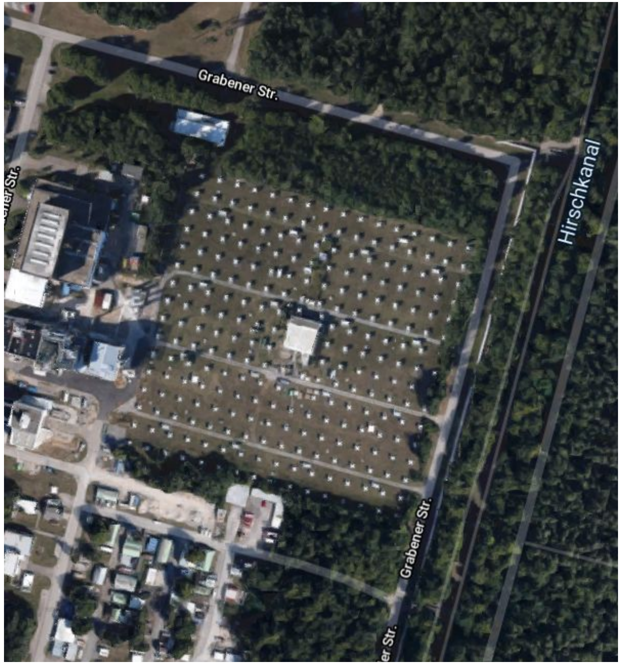
Towards mass composition study with KASCADE using deep neural networks

Speaker:
Vladimir Sotnikov
Co-authors:
M. Kuznetsov, N. Petrov

What is KASCADE?

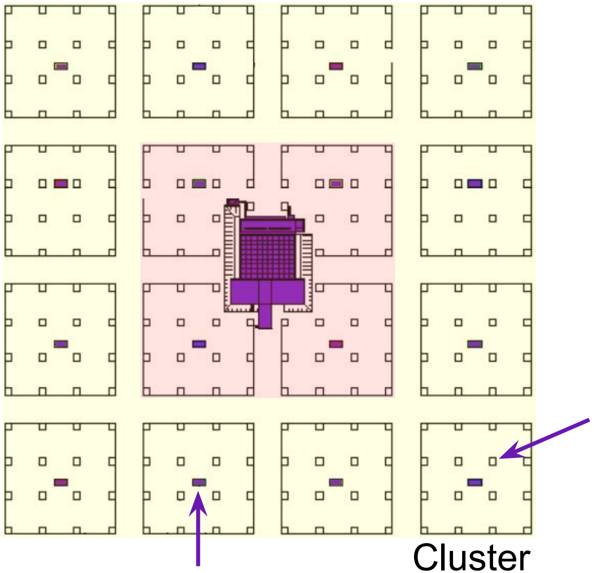
- KASCADE detector was operating for more than 15 years on the site of the Karlsruhe Institute of Technology, Germany
- Its detectors are aligned in a square 16 by 16 grid
- These detectors measure both hadronic and electromagnetic components of air-showers



Nucl.Instr. and Methods, A513 (2003) 490-510
The Cosmic-Ray Experiment KASCADE

Astroparticle Physics 24 (2005) 1-25
KASCADE Measurements of energy spectra for elemental groups of cosmic rays: Results and open problems

Schematic view



Type-1 stations
detect **e/γ** and **muon** signals

Type-2 stations
detect **only e/γ** signals

Event is recorded when ≥ 1 cluster detects a signal > certain threshold

Run is a group of events

Approach

Input: Event

3×16×16 experimental features

9 reconstructed features

Predict

Target: Primary particle type

Categorical feature (p, He, C, Si, Fe)

Some of our models

- Random Forest classifier (baseline)
- CNN classifier
- Self-attention MLP

Semi-blind analysis

20% Unblinded — For tuning

80% Blinded — For testing

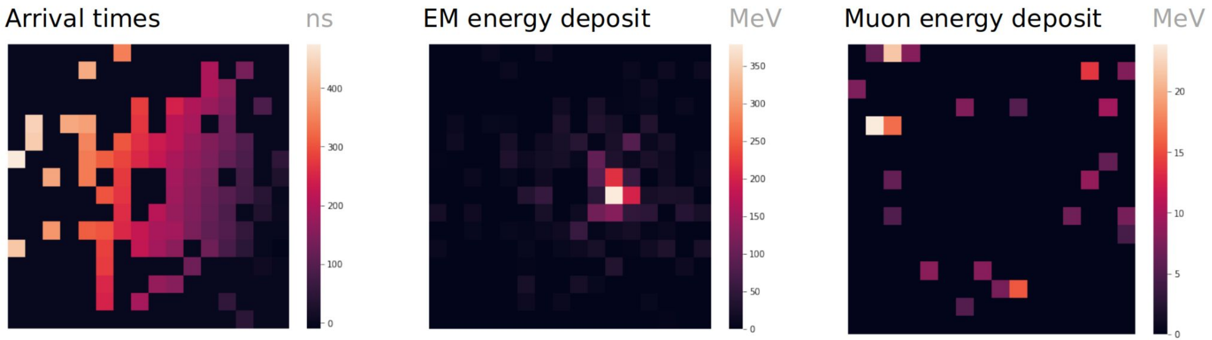
Training step	CORSIKA simulations
Validation step	Checking out predicted particles spectra with unblinded data
Testing step	Revealing the blinded part

https://kcdc.iap.kit.edu/simul/simgeneral/

Data

- Real-world archive data provided by KCDC contains over 400M air shower events with E > 10^{15} eV
- Our training dataset consists of over 2M simulated events provided by the latest hadronic interaction models: EPOS-LHC, QGSJet II-04, Sybill 2.3

Example of Monte-Carlo event



- We apply the following cuts:
  - Ze < 40
  - Ne > 4.8
  - Nmu > 3.6
  - 0.2 < Age < 1.48

A.Haungs et al; *Eur. Phys. J. C (2018) 78:741;*

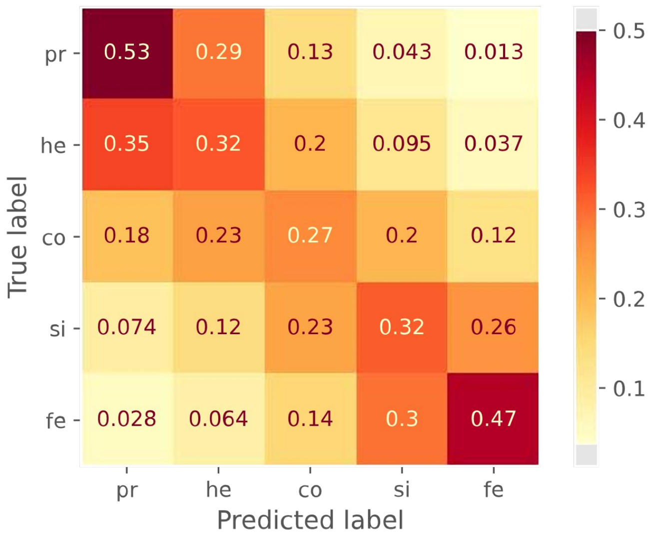
"The KASCADE Cosmic ray Data Centre KCDC: granting open access to astroparticle physics research data";

(doi: 10.1140/epjc/s10052-018-6221-2)

Random Forest: accuracy and predicted spectra

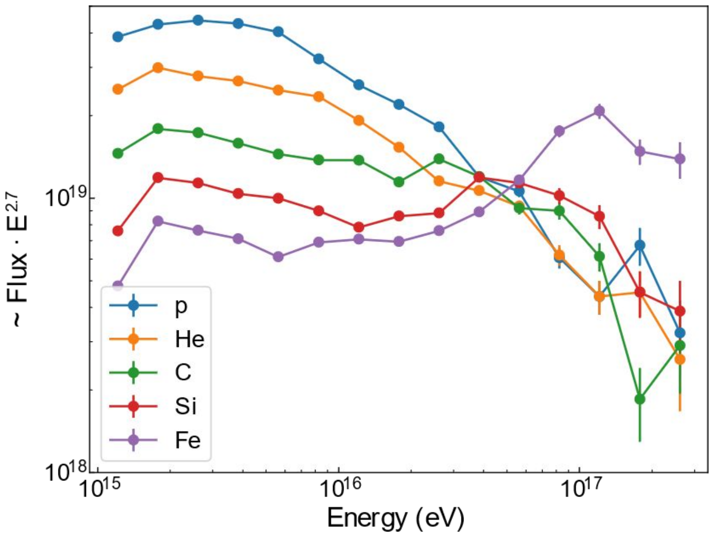
Confusion matrix

Simulated data (EPOS-LHC)

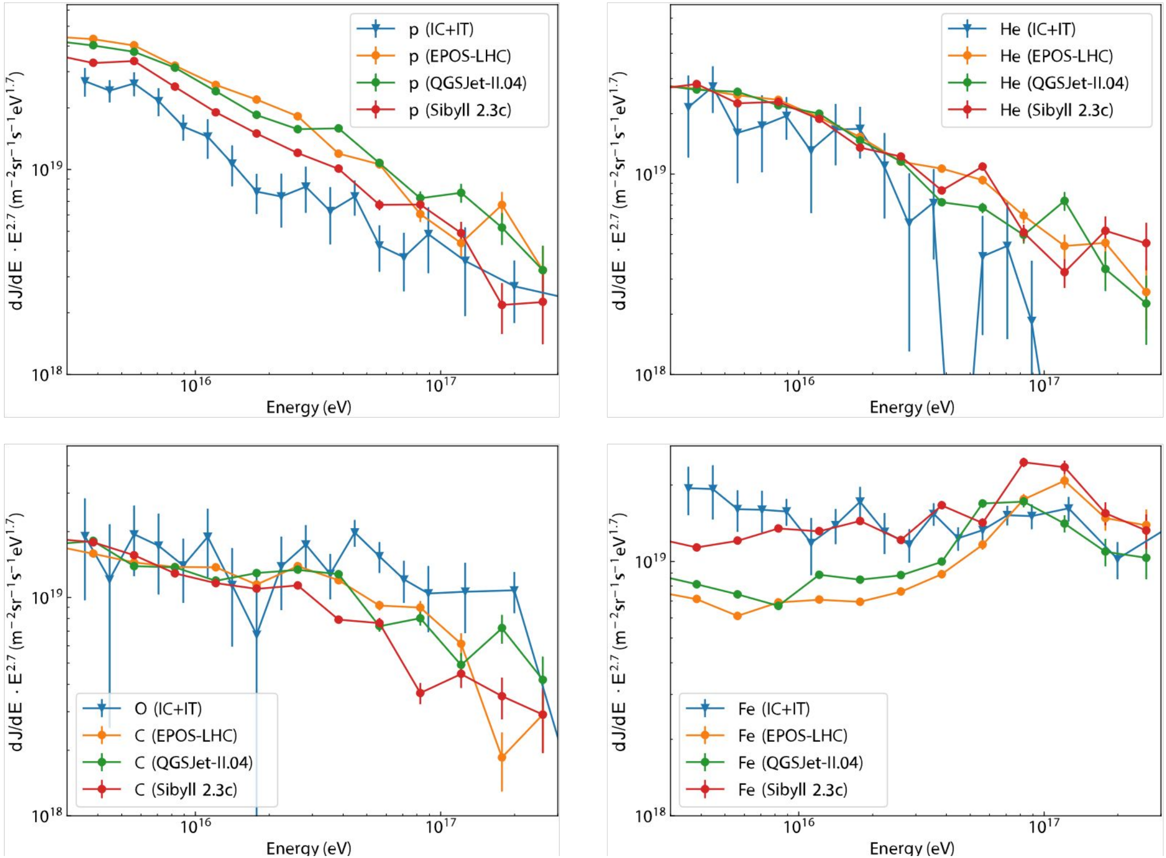


Spectra

Experimental (unblinded) data



Random Forest: comparison with IceCube collaboration*



Why we compare to IC+IT:

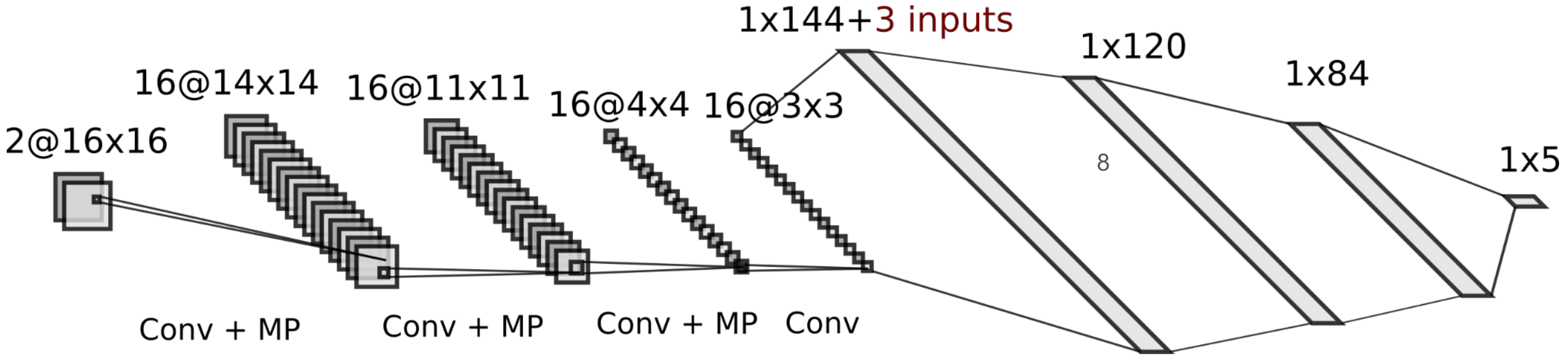
- They used Sibyll model
- Particles are divided into 4 mass groups
- ML approach
- Same energy range

IceCube Collaboration, *Cosmic ray spectrum and composition from PeV to EeV using 3 years of data from IceTop and IceCube*, Phys. Rev. D 100 (2019) no.8, 082002

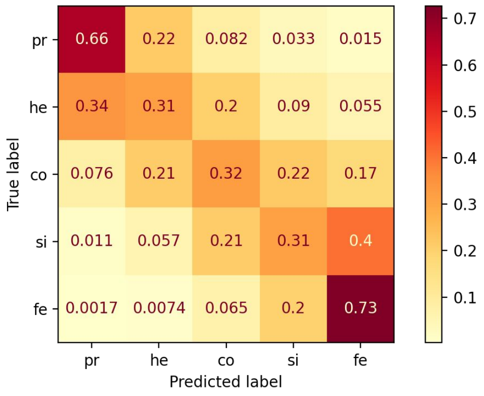
D. Kostunin et al. *New insights from old cosmic rays: A novel analysis of archival KASCADE data*, ICRC2021, https://arxiv.org/abs/2108.03407

Convolutional neural network (CNN)

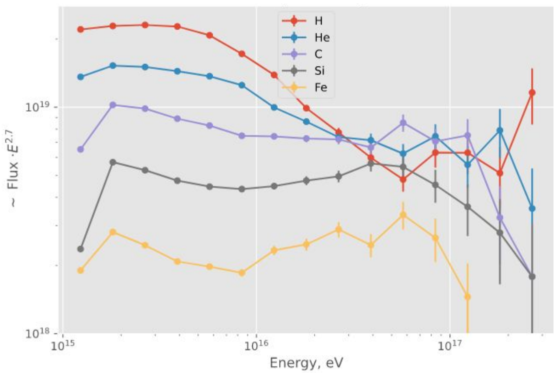
- Input: energy deposits per station (2, 16, 16) + 3 reconstructed features (Age, log10 Ne, log10 Nμ)
- Augmentations: rotations by a multiple of 90° + flips



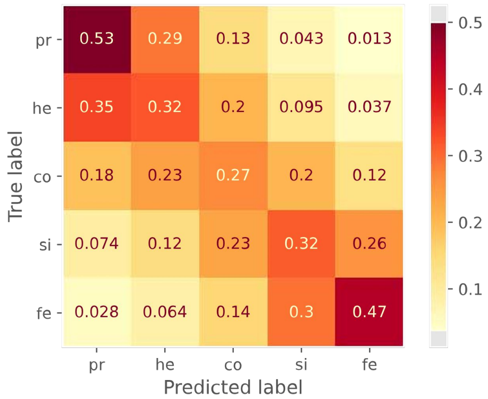
CNN: performance and comparison to Random Forest



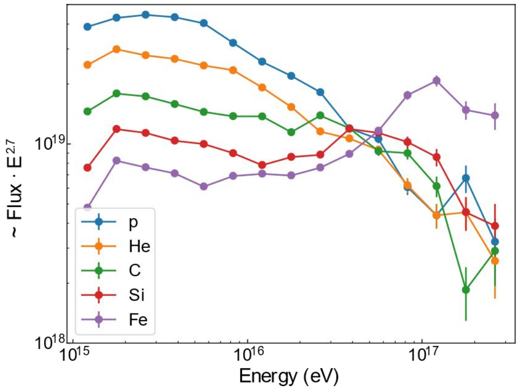
Confusion matrix for CNN (QGSJet II-04)



Spectra for CNN



Confusion matrix for Random Forest (EPOS-LHC)

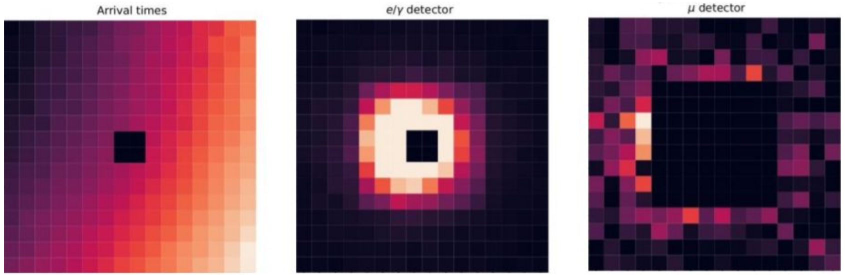


Spectra for Random Forest

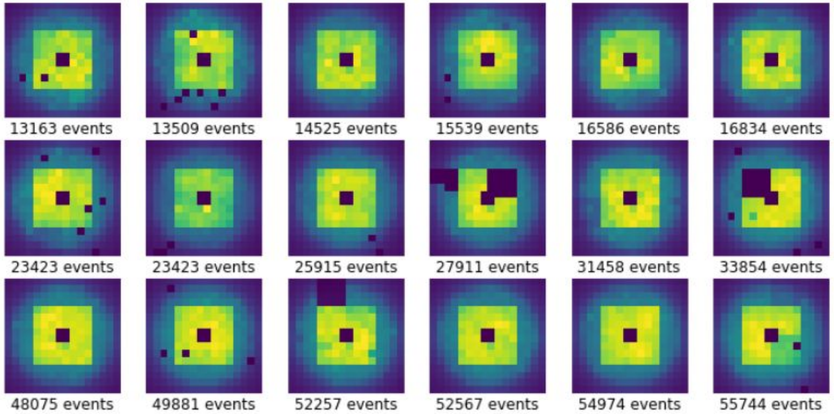
Preliminary results

CNN: motivation behind quality cuts

In a simulated event
all detectors have 100% uptime

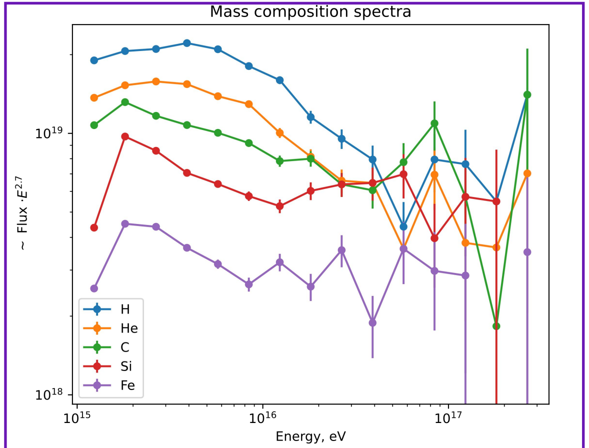


In a real event
some detectors might go down



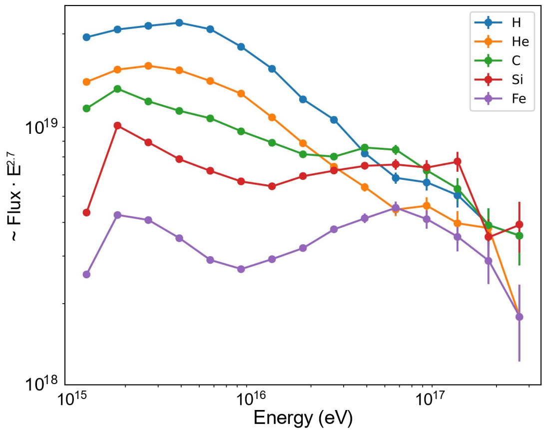
Each square shows sum of EM energy deposits for some run

CNN: performance with quality cuts



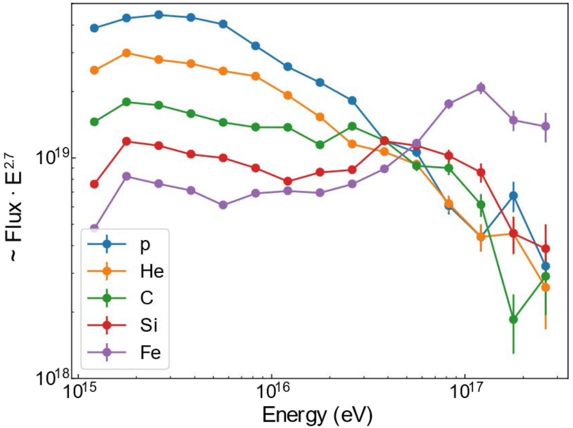
Spectra for CNN

Data with quality cuts



Spectra for CNN

No quality cuts



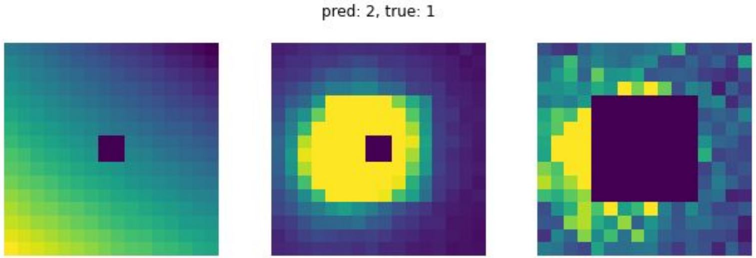
Spectra for RF

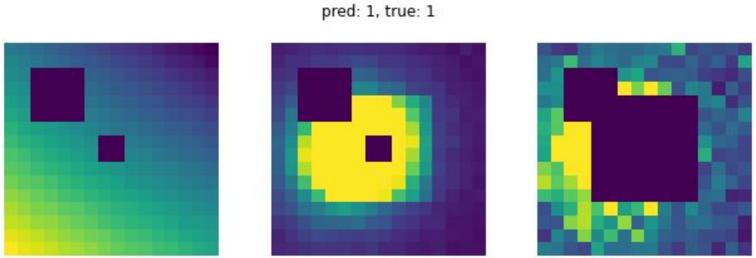
No quality cuts

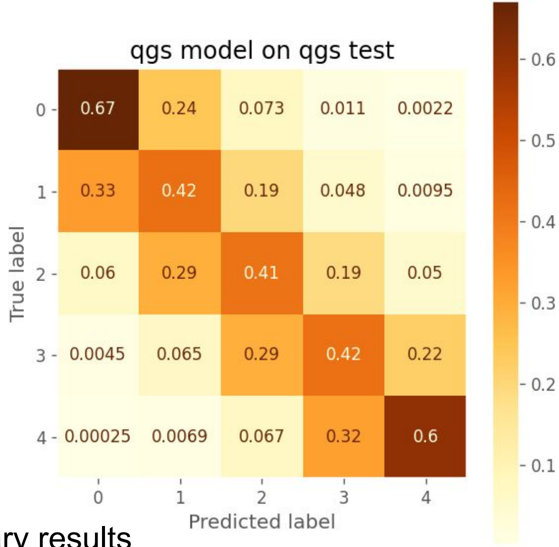
1/5 of the full dataset

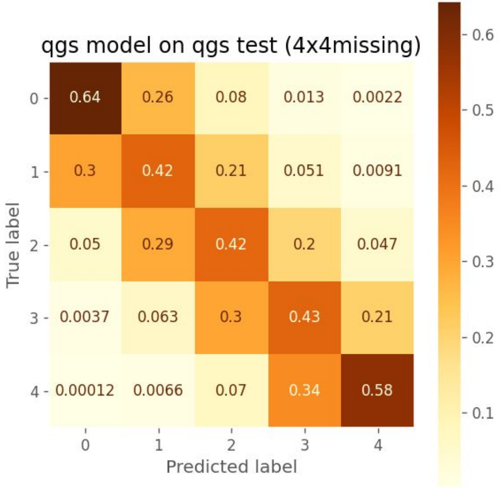
Preliminary results

CNN: estimating robustness





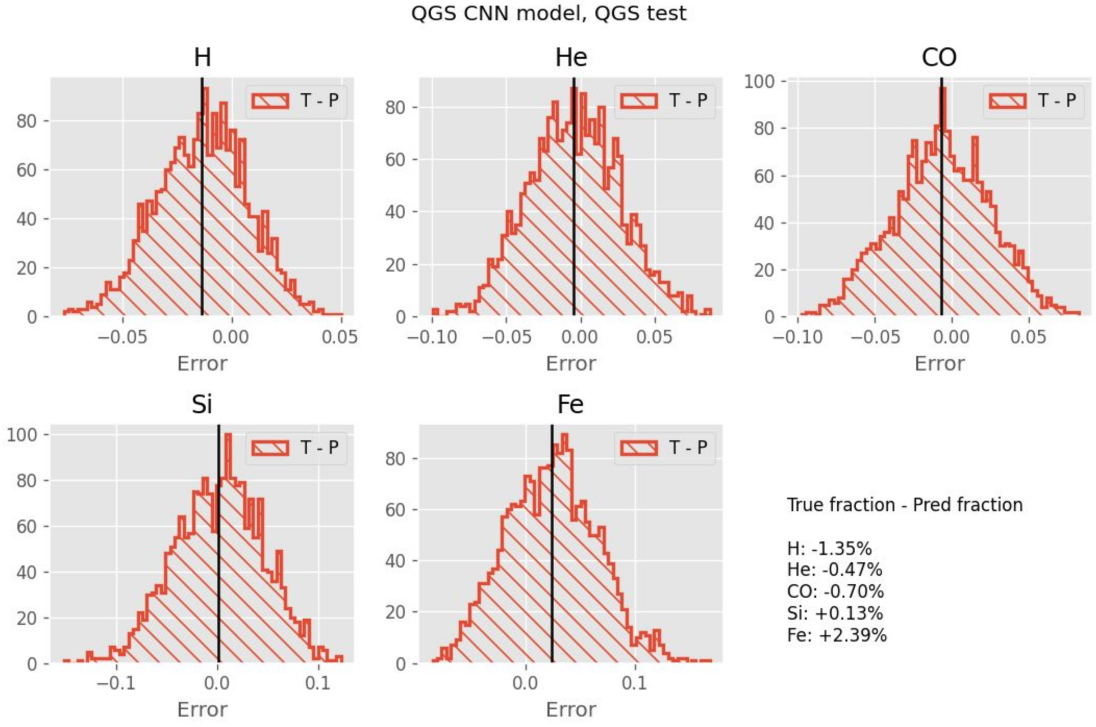




Preliminary results

CNN: estimating mass composition errors

- We've generated 2000 random ensembles containing 5000 events in each
- We evaluate the model on each of ensembles, each one has its own true mass composition
- Such an approach allows us to measure accuracy of mass composition predictions

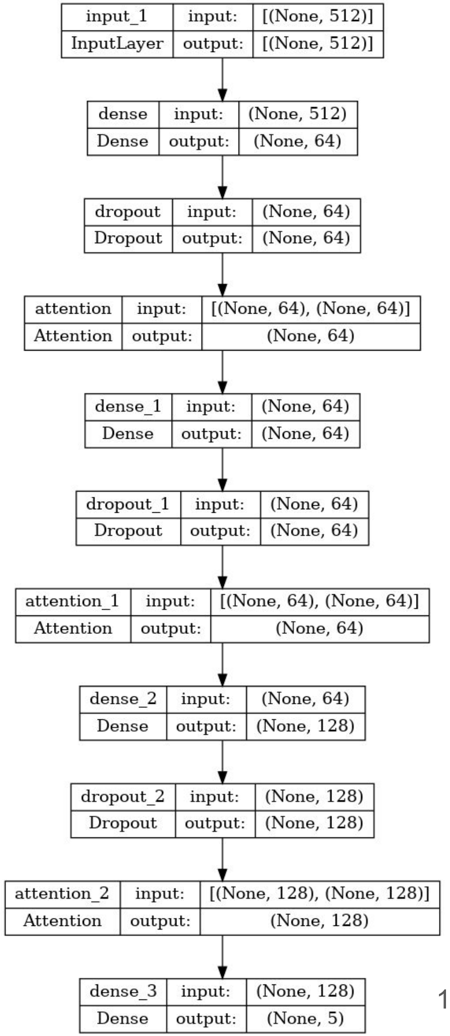


Error distribution for elements in ensembles

Preliminary results

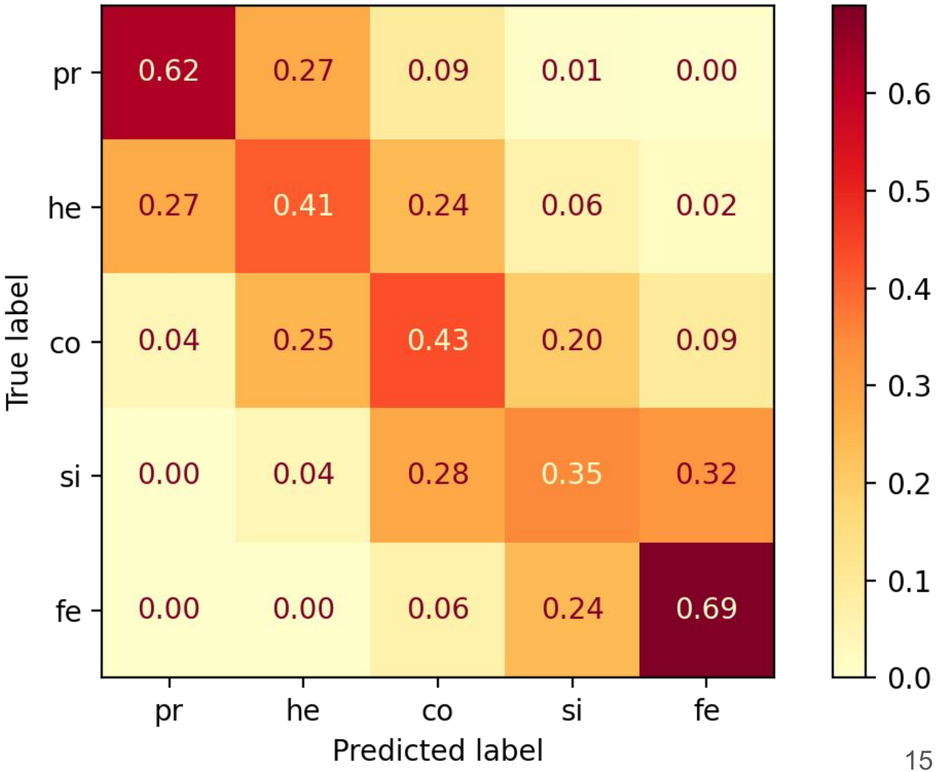
Self-attention MLP

- Our data isn't spatially invariant (due to cutouts in the center)
- To exploit the spatial-specific information, we trained a self-attention feedforward network



Self-attention MLP

- The model appears to be more accurate than deep CNNs but more careful evaluation is needed



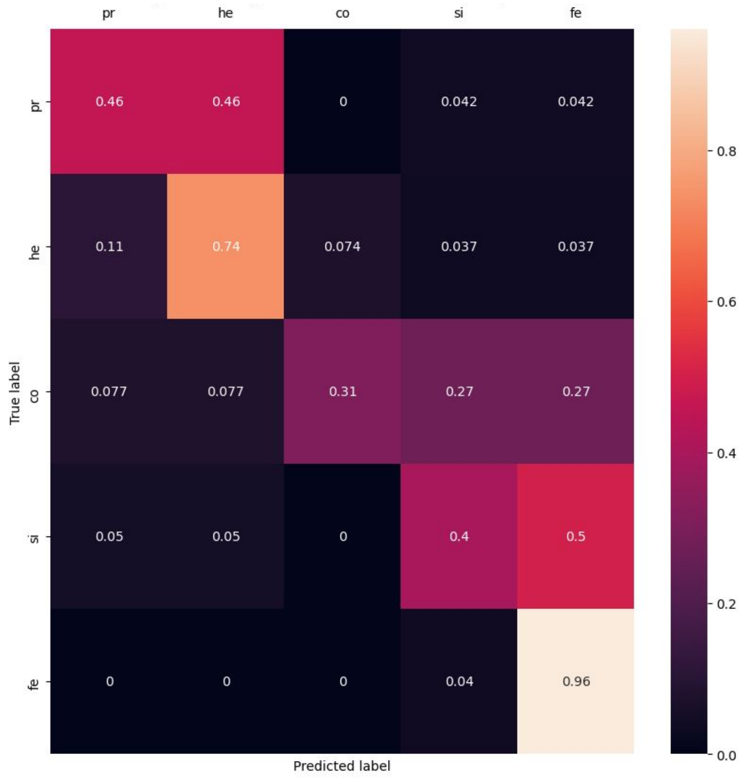
Preliminary results

CONCLUSION

- We have developed multiple deep neural networks for analyzing CR mass composition
- We have calculated an estimate of CNN’s performance on the downstream task of predicting CR mass composition
- CNNs appear to be robust to data artifacts (e.g. broken detectors)
- Self-attention MLP seems to even outperform CNN models but requires additional sanity checks

Supplementary: RF accuracy on high energies

17 < E < 17.5



15 < E < 15.5