Composition of 100 TeV - 100 PeV Cosmic Rays with IceCube and IceTop using Boosted Decision Trees

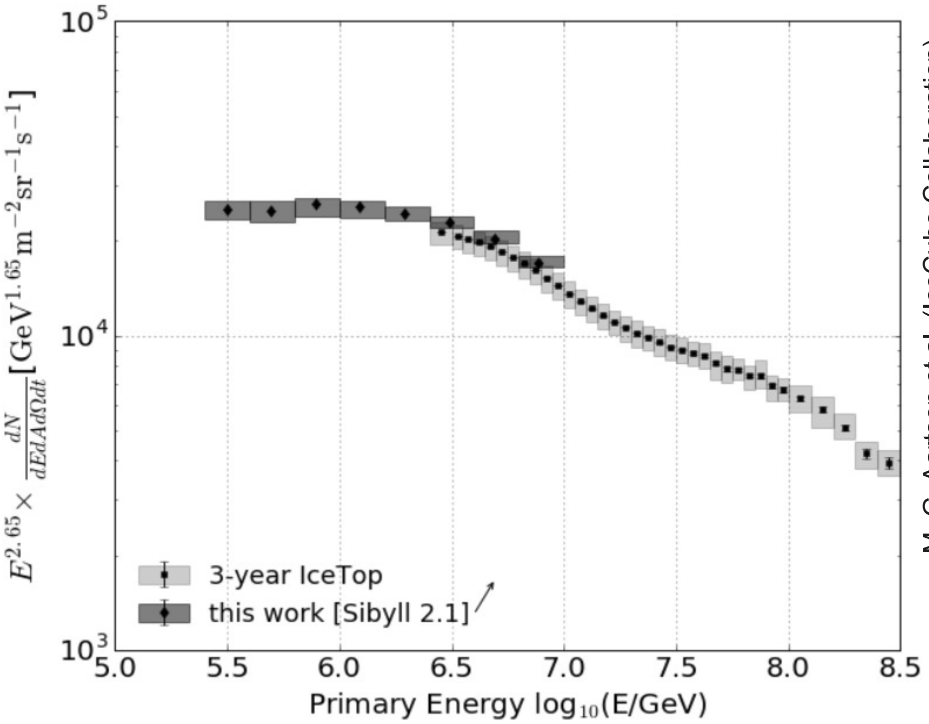
Julian Saffer - February 1st, 2022

Workshop on Machine Learning for Cosmic-Ray Air Showers, Newark, DE, USA

julian.saffer@kit.edu

Low-Energy Cosmic-Ray Events at IceCube

- Low-energy trigger used for CR spectrum down to 250 TeV
- Main contribution to uncertainty: composition
- Idea: improve this previous Random Forest analysis to other techniques and include in-ice signature

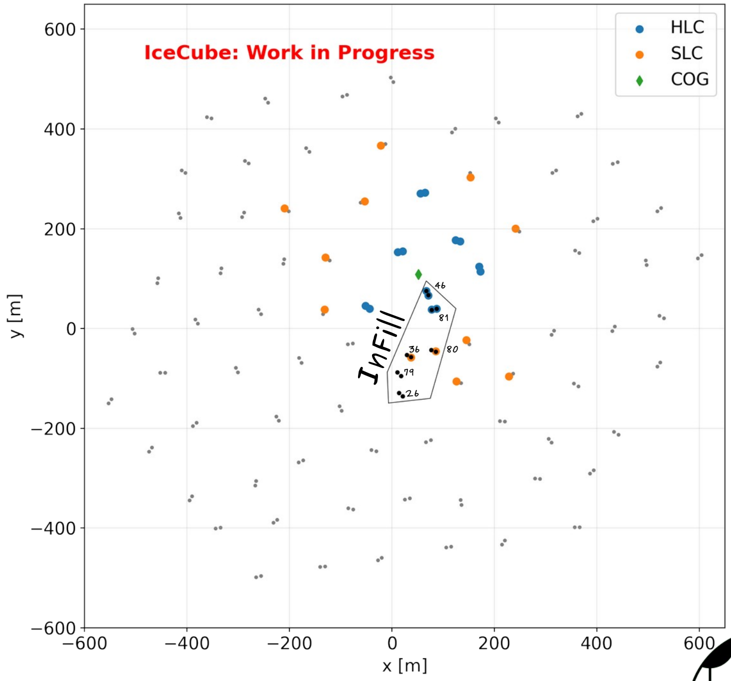


M. G. Aartsen et al. (IceCube Collaboration)
Phys. Rev. D 102, 122001

Low-Energy Cosmic-Ray Events at IceCube

Karlsruher Institut für Technologie

- CR-induced air showers reaching down to $\mathcal{O}(10^5$ GeV) primary energy can trigger surface station pair(s) in dense center (*InFill*)



81 IceTop stations (2 tanks each)

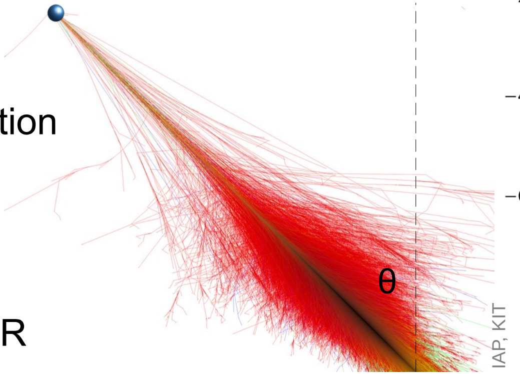
4 pairs of nearby stations in the InFill

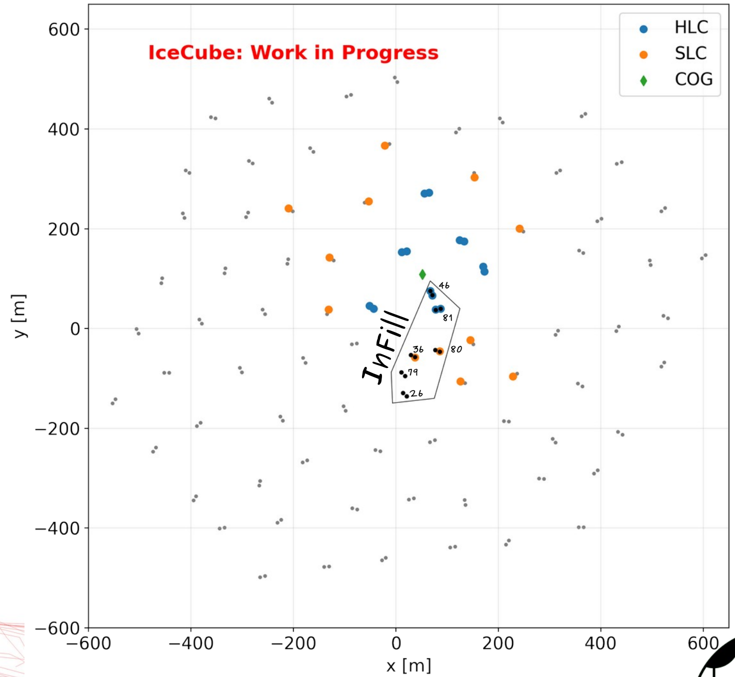
ICECUBE

Low-Energy Cosmic-Ray Events at IceCube

- CR-induced air showers reaching down to $\mathcal{O}(10^5\ \text{GeV})$ primary energy can trigger surface station pair(s) in dense center (*InFill*)

- Aim reconstruction of
  - shower core position
  - zenith angle θ
  - primary energy
  - type of primary CR



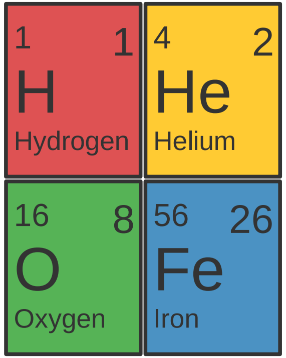


81 IceTop stations (2 tanks each)

4 pairs of nearby stations in the InFill

Data Used

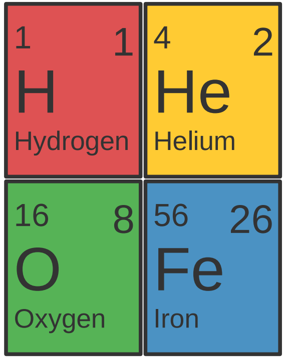
- CORSIKA simulations of 4 primary types: Proton, Helium, Oxygen and Iron
- Sibyll 2.1 interaction model
- Energy range $5.0 \leq \log_{10}(E/\mathrm{GeV}) \leq 8.0$
- Amount of events:
  - H: 3432
  - He: 3479
  - O: 3180
  - Fe: 2993
  - Σ: 13084



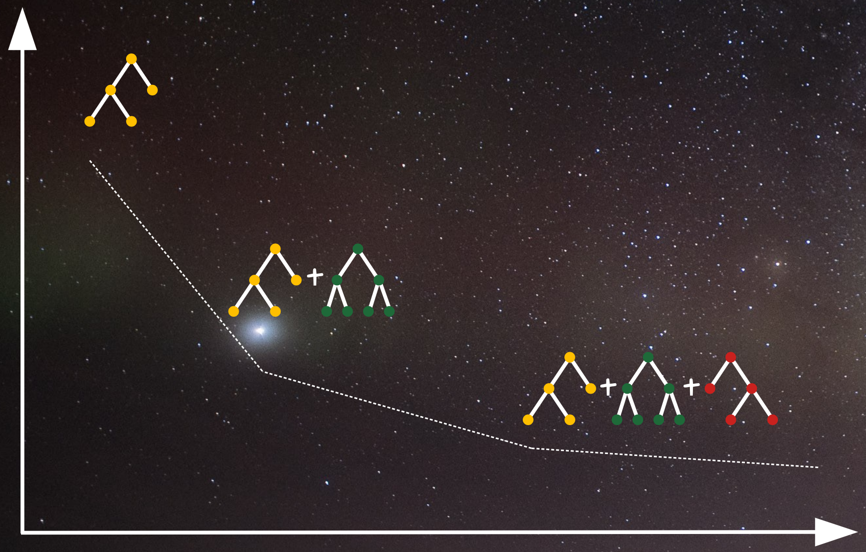
Data Used

- CORSIKA simulations of 4 primary types: Proton, Helium, Oxygen and Iron
- Sibyll 2.1 interaction model
- Energy range $5.0 \leq \log_{10}(E/\mathrm{GeV}) \leq 8.0$
- Amount of events:

H:	3432	(in-ice: 1877)
He:	3479	(in-ice: 1967)
O:	3180	(in-ice: 1899)
Fe:	2993	(in-ice: 1816)
Σ:	13084	(in-ice: 7559)

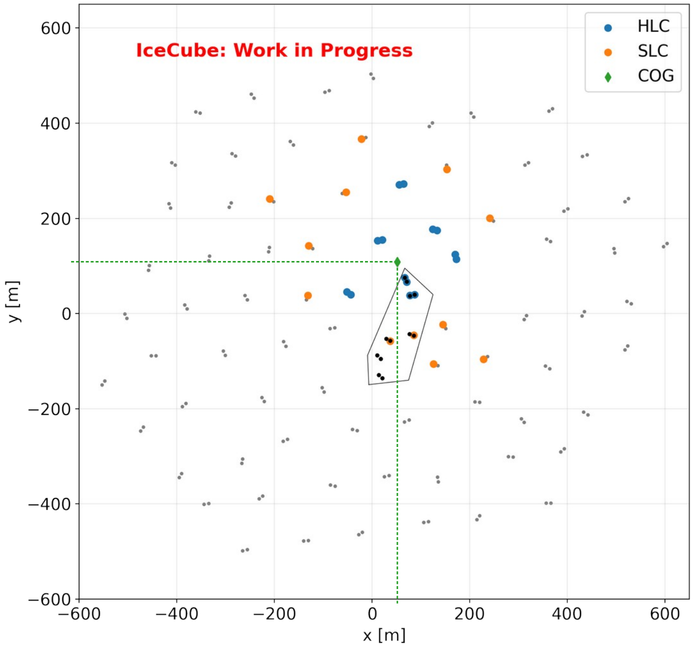


Boosted Decision Trees



BDT for Shower Core

- Two independent models for reconstruction of x- and y-coordinate
- Input features:
  - x-coordinate of center-of-gravity (COG)
  - y-coordinate of COG
  - *cos* of zenith from plane-front fit
  - *log* of number of stations with HLC hits
- Target: Monte-Carlo x resp. y



BDT for Shower Core

- `GradientBoostingRegressor`
- Test size of 40%

BDT for Shower Core

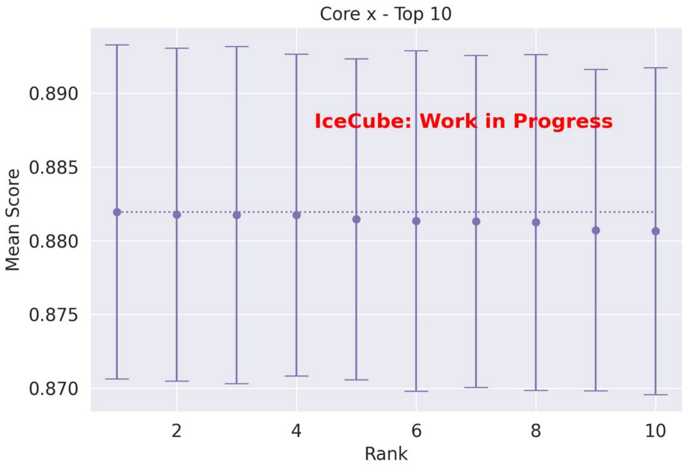
- `GradientBoostingRegressor`
- Model hyperparameters:
  - Loss: least squares
  - *sqrt*(4) = 2 features considered at each split
  - early stopping (when loss improvement < 1e-5 for 20 iterations)
  - subsample of 90% for fitting
- Test size of 40%

- Randomized search (5-fold cross-validation, 100 parameter combinations) for
  - Learning rate (`learning_rate`) 0.001 – 0.1
  - Number of trees (`n_estimators`) 100 – 2000
  - Maximal tree depth (`max_depth`) 1 – 15
  - Minimal number of samples required for split (`min_samples_split`) 2 – 20

BDT for Shower Core

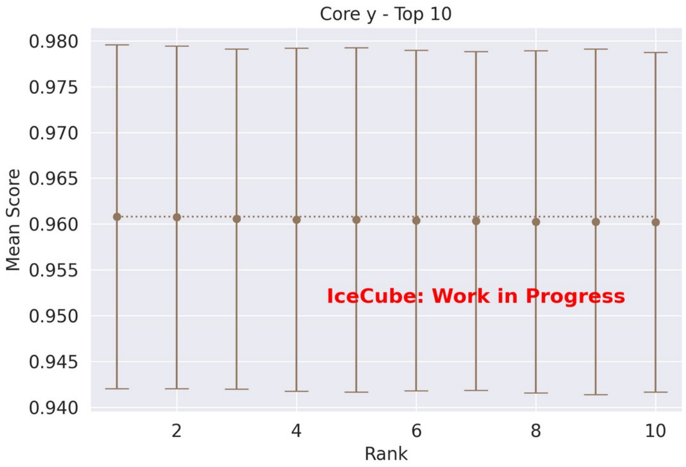
- The top 10 highest CV-scores with standard deviation
- Score: coefficient of determination

$$R^2 = 1 - \frac{\sum_i \left(y_{\mathrm{true},i} - y_{\mathrm{pred},i}\right)^2}{\sum_j \left(y_{\mathrm{true},j} - \langle y_{\mathrm{true}} \rangle\right)^2}$$



Top BDT:
learning rate: 0.0178
max depth: 7
min sam. split: 15
trees: 1795

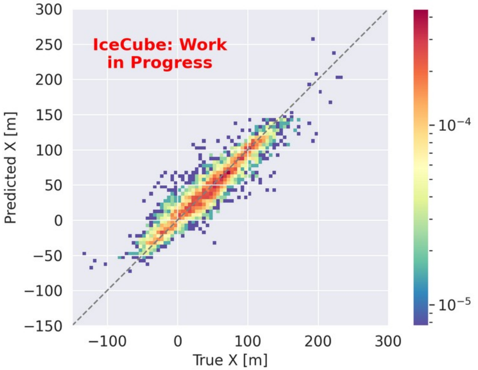
Train score: 92.15%
Test score: 82.22%



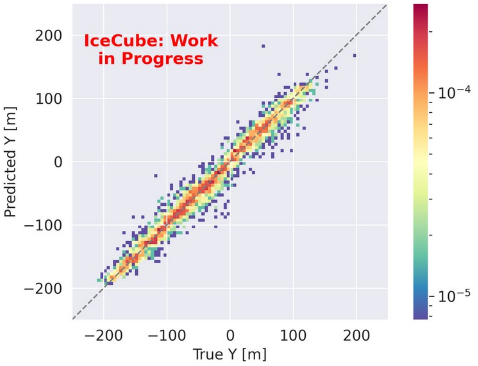
Top BDT:
learning rate: 0.0057
max depth: 6
min sam. split: 18
trees: 1731

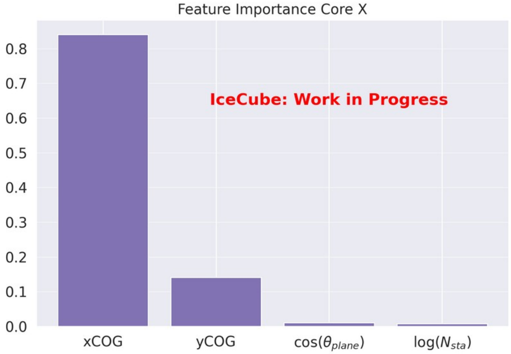
Train score: 97.45%
Test score: 96.50%

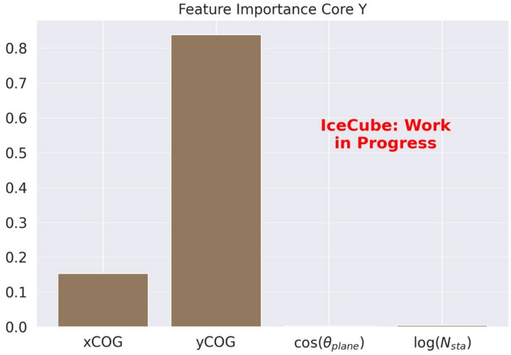
BDT for Shower Core



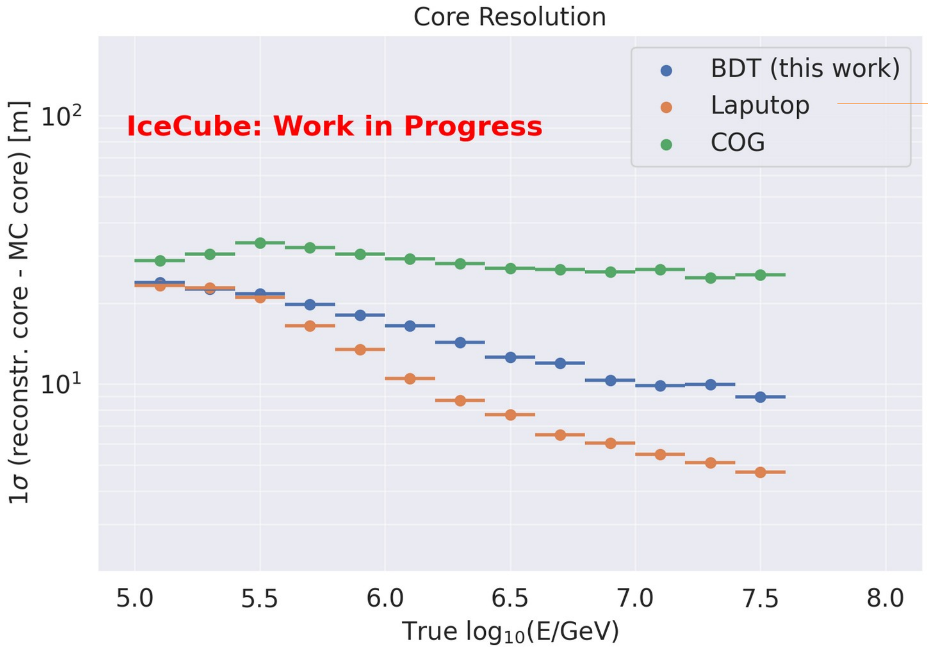
Test set







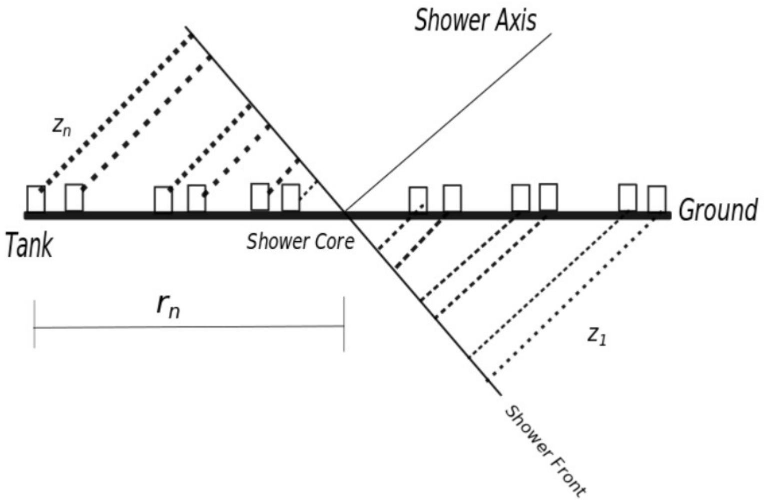
BDT for Shower Core



standard IceTop reco

BDT for Zenith Angle

- Input features:
  - x-coordinate of center-of-gravity (COG)
  - y-coordinate of COG
  - zenith from plane-front fit
  - azimuth from plane-front fit
  - average z in shower coordinates (ZSC_avg)
  - *log* of number of stations with HLC hits



R. Koirala

- Target: Monte-Carlo zenith

BDT for Zenith Angle

Karlsruher Institut für Technologie

- `GradientBoostingRegressor`
- Model hyperparameters:
  same as for shower core (*sqrt*(6) = 2)
- Test size: same
- Randomized search: same

Top BDT:

learning rate:	0.0322
max depth:	6
min sam. split:	11
# trees:	1421

Train score:	90.54%
Test score:	87.21%

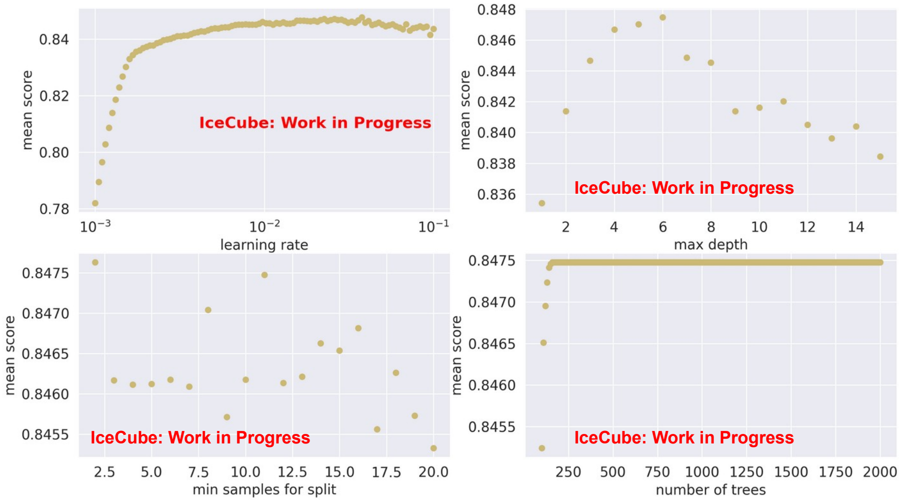
IceCube

BDT for Zenith Angle

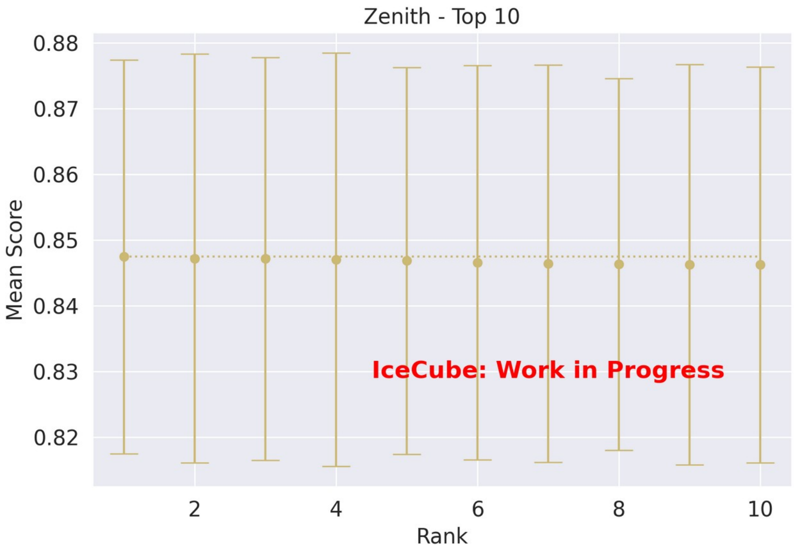
- `GradientBoostingRegressor`
- Model hyperparameters:
  same as for shower core (*sqrt*(6) = 2)
- Test size: same
- Randomized search: same

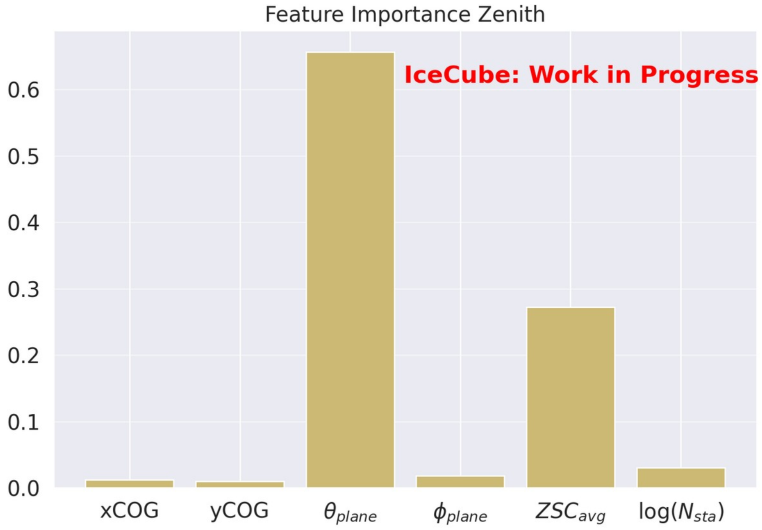
Top BDT:
learning rate: 0.0322
max depth: 6
min sam. split: 11
trees: 1421

Train score: 90.54%
Test score: 87.21%



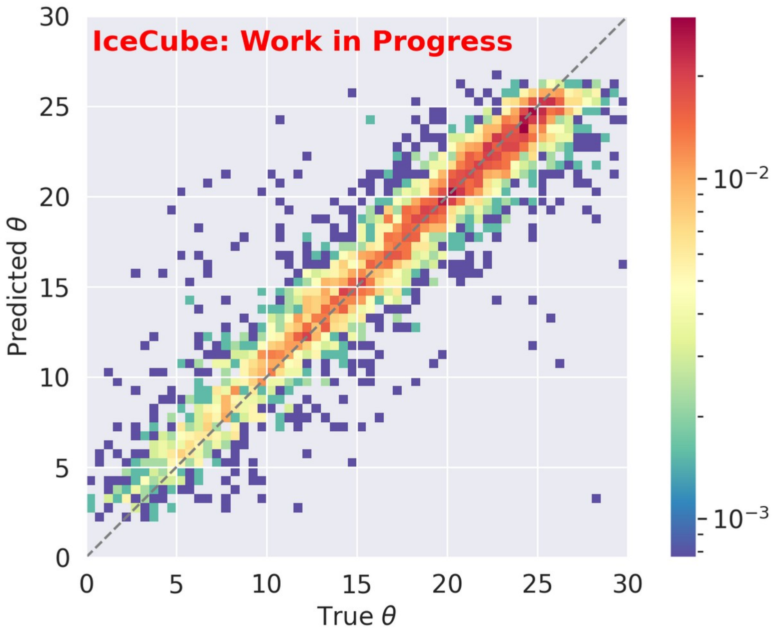
BDT for Zenith Angle

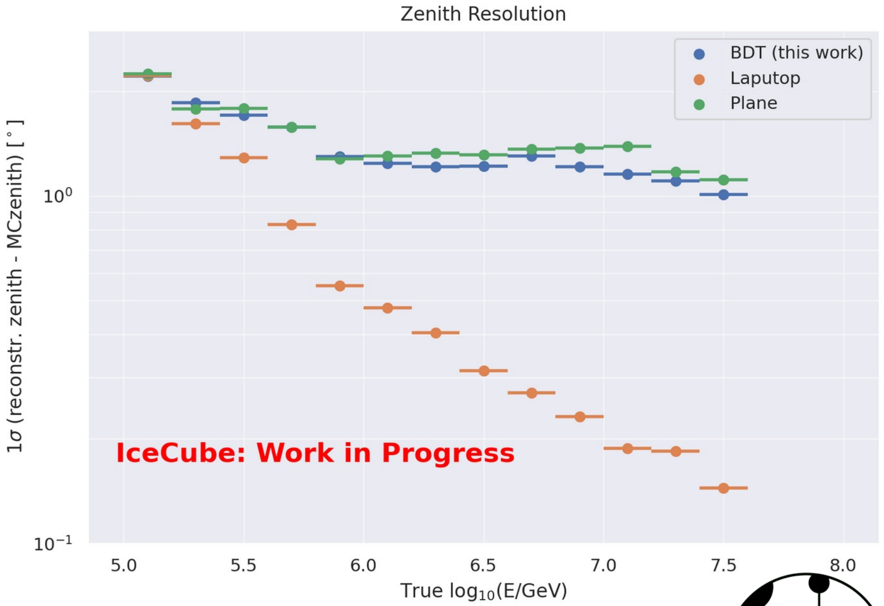




BDT for Zenith Angle

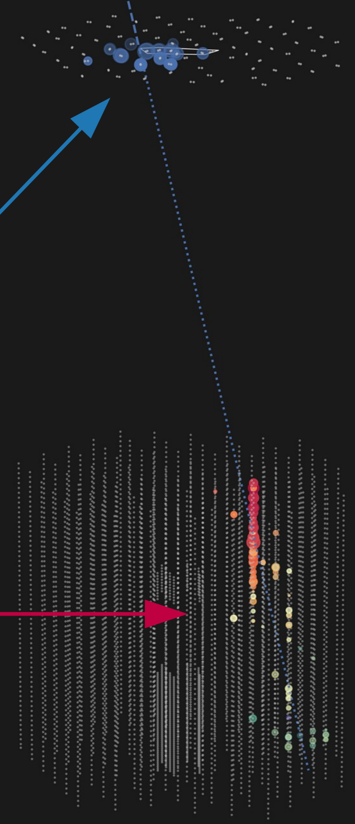
Test set





BDT for Primary Energy

- Input features:
  - *cos* of previously reconstructed zenith θ_{reco}
  - *log* of number of stations with HLC hits
  - *log* of sum of all HLC charges
  - *log* of sum of 2 highest HLC charges
  - mean distance of hit tanks from reconstructed shower core (R_{mean})
  - R_{mean} weighted with corresponding tank charges
  - *log* of number of hit in-ice DOMs



- Target: Monte-Carlo energy

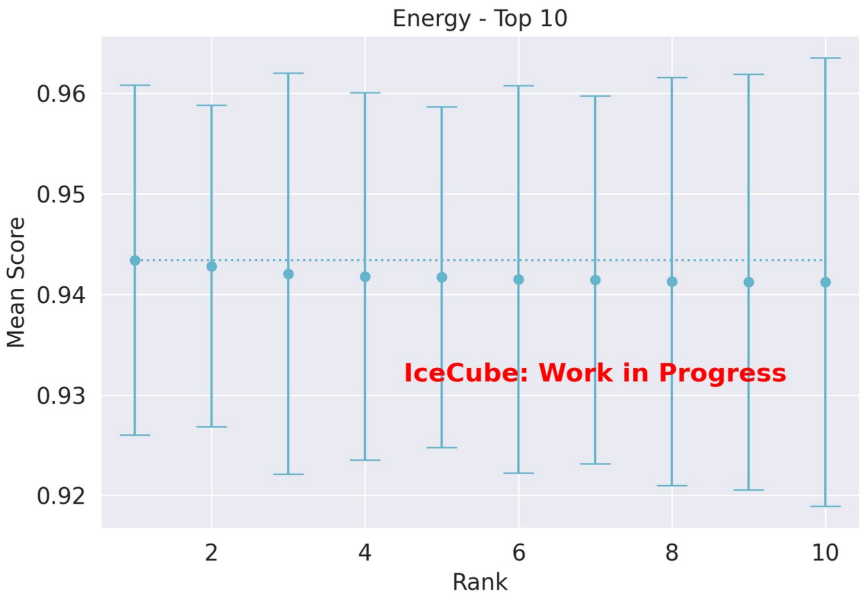
BDT for Primary Energy

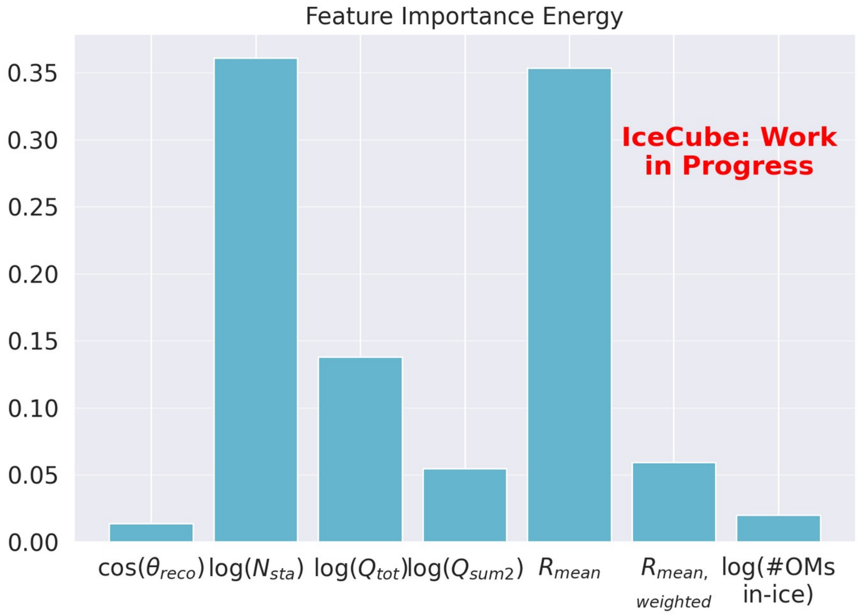
- `GradientBoostingRegressor`
- Model hyperparameters:
  same as for shower core (*sqrt*(7) = 2)
- Test size: same
- Randomized search: same

BDT for Primary Energy

Top BDT:
learning rate: 0.0307
max depth: 9
min sam. split: 8
trees: 374

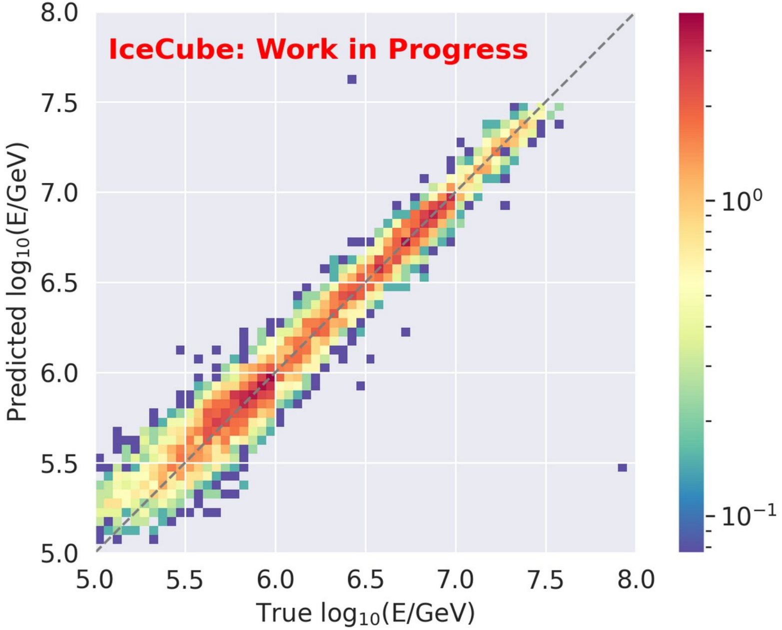
Train score: 99.28%
Test score: 89.77%

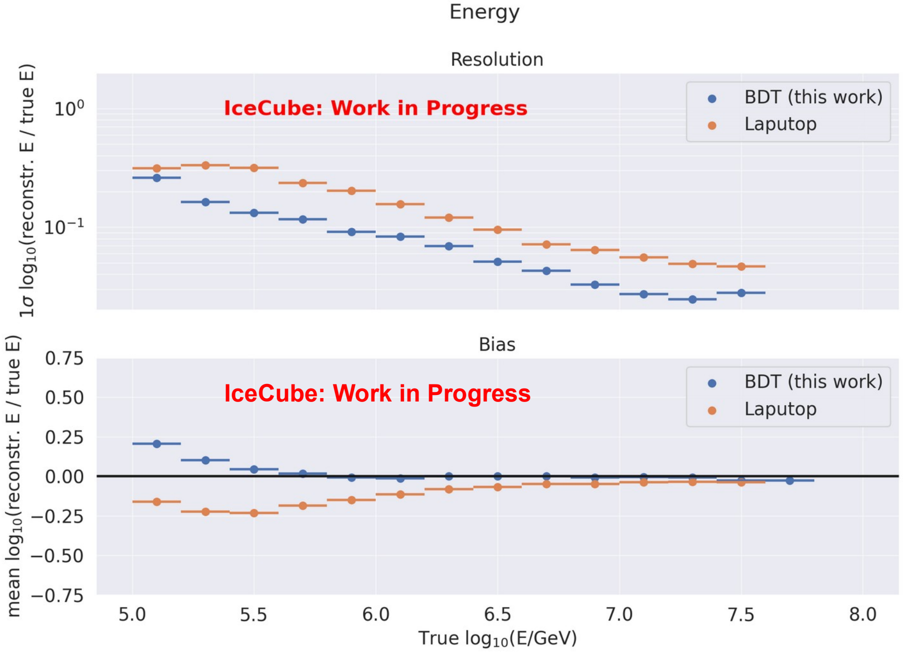




BDT for Primary Energy

Test set





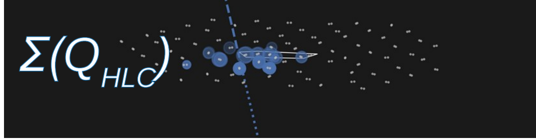
BDT for Primary Type (Classification)

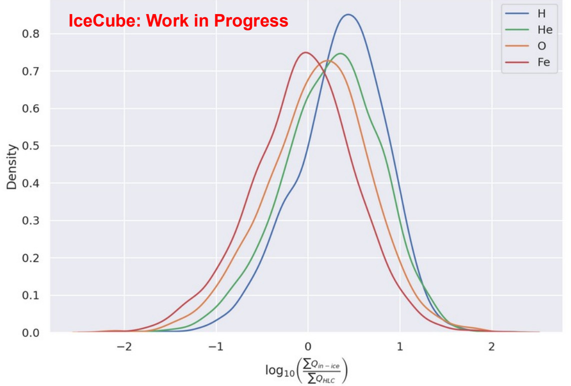
- Input features:
  selected partly according to high figure-of-merit (FOM) value

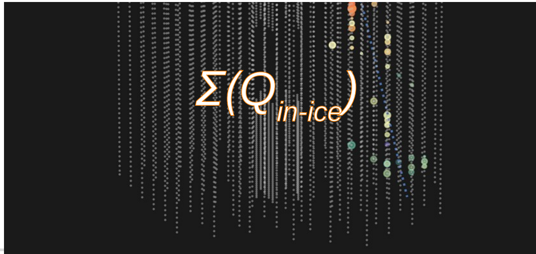
$$\mathrm{FOM}_{i,j} = \frac{|\mu_i - \mu_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}}$$

  computed for all potential features and primary pair combinations

- Target: Monte-Carlo particle

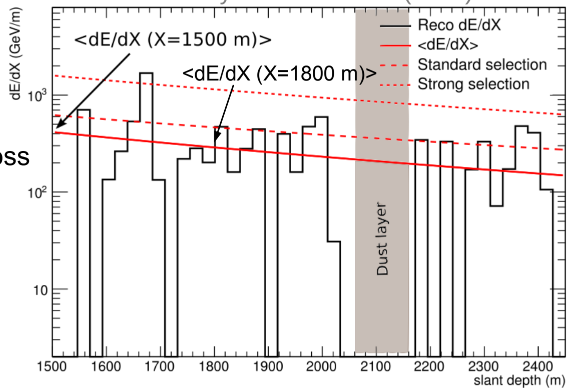


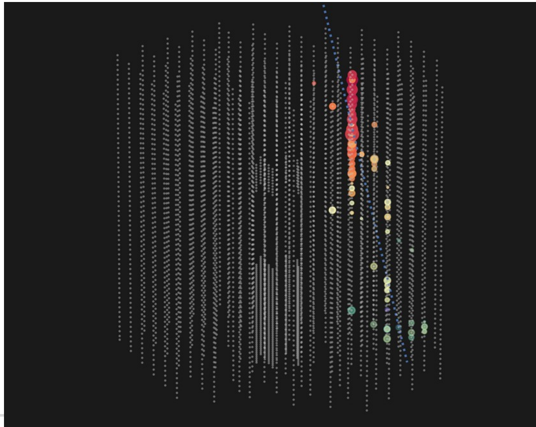




BDT for Primary Type (Classification)

Phys. Rev. D 100.8 (2019): 082002

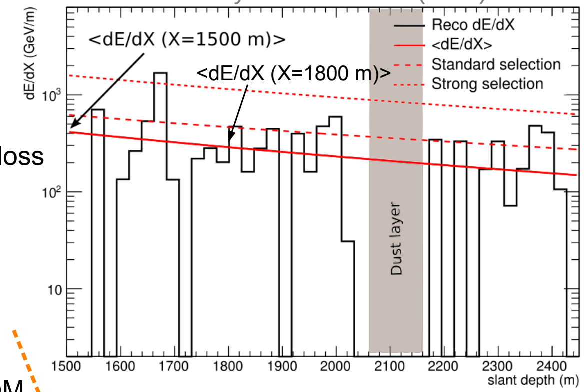




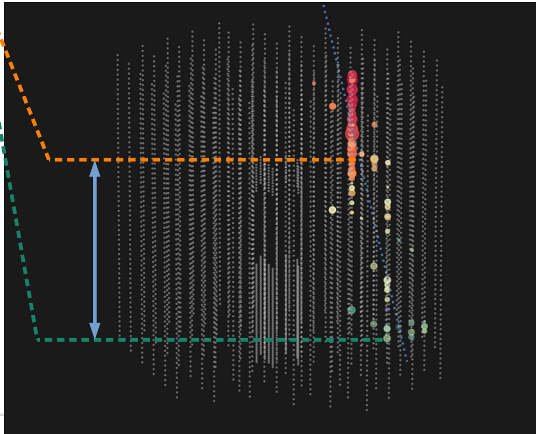
- Input features:
  - *log* of millipede energy loss dE/dX at 1500 m depth
  - *log*(dE/dX$_{1800\ \mathrm{m}}$)
  - *log*(dE/dX$_{1800\ \mathrm{m}}$ – dE/dX$_{1500\ \mathrm{m}}$)
  - *log* of the highest stochastic energy loss
  - *log* of the average stochastic energy loss
  - *log* of the total stochastic energy loss
  - *log* of the difference in total stochastic energy loss for standard and strong selection
  - *log* of the difference in highest stochastic energy loss for standard and strong selection
  - difference in average stochastic energy loss for standard and strong selection
  - difference in average stochastic loss depth for standard and strong selection

BDT for Primary Type (Classification)

Phys. Rev. D 100.8 (2019): 082002

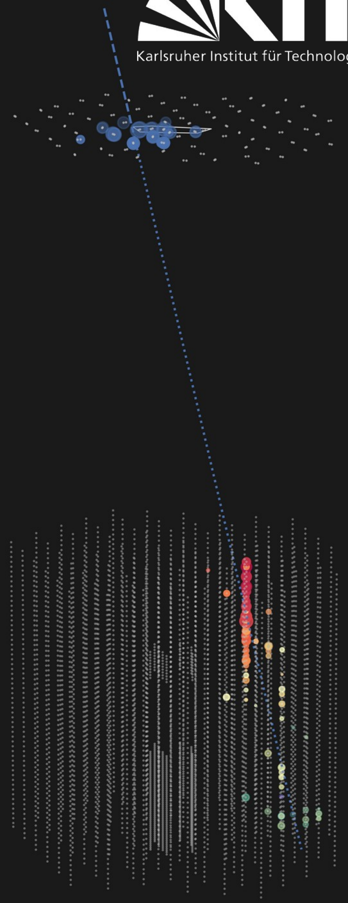


- Input features:
  - *log* of millipede energy loss dE/dX at 1500 m depth
  - *log*(dE/d$X_{1800\ m}$)
  - *log*(dE/d$X_{1800\ m}$ – dE/d$X_{1500\ m}$)
  - *log* of the highest stochastic energy loss
  - *log* of the average stochastic energy loss
  - *log* of the total stochastic energy loss
  - *log* of the difference in total stochastic energy loss for standard and strong selection
  - *log* of the difference in highest stochastic energy loss for standard and strong selection
  - difference in average stochastic energy loss for standard and strong selection
  - difference in average stochastic loss depth for standard and strong selection
  - *log* of number of hit in-ice DOMs
  - z-coordinate of the in-ice COG
  - z-coordinate of the lowest hit DOM
  - difference of z-coordinated of COG and lowest hit DOM
  - ratio of the *log*s of total detected charge in-ice and on the surface
  - *log* of the ratio of total detected charge in-ice and on the surface
  - previously reconstructed energy

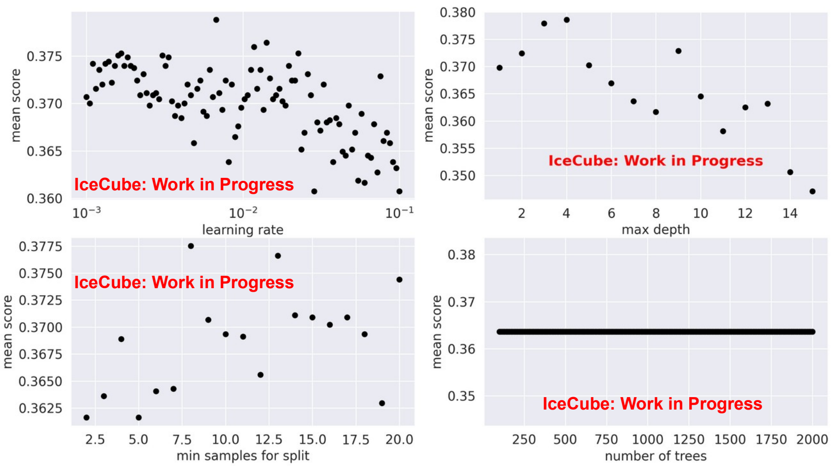


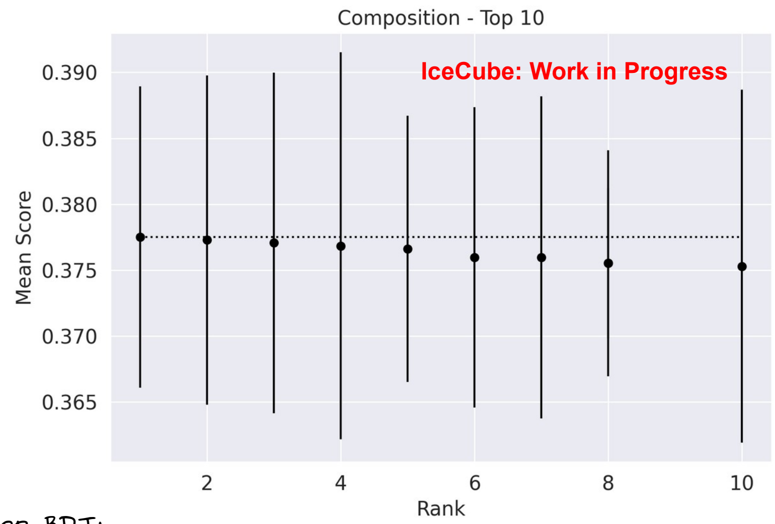
BDT for Primary Type (Classification)

- `GradientBoostingClassifier`
- Model hyperparameters:
  same except:
  - Loss: deviance
  - *sqrt*(17) = 4 features considered at each split
- Test size of 40%
- Randomized search: same except stratified 5-fold CV



BDT for Primary Type (Classification)

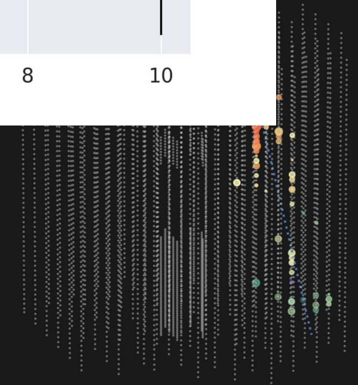




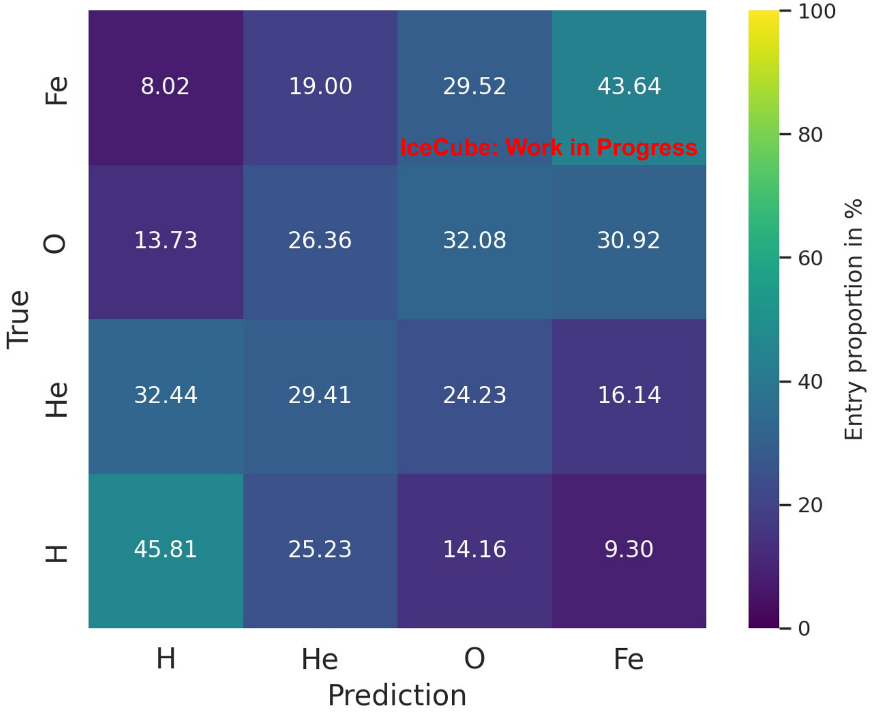
<u>Top BDT:</u>

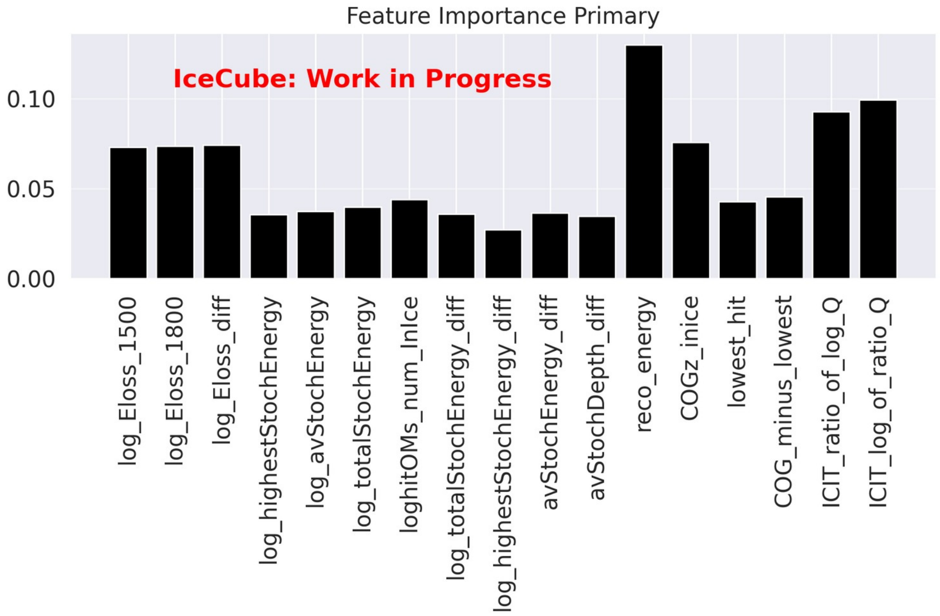
learning rate: 0.0492
max depth: 7
min sam. split: 3
trees: 1620

Train score: 86.68%
Test score: 37.90%



BDT for Primary Type (Classification)

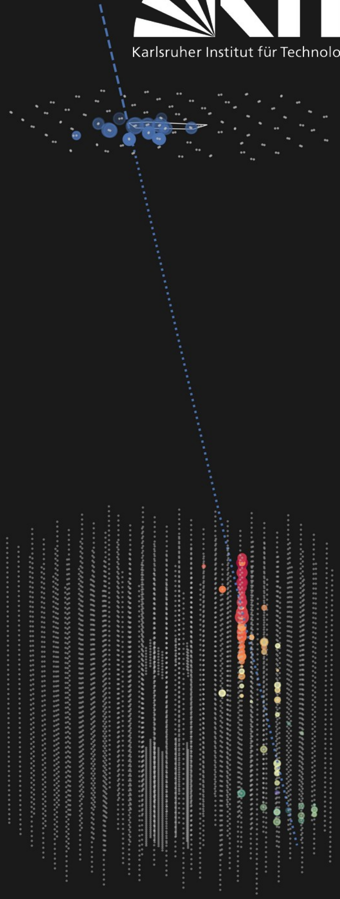




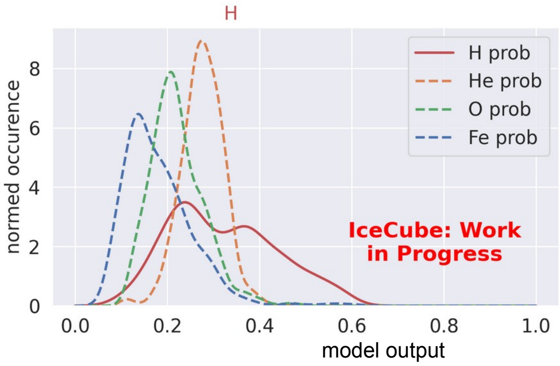
BDT for Primary Type (Classification)

Model output ('probability') for assignment as H, He, O or Fe for protons (KDE)

	H_proba	He_proba	O_proba	Fe_proba
0	0.82	0.11	0.04	0.03
1	0.70	0.19	0.06	0.05
2	0.75	0.10	0.08	0.07
3	0.68	0.17	0.06	0.09
4	0.86	0.07	0.02	0.05
5	0.77	0.08	0.06	0.09

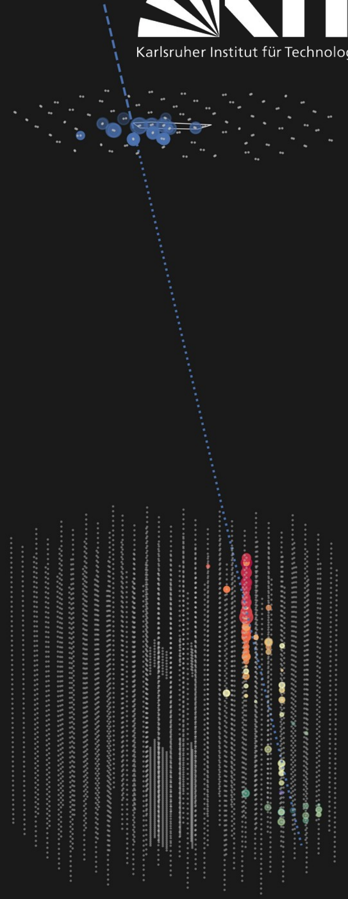


BDT for Primary Type (Classification)

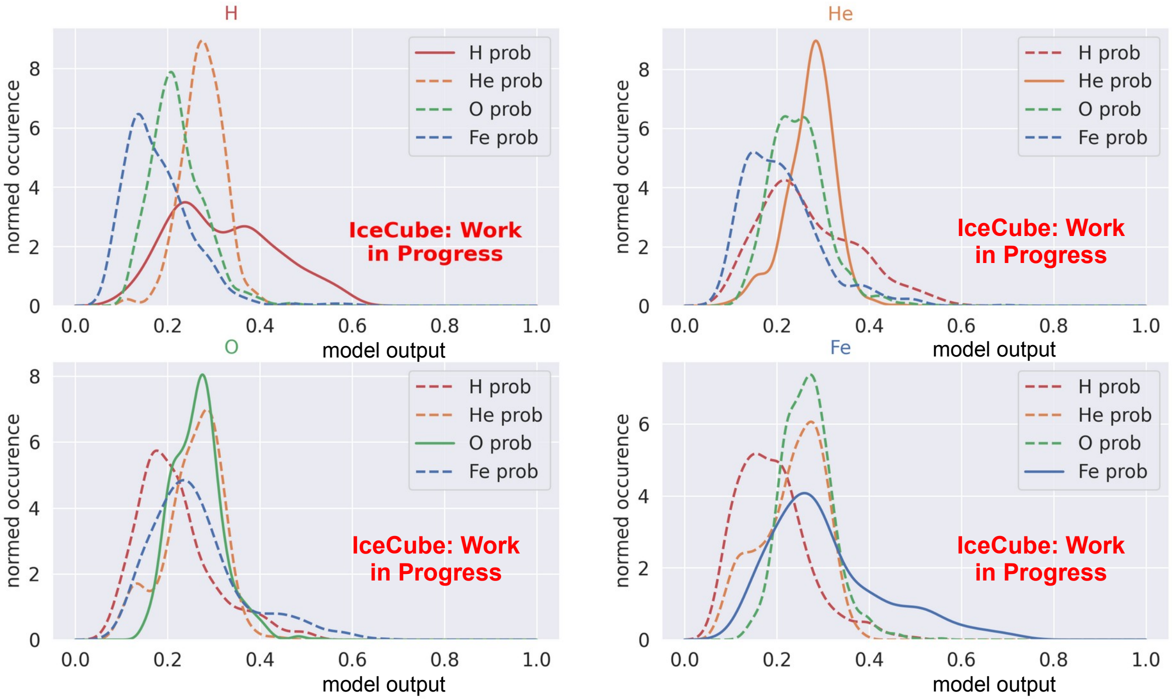


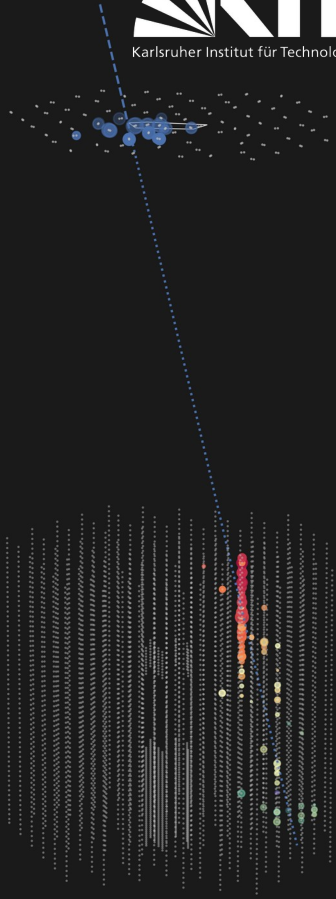
Model output ('probability') for assignment as H, He, O or Fe for protons (KDE)

	H_proba	He_proba	O_proba	Fe_proba
0	0.42	0.27	0.14	0.17
1	0.57	0.18	0.02	0.23
2	0.23	0.12	0.40	0.25
3	0.08	0.49	0.33	0.10
4	0.36	0.32	0.29	0.03
5	0.22	0.26	0.41	0.11



BDT for Primary Type (Classification)

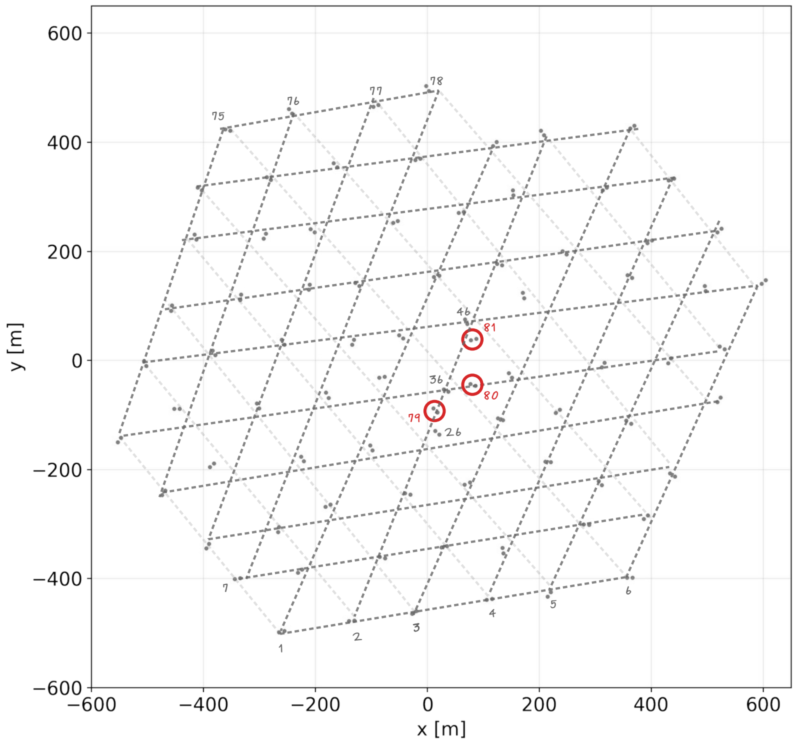


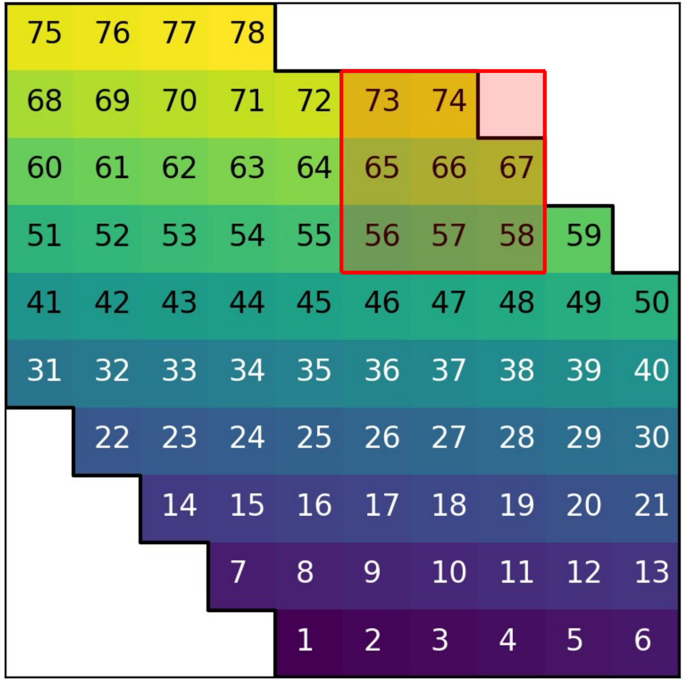


Karlsruher Institut für Technologie

Convolutional Neural Networks

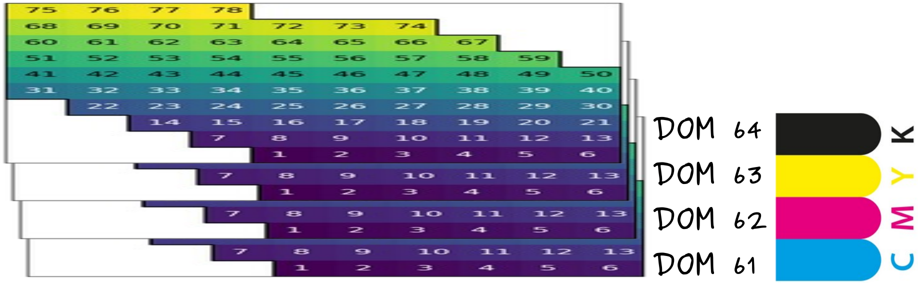
Arranging IceCube in a CNN





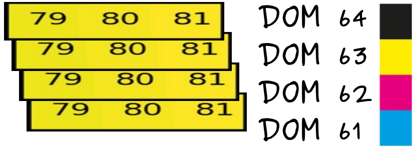
Arranging IceCube in a CNN

IceTop



- **2D** Convolution (10x10, 4 „color“ channels, kernel 3x3, stride 1, padding 1 → 10x10)
- ReLU activation
- Batch Normalization
- Max Pooling (kernel size 2x2, stride 2)
- Dropout

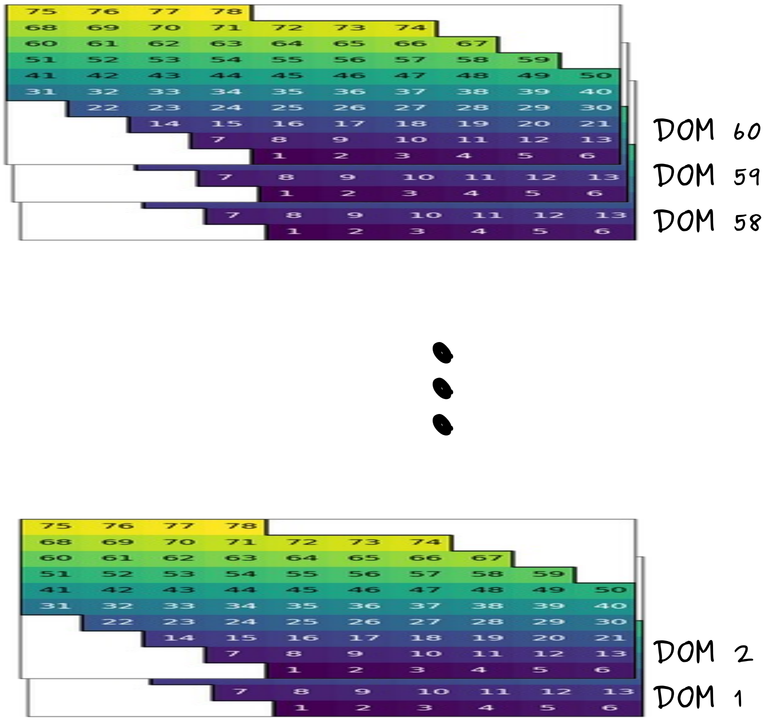
IceTop InFill



- **1D** Convolution (in_shape 3, 4 „color“ channels, kernel size 2, stride 1, padding 1 → out_shape 4)
- ReLU activation, Batch Normalization
- 1D Convolution (in_shape 4, kernel size 3, stride 1, padding 1 → out_shape 4)
- ReLU activation, Batch Normalization
- Dropout

Arranging IceCube in a CNN

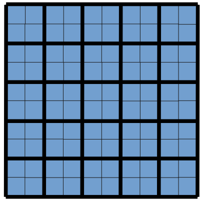
in-ice



- DeepCore not included
- **3D** Convolution (60x10x10, 1 „color“ channel, kernel 3x3x3, stride 1, padding 1 → 60x10x10)
- ReLU activation
- Batch Normalization
- Max Pooling (kernel size 2x2x2, stride 2)
- Dropout

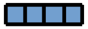
IceCube in a CNN

IceTop



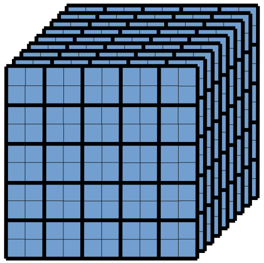
Ax5x5 → flatten

IceTop InFill



Bx4 → flatten

in-ice



Cx30x5x5 → flatten

concatenate
→ fully-connected (in_nodes: D, out_nodes: 64)
→ ReLU, dropout
→ fully-connected(in_nodes: 64, out_nodes: 32)
→ ReLU
→ fully-connected(in_nodes: 32, out_nodes: 1)

e.g. θ or $core_x$

Summary

- BDT models are fast to train and stable (little variation in top 10)
- Primary energy prediction works good
- Hope on CNN for better core, zenith and mass estimation

Outlook

- Way more MC necessary
(currently only ~7500 coincident events)

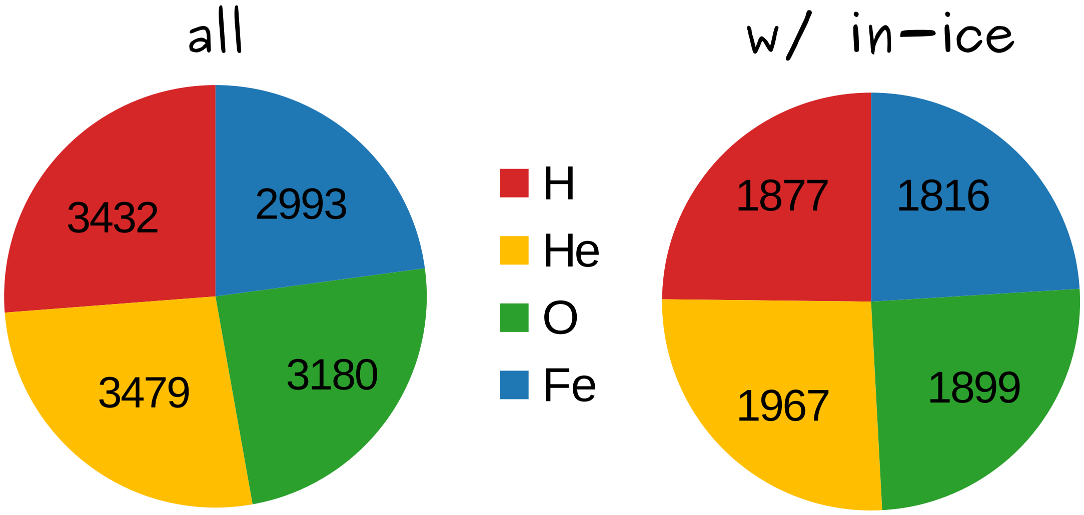
- CORSIKA simulations ongoing (26680 events in range $4.0 \leq \log_{10}(E/\mathrm{GeV}) \leq 8.0$, Sibyll 2.3c and FLUKA)

- Detector (surface and in-ice) response simulation pending

- Improvement of CNN structure and better training with more data

Backup

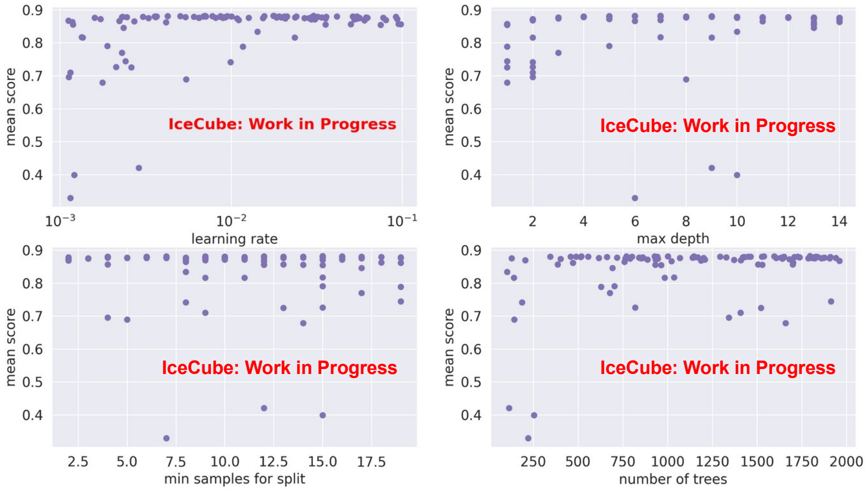
Data used



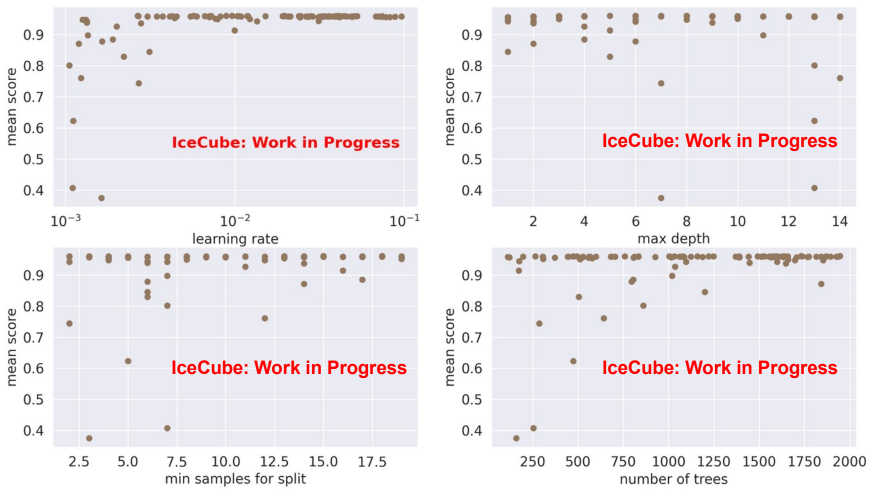
BDT for Shower Core

Takes ~ 8 min on 4 CPUs (3800 MHz, cobalt)

x (from RandomSearch)

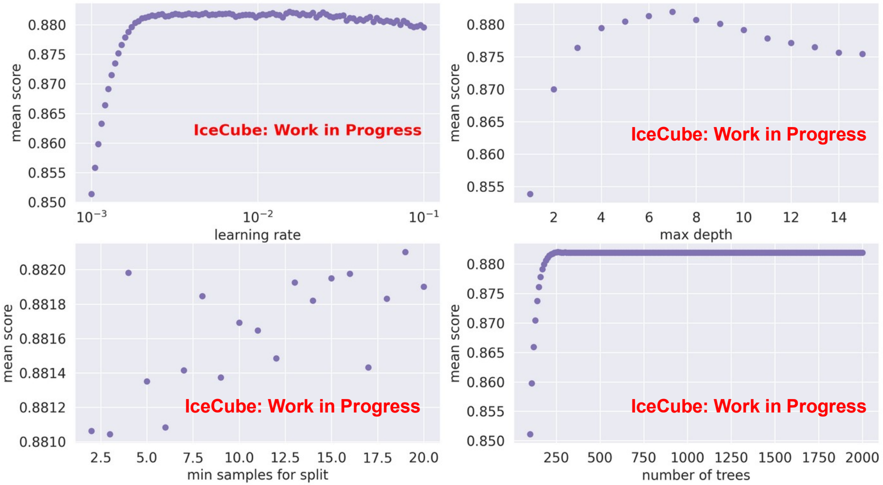


y (from RandomSearch)

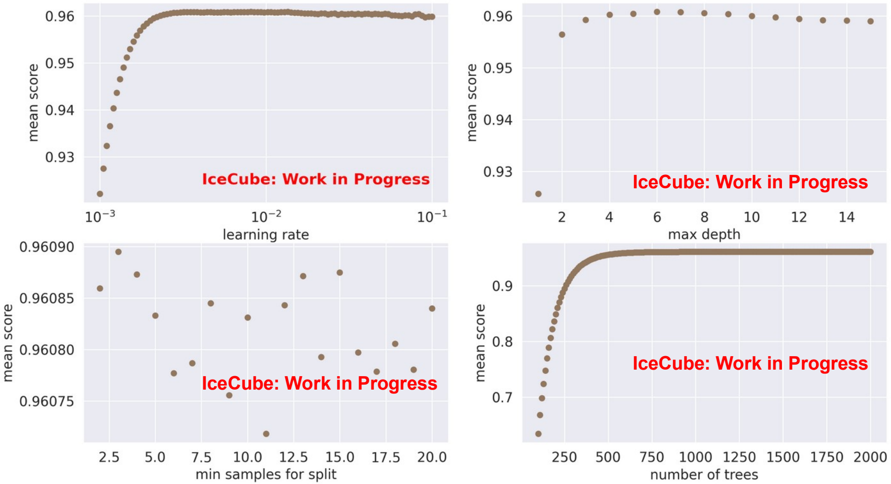


BDT for Shower Core

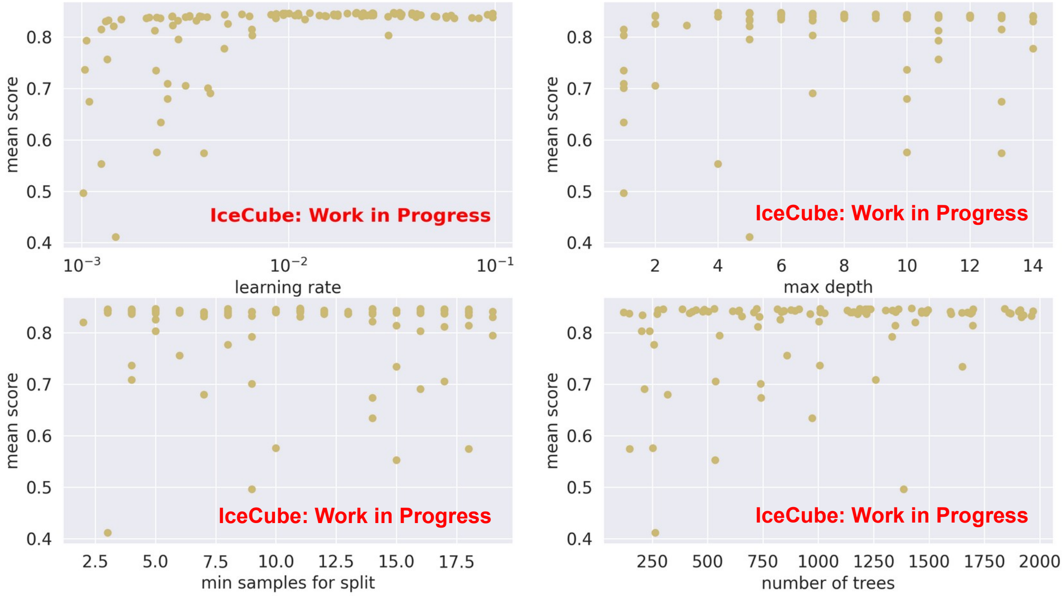
x



y

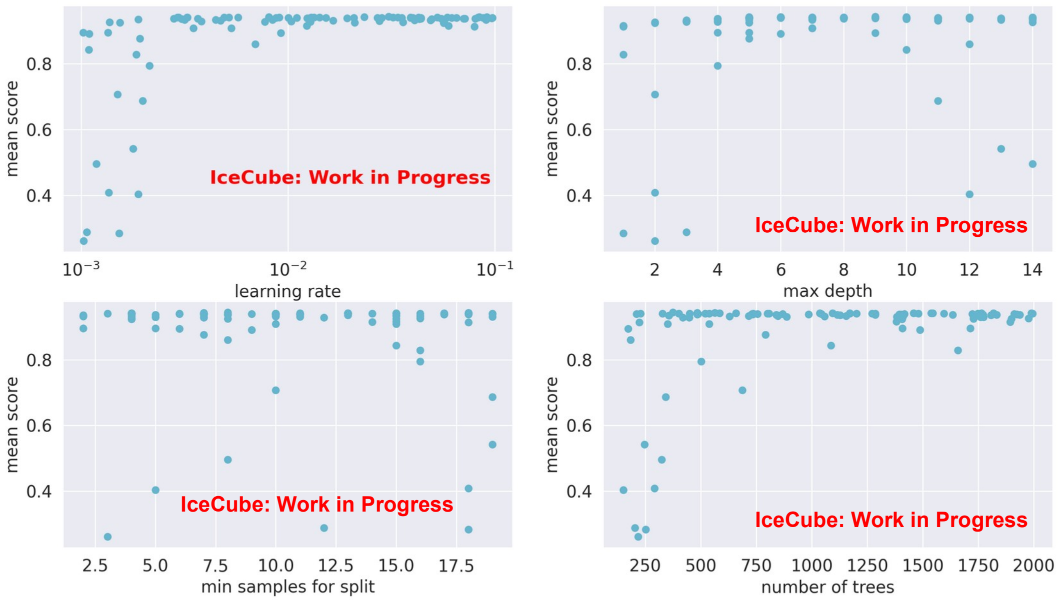


BDT for Zenith Angle



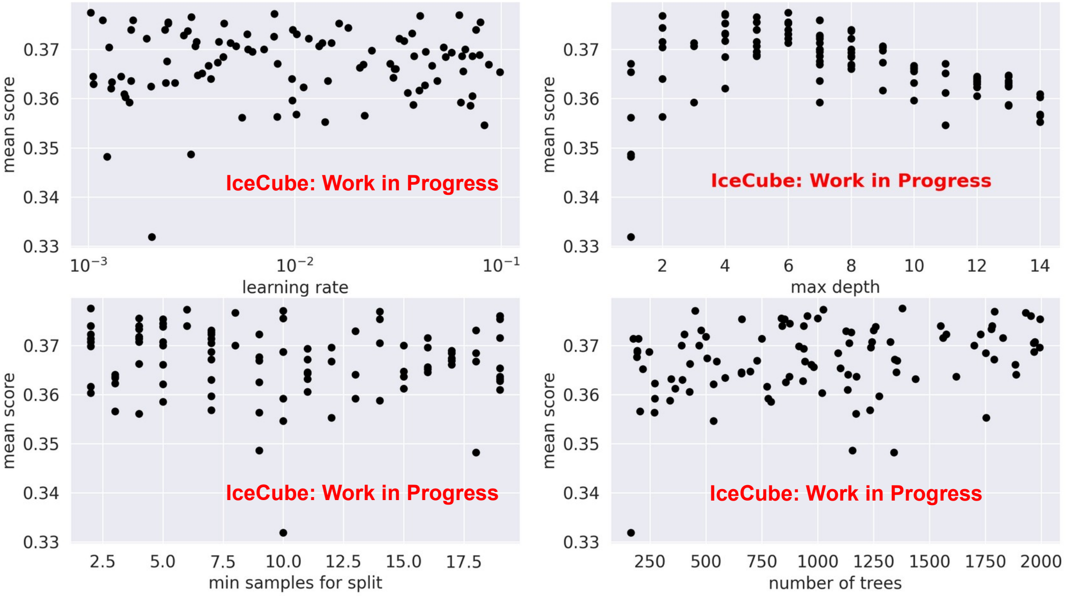
Takes ~ 7 min on 4 CPUs (cobalt)

BDT for Primary Energy



Takes ~ 7 min on 4 CPUs (cobalt)

BDT for Primary Type



Takes ~ 40 min on 4 CPUs (cobalt)

BDT for Primary Type

Karlsruher Institut für Technologie

FOM weighting:

$$\mathrm{FOM}^{(f)} = \frac{\sum_{i,j\in[\mathrm{H},\mathrm{He},\mathrm{O},\mathrm{Fe}]} \frac{\mathrm{FOM}_{i,j}^{(f)}}{|\ln(A_i)-\ln(A_j)|}}{\sum_{i,j\in[\mathrm{H},\mathrm{He},\mathrm{O},\mathrm{Fe}]} \frac{1}{|\ln(A_i)-\ln(A_j)|}}$$