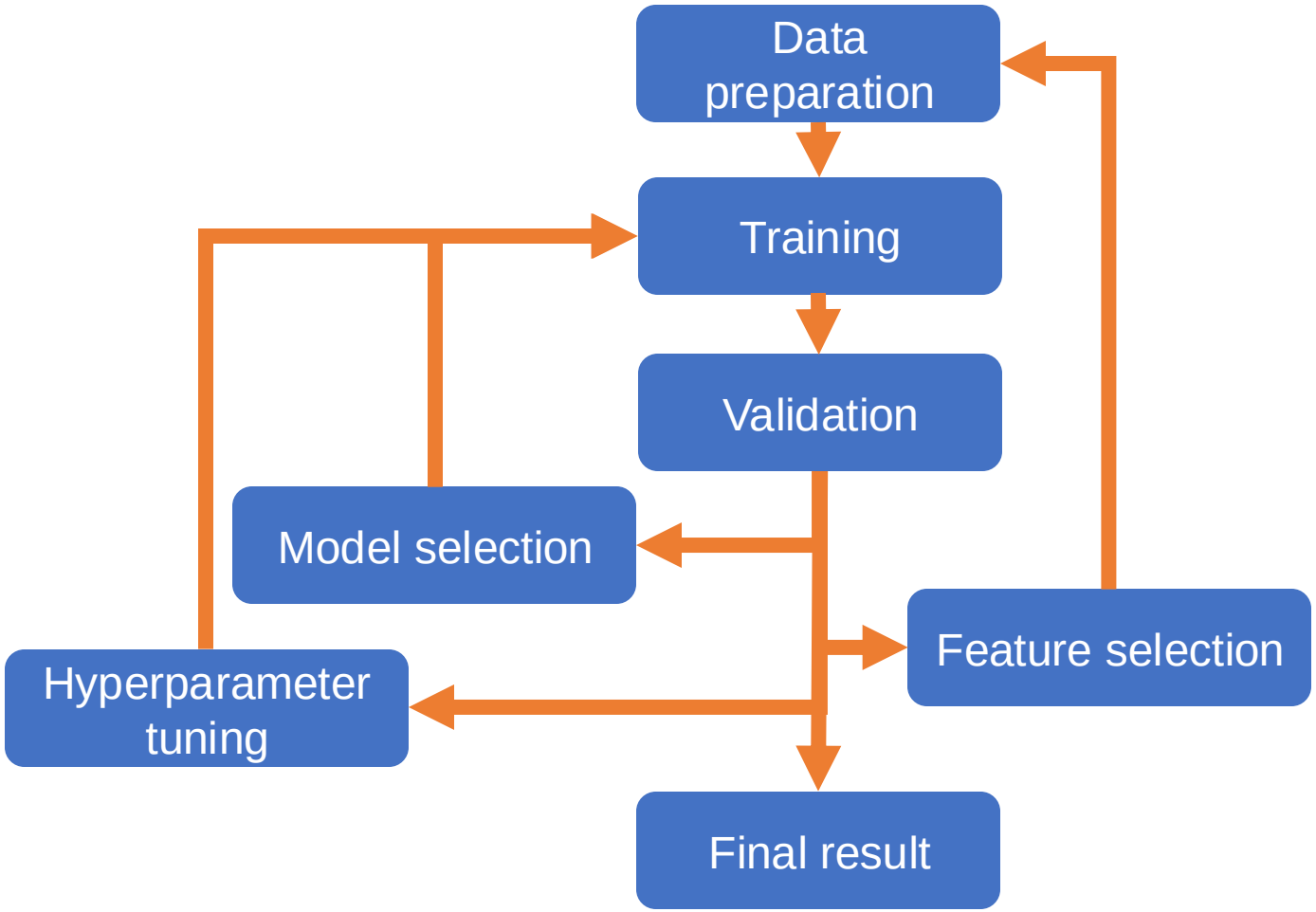
OPTUNA

scikit learn

LightGBM

dmlc XGBoost

My workflow:



My reco: data preparation

Read data

features
weights
targets

Scale weights

(in case the training performs better with scaled weights; depends on the model and target)

Scale features

one dataset-wide scaling:

$$x_{\text{scaled}} = \frac{(x - x_{\text{mean}})}{x_{\text{std}}}$$

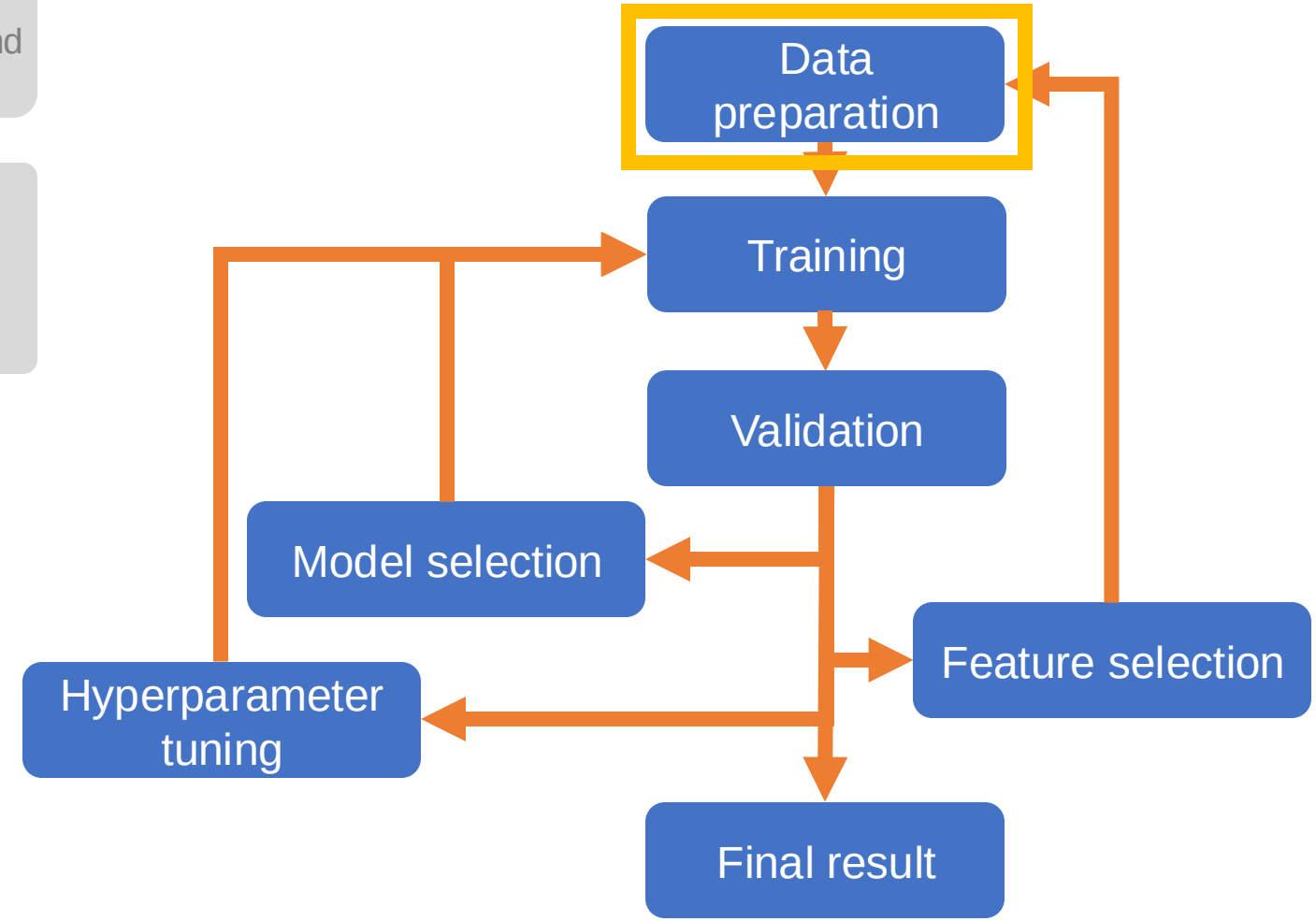
64% → Train data

16% → Validation data

20% → Test data

90% of success: feature engineering & preprocessing!

My workflow:



My reco: data summary

Read data → features, weights, targets

Scale weights → (in case the training performs better with scaled weights; depends on the model and target)

Scale features → one dataset-wide scaling:

$$x_{\text{scaled}} = \frac{(x - x_{\text{mean}})}{x_{\text{std}}}$$

64% → Train data

16% → Validation data

20% → Test data

Exp data

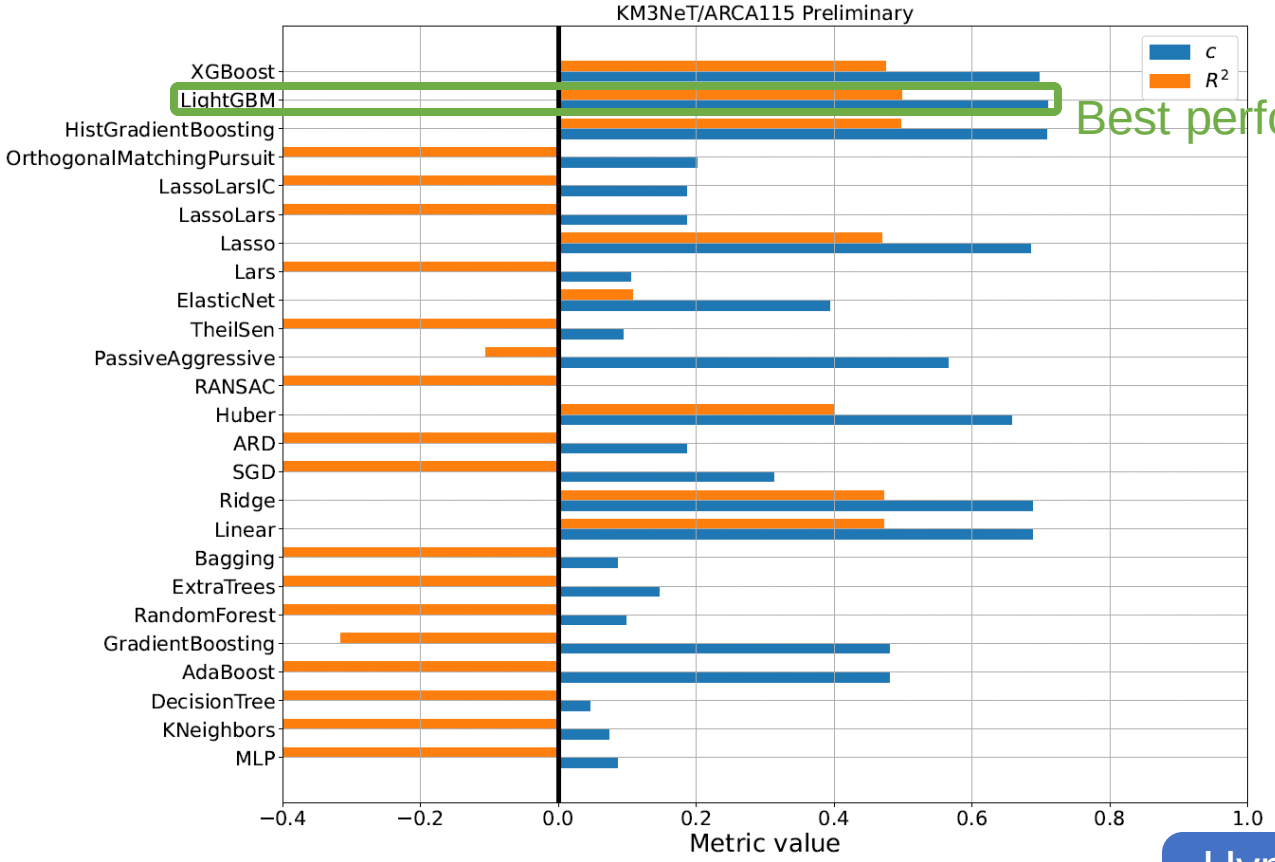
	ARCA115	ARCA6	ORCA115	ORCA6
Train data	14 374 415	3 894 293	15 439 946	6 154 895
Validation data	3 598 133	974 371	3 857 715	1 539 988
Test data	4 489 177	1 215 552	4 825 073	1 923 475
Exp data	X	3 013 982	X	4 574 887

official KM3NeT CORSIKA simulation (by me)

real data from 2020-2021

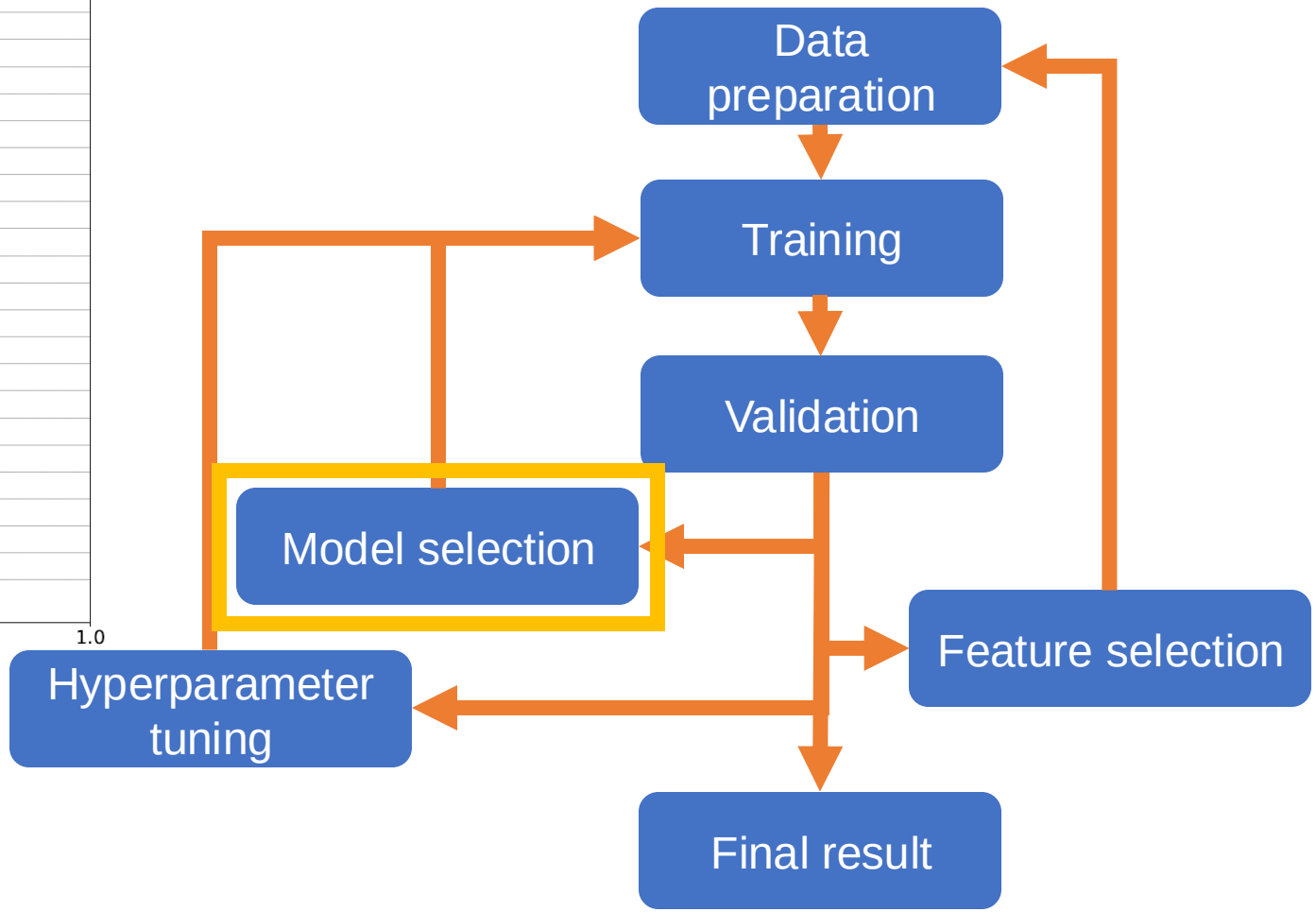
My reco: model selection

Performance comparison
(on a fraction (50k events) of the training dataset):



Best performance: LightGBM

My workflow:

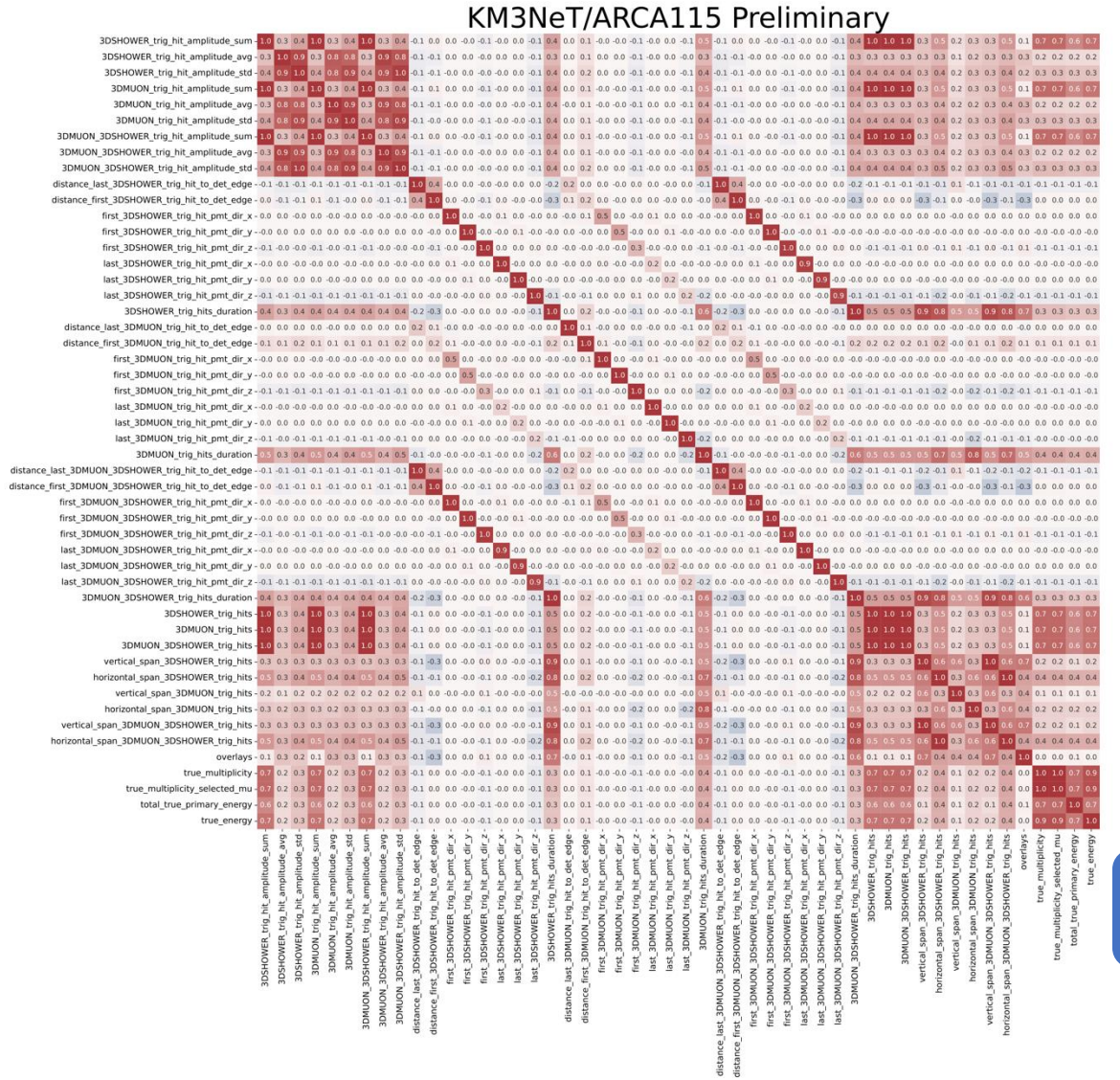


Metrics:

- weighted R^2-score
- weighted Pearson correlation coefficient

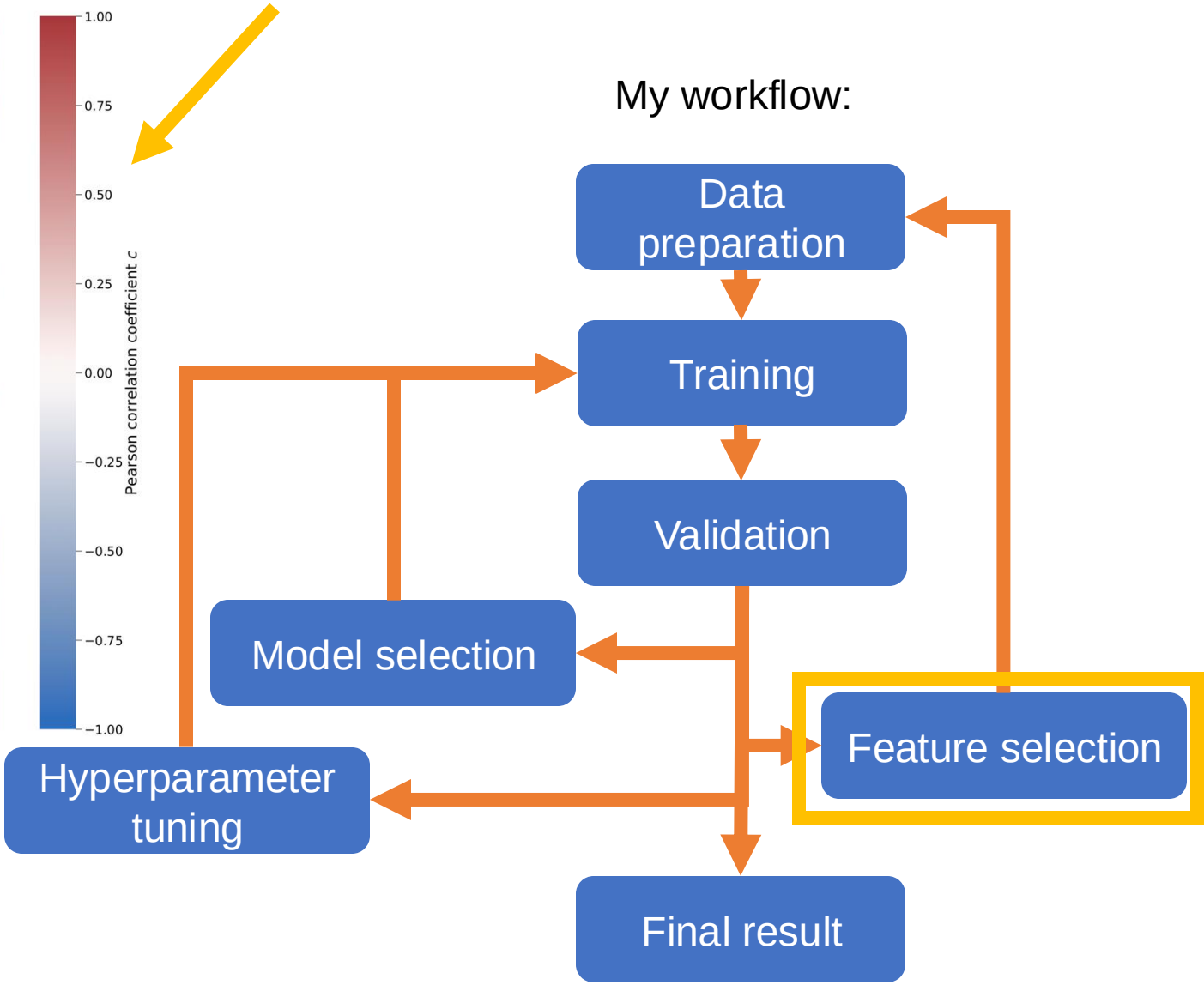
My reco: feature selection

Correlation matrix of all 46 features and 4 targets:

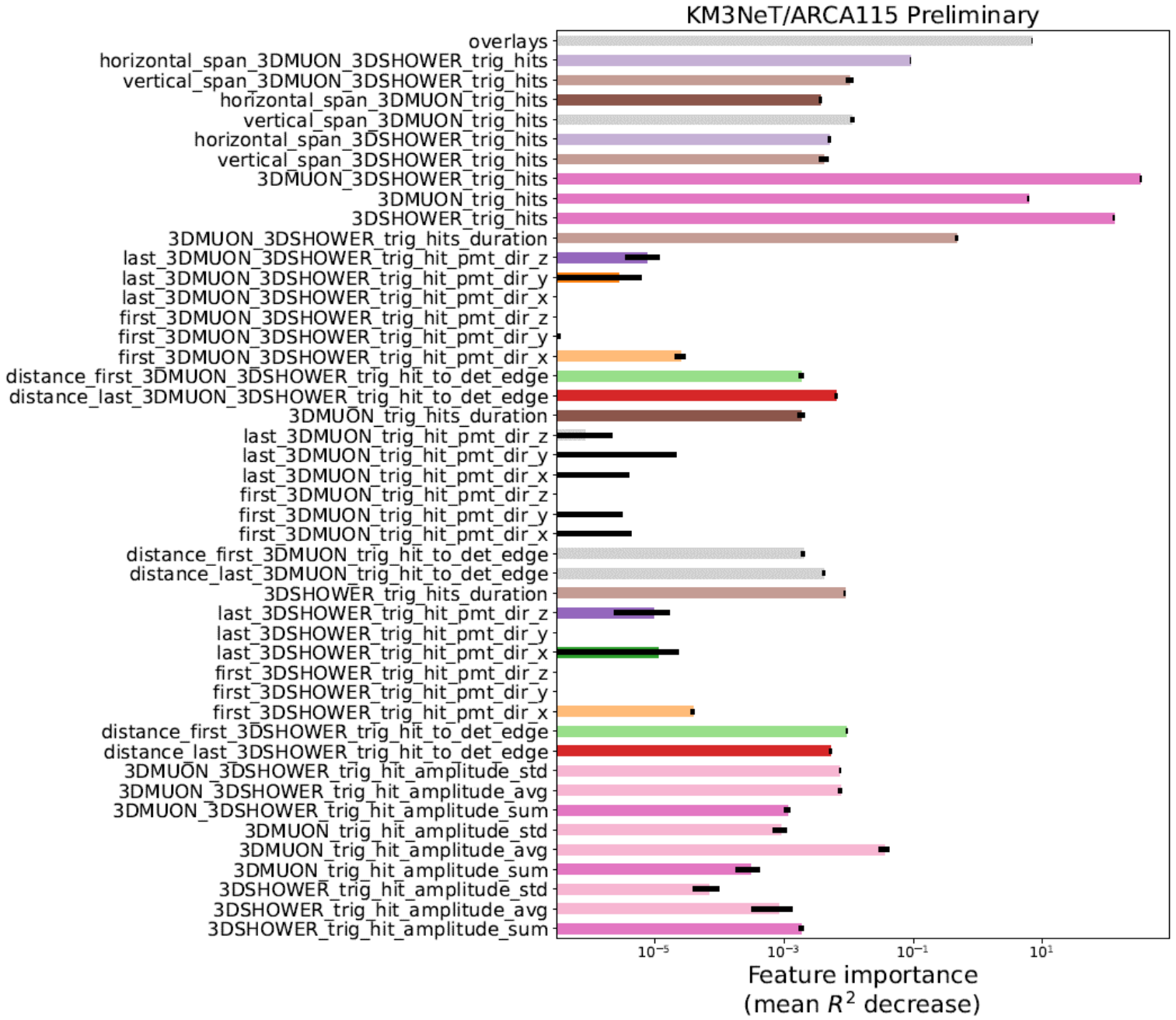


NOT equally important, maybe not all needed?

My workflow:



My reco: feature selection



Colors not random!
They match the feature clustering (see backup)

The idea:
Try to select only the most important feature in each cluster?

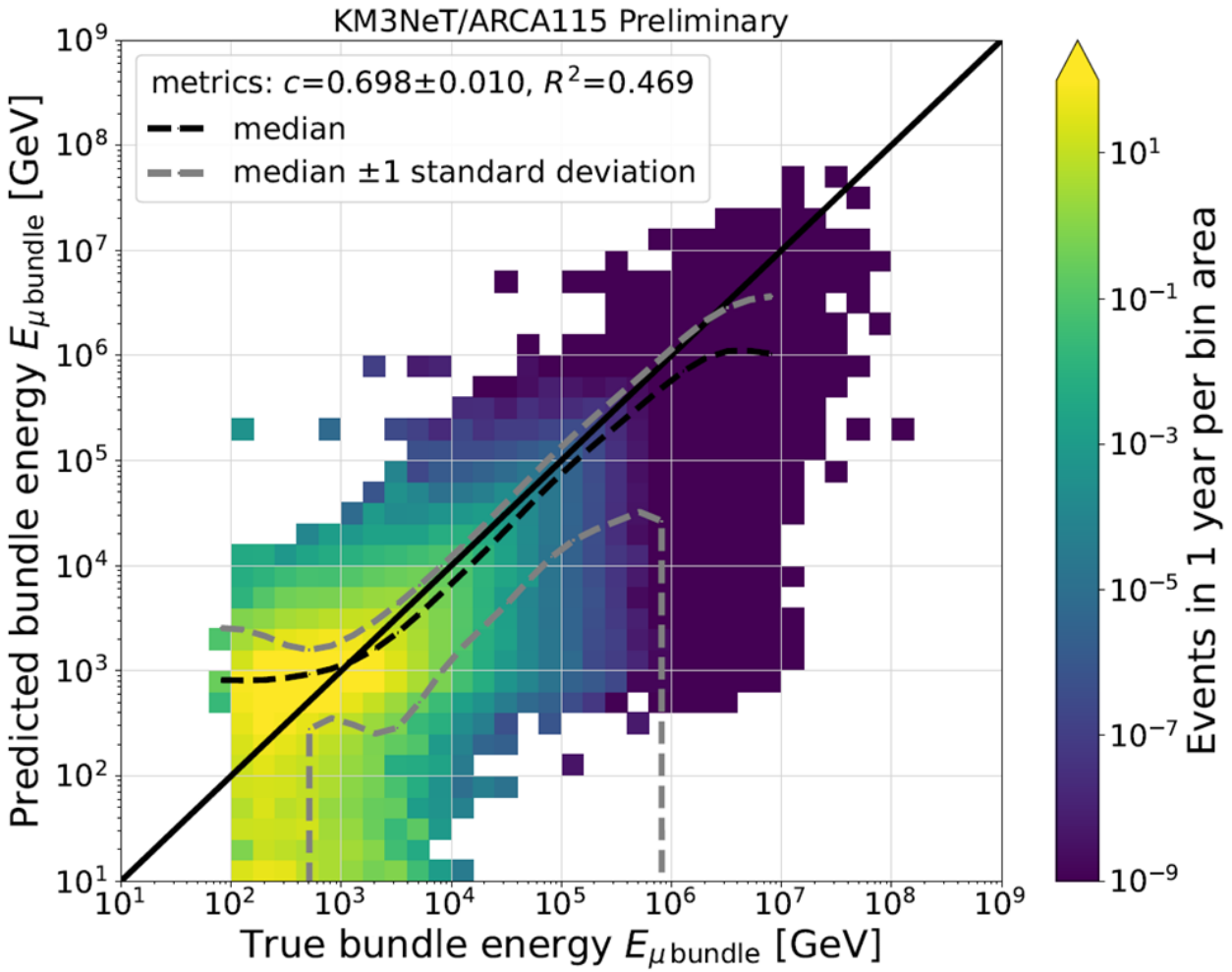
Feature	Importance
3DMUON_3DSHOWER_trig_hits	325
overlays	6.84
3DMUON_3DSHOWER_trig_hits_duration	0.47
horizontal_span_3DMUON_3DSHOWER_trig_hits	0.0900
3DMUON_trig_hit_amplitude_avg	0.0359
vertical_span_3DMUON_trig_hits	0.0114
distance_first_3DSHOWER_trig_hit_to_det_edge	0.0093
distance_last_3DMUON_3DSHOWER_trig_hit_to_det_edge	0.0063
distance_last_3DMUON_trig_hit_to_det_edge	0.0041
horizontal_span_3DMUON_trig_hits	0.0036
distance_first_3DMUON_trig_hit_to_det_edge	0.0020

My reco: 4 scenarios of feature selection

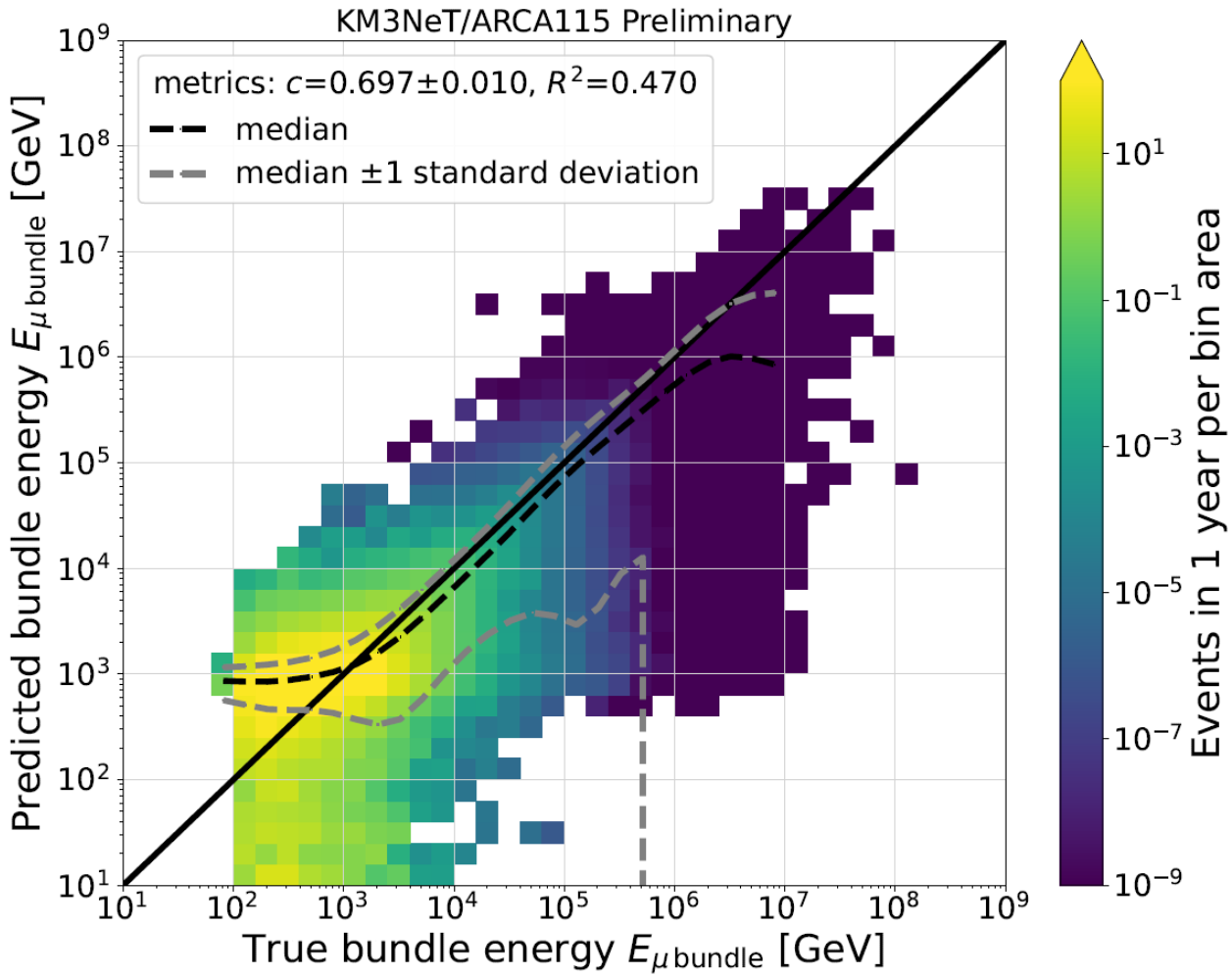
I considered 4 options:

1. All features
2. Features with importance>0 & only one per cluster
3. The most important feature only
4. Features with importance>0

1. All features



2. importance>0 & clustering

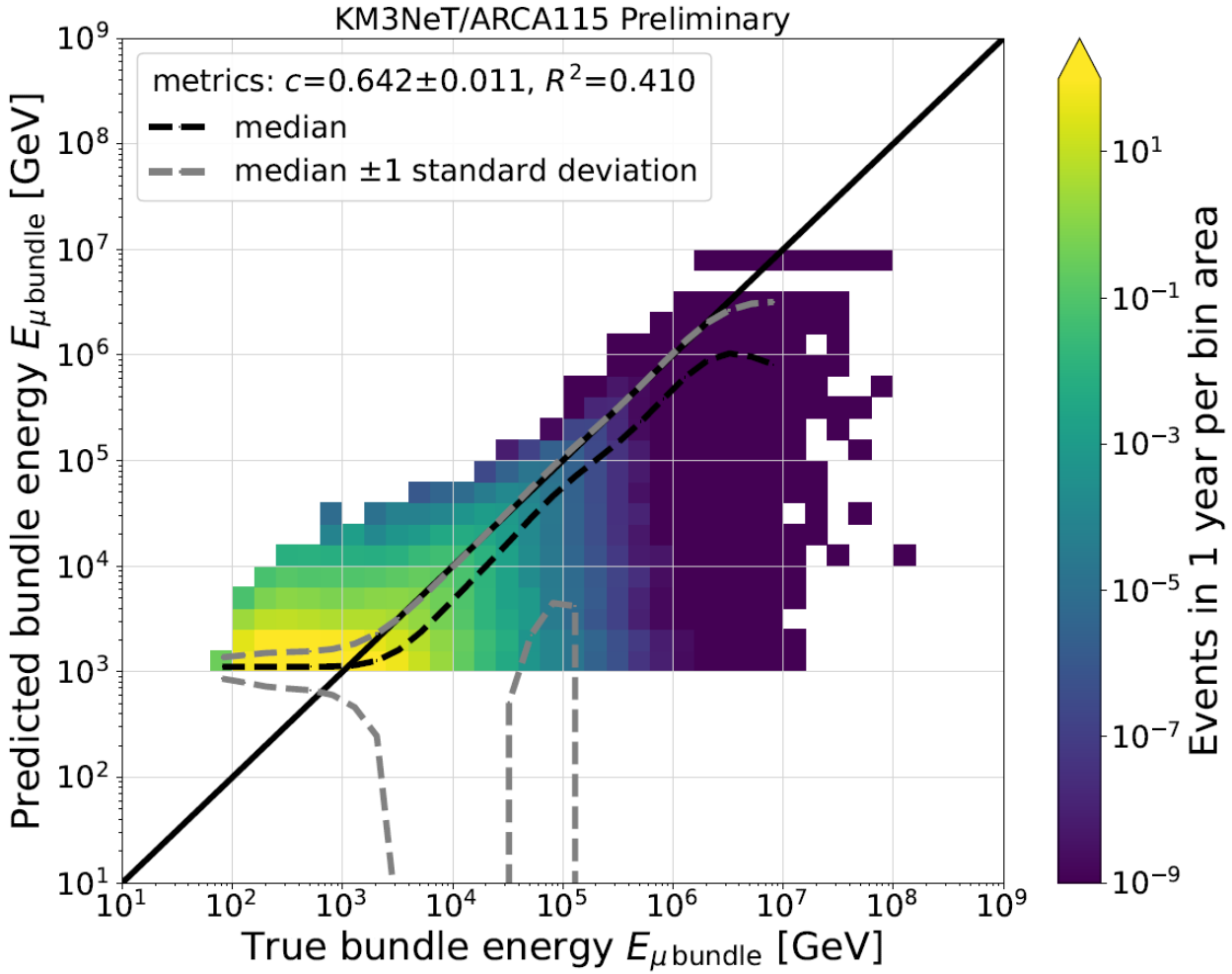


My reco: 4 scenarios of feature selection

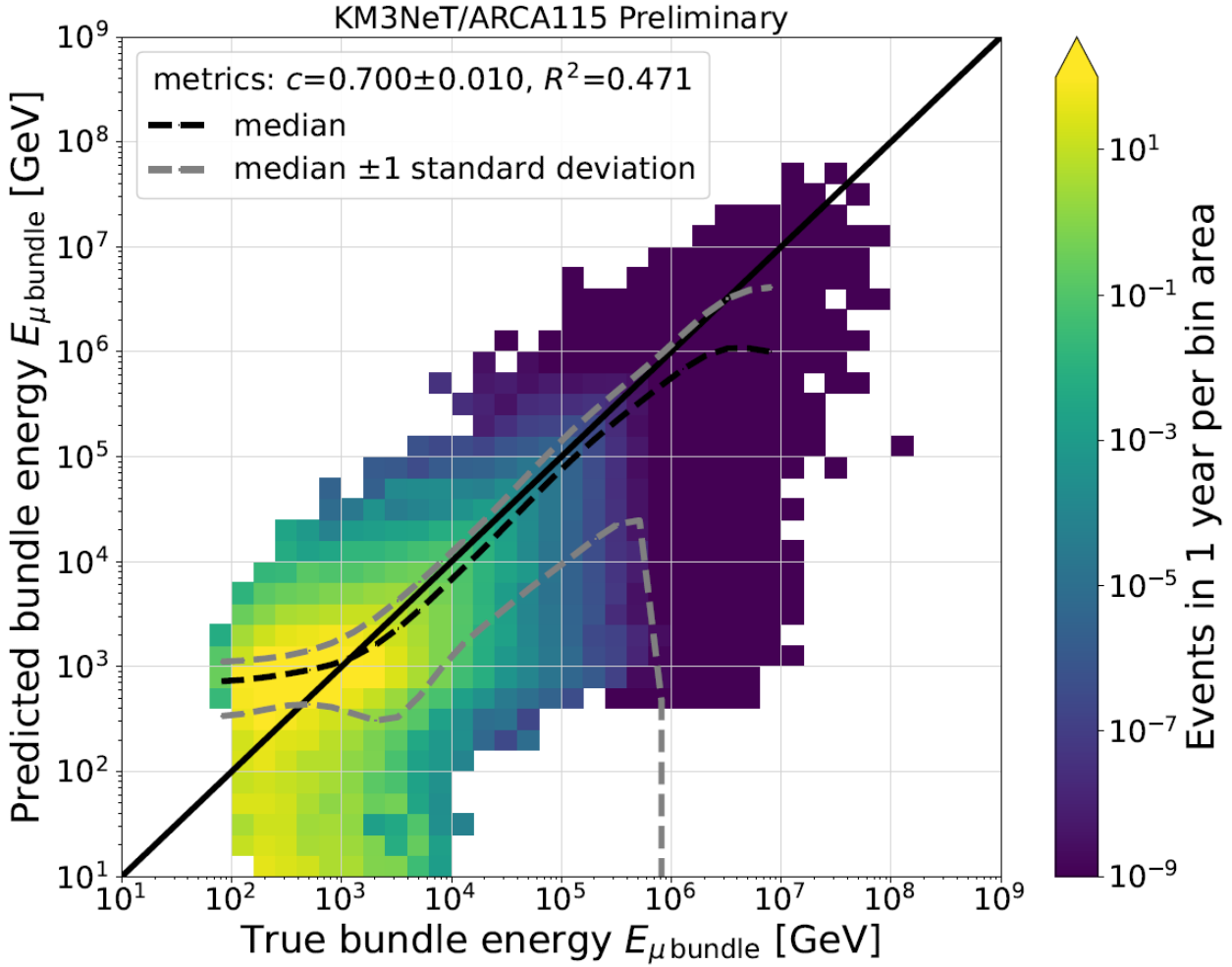
I considered 4 options:

1. All features
2. Features with importance>0 & only one per cluster
3. The most important feature only
4. Features with importance>0

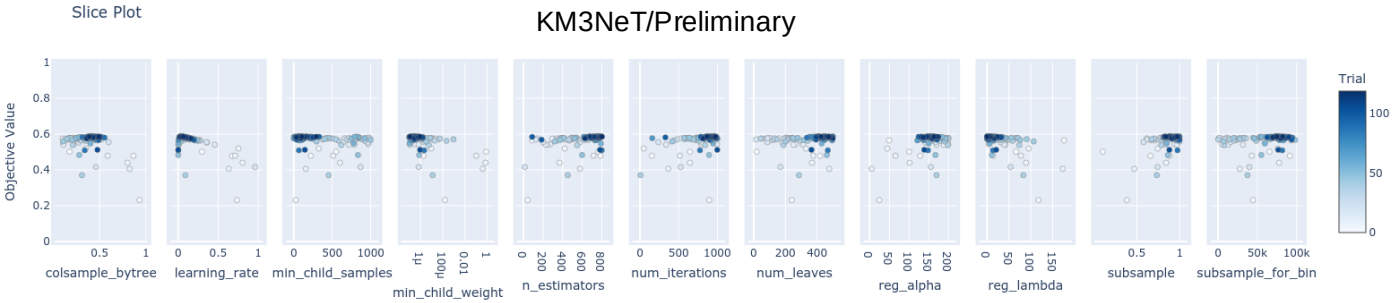
3. 3DMUON_3DSHOWER_trig_hits only

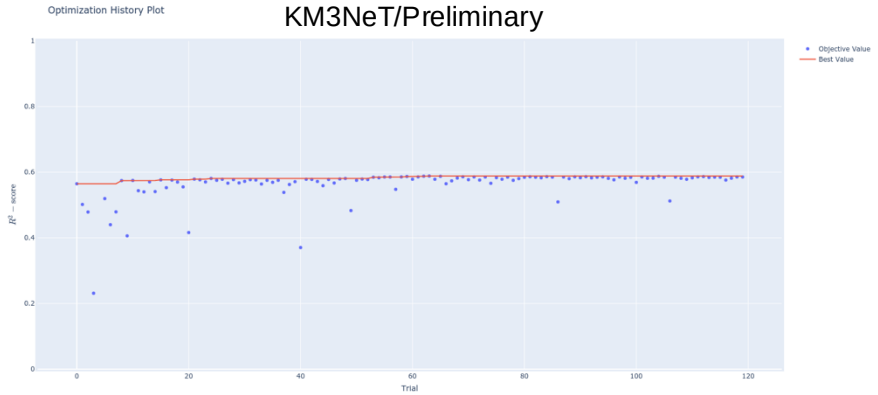


4. importance>0



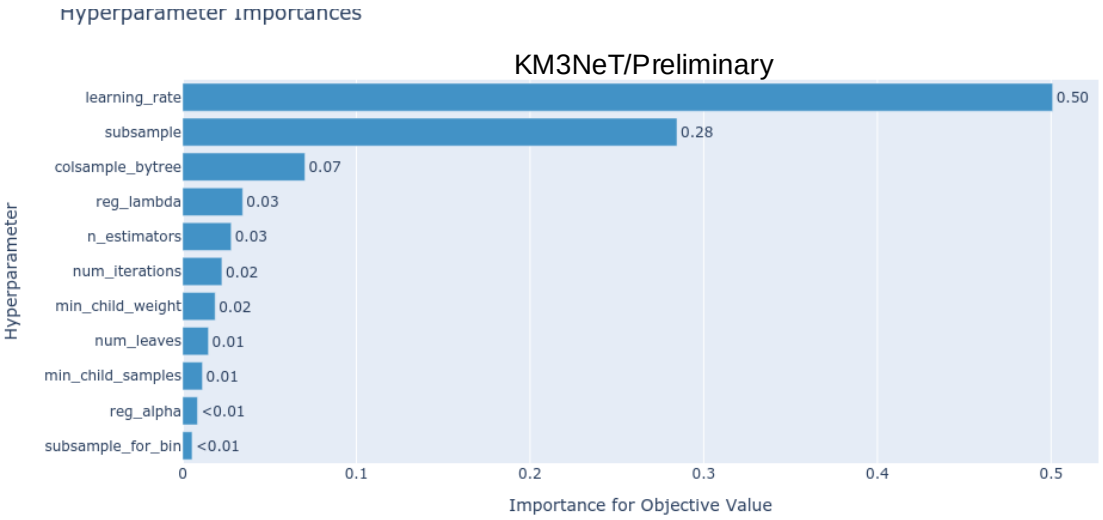
My reco: hyperparameter tuning



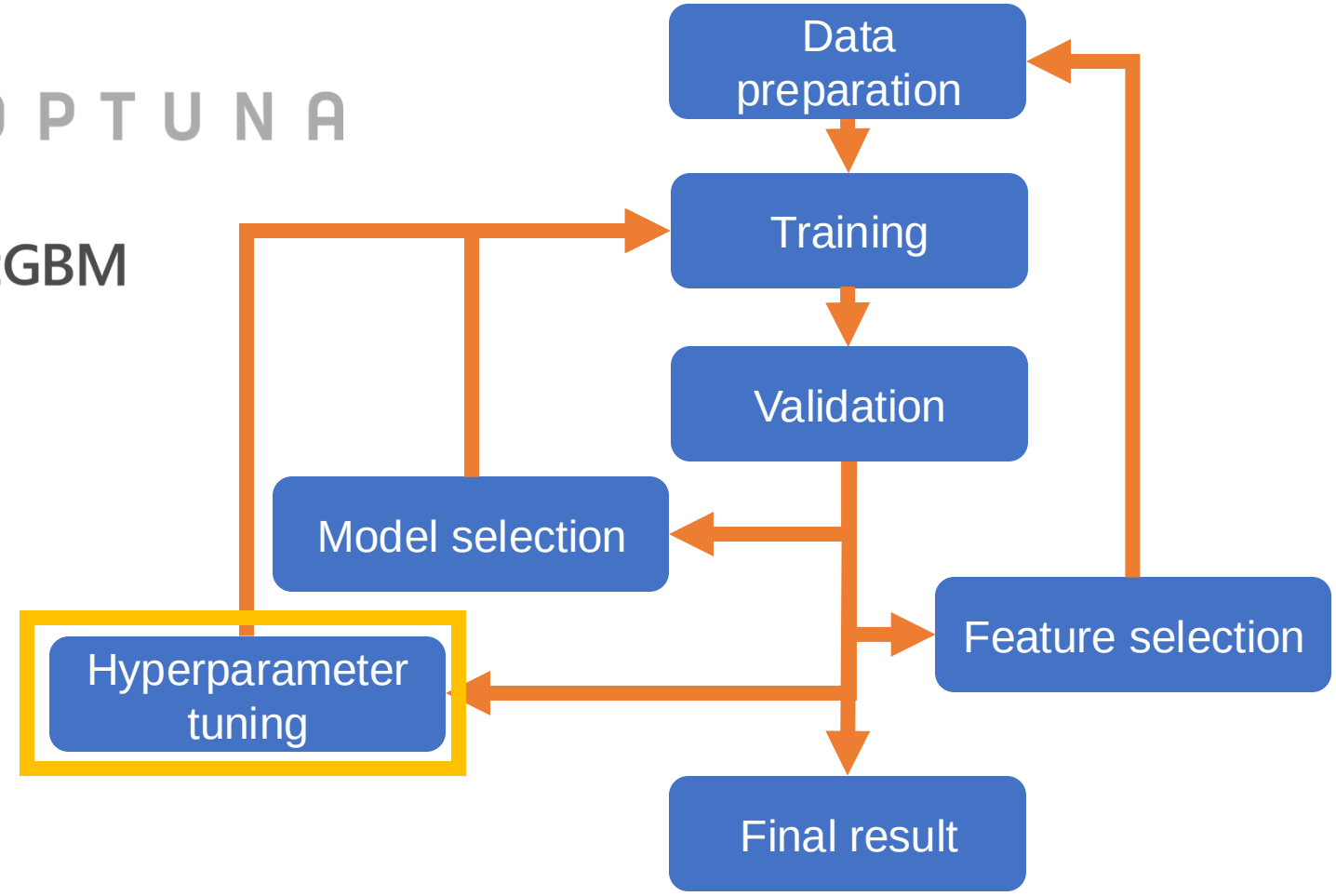


OPTUNA

LightGBM



My workflow:

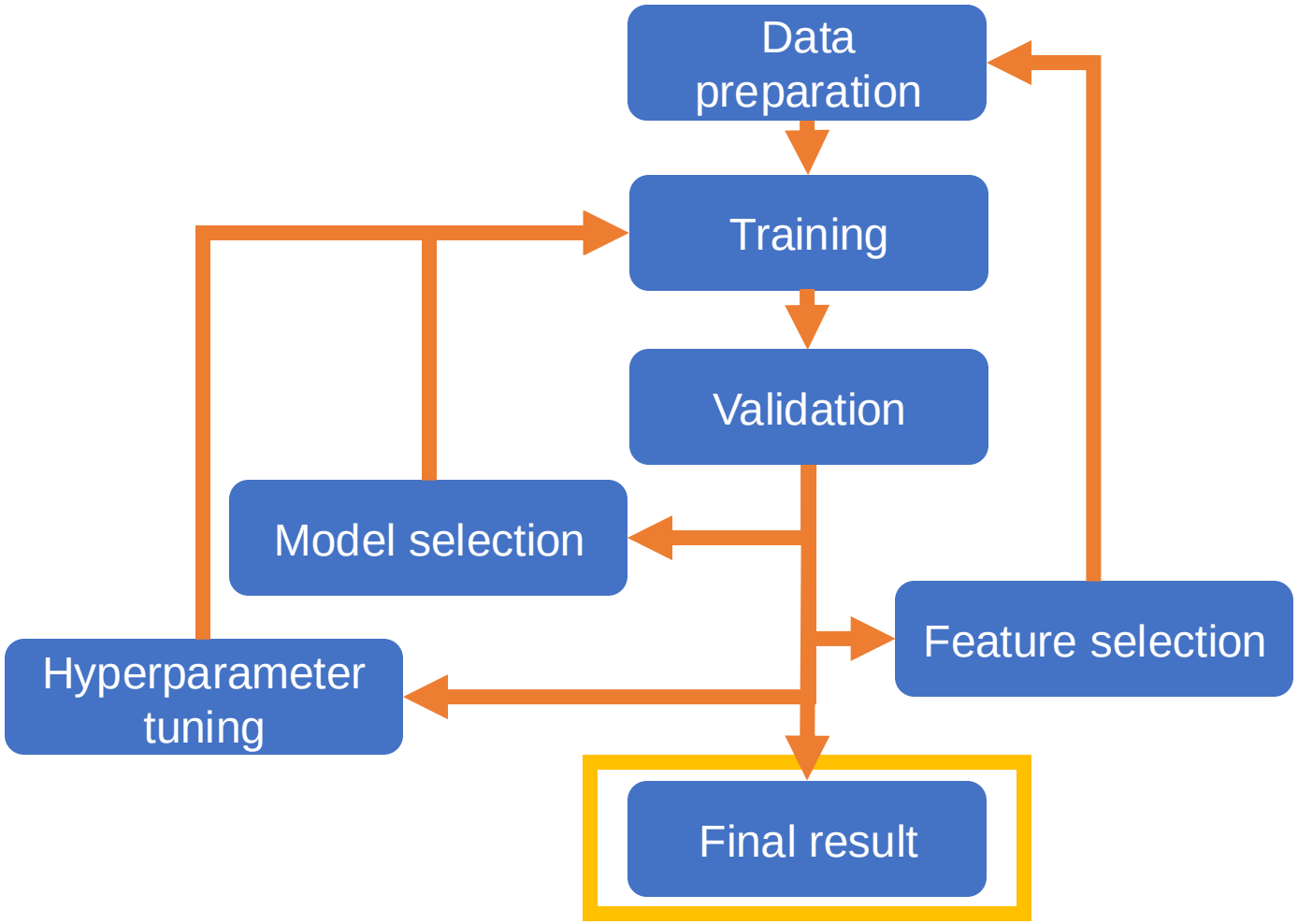


My reco: Final results

Note:

I show only some of the results!
(because there is too much)

My workflow:

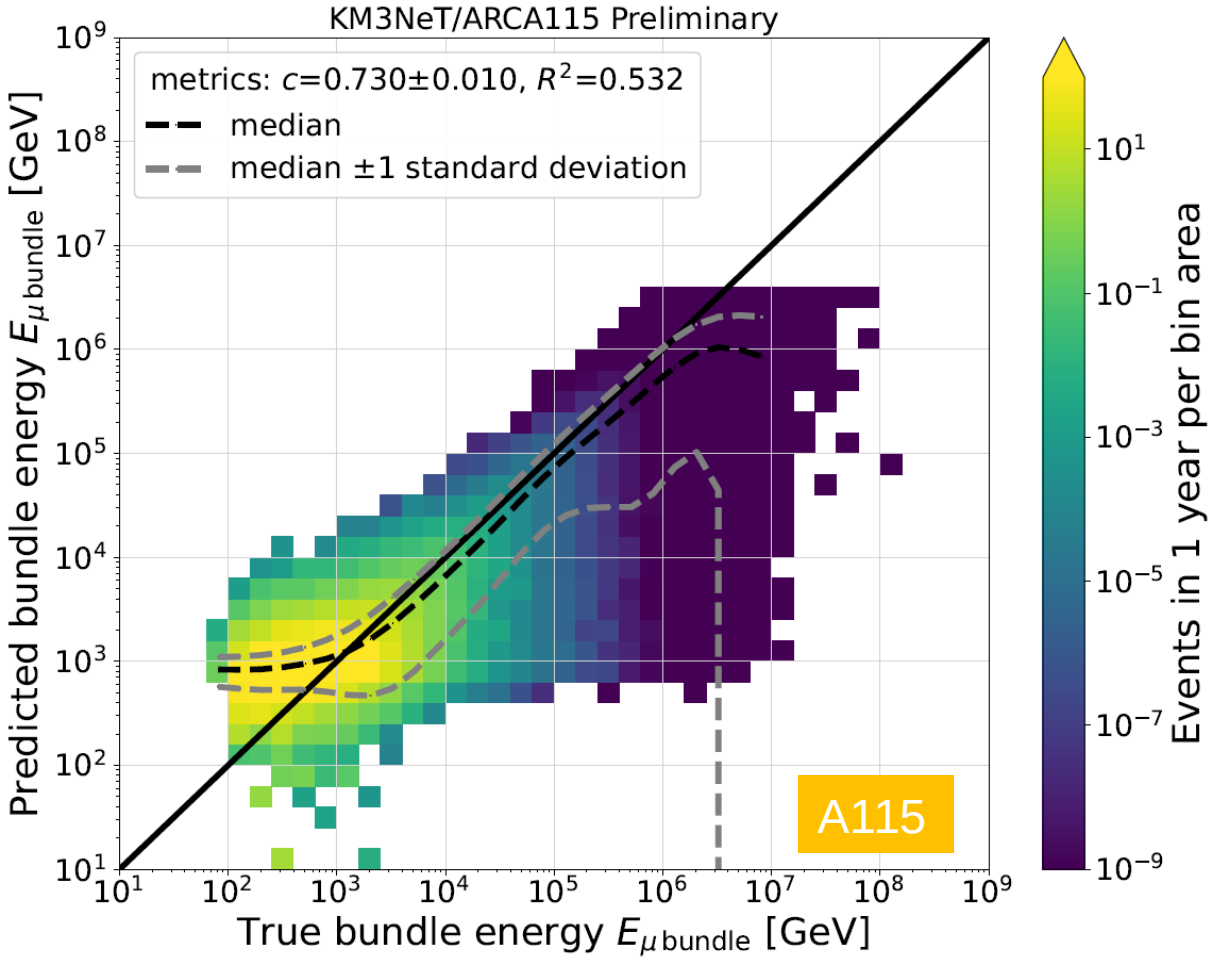


Bundle Energy reconstruction

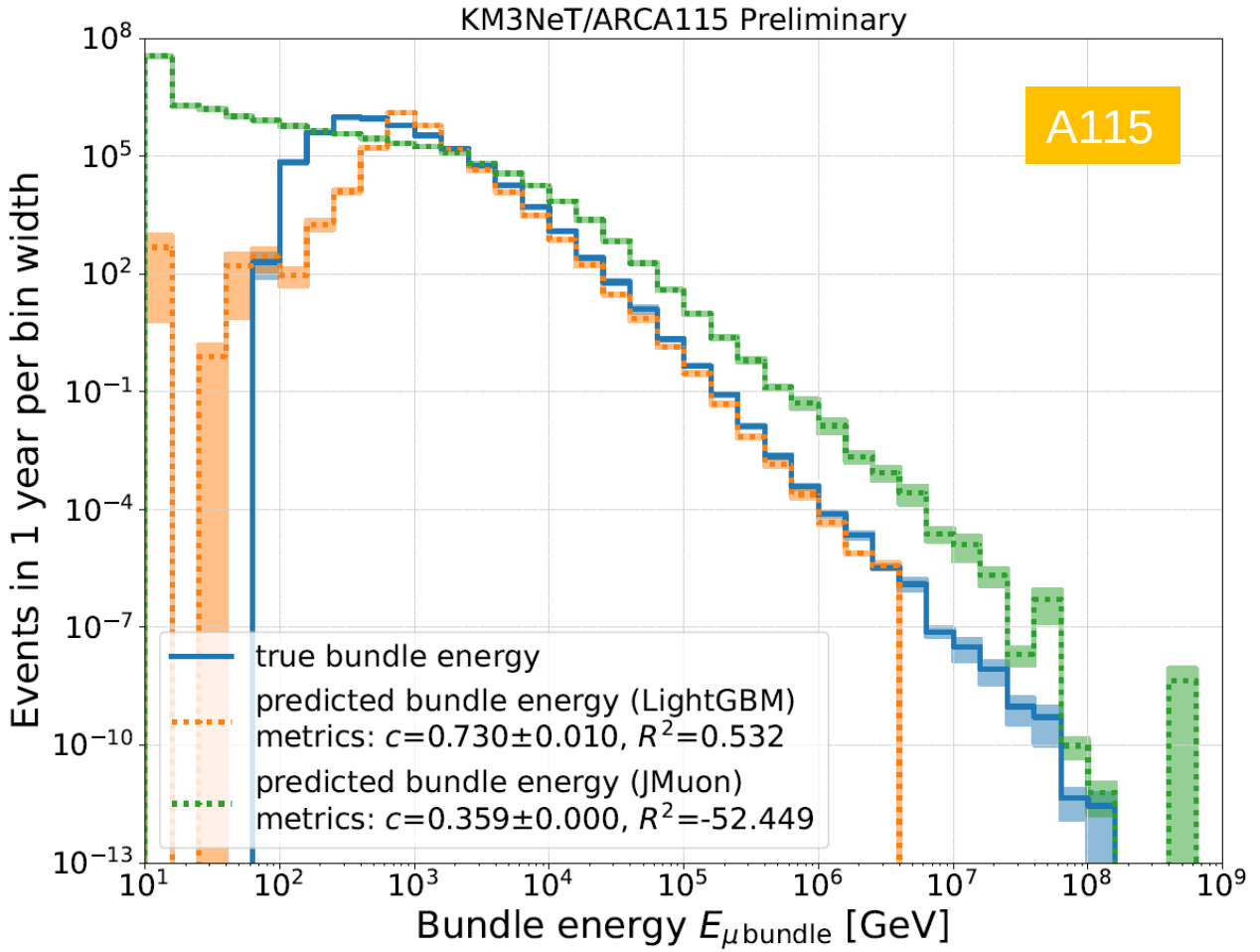
Analogous results for ARCA6, ORCA115 and ORCA6

Results for ARCA115:

2D: reco vs true



1D histograms

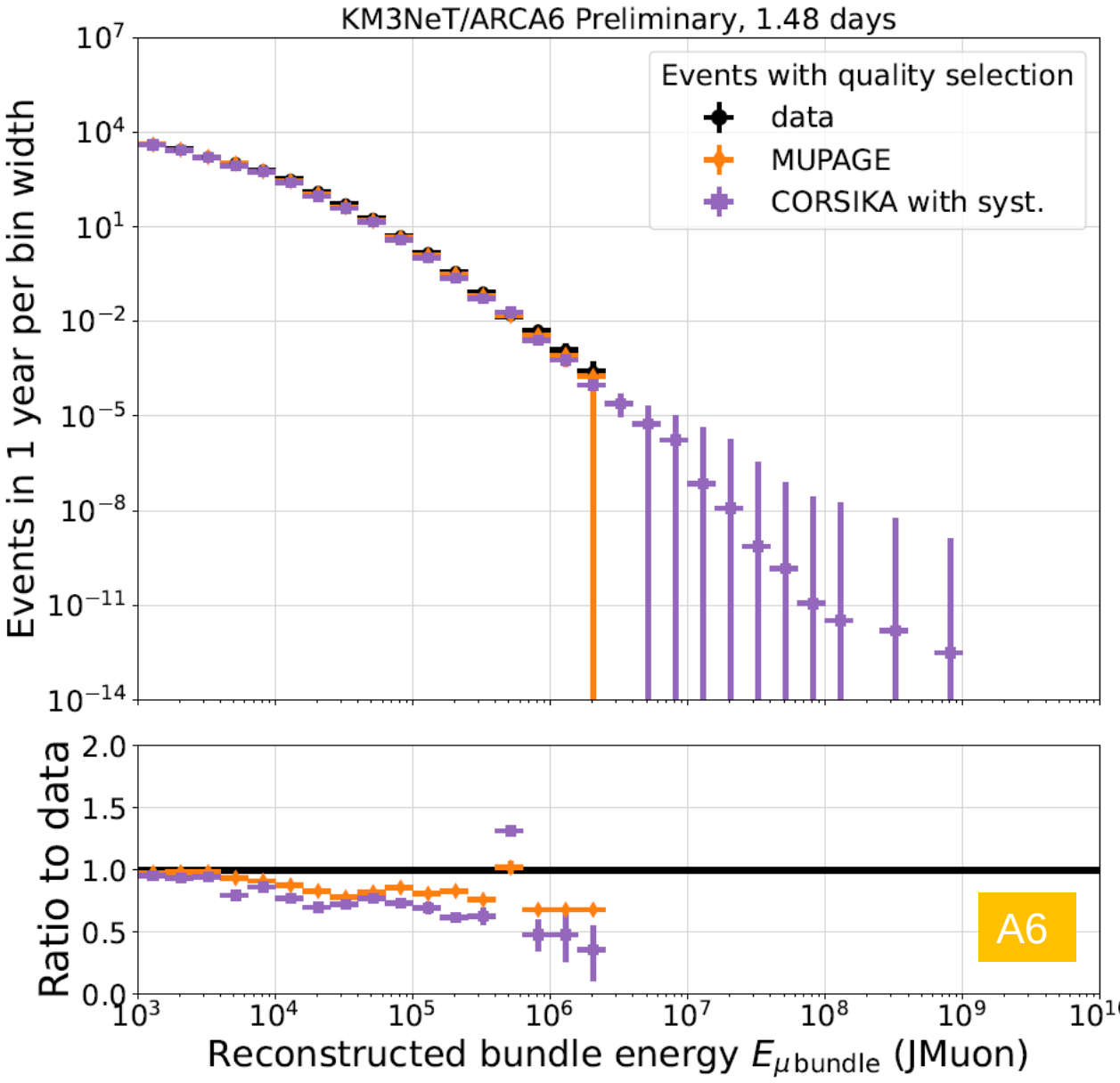


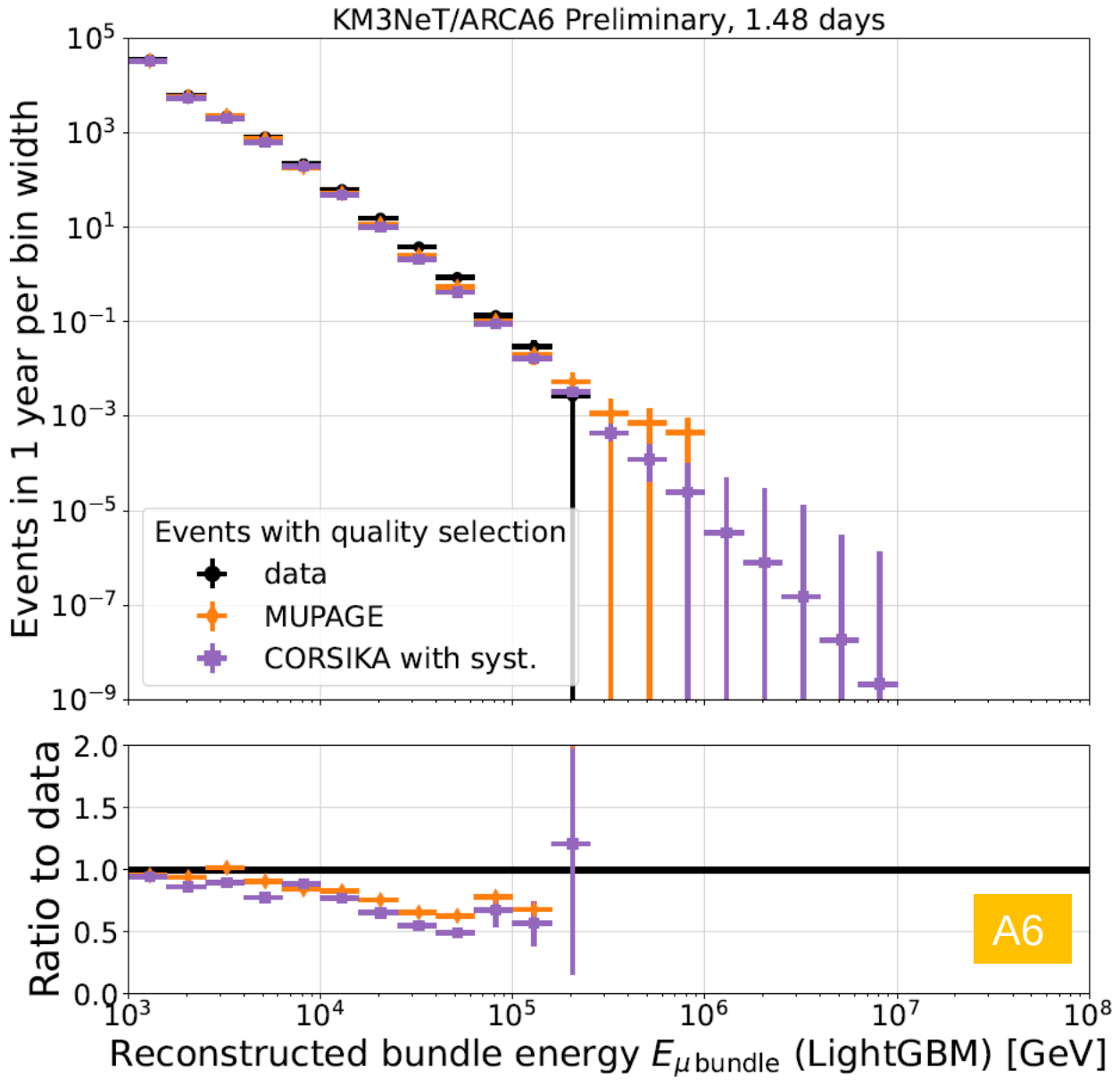
Bundle energy reco applied on ARCA6 data

Jmuon (standard reco)

Analogous results for ORCA6

LightGBM



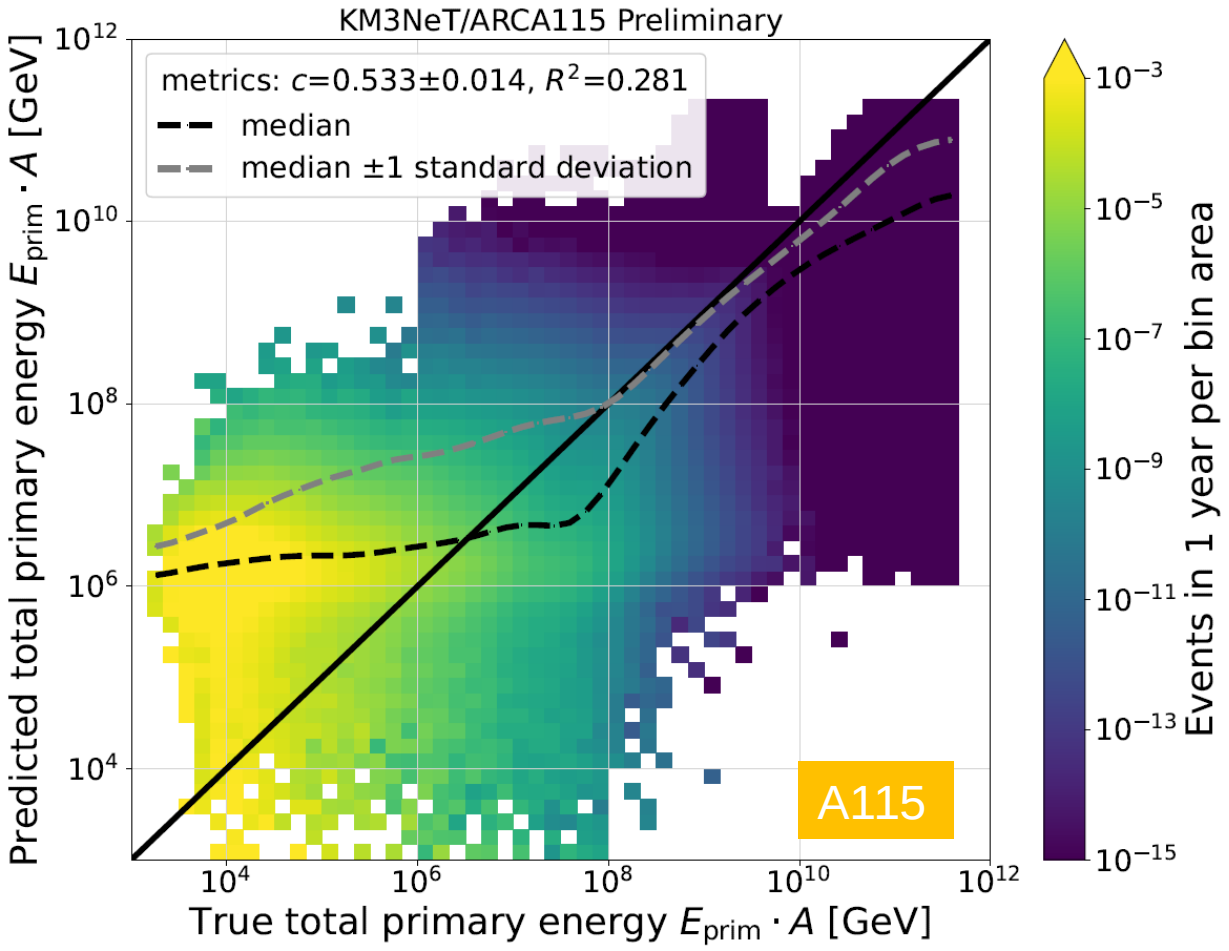


Primary CR energy reco

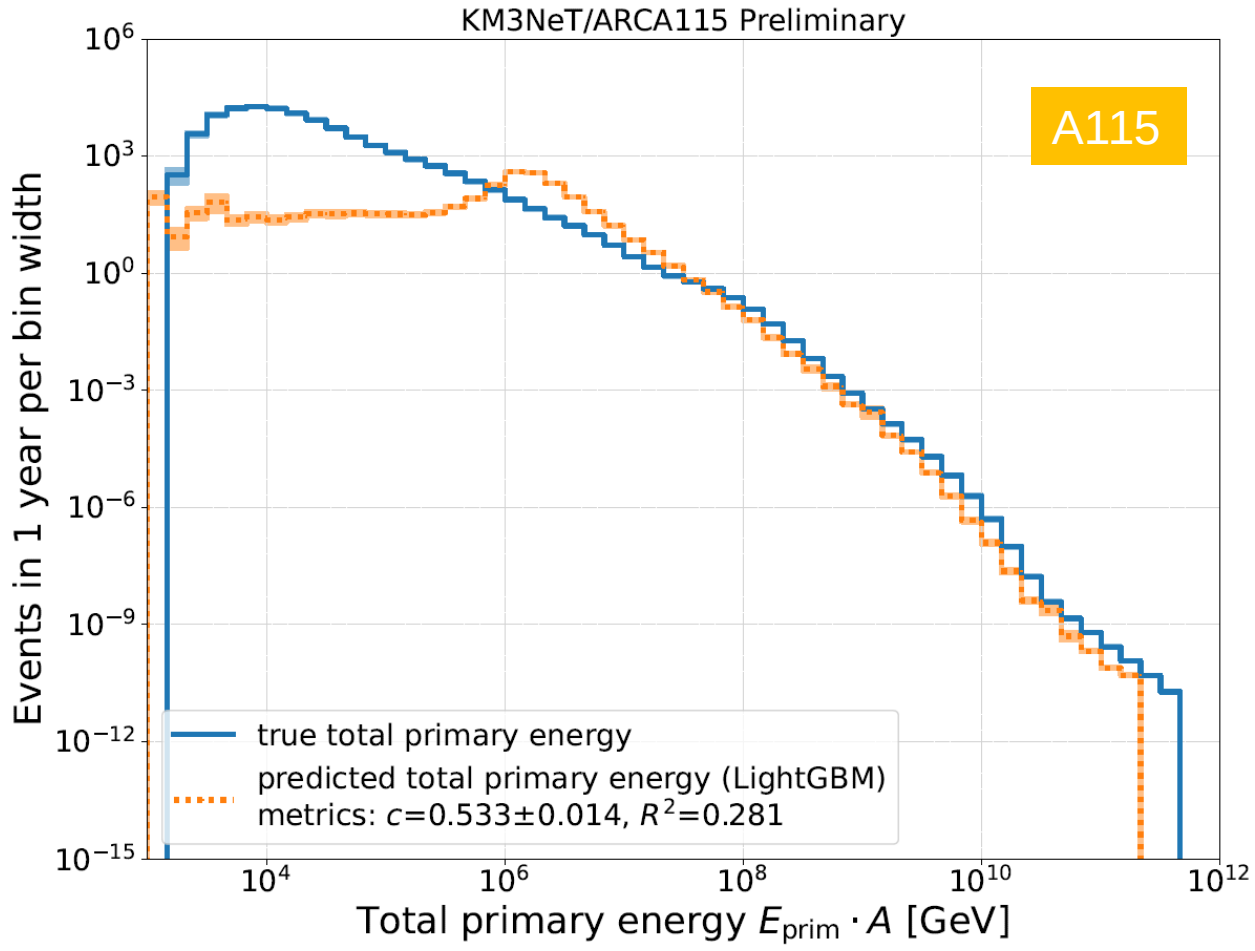
Analogous results for ARCA6, ORCA115 and ORCA6

Results for ARCA115:

2D: reco vs true

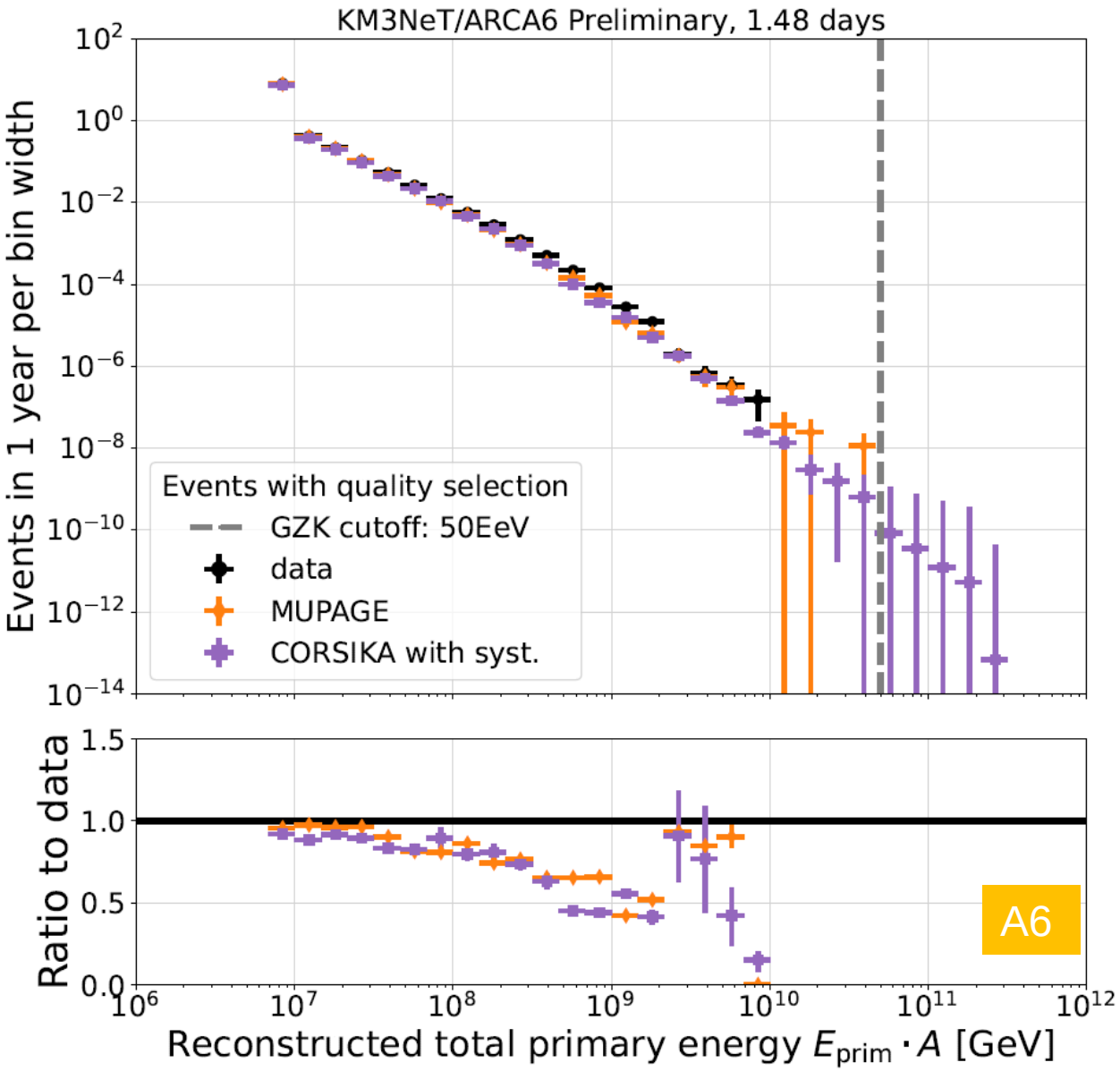


1D histograms



Primary CR energy reco applied on ARCA6 data

LightGBM



Analogous results for ORCA6

Results obtained with ARCA6/ORCA6:

- First measurement of primary CR energy with KM3NeT detectors ever
- Simulations underestimate the flux at high energies
  (similarly as for bundle energy; expected)
- The Greisen–Zatsepin–Kuzmin (GZK) cutoff cannot be confirmed/excluded (yet!)

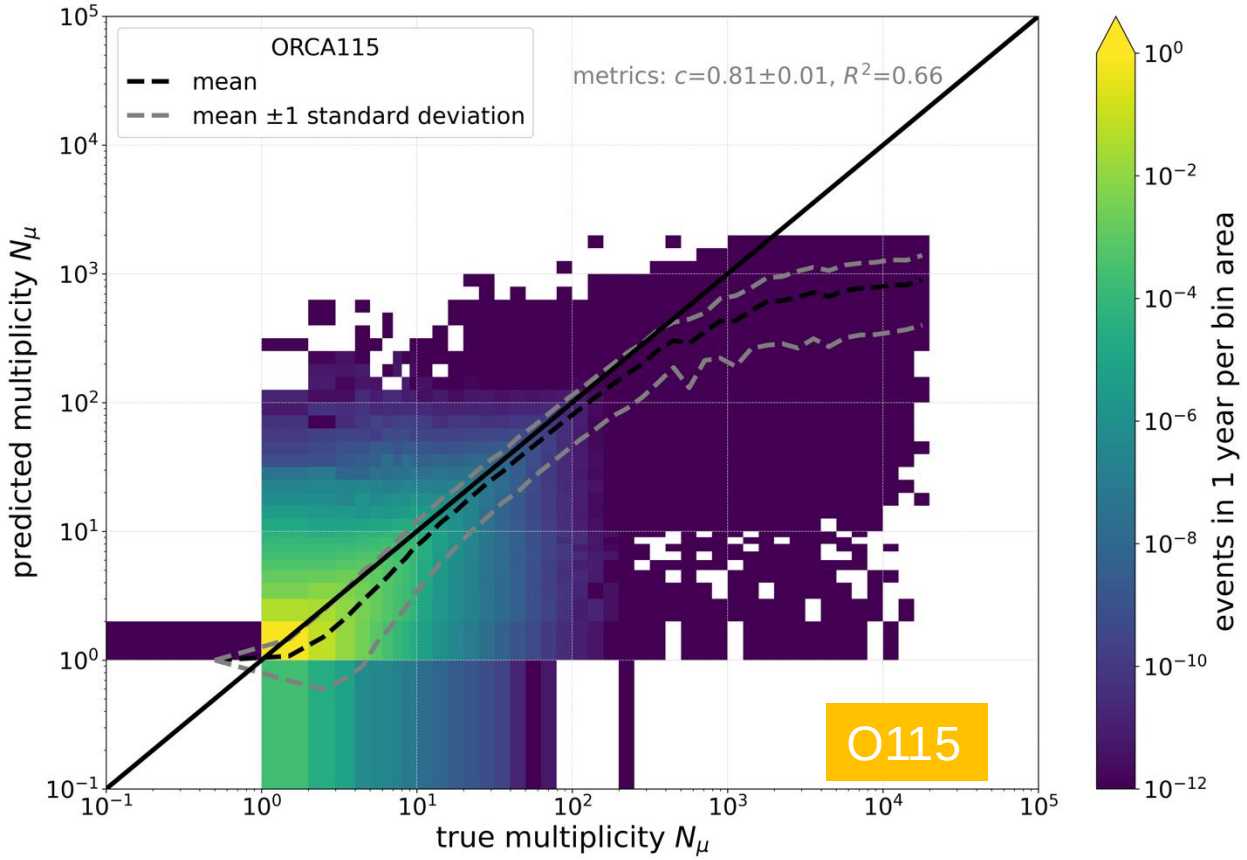
*GZK cutoff – theoretical upper bound on possible CR energy, due to interactions with the cosmic microwave background radiation (CMB)

Multiplicity reconstruction

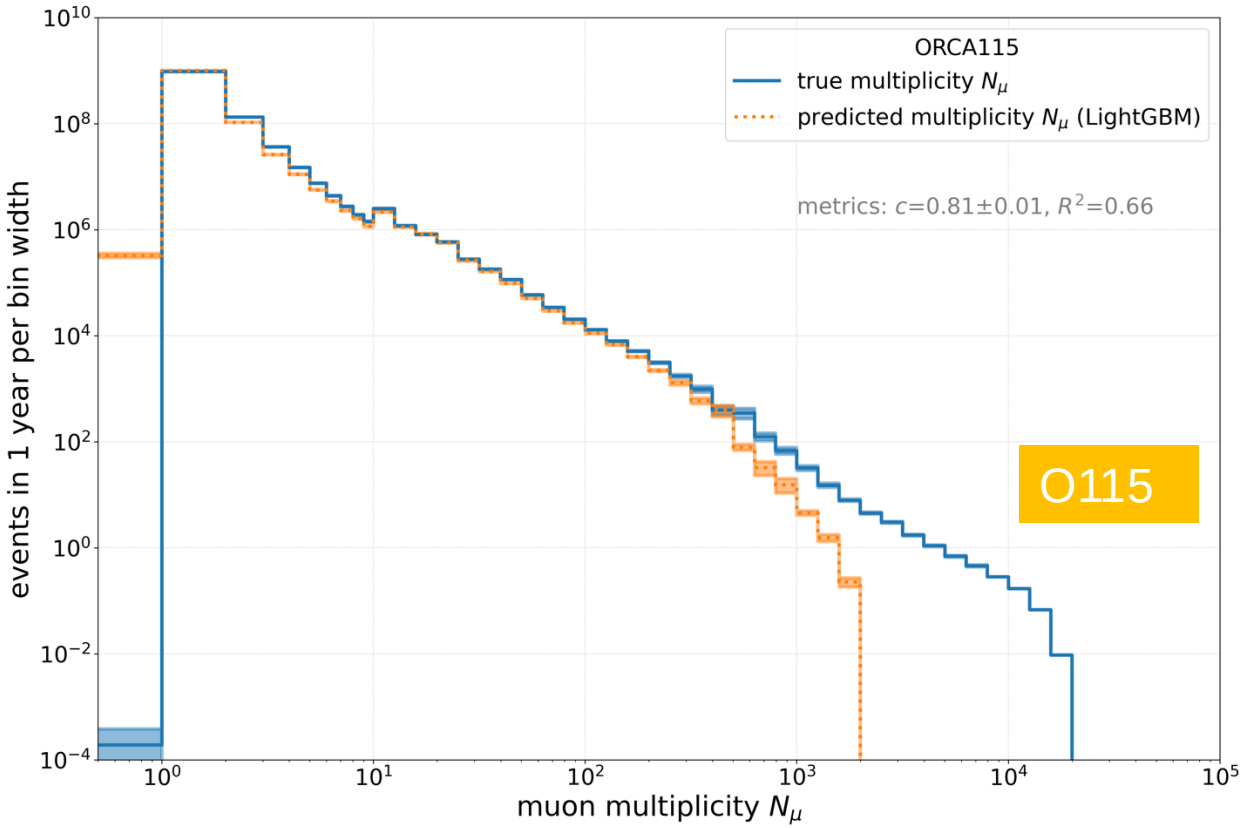
Analogous results for ARCA6, ARCA115 and ORCA6

Results for ORCA115:

2D: reco vs true

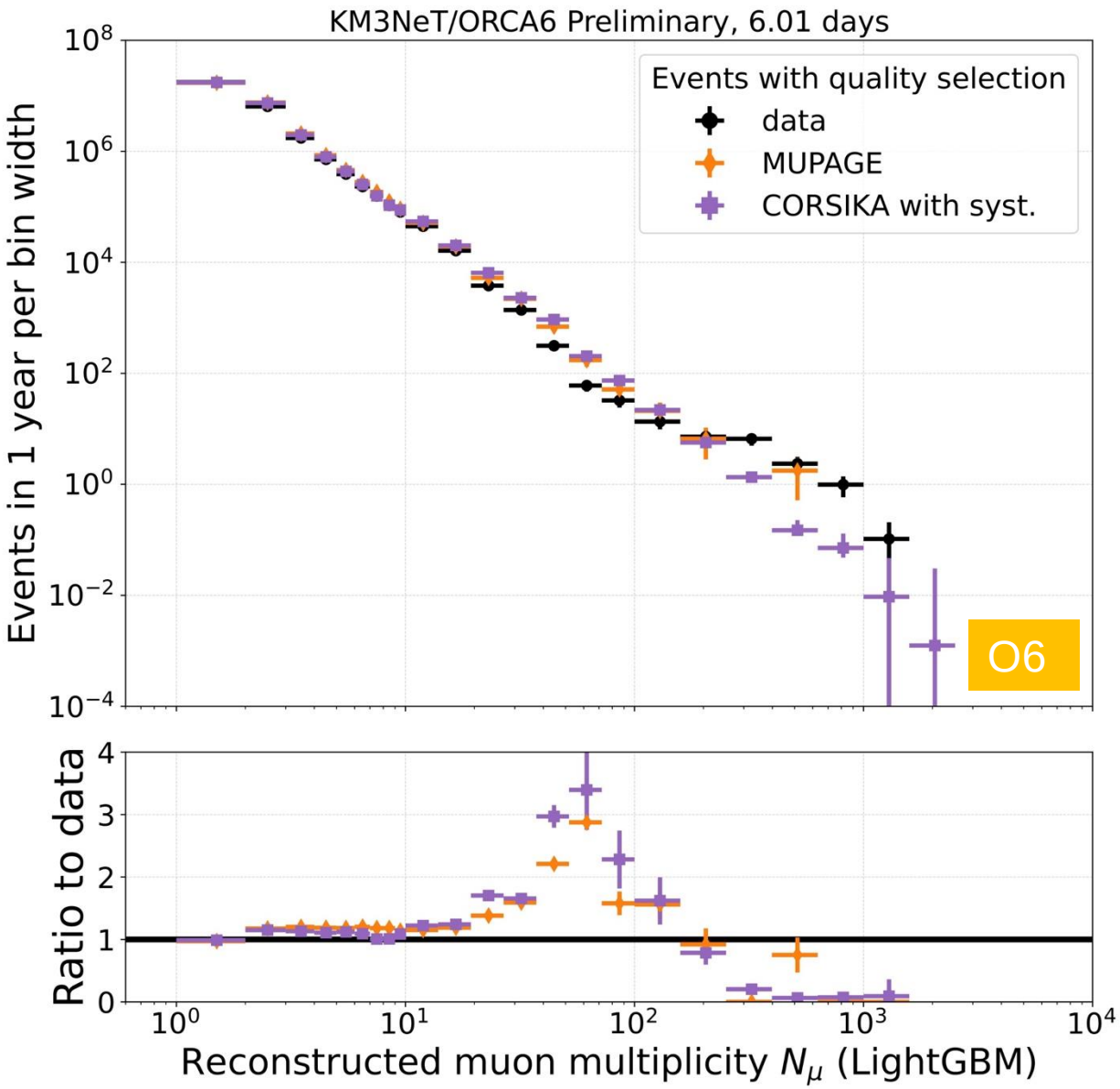


1D histograms



Multiplicity reconstruction applied to ORCA6 data

LightGBM



Analogous results for ARCA6

Results obtained with ARCA6/ORCA6 data:

- First muon multiplicity measurement in KM3NeT:
  - Ex aequo with S. Reck (using ORCA4 data)
  - First ever for ARCA
- Mid-multiplicities overestimated in simulations
- High multiplicities underestimated in simulations

Outline

Introduction

Muon event reconstruction

Summary

More in my Thesis: arXiv:2402.02620v1
and upcoming gSeaGen paper …

Summary:

- KM3NeT detectors under construction, but already collecting valuable data
- Accurate muon simulations are crucial
- Reconstructing N_μ, E_{bundle}, E_{prim} possible

Outlook:

- Even lower level information as features?
- Reconstruction of individual muon tracks?

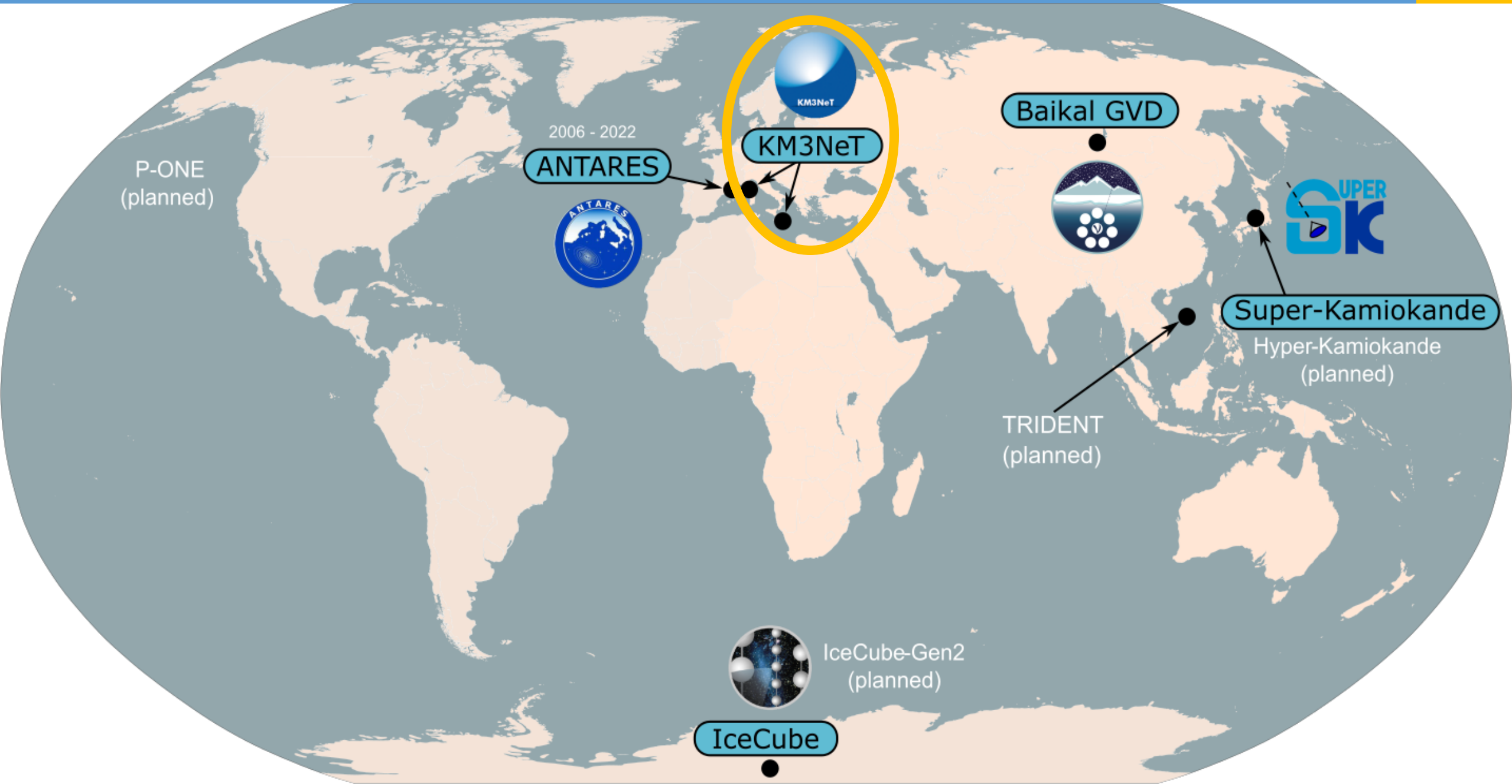
Thank you for your attention!

KM3NeT

European Funds Smart Growth
Republic of Poland
FNP Foundation for Polish Science
European Union European Regional Development Fund

Backup

Water Cherenkov ν telescopes



Digital Optical Module (DOM)

acrylic glass sphere with:

- 31 3” PMTs,
- readout electronics,
- pressure gauge,
- acoustic sensonrs,
- ...

2022 JINST 17 P0703

JATIS 7(1), 016001 (2021)

Photomultiplier Tube (PMT)

converts light into electric signal

JINST13 (2018) P05035

DOM arrangement

Detection Unit (DU):

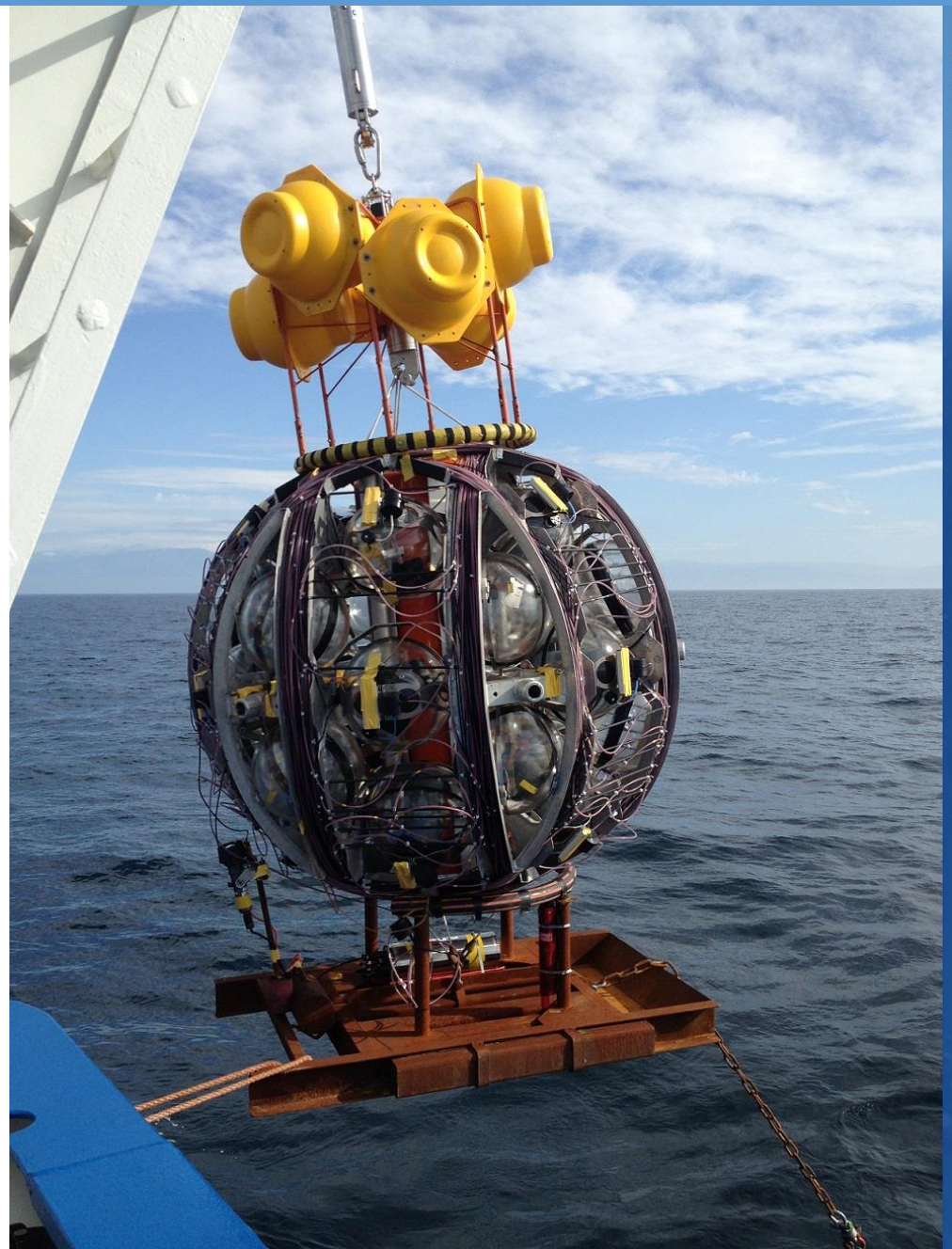
vertical **string** with 18 DOMs

Eur. Phys. J. C 76 (2016) 76:54

Naming:

ORCA6 ↔ ORCA with 6 strings
ARCA2 ↔ ARCA with 2 strings
etc.

2020 JINST 15 P11027



DOM arrangement

Detection Unit (DU):

vertical **string** with 18 DOMs

Eur. Phys. J. C 76 (2016) 76:54

Naming:

ORCA6 ↔ ORCA with 6 strings
ARCA2 ↔ ARCA with 2 strings
etc.

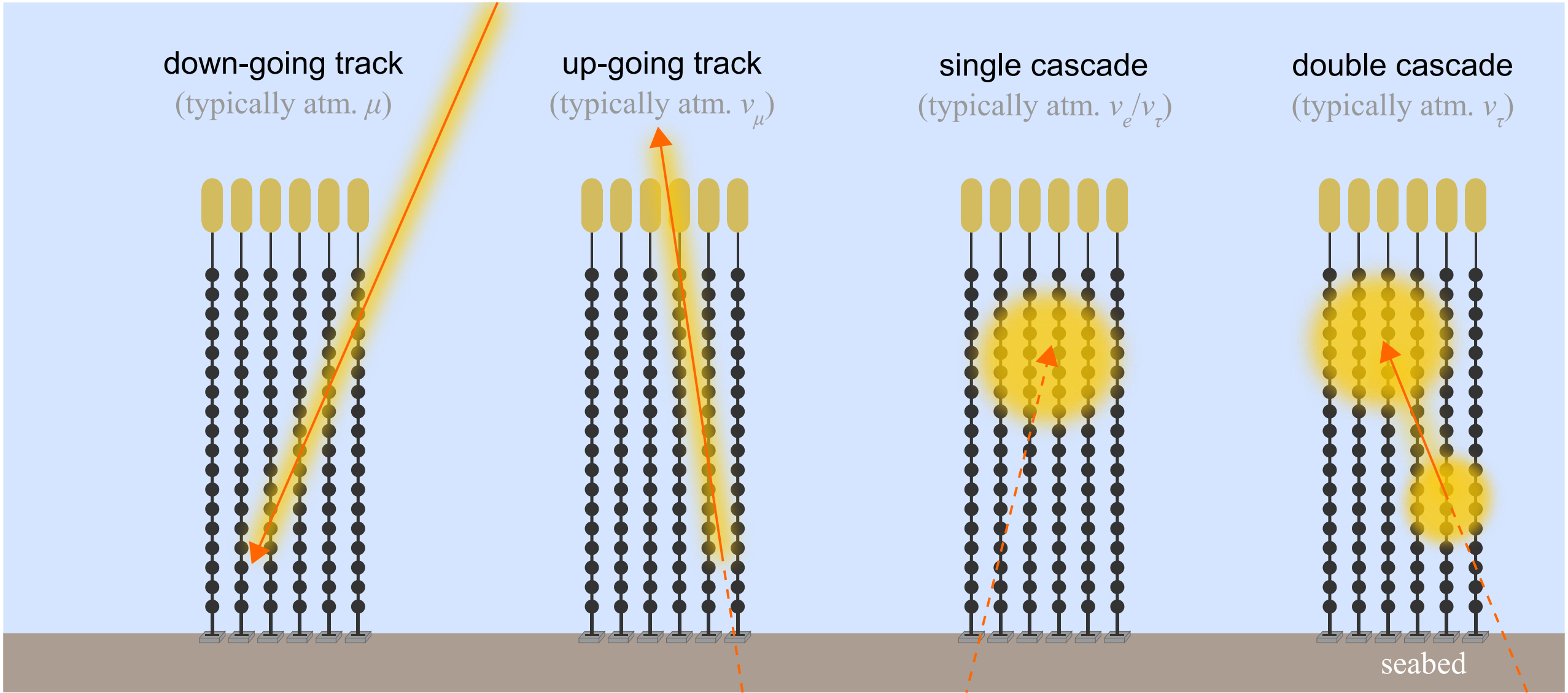
2020 JINST 15 P11027

Event topologies

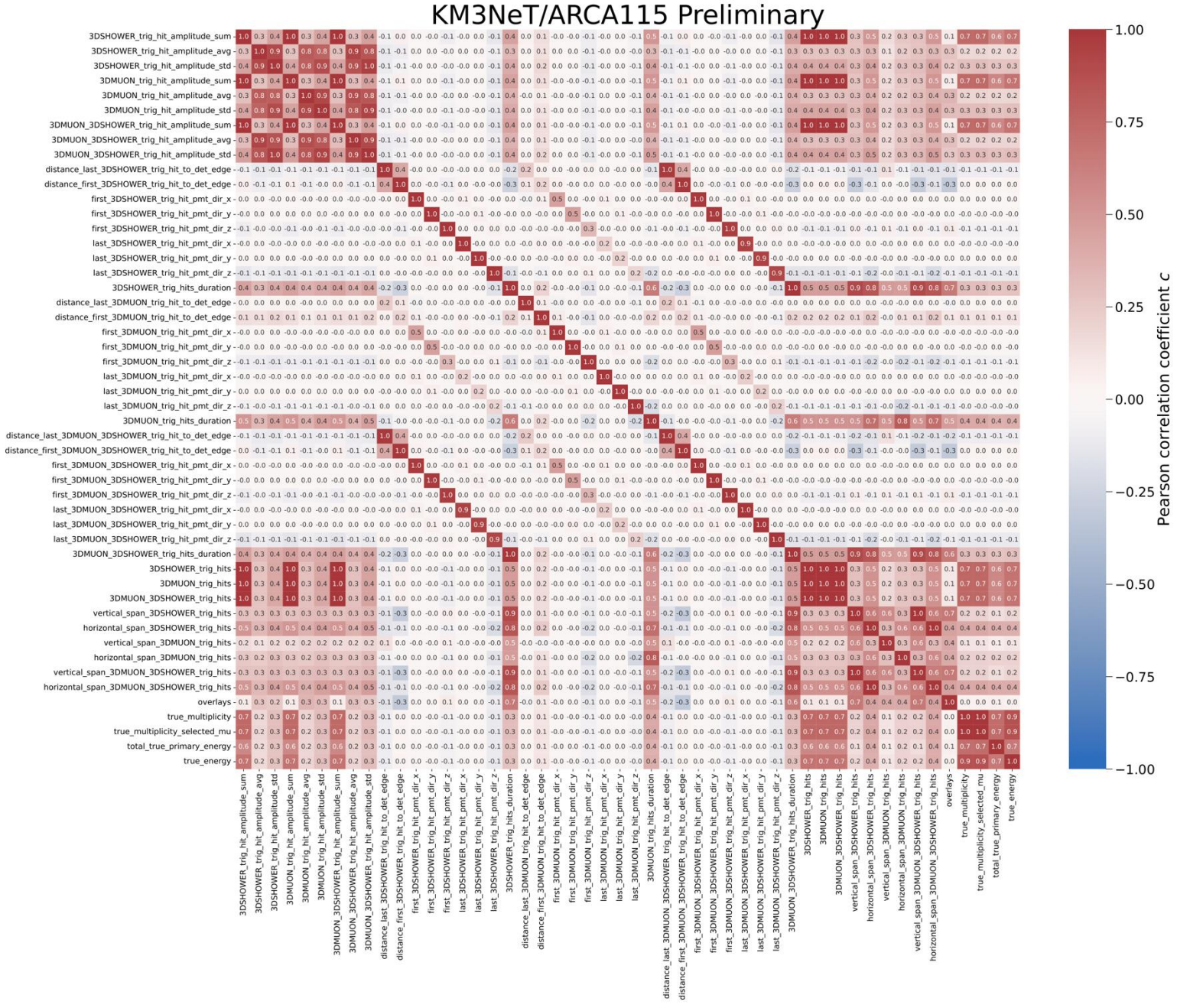
Examples of basic event topologies:

Classes based on combinations of:

- Direction
- Shape



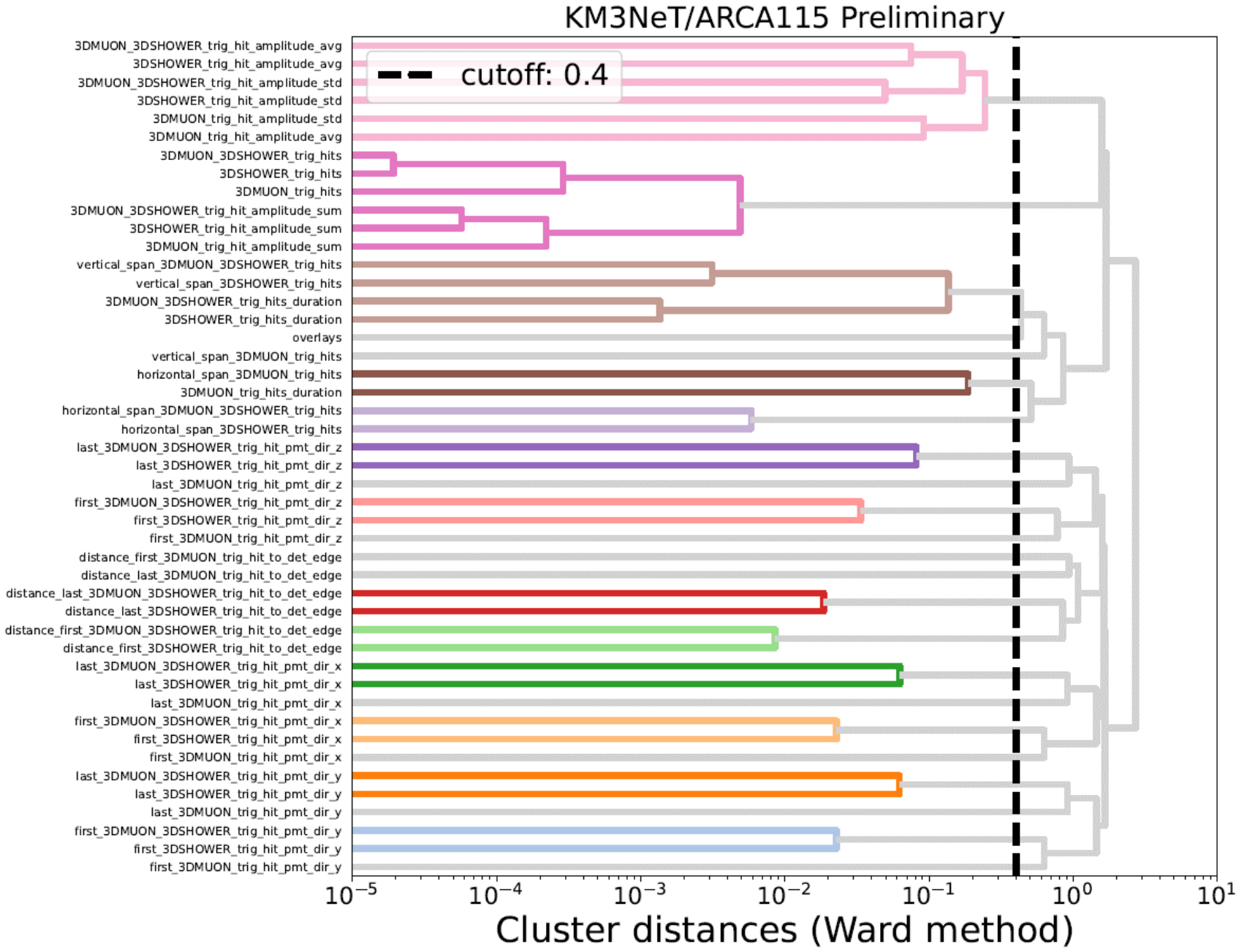
ML reco: features used for reconstruction



In total: 46 features (+4 targets)

Example for ARCA115
(the same was done for ARCA6, ORCA115 and ORCA6)

ML reco: feature clustering



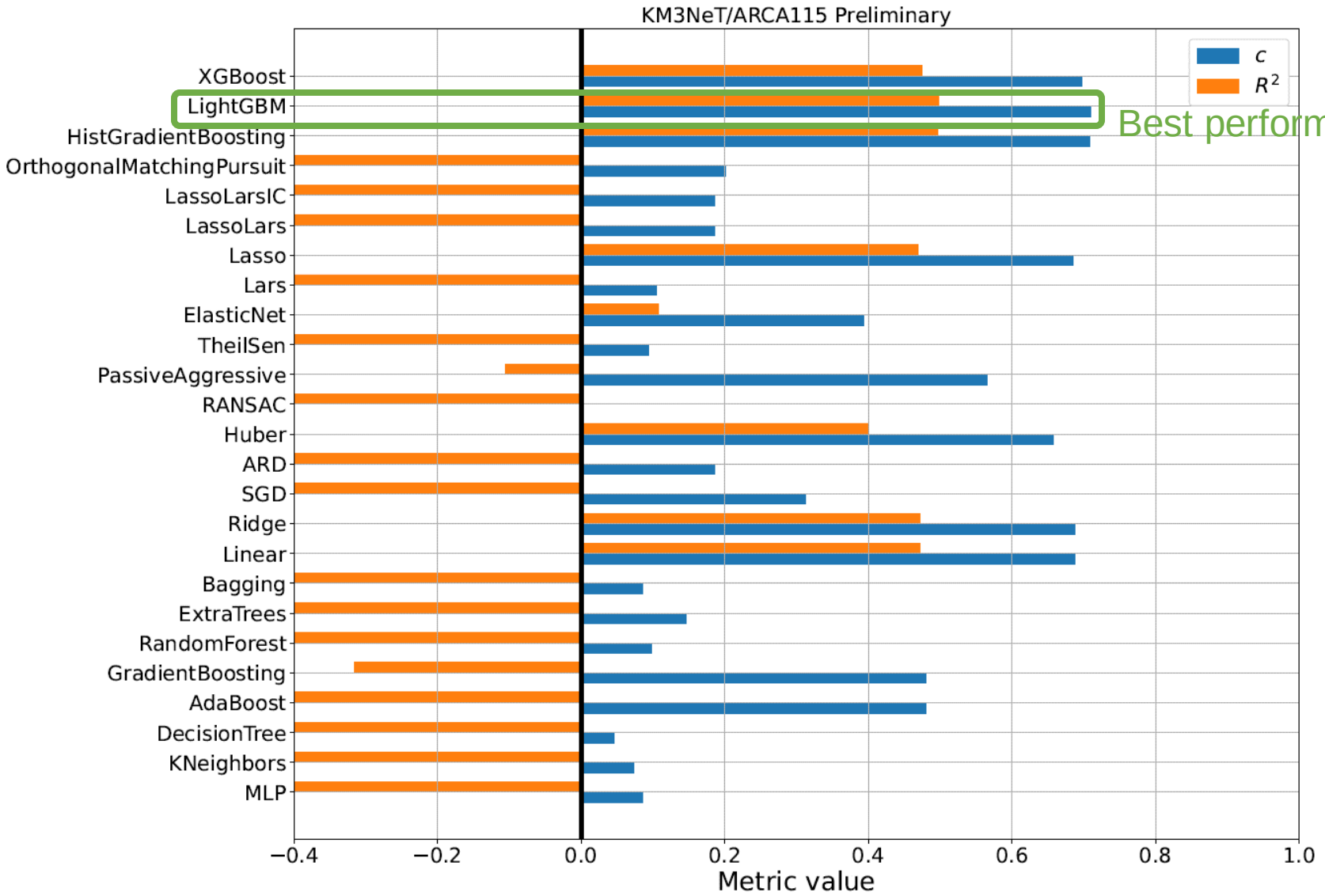
Cluster distance cutoff is arbitrary

Clusters are marked by different colors

Example for ARCA115
(the same was done for ARCA6, ORCA115 and ORCA6)

Bundle energy reco: best ML model selection

Performance comparison on a fraction (50k events) of the training dataset:



Best performance

The (weighted) Pearson correlation coefficient:

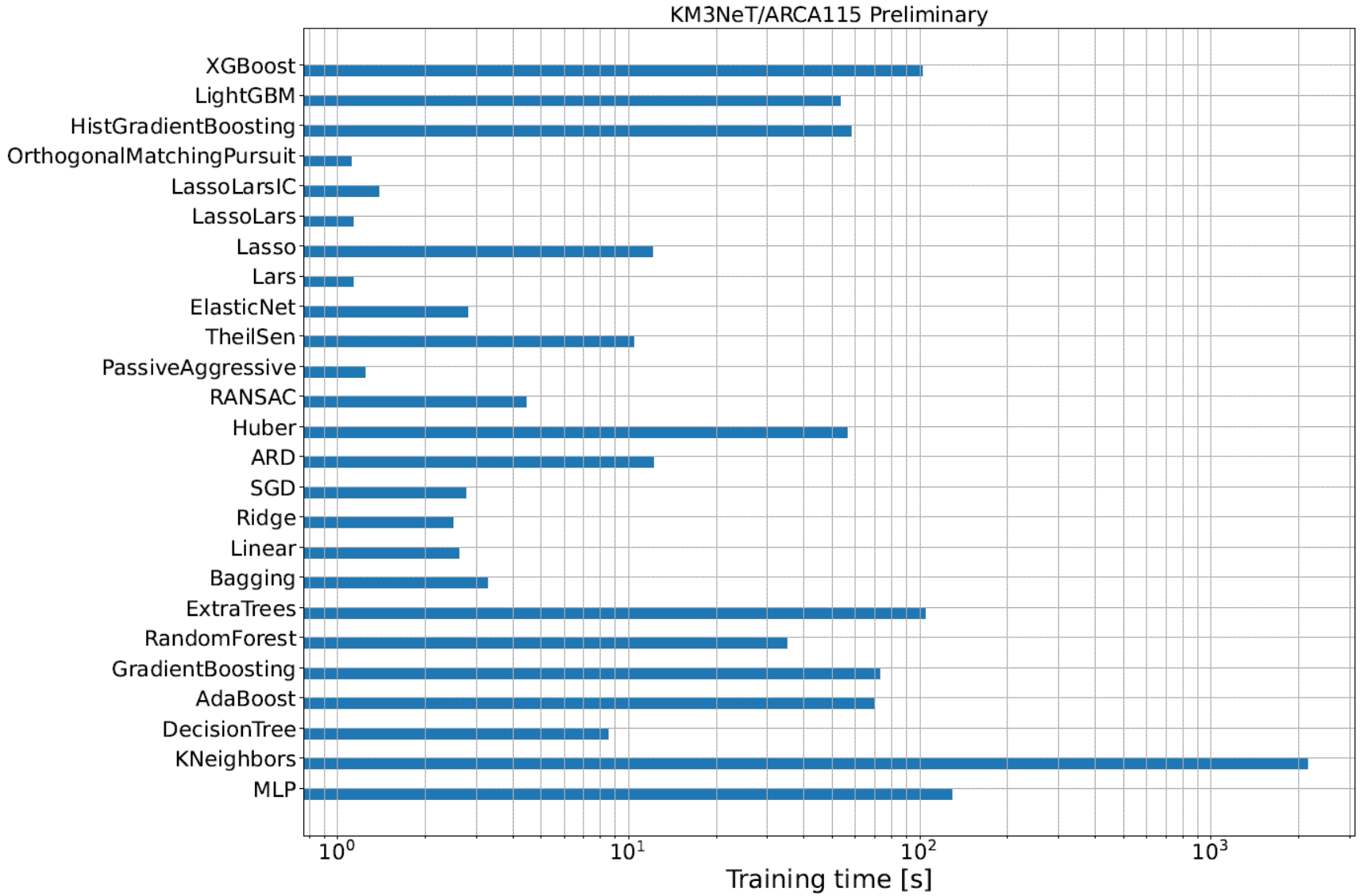
$$c(x,y) = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i w_i (x_i - \bar{x})^2 \sum_i w_i (y_i - \bar{y})^2}}$$

The (weighted) R2-score,
called the coefficient of determination:

$$R^2(y_{\text{true}}, y_{\text{pred}}) = 1 - \frac{\sum_i w_i (y_{\text{true}} - y_{\text{pred}})^2}{\sum_i w_i (y_{\text{true}} - \bar{y})^2}$$

Bundle energy reco: best ML model selection

Speed comparison on a fraction (50k events) of the training dataset:

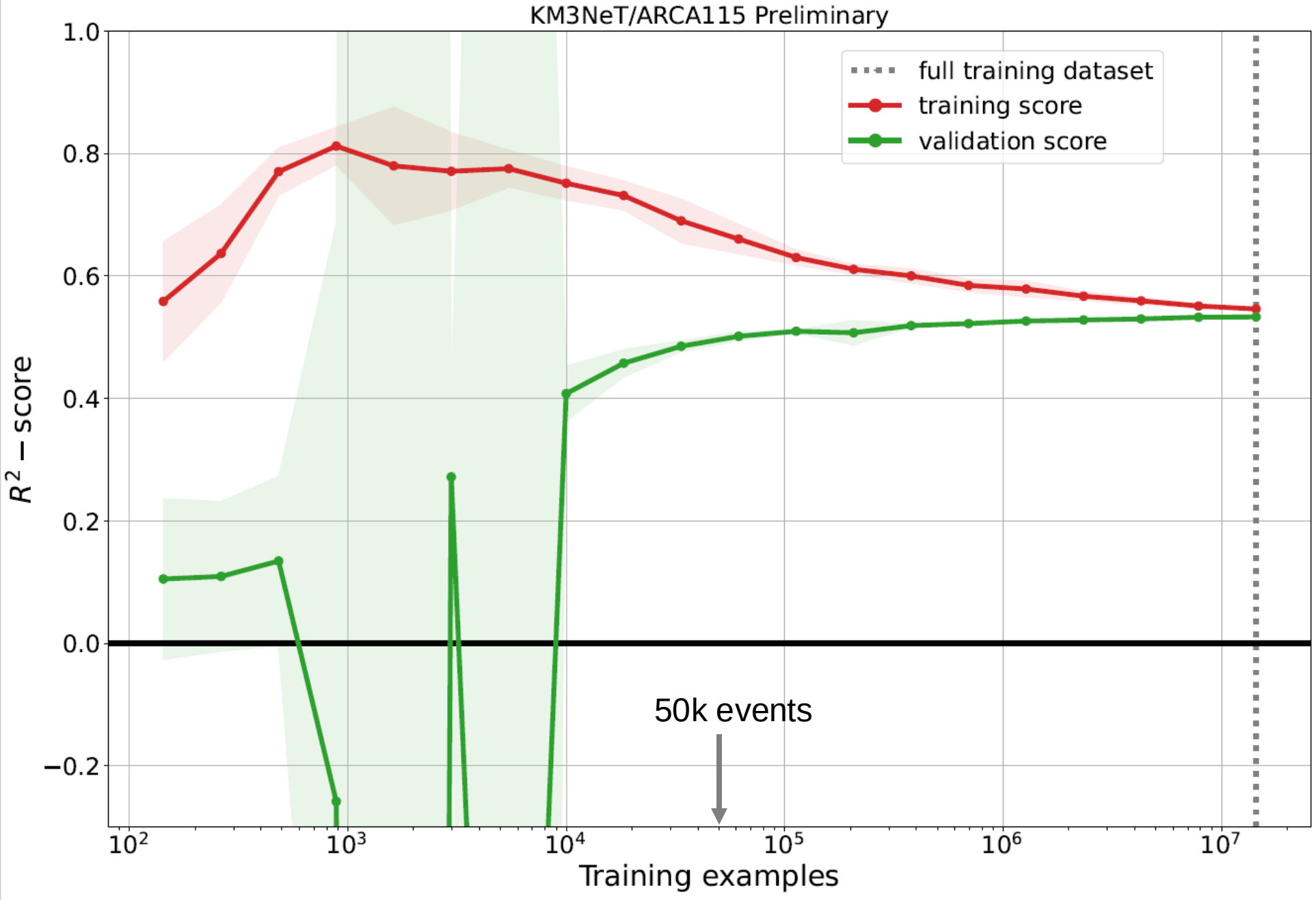


LightGBM:

- ❖ not the fastest, but still very decent
- ❖ + it turned out to scale up very well (entire dataset is orders of magnitude larger)

These times were obtained running with 20 CPU cores in parallel

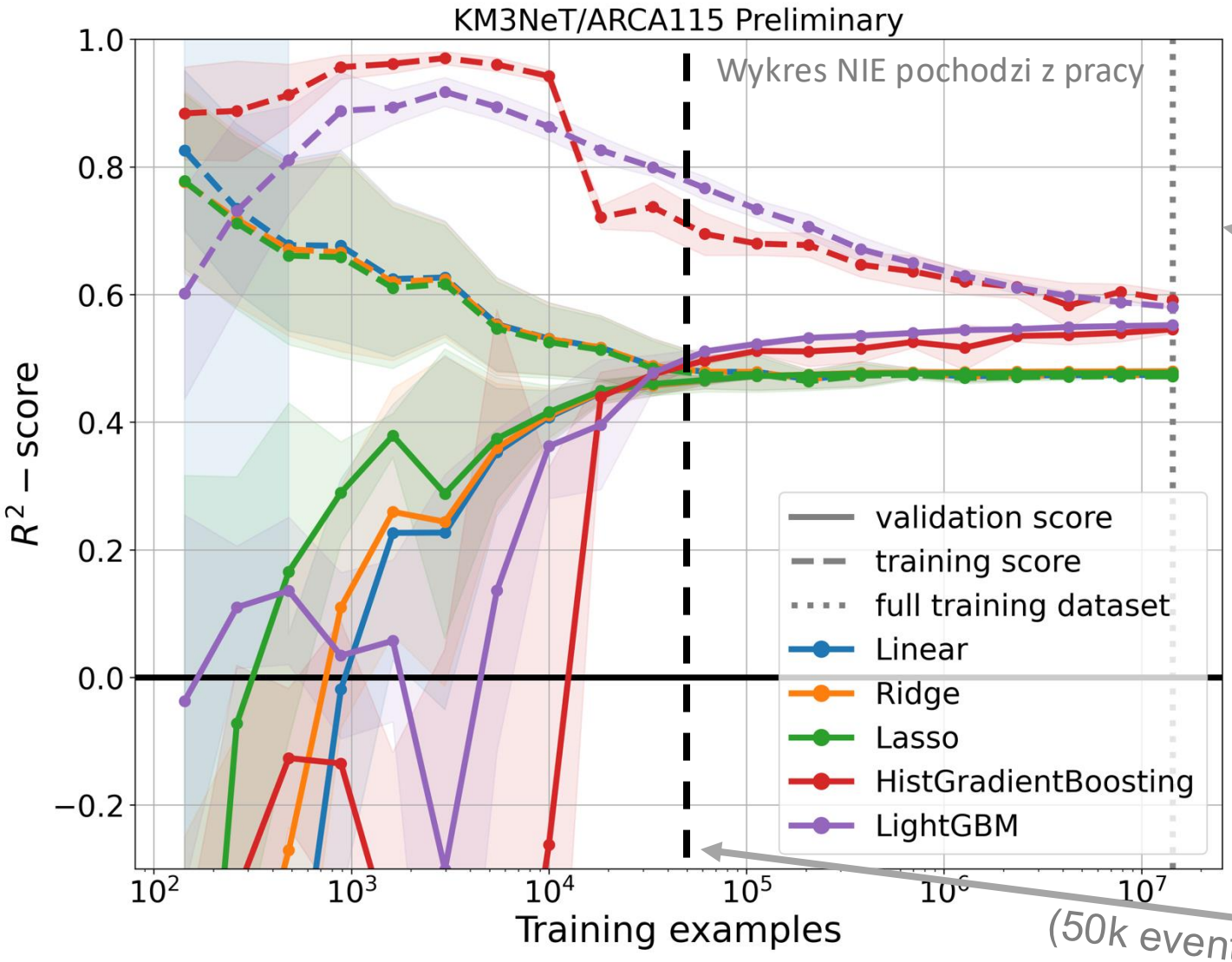
Bundle energy reco: learning inspection



Here we see why 50k events were fine for testing (but e.g. 5k would not be)

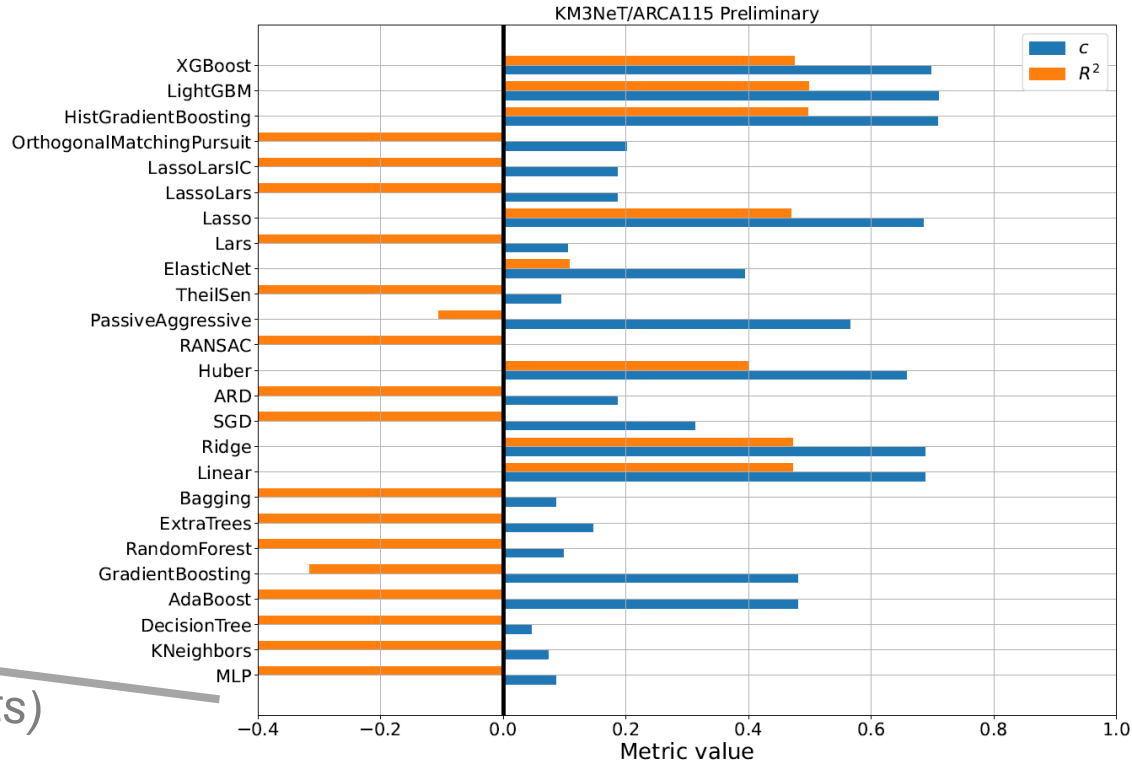
Why not linear model?

Linear does <u>almost</u> as good as LightGBM:



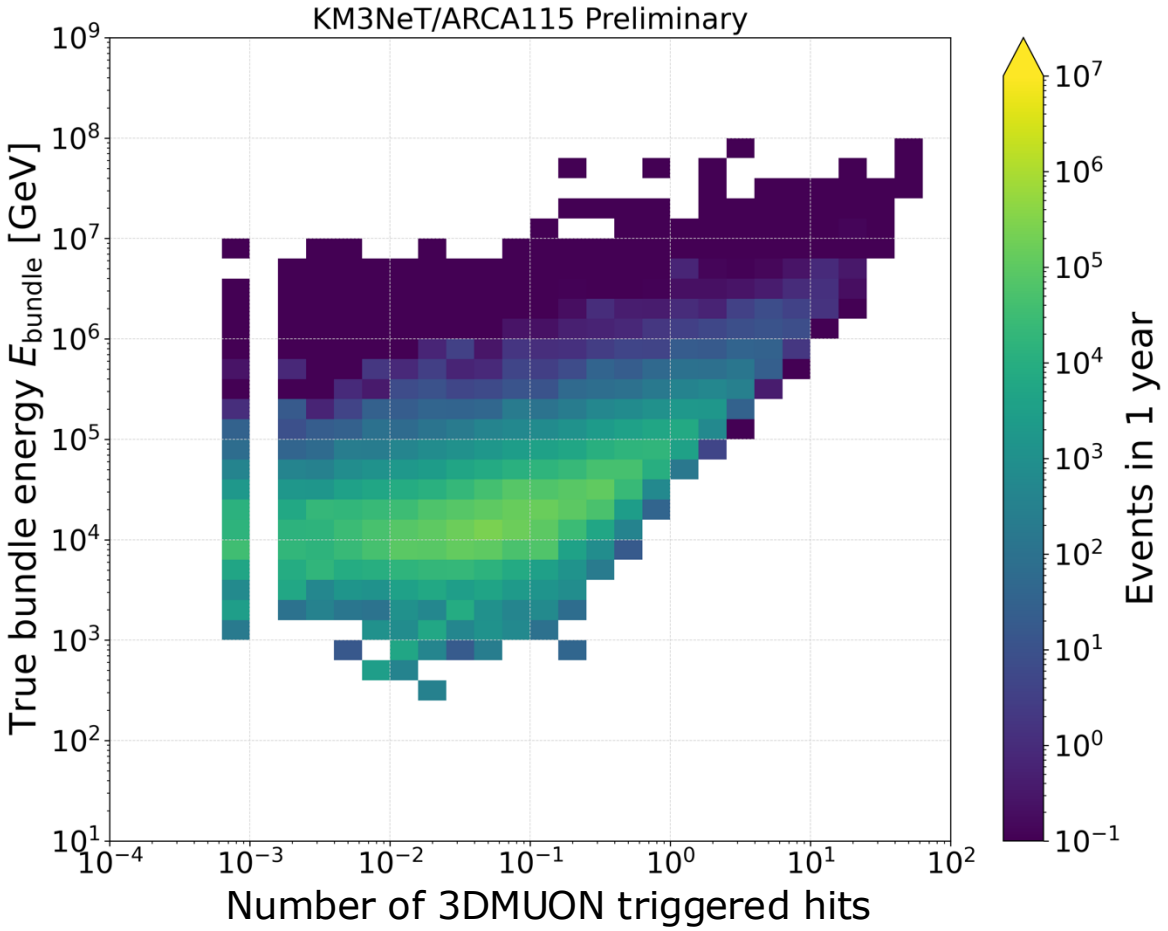
same hyperparameters

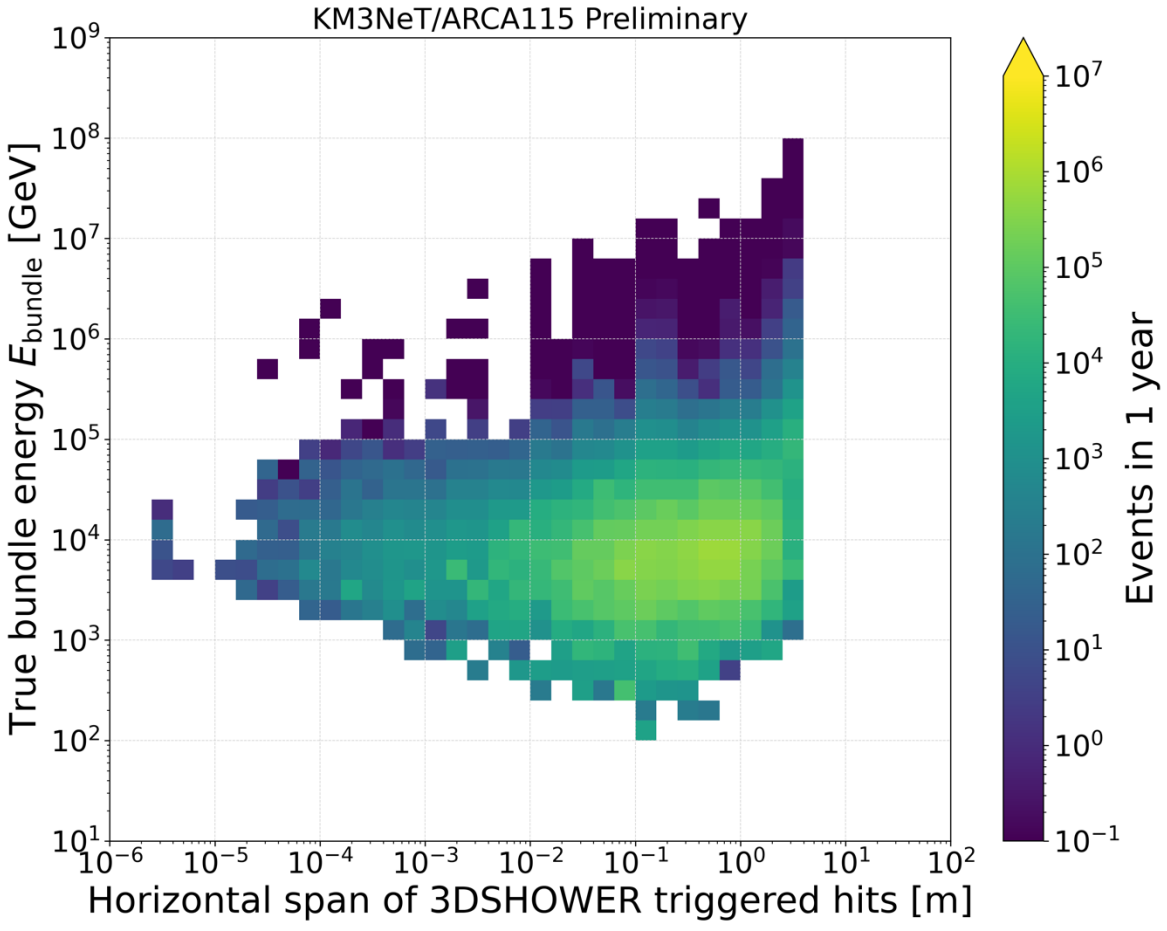
(50k events)



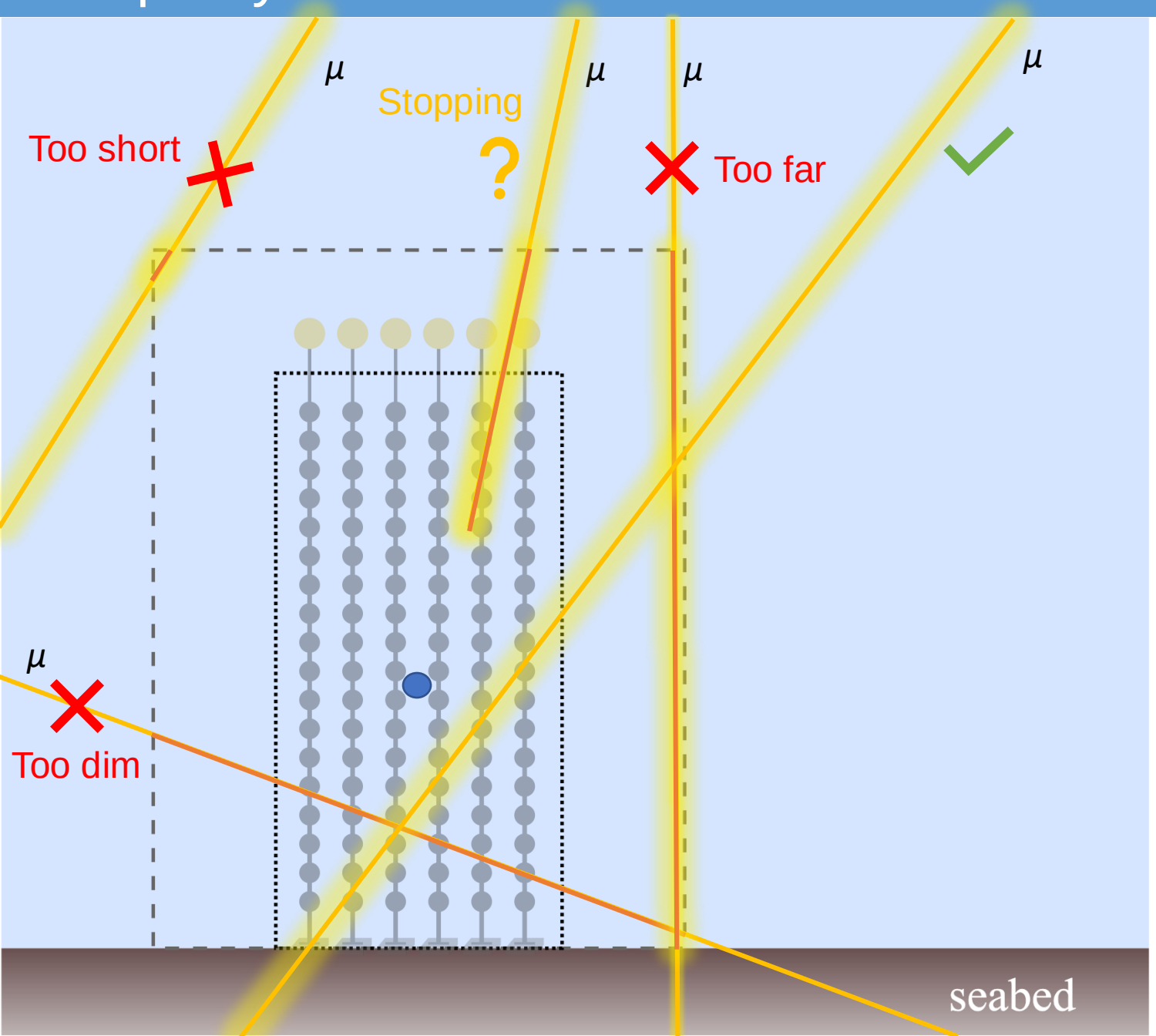
Why not linear model?

- Performance of Linear could not be further improved:
  - Flat learning curve (NOT for LightGBM!)
  - No hyperparameters to tune
- Rekonstructed observables are non-linear functions of features:
  (so Linear has no chance to fully describe them)





Multiplicity reco: muon selection



We want to exclude muons, which:

- Are too far from the detector
- Have too short pathlength inside the volume of interest
- Emit too faint light (have too low Energy)
- Basically are not visible or would be poorly reconstructed

How?

- Check the JMuon* likelihood $\mathcal{L}$ for single muon events against:
  - distance of muon from the DET center (●) for vertical muons → pick an optimal volume by shrinking the can by x as:
    $r_{\mathrm{can}} - x,\ h_{\mathrm{can}} - x$
  - muon pathlength L but for shrinked can
  - muon energy → E cut

Multiplicity reco: muon selection

Summary of the selction:

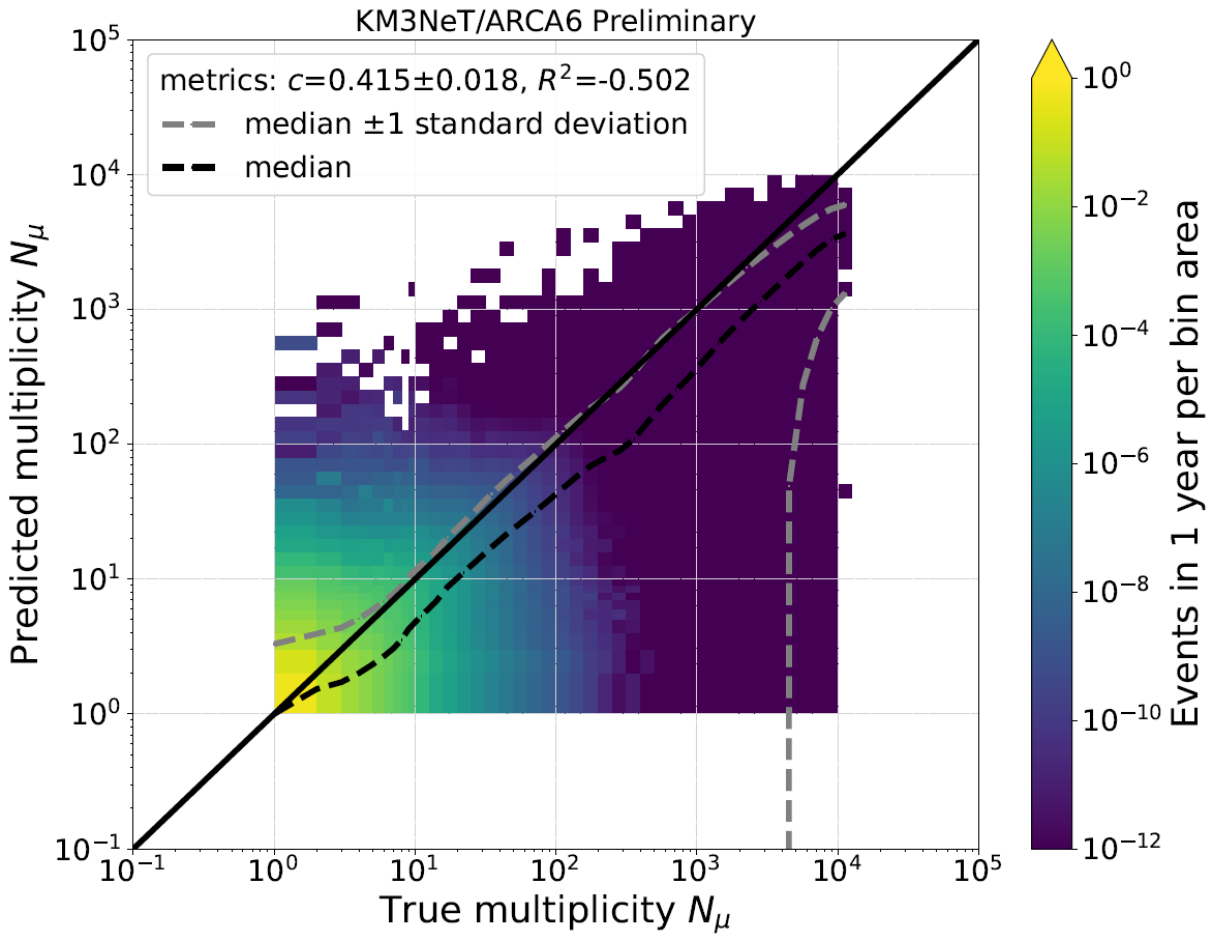
Detector	Minimal E_μ [GeV]	$d_{\max}$ [m]	minimal L_μ [m]
ARCA115	120	-	-
ARCA6	120	269.4	240
ORCA115	1	-	-
ORCA6	1	-	-

This selection is used for further multiplicity results

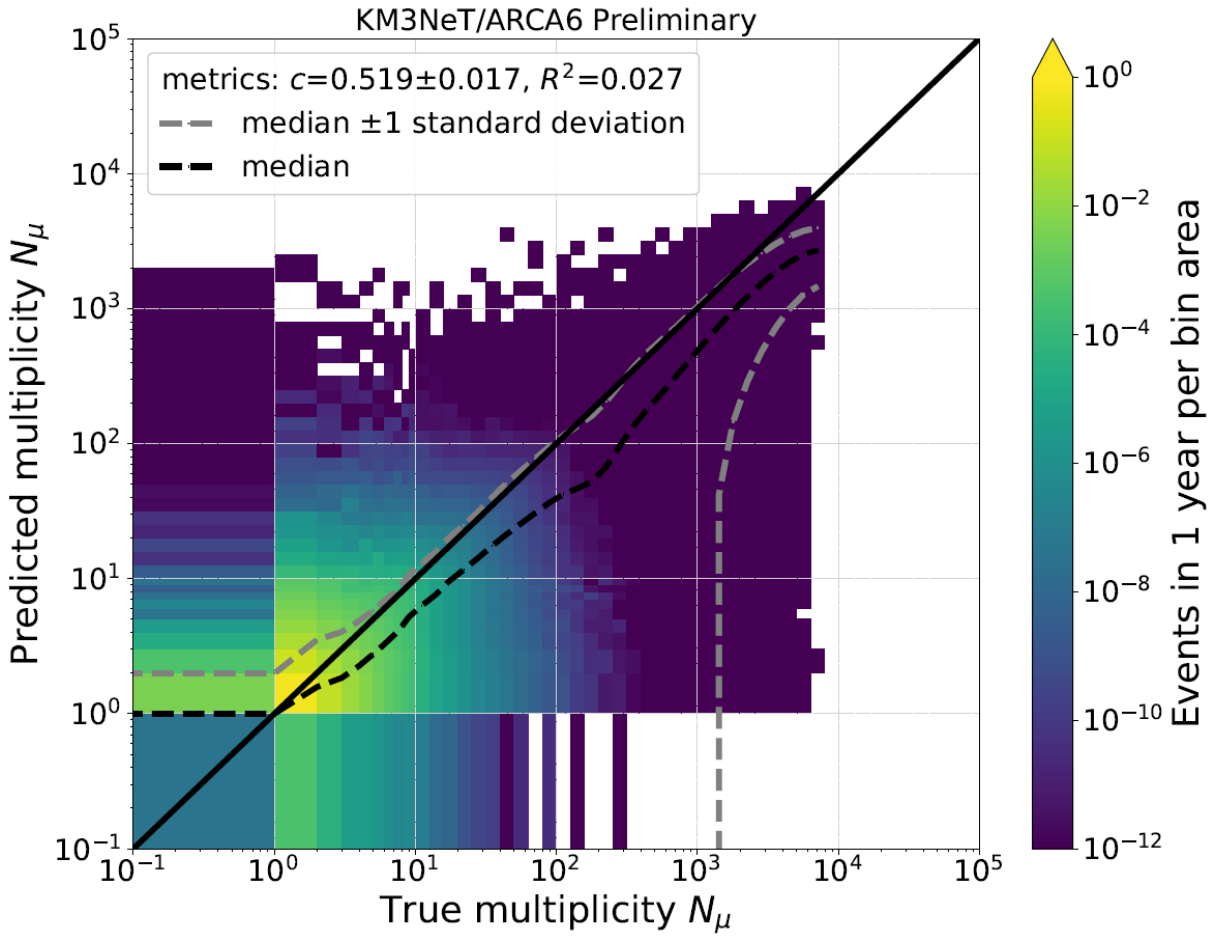
Multiplicity reco: muon selection check

Example of ARCA6, for which the effect is the most pronounced

All muons



Selected muons

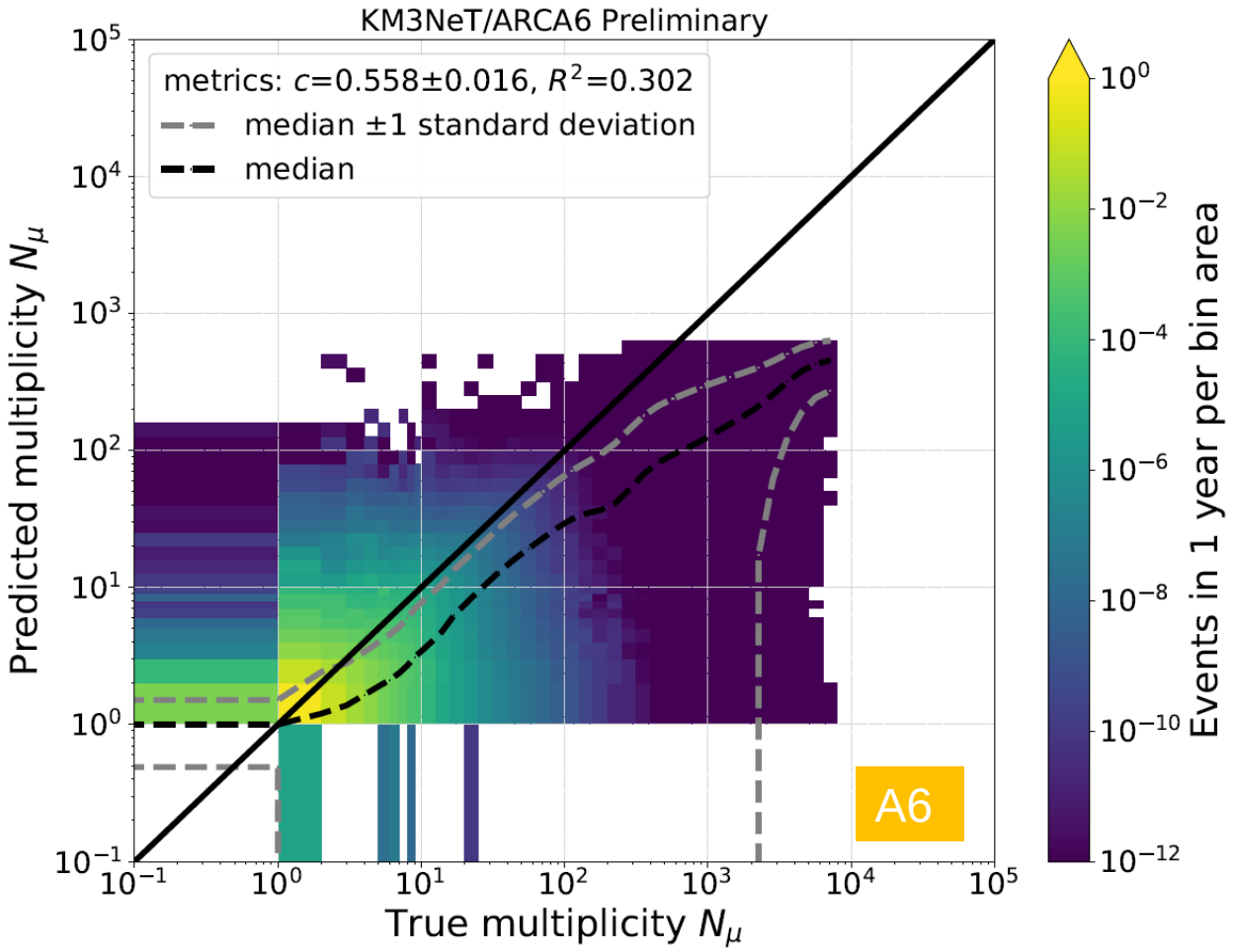


Multiplicity reco: final results

Analogical results obtained for ARCA115, ORCA115 and ORCA6

Example of the results for ARCA6:

2D: pred vs true



1D histograms