Mass composition study with machine learning on KASCADE archival data

Speaker: Nikita Petrov (NSU, INR RAS)

Outline

1. KASCADE experiment
2. Mass composition reconstruction
    a. ML methods in detail
    b. Unfolding
3. Results & Conclusion

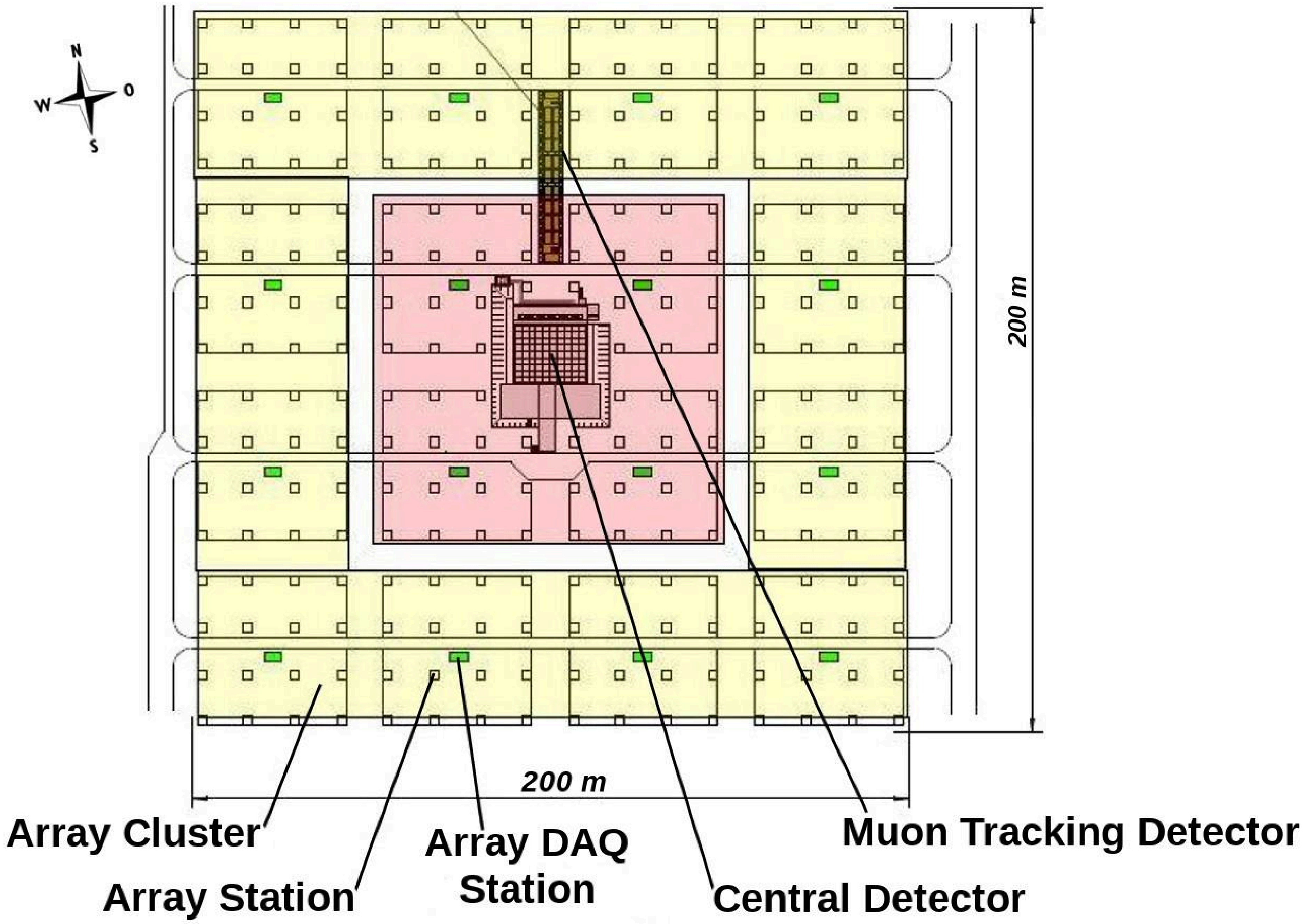
Kuznetsov, M. et al (2024). Methods of machine learning for the analysis of cosmic rays mass composition with the KASCADE experiment data. Journal of Instrumentation, 19(01), P01025. doi:10.1088/1748-0221/19/01/p01025
Kuznetsov, M. et al (2024). Energy spectra of elemental groups of cosmic rays with the KASCADE experiment data and machine learning. Journal of Cosmology and Astroparticle Physics, 2024(05), 125. doi:10.1088/1475-7516/2024/05/125

KASCADE

KASCADE is an extensive air shower experiment that was located in KIT Campus, Karlsruhe, Germany (1996 - 2013)

KASCADE array: 252 scintillator detectors placed in a rectangular grid at 13 m intervals and covering a total area of 200 × 200 m² in total.

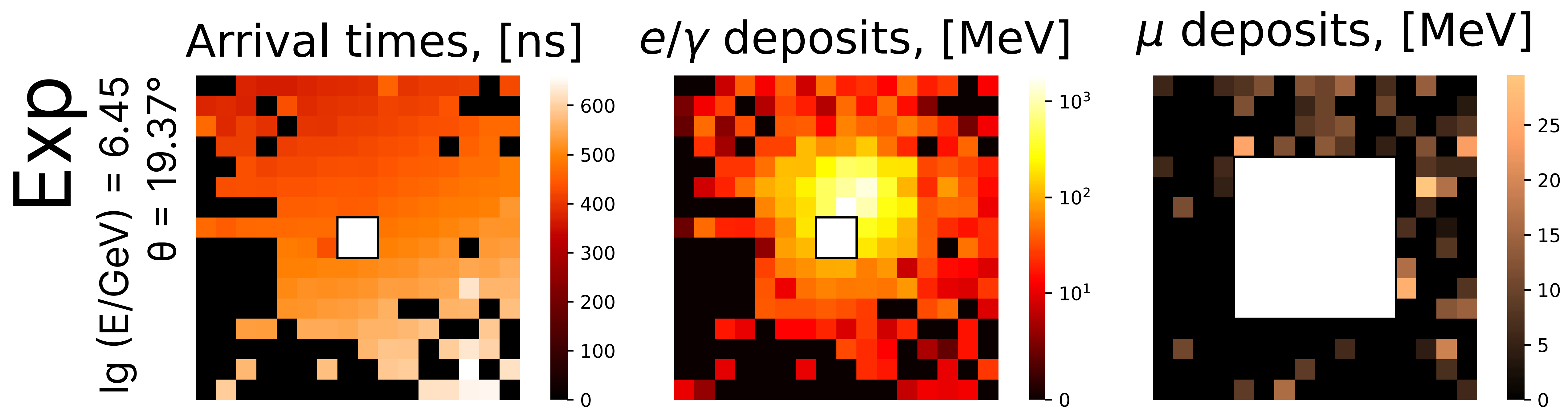
Energy range: ~ 500 TeV — 100 PeV



KASCADE scheme
(green – e/γ and μ detectors, red – e/γ only)

Experimental data & Monte Carlo

provided by KCDC*



Experimental event example

- $\theta < 18°$
- $\log_{10} Ne > 4.8$
- $\log_{10} N\mu > 3.6$
- $\sqrt{(x^2 + y^2)} < 91$ m
- $0.2 < s < 1.48$

Quality cuts (for data and MC)

Event structure

3 arrays 16x16 shape (arrival times; e/γ, μ deposits)

reconstructed features (E, θ, ϕ, x, y, Ne, Nμ, s)

---

* 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

Datasets

Experimental dataset

Unblind: 20%

Blind: 80%

~ 8.5M events in total (after quality cuts)

Monte Carlo datasets (protons, He, C, Si, Fe)

CORSIKA + detector simulation

QGSJet-II.04 ~ 180k events

Sibyll 2.3c

EPOS-LHC

QGSJet-II.02

Mass composition reconstruction

Main stages:

1. Event-by-event classification (particle type: p, He, C, Si, Fe)

☐ **`Random Forest`**

baseline model

input: x, y, E, Ne, Nμ, θ, ϕ, s

☐ **`Multi-Layer Perceptron (MLP)`**

exploits spacial-specific info

input: deposit arrays [flatten] + θ, ϕ

★ **`Convolutional NN (CNN)`**

inspired by LeNet-5 (~30k parameters)

input: deposit arrays [2x16x16] + Ne, Nμ, θ, s

☐ **`EfficientNet v2`**

common standard architecture

input: deposit arrays [2x16x16] + θ, ϕ

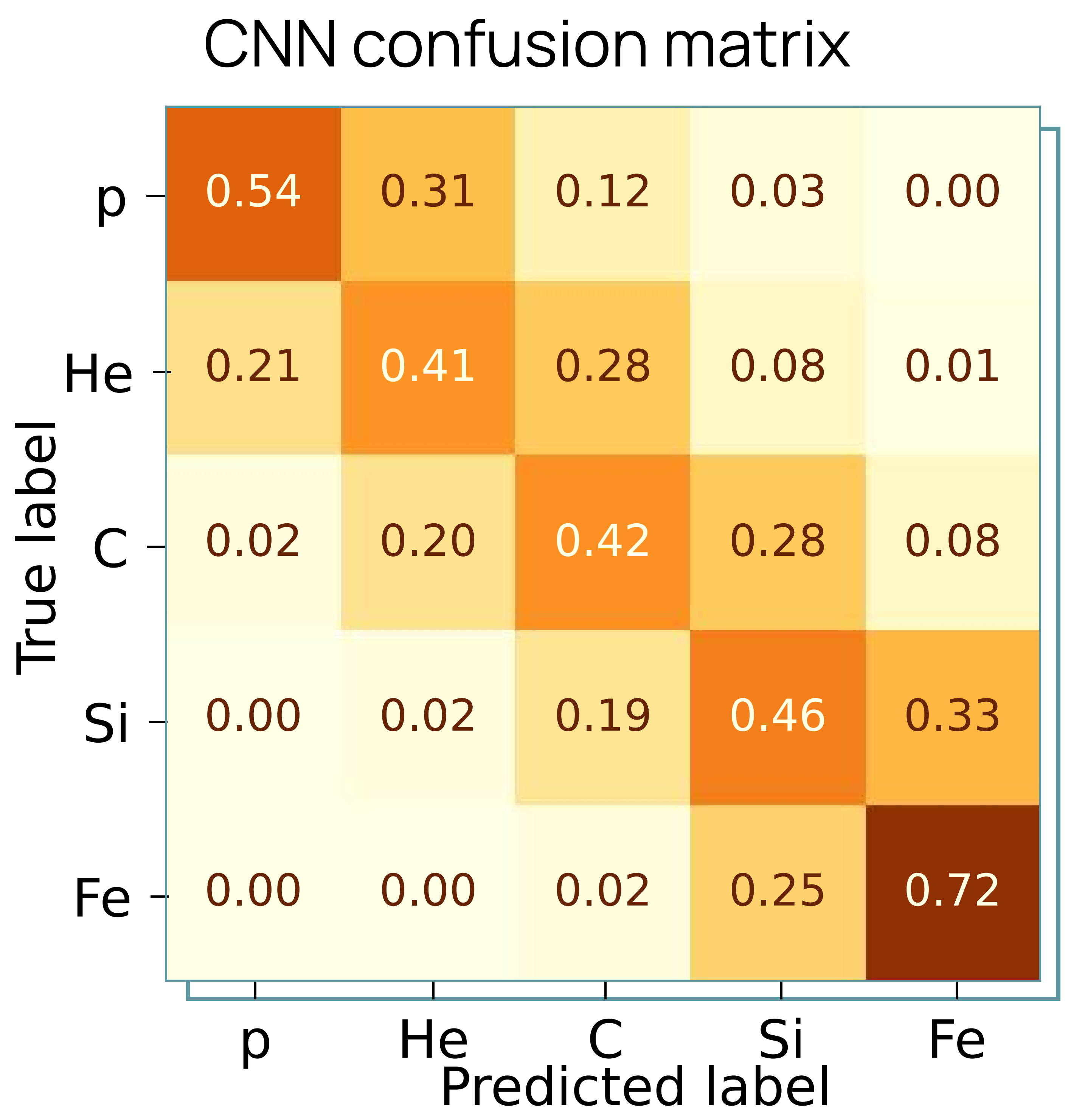
2. Unfolding (particle and energy)

`Bayesian iterative approach*`

★ means selected classifier

* G. D'Agostini. A Multidimensional unfolding method based on Bayes' theorem. Nucl. Instrum. Meth. A, 362:487–498, 1995. doi:10.1016/0168-9002(95)00274-X.

Event-by-event classification



	p	He	C	Si	Fe
p	0.54	0.31	0.12	0.03	0.00
He	0.21	0.41	0.28	0.08	0.01
C	0.02	0.20	0.42	0.28	0.08
Si	0.00	0.02	0.19	0.46	0.33
Fe	0.00	0.00	0.02	0.25	0.72

for QGSJet-II.04 hadronic interaction model
(here and another extra cut at $\log_{10}(E/\mathrm{GeV}) > 6.15$)

Training

Normalize features

Maximize train sample

- Expand selections: $\theta < 30°$
- Augment data: rotations

Quality

Estimate the performance of the ML classifier using the confusion matrix

- The more diagonal, the better
- 0.2 in each cell is a random guess

Ablation study

Impact of the individual input features

Train and test CNN with deposits only and reconstructed features only

CNN is stable with exclusion features except for the zenith angle.

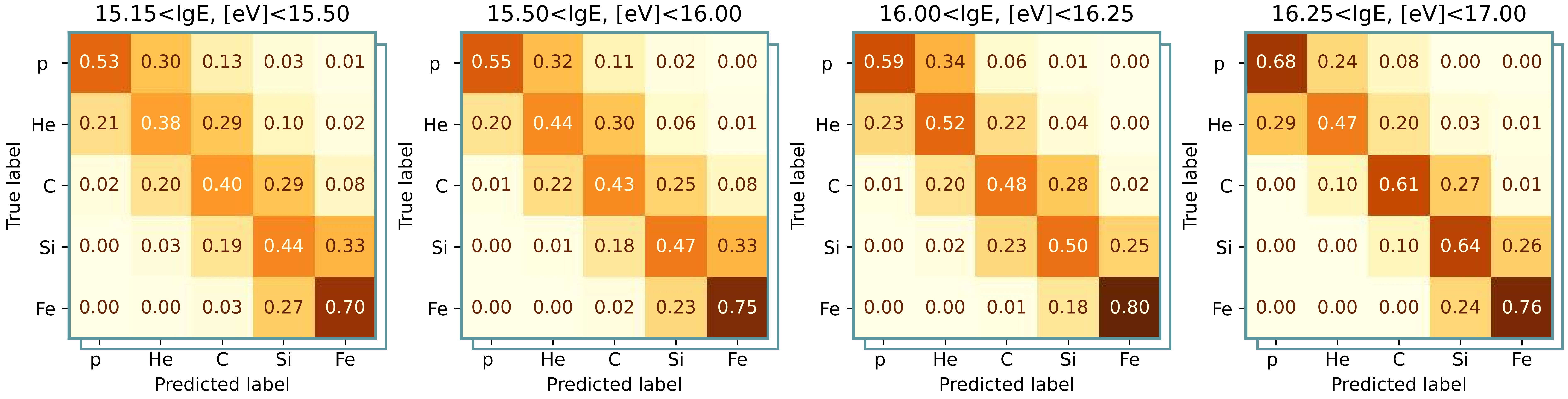
Missing detectors study

Compare CNN performance on default and “corrupted” datasets

Decrease of diagonal cells of the confusion matrices by up to 4%

Energy dependence

The more energetic showers are better classified



15.15<lgE, [eV]<15.50

True label \ Predicted label	p	He	C	Si	Fe
p	0.53	0.30	0.13	0.03	0.01
He	0.21	0.38	0.29	0.10	0.02
C	0.02	0.20	0.40	0.29	0.08
Si	0.00	0.03	0.19	0.44	0.33
Fe	0.00	0.00	0.03	0.27	0.70

15.50<lgE, [eV]<16.00

True label \ Predicted label	p	He	C	Si	Fe
p	0.55	0.32	0.11	0.02	0.00
He	0.20	0.44	0.30	0.06	0.01
C	0.01	0.22	0.43	0.25	0.08
Si	0.00	0.01	0.18	0.47	0.33
Fe	0.00	0.00	0.02	0.23	0.75

16.00<lgE, [eV]<16.25

True label \ Predicted label	p	He	C	Si	Fe
p	0.59	0.34	0.06	0.01	0.00
He	0.23	0.52	0.22	0.04	0.00
C	0.01	0.20	0.48	0.28	0.02
Si	0.00	0.02	0.23	0.50	0.25
Fe	0.00	0.00	0.01	0.18	0.80

16.25<lgE, [eV]<17.00

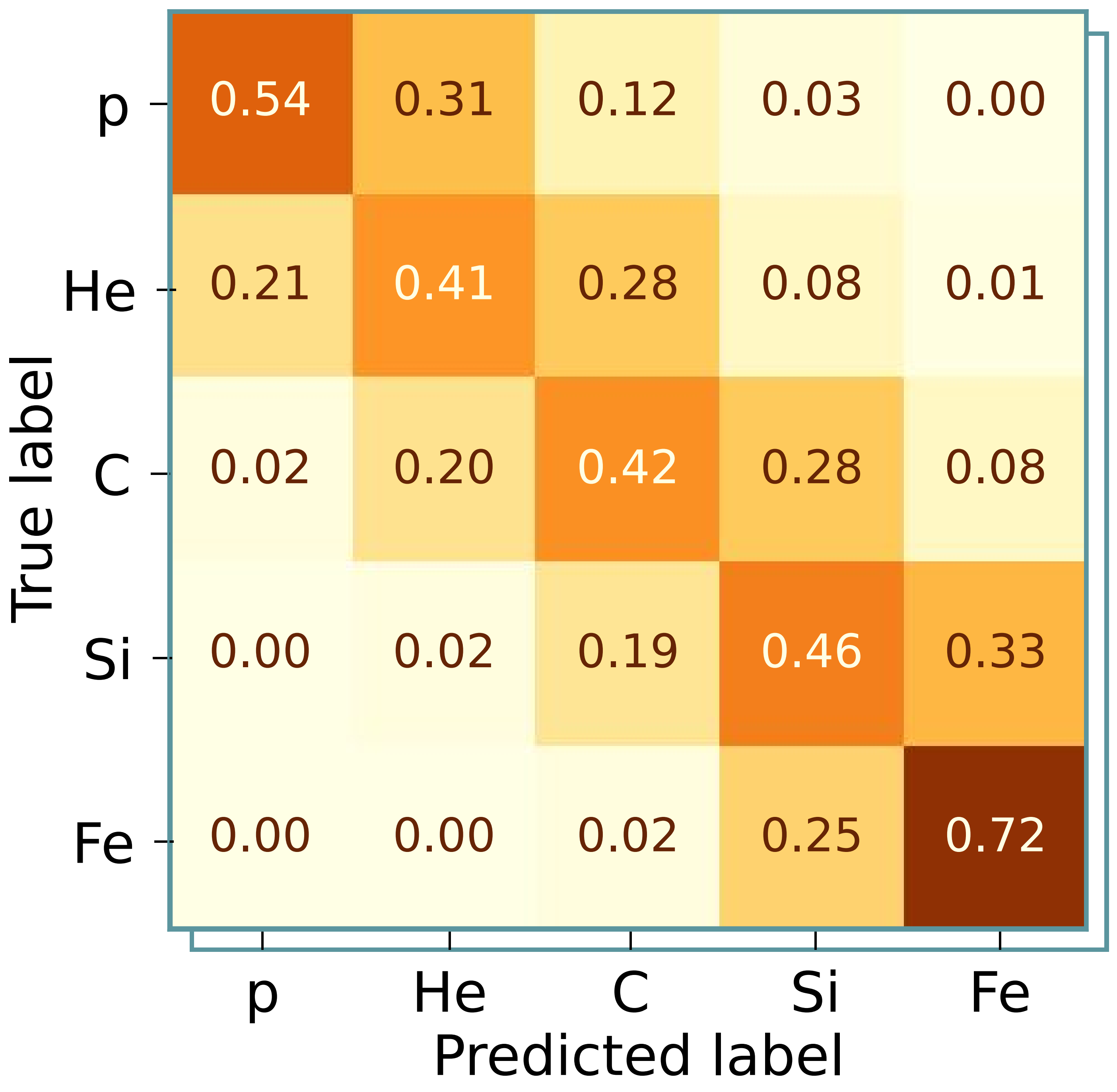
True label \ Predicted label	p	He	C	Si	Fe
p	0.68	0.24	0.08	0.00	0.00
He	0.29	0.47	0.20	0.03	0.01
C	0.00	0.10	0.61	0.27	0.01
Si	0.00	0.00	0.10	0.64	0.26
Fe	0.00	0.00	0.00	0.24	0.76

Cross-hadronic reconstruction

Test the same CNN (trained on QGSJet-II.04) on different hadronic models

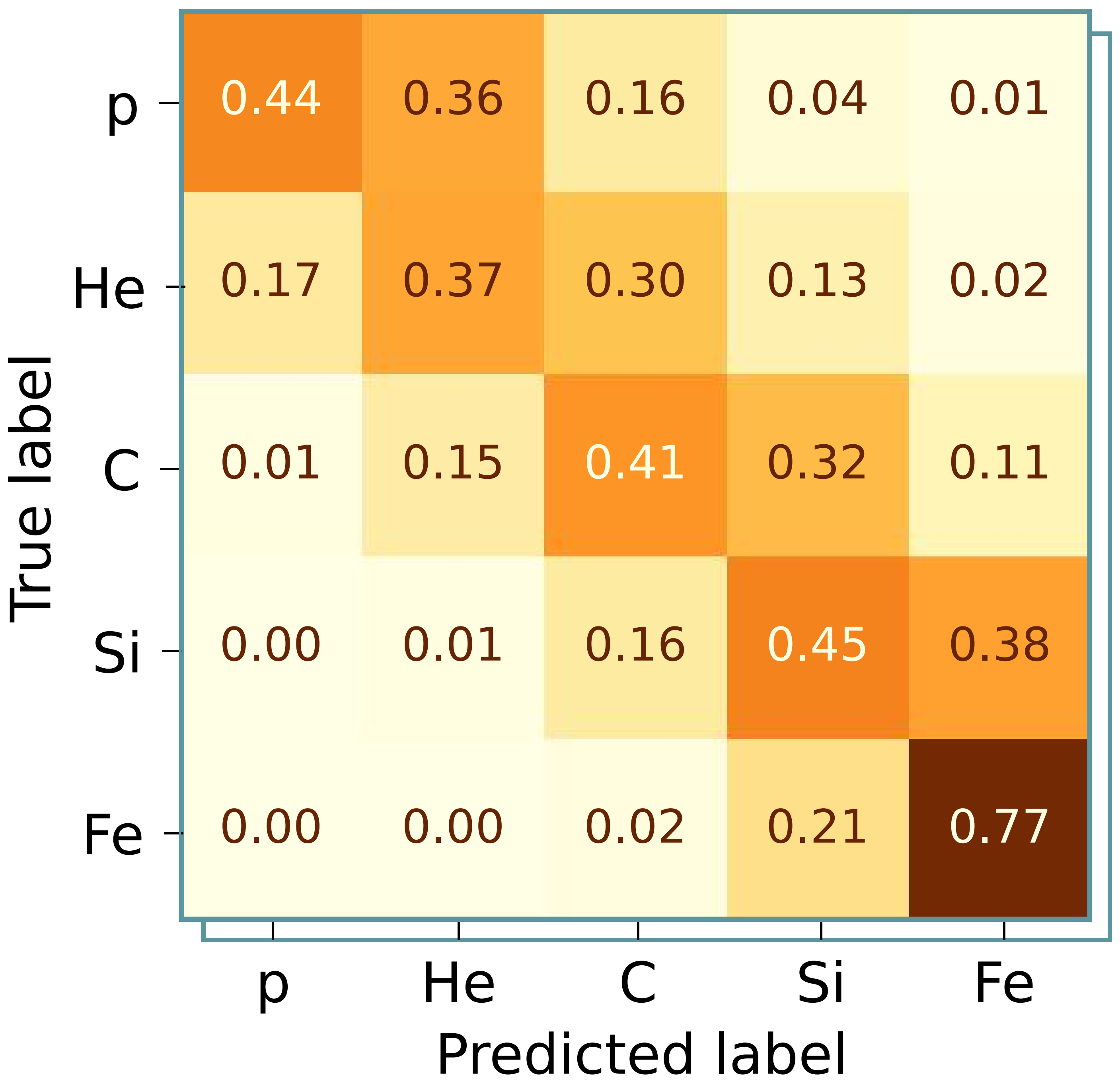
EPOS-LHC predicts “lighter” composition (vs QGSJet-II.04), Sibyll 2.3c → “harder”

Test: QGSJet-II.04



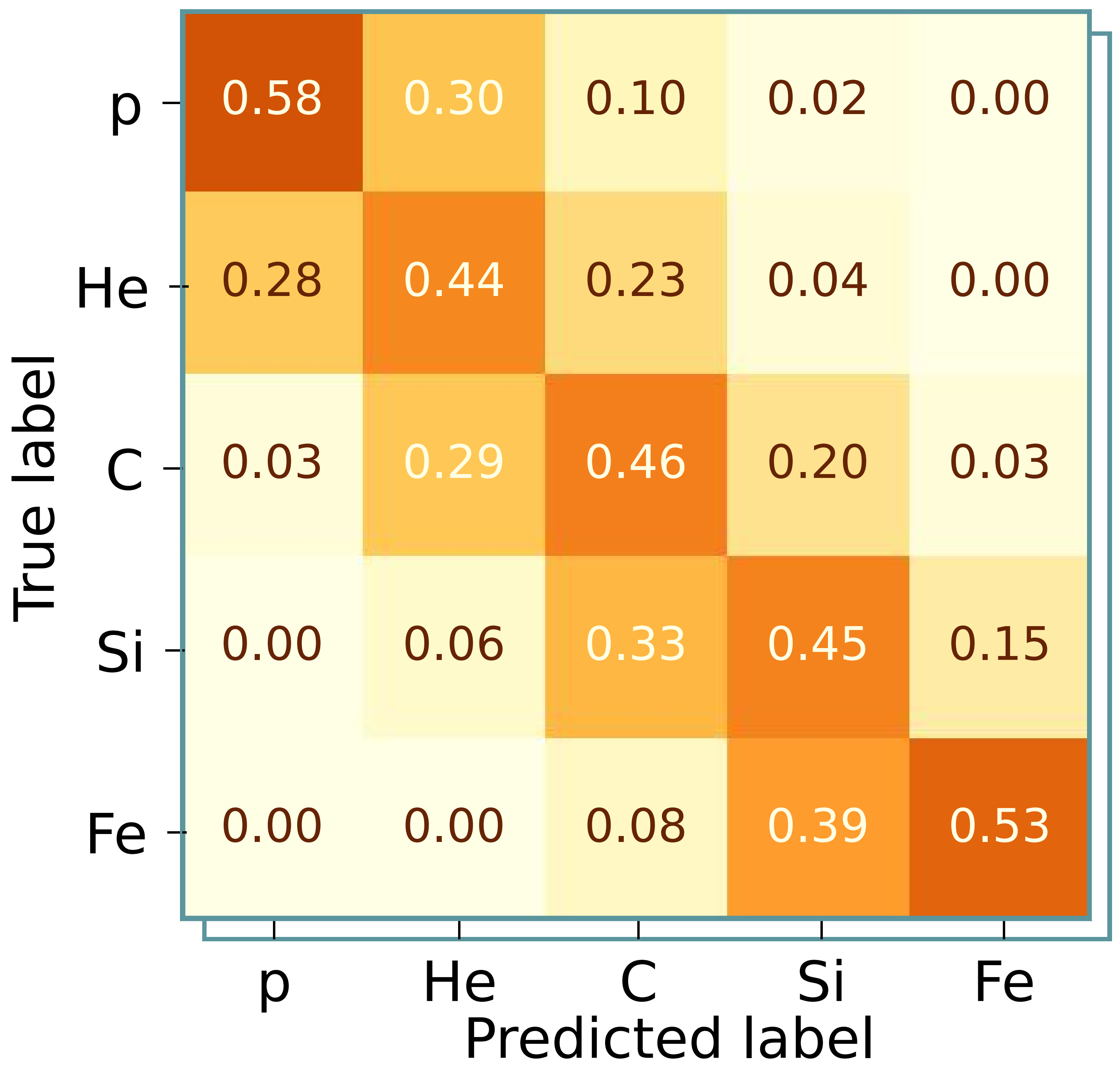
	p	He	C	Si	Fe
p	0.54	0.31	0.12	0.03	0.00
He	0.21	0.41	0.28	0.08	0.01
C	0.02	0.20	0.42	0.28	0.08
Si	0.00	0.02	0.19	0.46	0.33
Fe	0.00	0.00	0.02	0.25	0.72

Test: EPOS-LHC



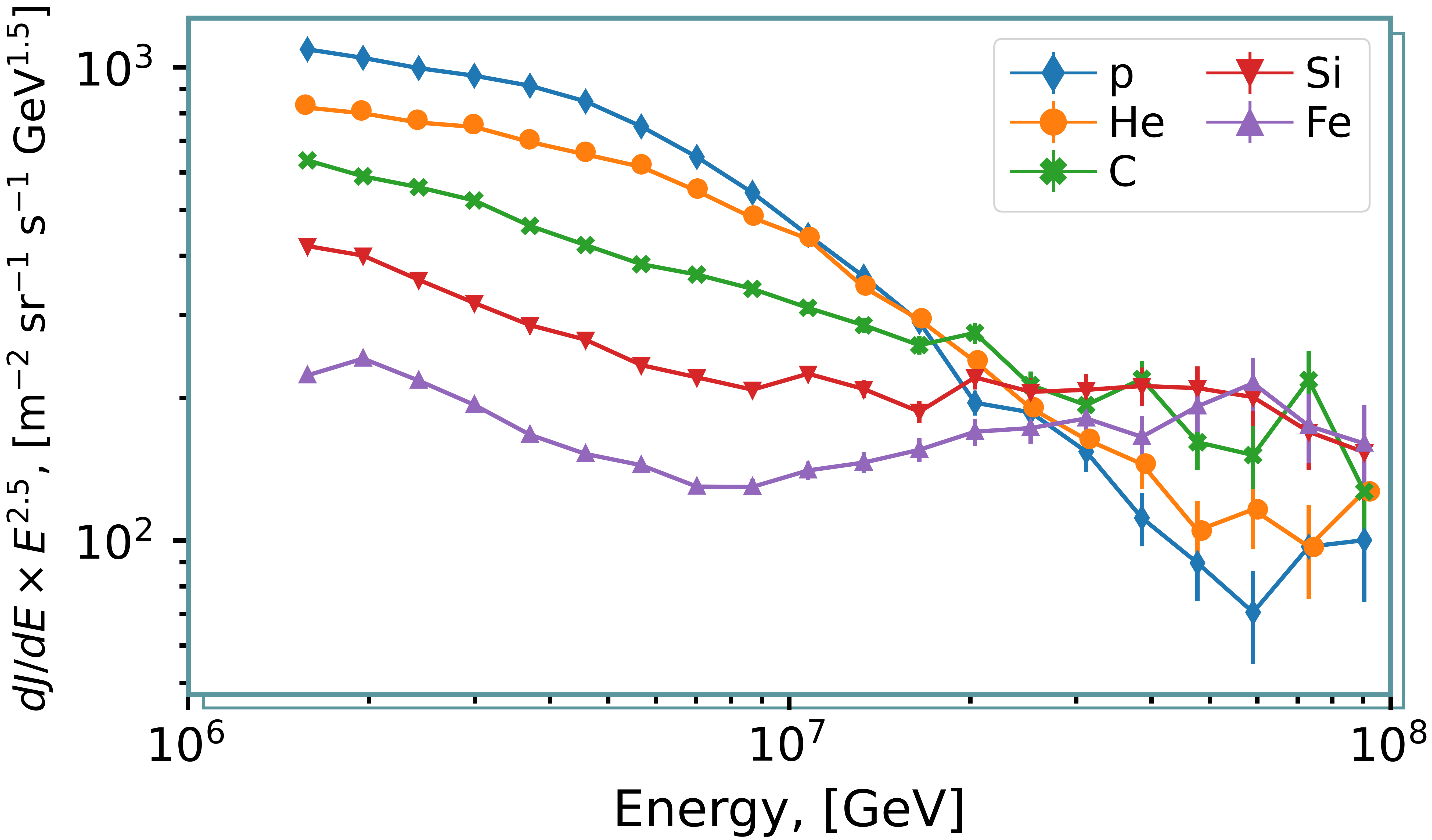
	p	He	C	Si	Fe
p	0.44	0.36	0.16	0.04	0.01
He	0.17	0.37	0.30	0.13	0.02
C	0.01	0.15	0.41	0.32	0.11
Si	0.00	0.01	0.16	0.45	0.38
Fe	0.00	0.00	0.02	0.21	0.77

Test: Sybill 2.3c



	p	He	C	Si	Fe
p	0.58	0.30	0.10	0.02	0.00
He	0.28	0.44	0.23	0.04	0.00
C	0.03	0.29	0.46	0.20	0.03
Si	0.00	0.06	0.33	0.45	0.15
Fe	0.00	0.00	0.08	0.39	0.53

Folded energy spectra



Folded energy spectra, unblind experimental data (CNN, trained with QGSJet-II.04)

Folded spectra means the spectra obtained by the direct predictions of the classifier

Unblind set is 20% of the total experimental data

Unfolding

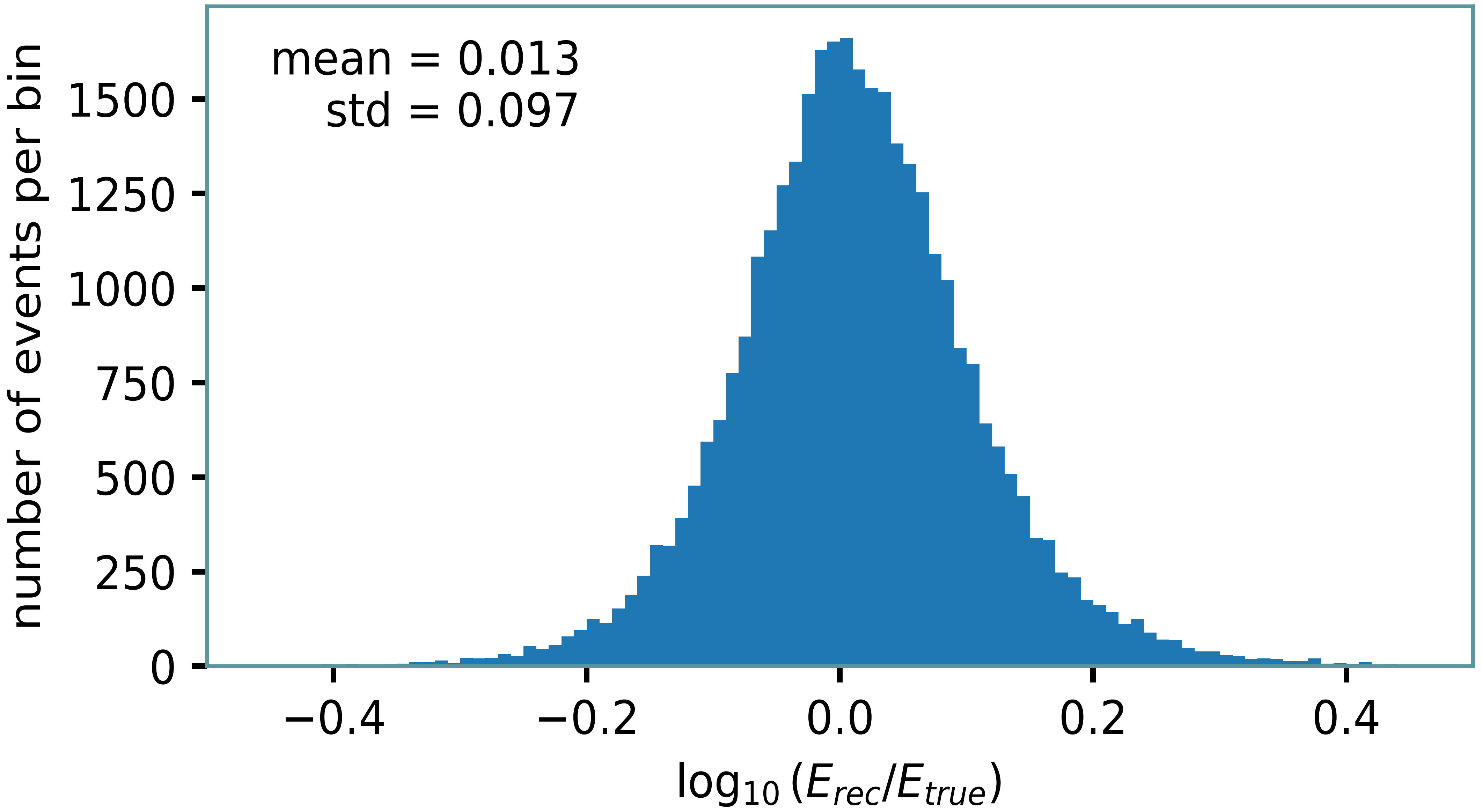
a correction to the confusion matrix

We reconstruct mass composition spectra with unfolding procedure

We apply consequently two unfoldings:

- energy unfolding
- particle type unfolding

We use iterative bayesian unfolding method from pyunfold* package

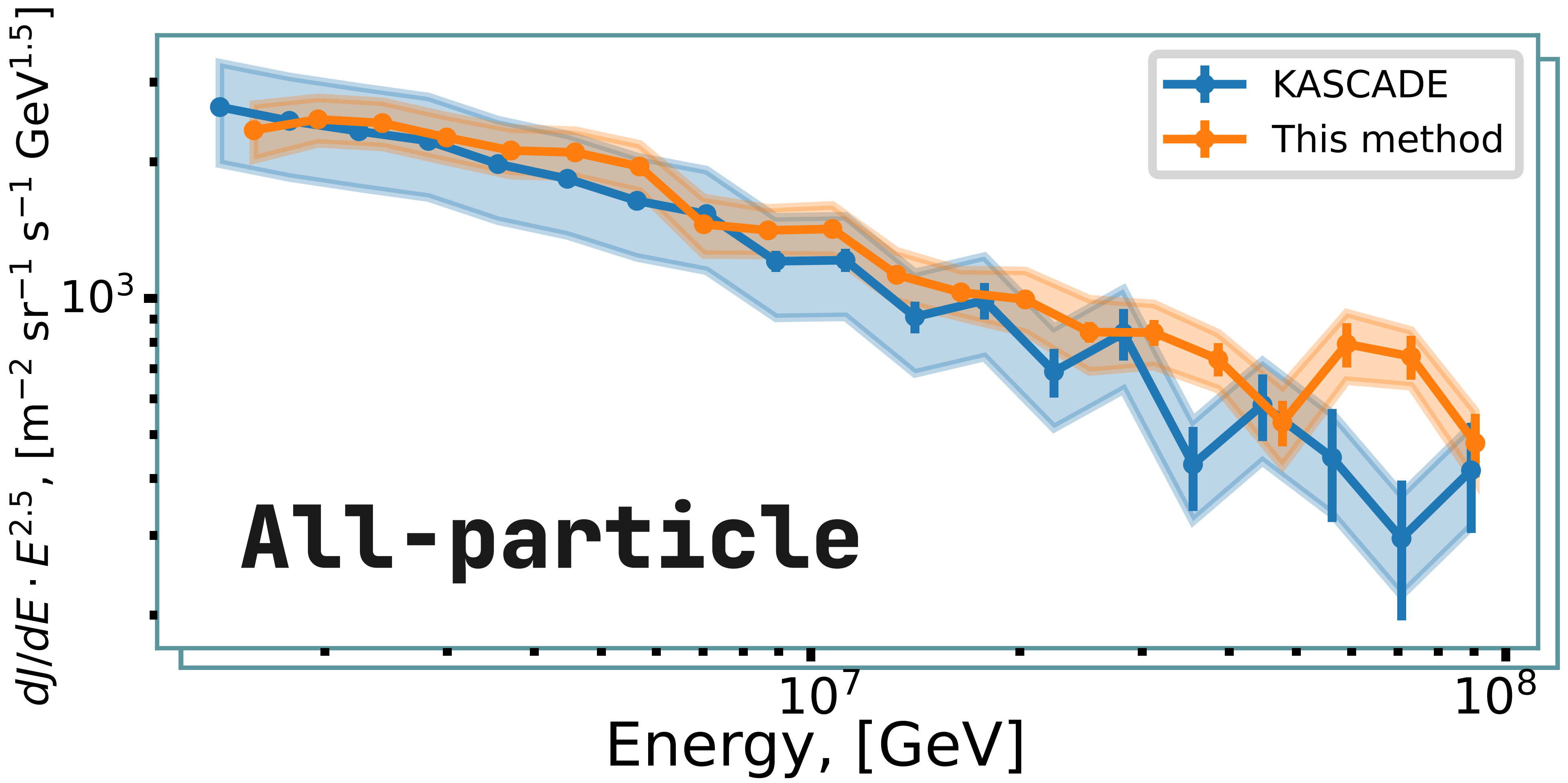


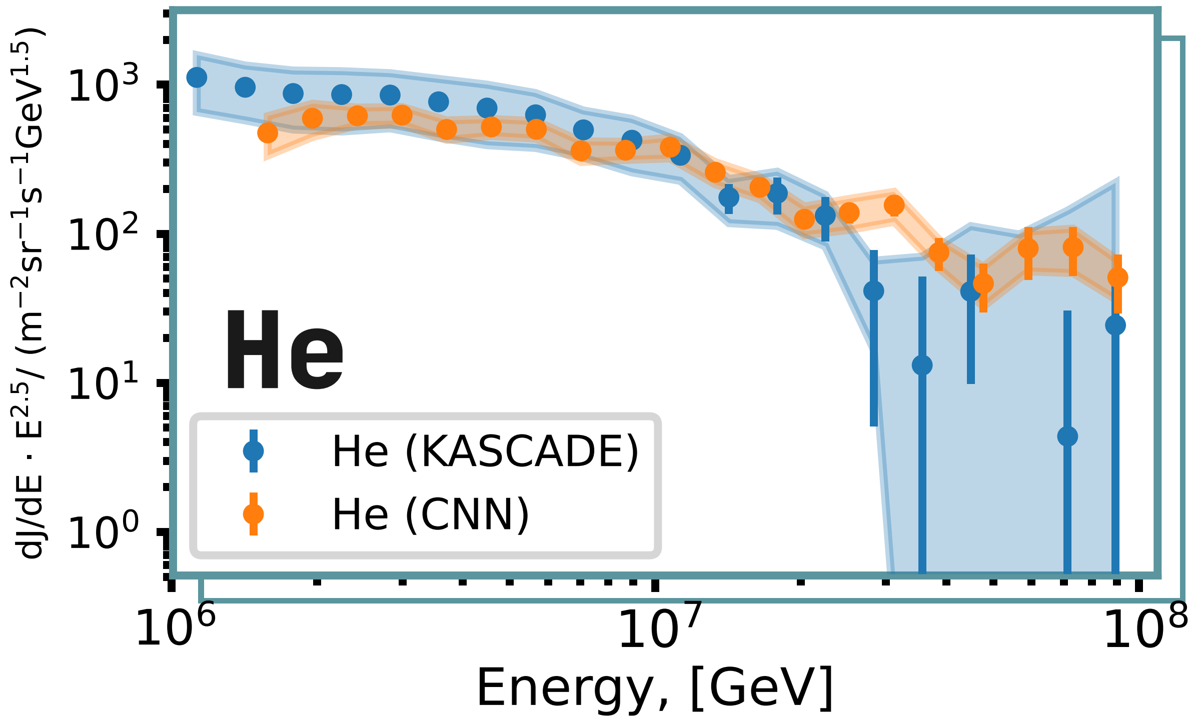
Energy resolution (default KASCADE reconstruction, QGSJet-II.02)

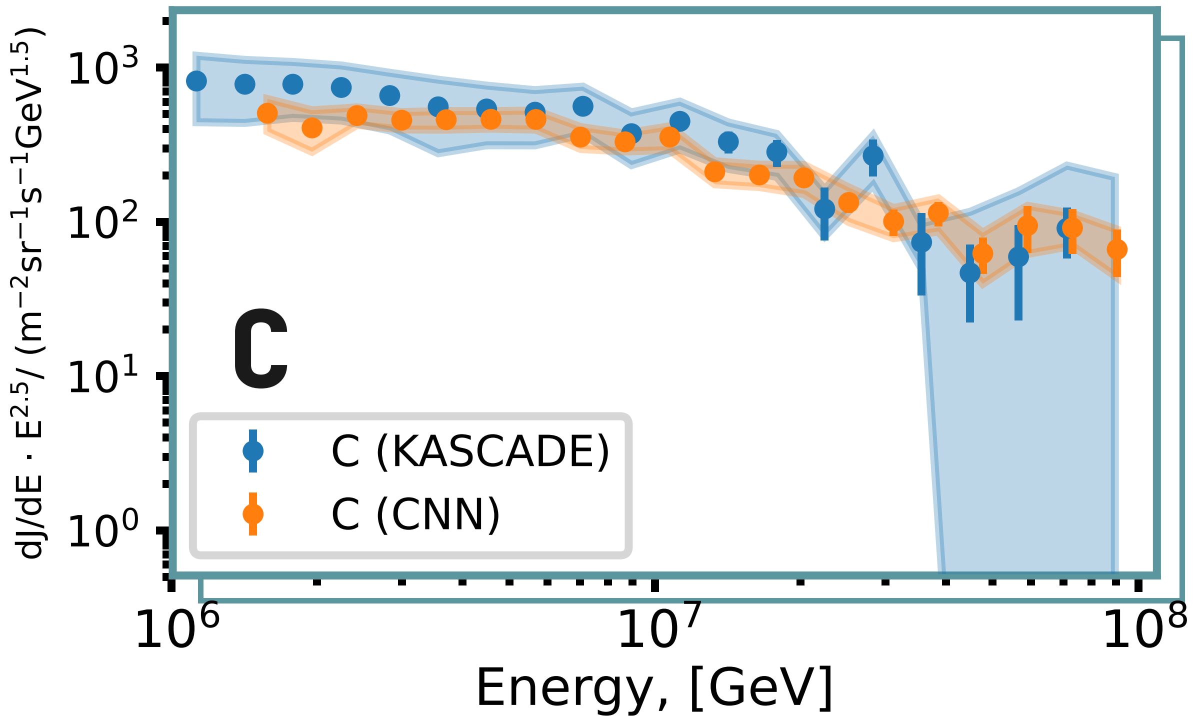
---

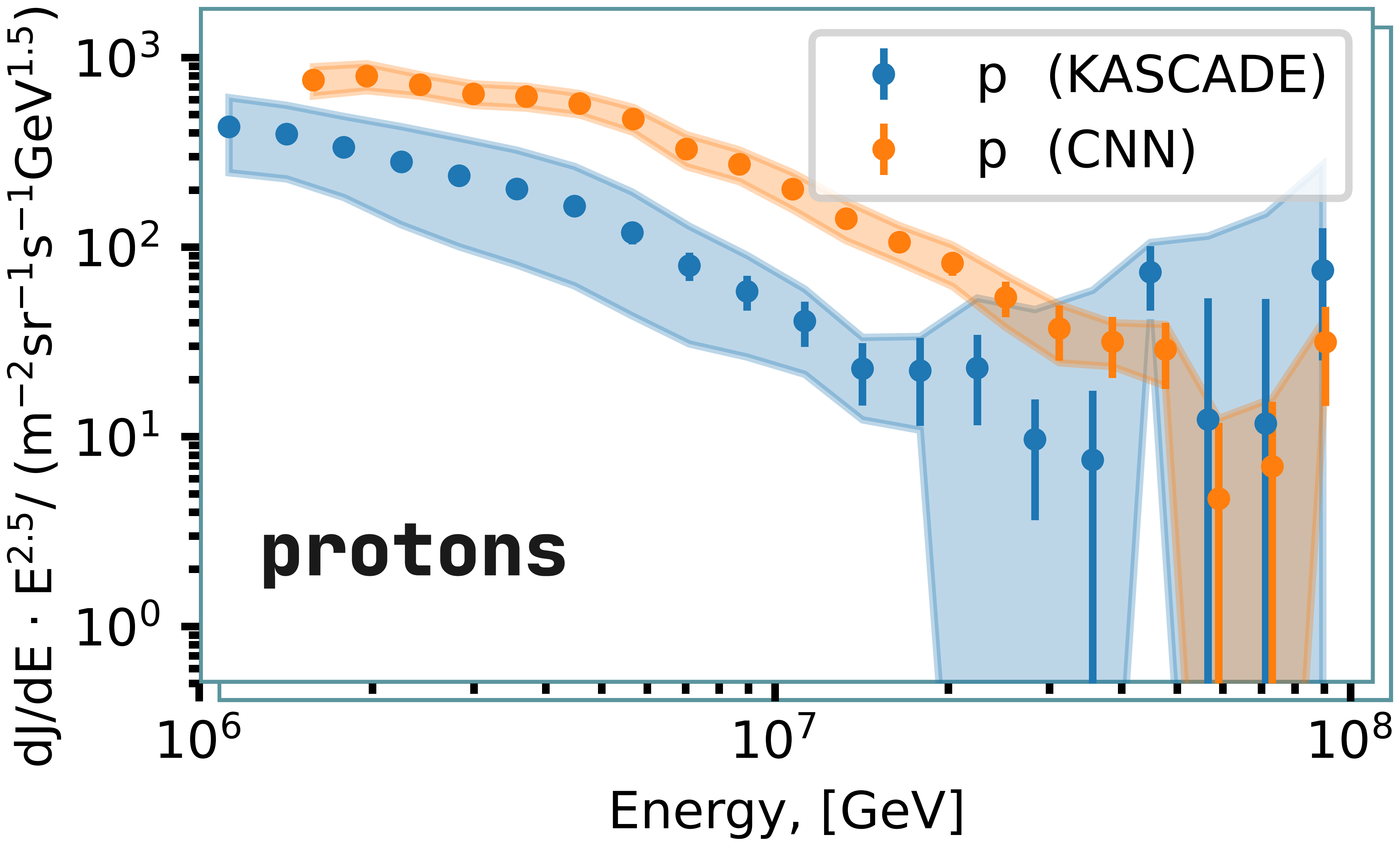
* James Bourbeau and Zigfried Hampel-Arias. Pyunfold: A python package for iterative unfolding. The Journal of Open Source Software, 3(26):741, June 2018. doi:10.21105/joss.00741

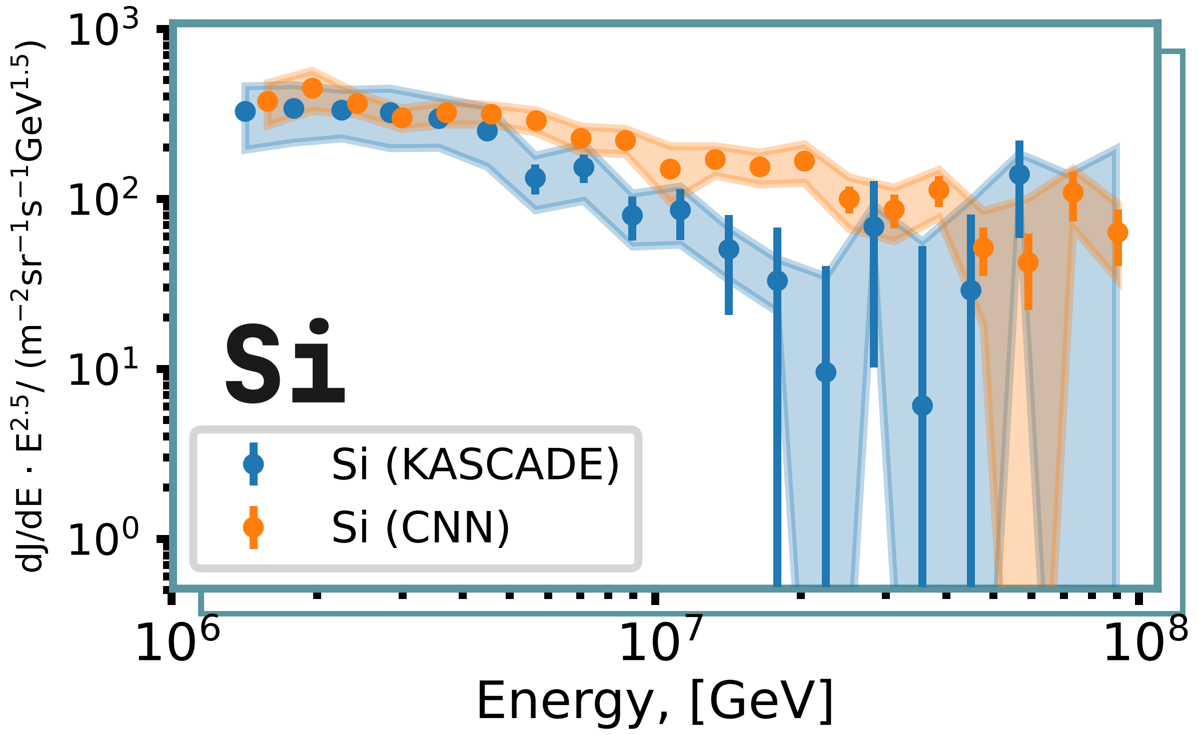
QGSJet-II.02 comparison

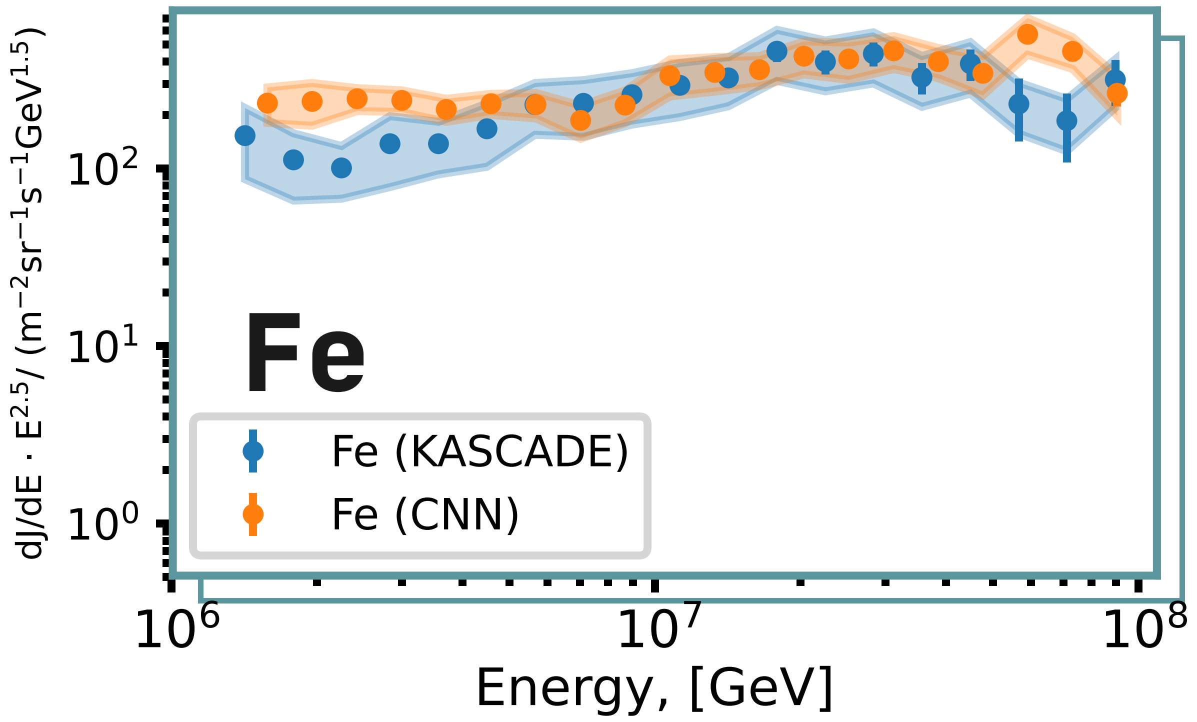








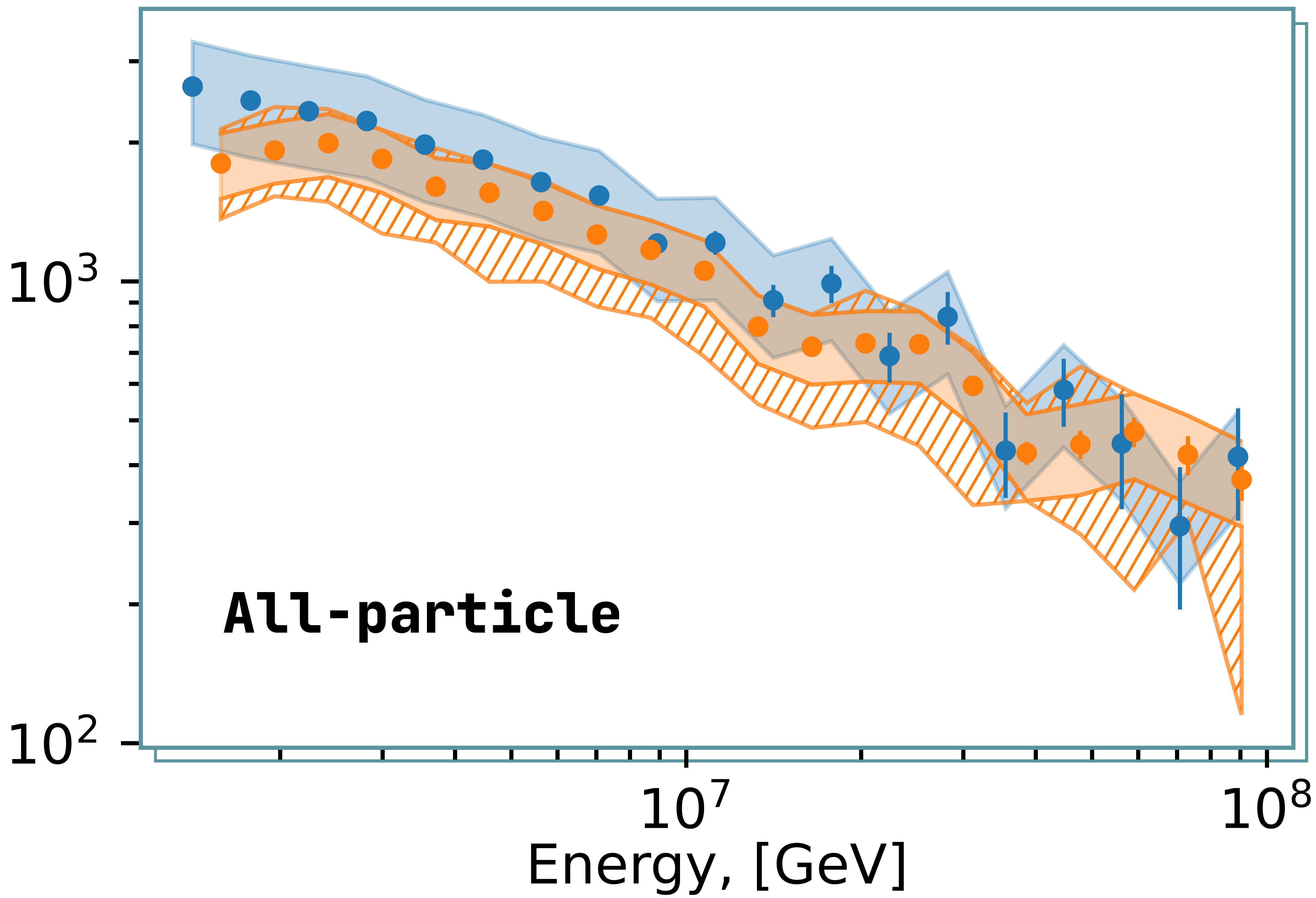




One-to-one comparison of this work (orange, unblind data) and original KASCADE spectra* (blue, QGSJet-II.02 hadronic interaction model)

* Apel, W. D. et al. (2013). KASCADE-Grande measurements of energy spectra for elemental groups of cosmic rays. Astropart. Physics, 47, 54–66. doi:10.1016/j.astropartphys.2013.06.004

Results (QGSJet-II.04, EPOS-LHC, Sibyll 2.3c)



Reconstructed all-particle energy spectrum in this (orange, blind data, QGSJet-II.04, EPOS-LHC, Sibyll 2.3c) and original KASCADE (blue, QGSJet-II.02)

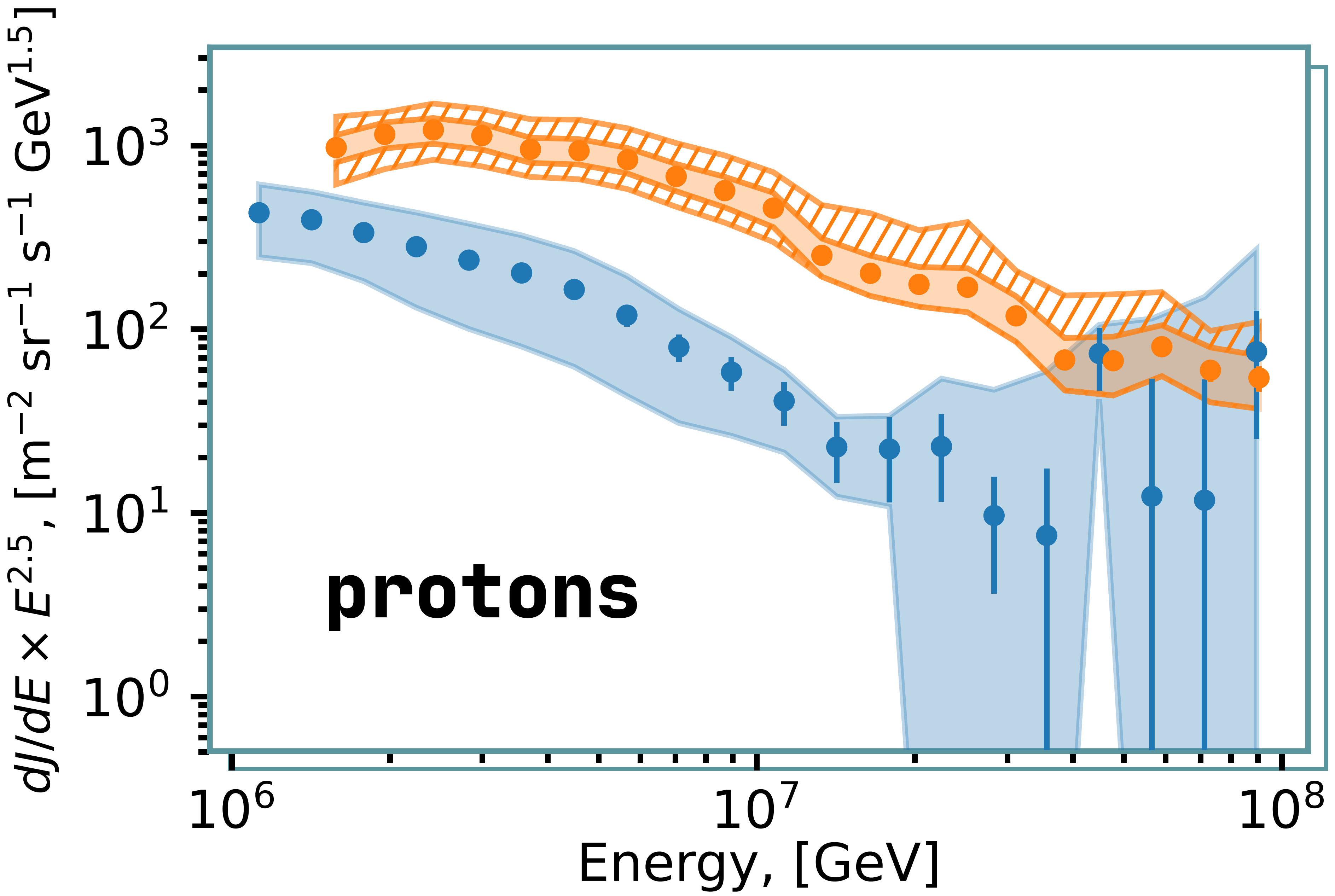
Our (orange) points, error bars, solid bands for QGSJet-II.04

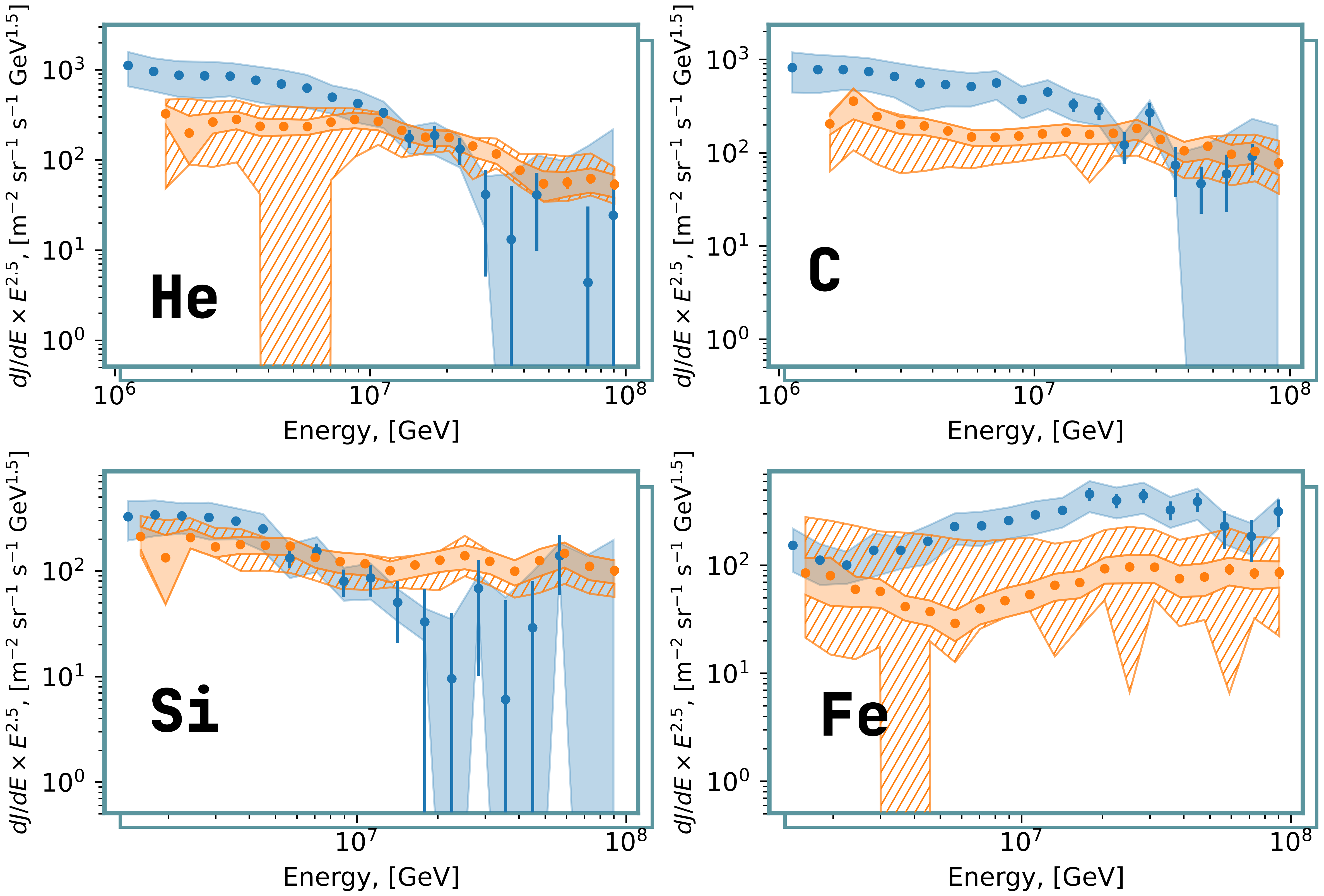
Theoretical uncertainties

A range between the minimum and maximum edges of the "basic" systematic uncertainty bands among all hadronic models used (hatches in fig.)

Results (QGSJet-II.04, EPOS-LHC, Sibyll 2.3c)

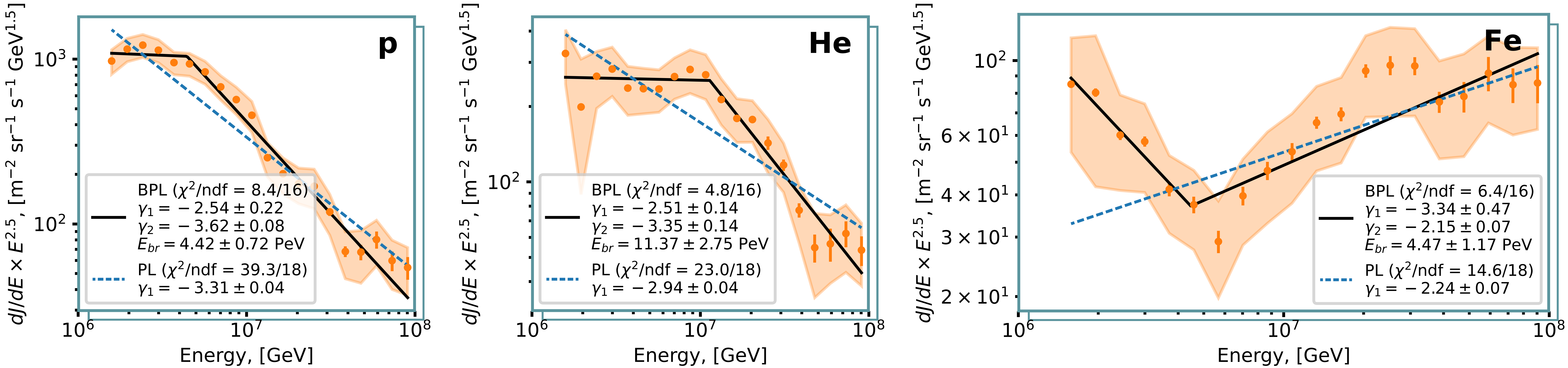
Orange: reconstructed spectra for QGSJet-II.04 on blind data with theoretical systematics (hatch)
Original KASCADE results (blue, QGSJet-II.02) for illustration purposes





Knee-like structure search

- Spectra of the proton and helium components show knee-like features (5.2σ and 3.9σ respectively)
- Iron component shows a hint (2.4σ) of the break at ~ 4.5 PeV
- No breaks are observed in the spectra of other components

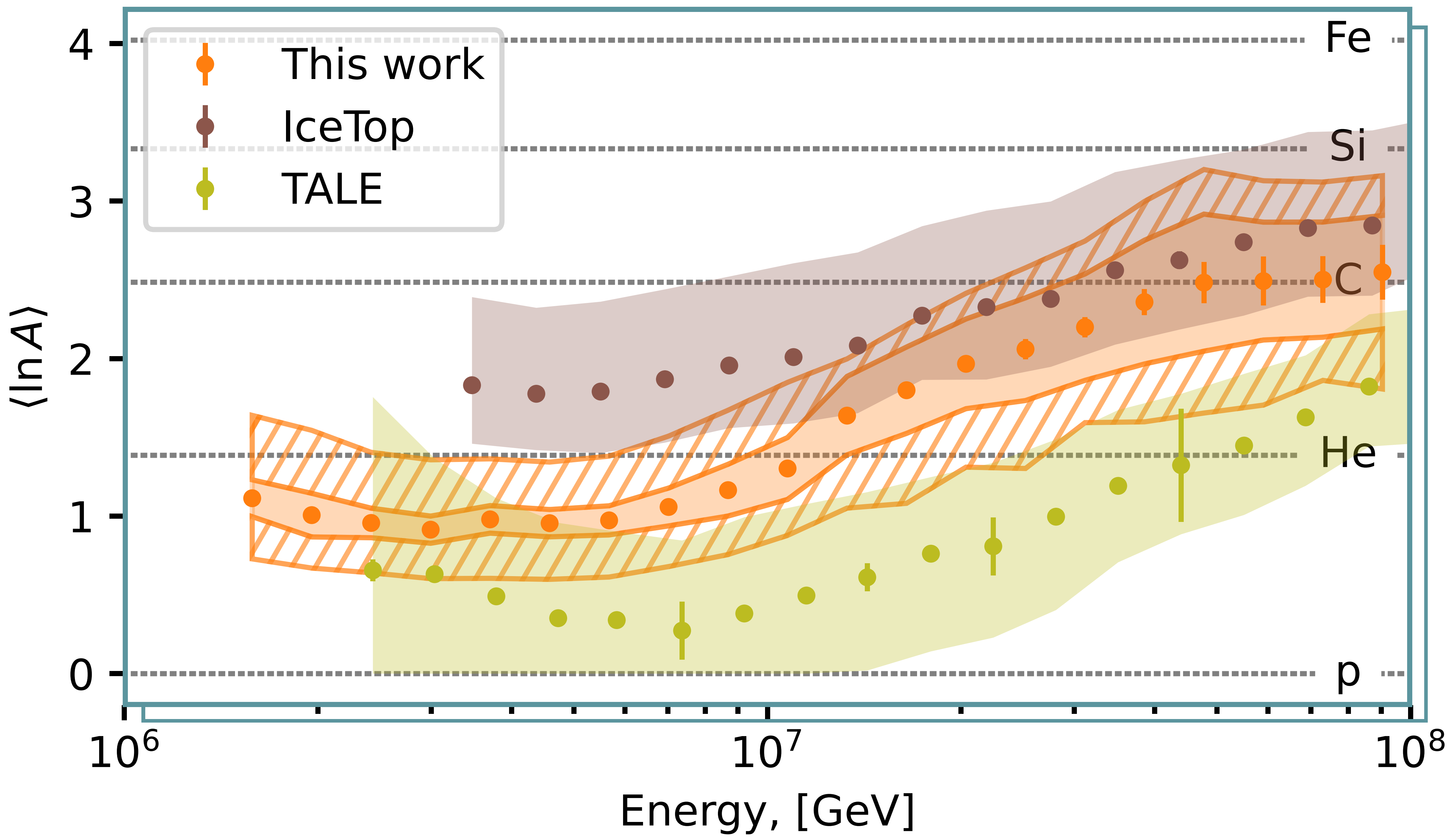


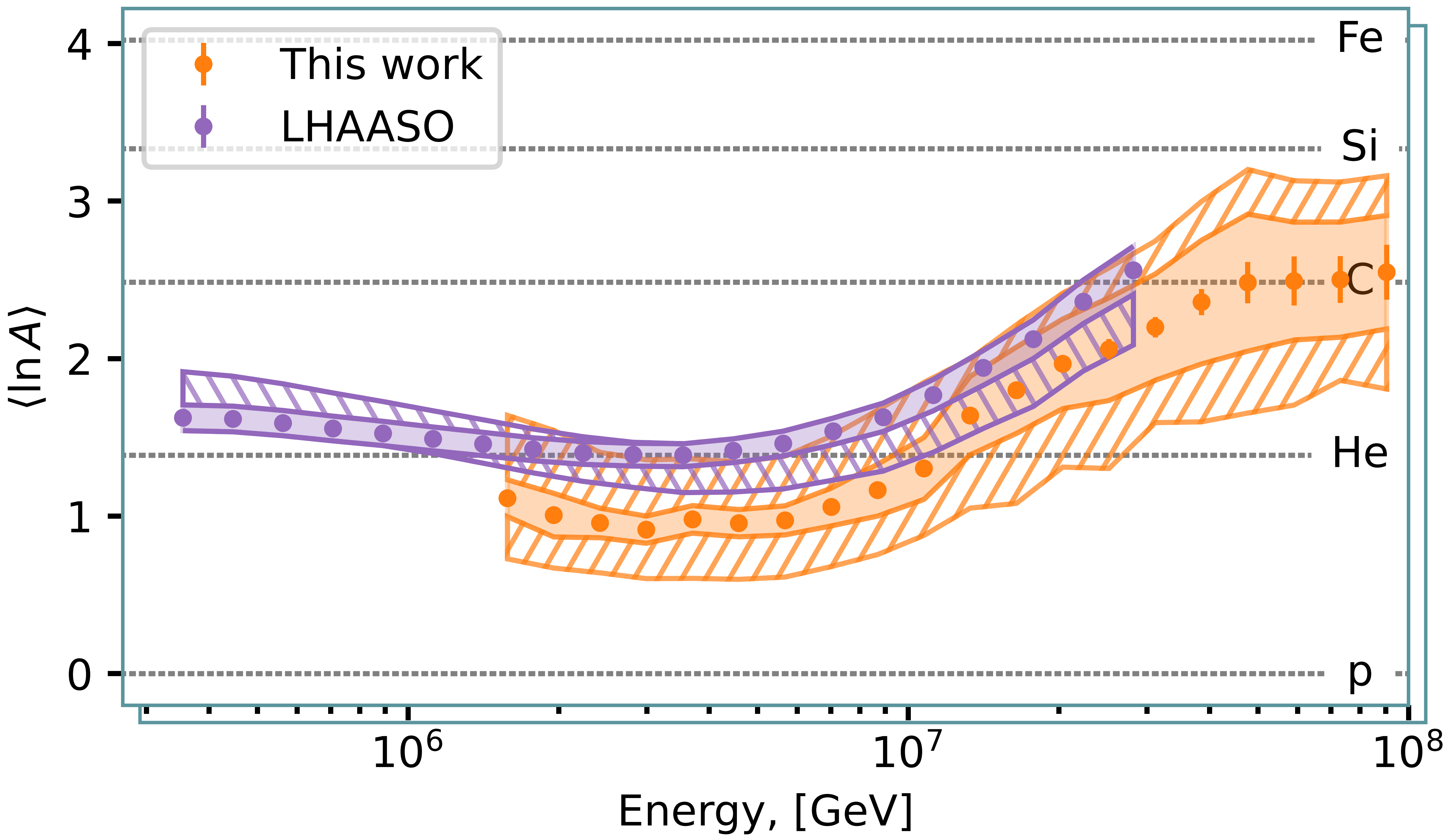
Individual mass component spectra. Power-law (PL, blue dash) and broken power-law (BPL, black solid) fits.

⟨lnA⟩ comparison

$$\langle \ln A \rangle = \sum_{i=1}^{5} f_i \ln A_i$$

These results are in partial agreement with IceTop and TALE
EPOS-LHC closer to TALE, a Sibyll 2.3c — to IceTop





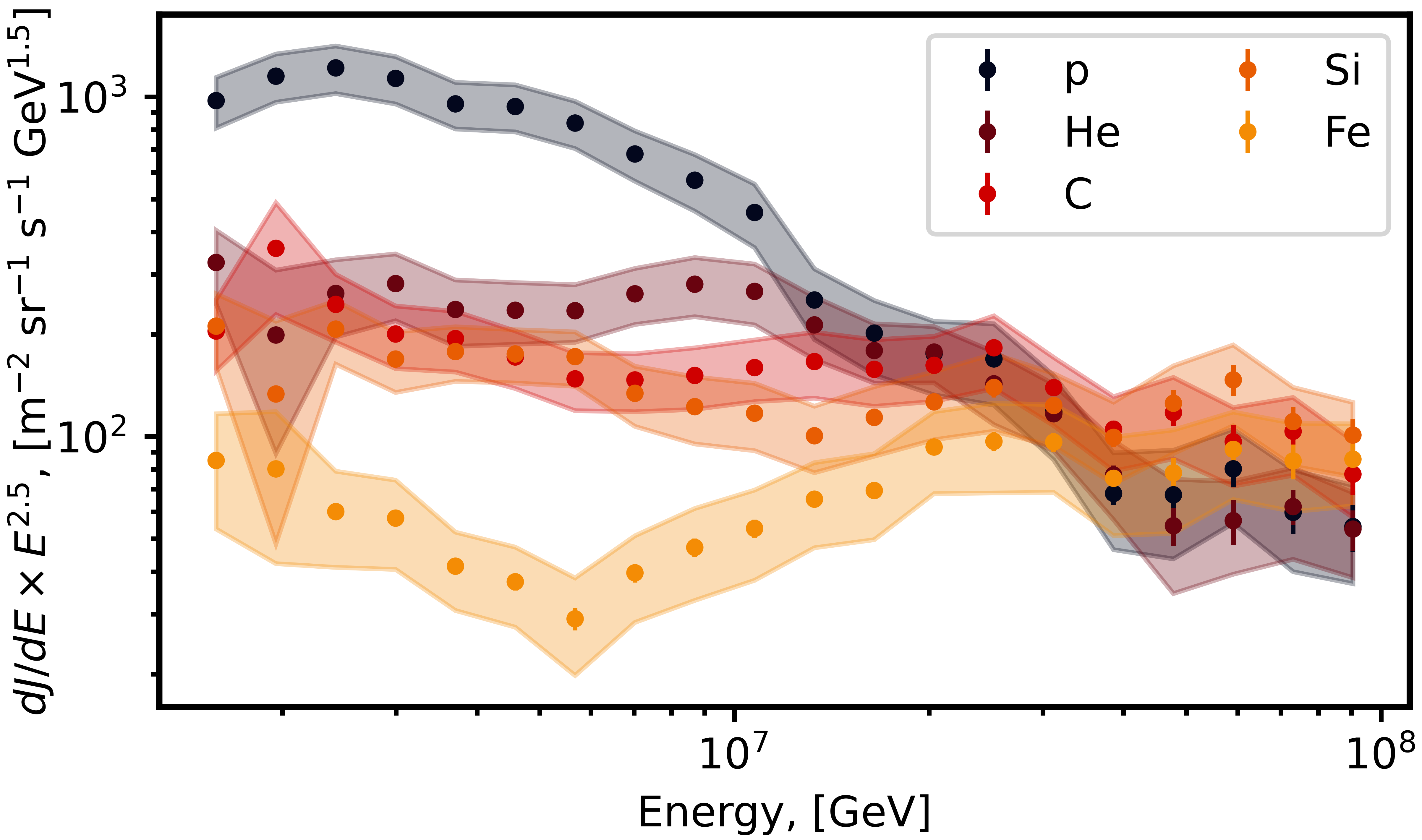
LHAASO collaboration, Phys. Rev. Lett. 132 (2024) 131002 [2403.10010]
Telescope Array collaboration, Astrophys. J. 909 (2021) 178 [2012.10372]
Aartsen, M. et al, Phys. Rev. D, 100(8), 082002.

Conclusion

- We reanalyzed data of KASCADE cosmic ray experiment
- We reconstructed cosmic ray mass components spectra for post-LHC hadronic interaction models (QGSJet-II.04, EPOS-LHC,Sibyll 2.3c) and took into account these systematics
- Basic uncertainties of the our method are much smaller than those of the standard KASCADE reconstruction
- We found a significant dominance of the proton component
- We found highly significant knee-like features in the proton and He individual spectra and a hint of the break in the iron spectrum.

Thanks for your attention!

QGSJet-II.04 results (only)

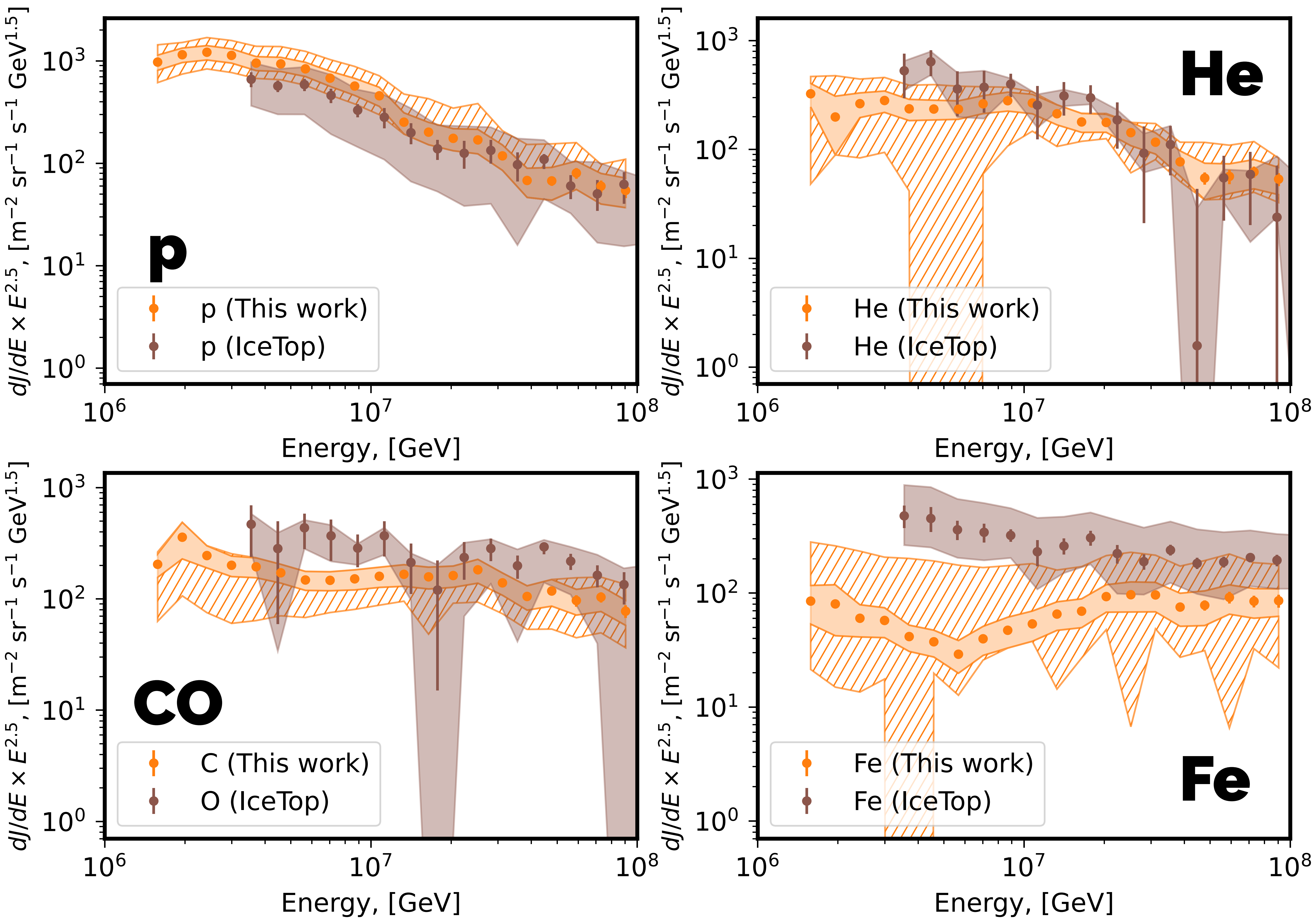


proton component dominates at energies < 10 PeV

Basic systematic uncertainties:

Missing detectors	5 – 18 %
MC mass composition	13 – 16 %
Limited MC	8 – 25 %
MC slope	up to 4 %
Unfolding regularization	1 – 24 %
Sequential unfolding	up to 8 %

IceTop comparison



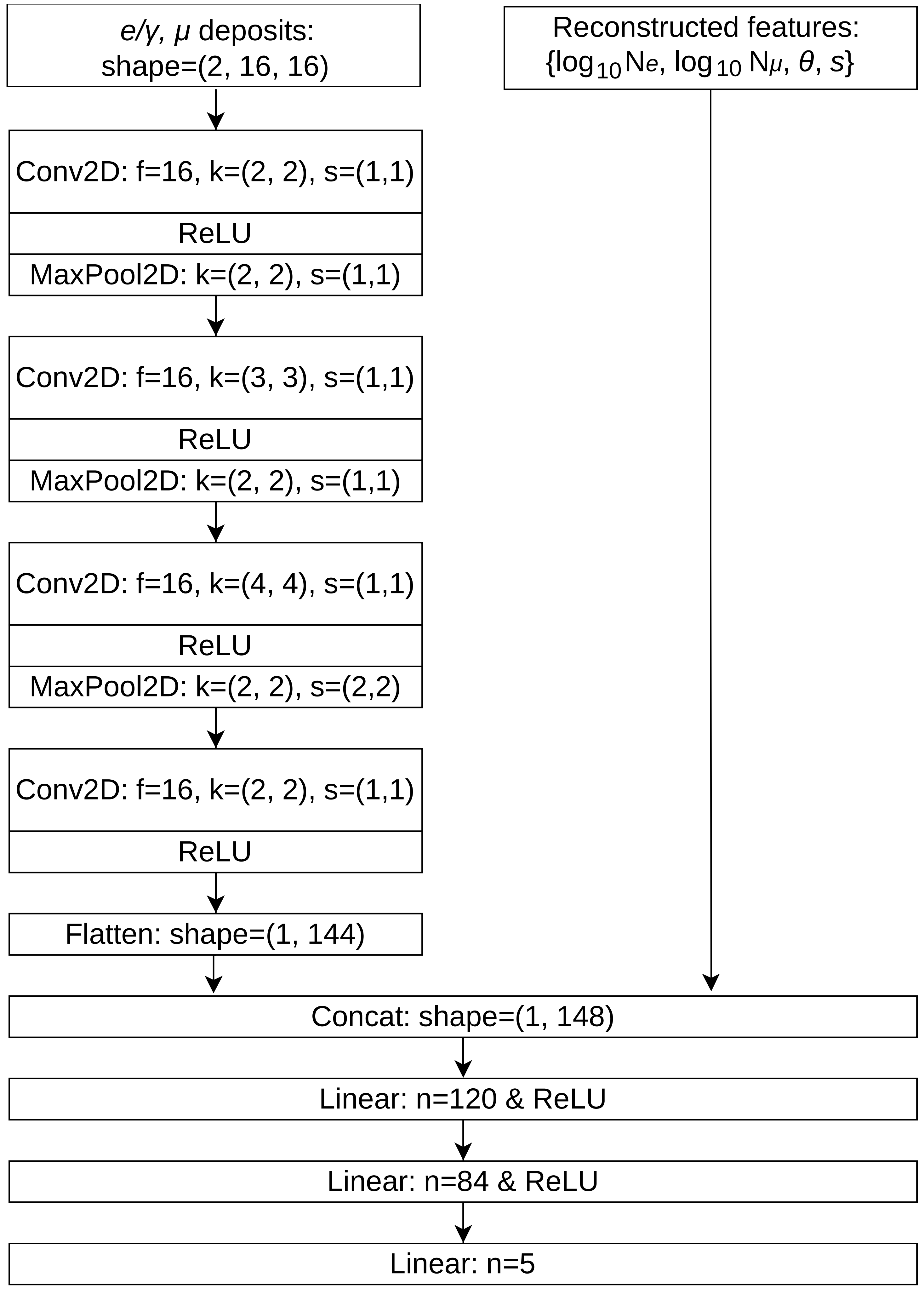
Orange: reconstructed spectra for QGSJet-II.04 hadronic interaction model on blind data with cross-hadronic model systematics

Brown: IceTop results*
(Sybill 2.1)

---

* Aartsen, M., & others (2019). Cosmic ray spectrum and composition from PeV to EeV using 3 years of data from IceTop and IceCube. Phys. Rev. D, 100(8), 082002.

Architectures



e/γ, μ deposits:
shape=(2, 16, 16)

Conv2D: f=16, k=(2, 2), s=(1,1)
ReLU
MaxPool2D: k=(2, 2), s=(1,1)

Conv2D: f=16, k=(3, 3), s=(1,1)
ReLU
MaxPool2D: k=(2, 2), s=(1,1)

Conv2D: f=16, k=(4, 4), s=(1,1)
ReLU
MaxPool2D: k=(2, 2), s=(2,2)

Conv2D: f=16, k=(2, 2), s=(1,1)
ReLU

Flatten: shape=(1, 144)

Reconstructed features:
$\{\log_{10} N_e, \log_{10} N_\mu, \theta, s\}$

Concat: shape=(1, 148)

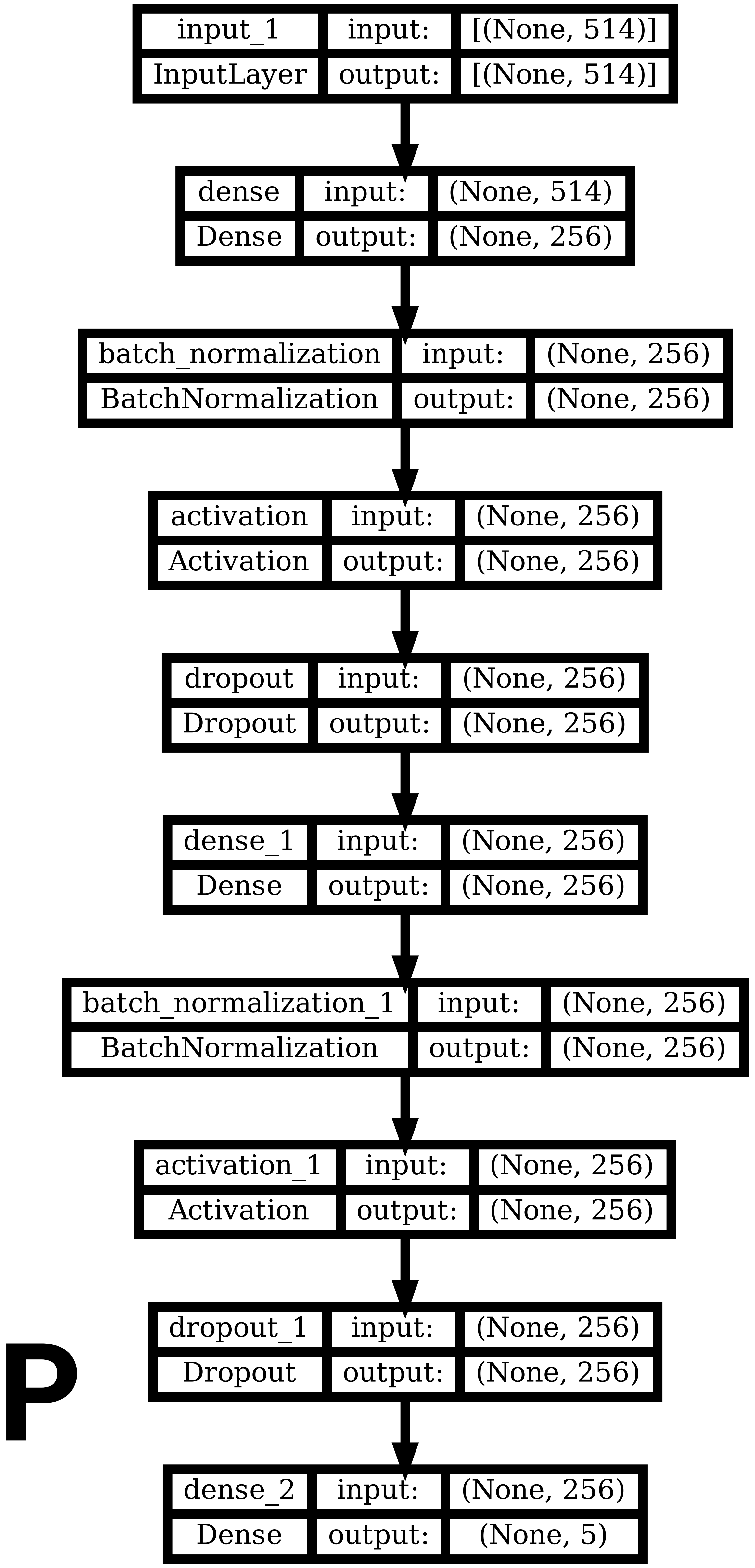
Linear: n=120 & ReLU

Linear: n=84 & ReLU

Linear: n=5

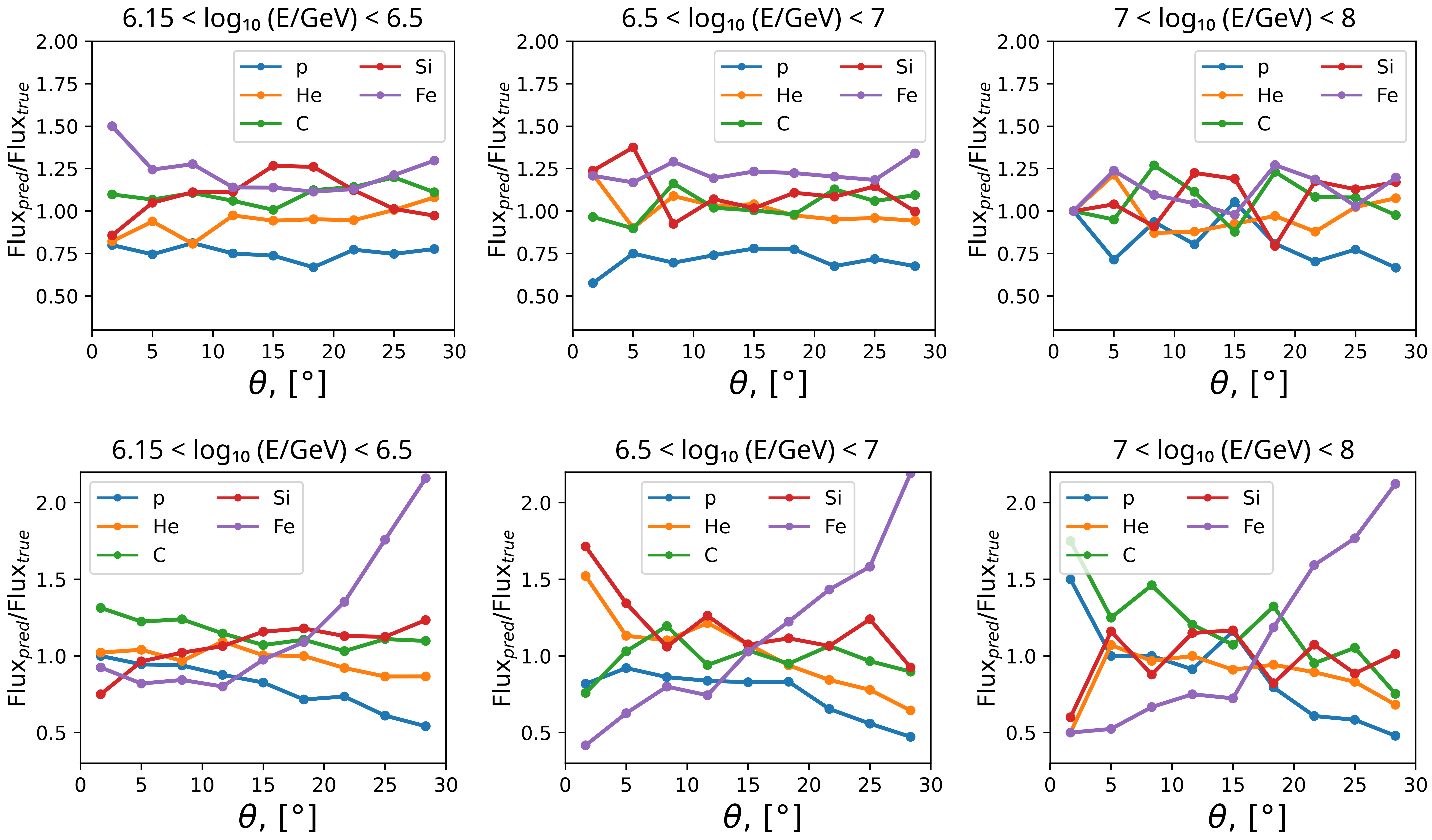
CNN

MLP



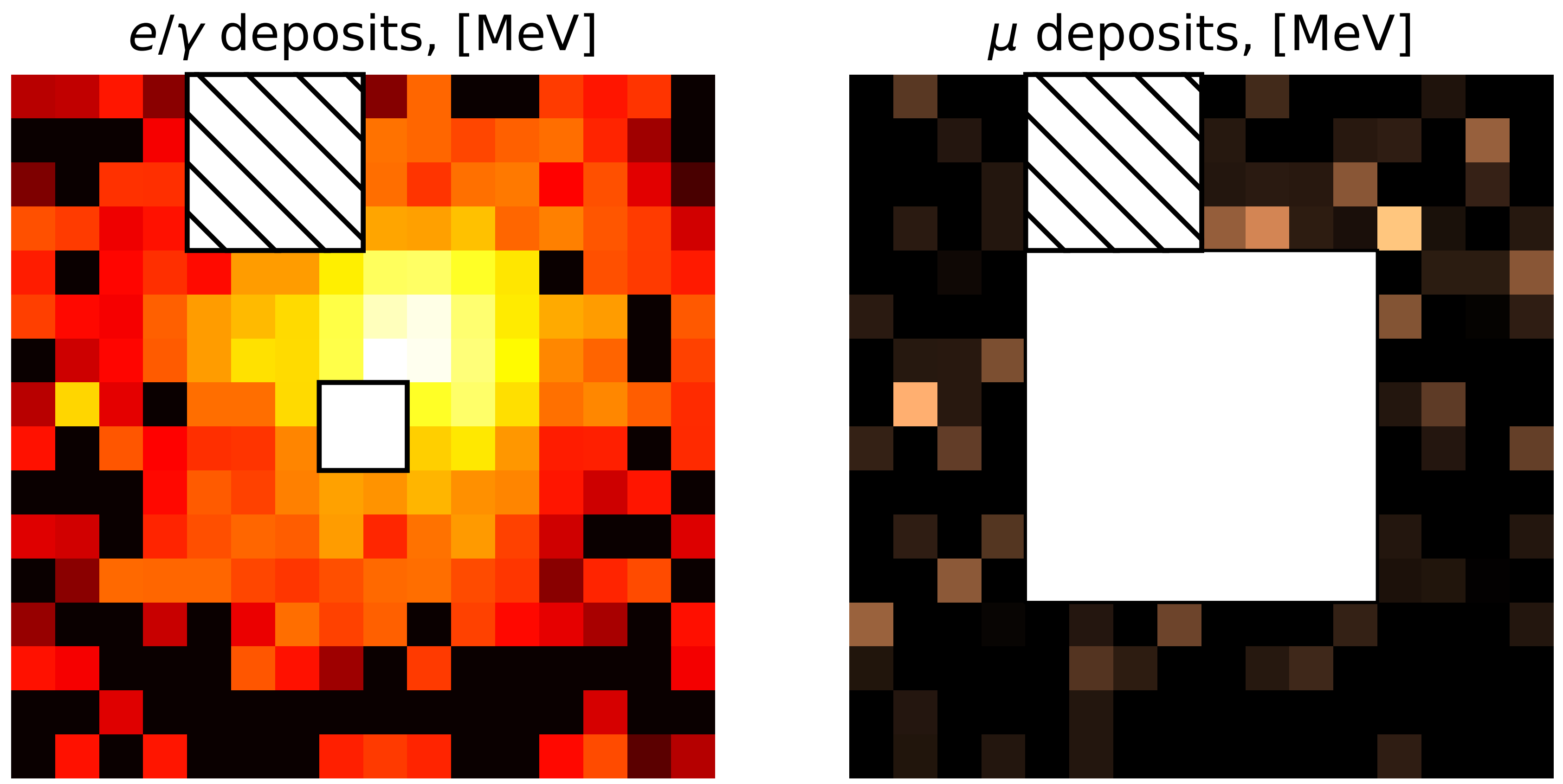
input_1	input:	[(None, 514)]
InputLayer	output:	[(None, 514)]
dense	input:	(None, 514)
Dense	output:	(None, 256)
batch_normalization	input:	(None, 256)
BatchNormalization	output:	(None, 256)
activation	input:	(None, 256)
Activation	output:	(None, 256)
dropout	input:	(None, 256)
Dropout	output:	(None, 256)
dense_1	input:	(None, 256)
Dense	output:	(None, 256)
batch_normalization_1	input:	(None, 256)
BatchNormalization	output:	(None, 256)
activation_1	input:	(None, 256)
Activation	output:	(None, 256)
dropout_1	input:	(None, 256)
Dropout	output:	(None, 256)
dense_2	input:	(None, 256)
Dense	output:	(None, 5)

Zenith angle dependence

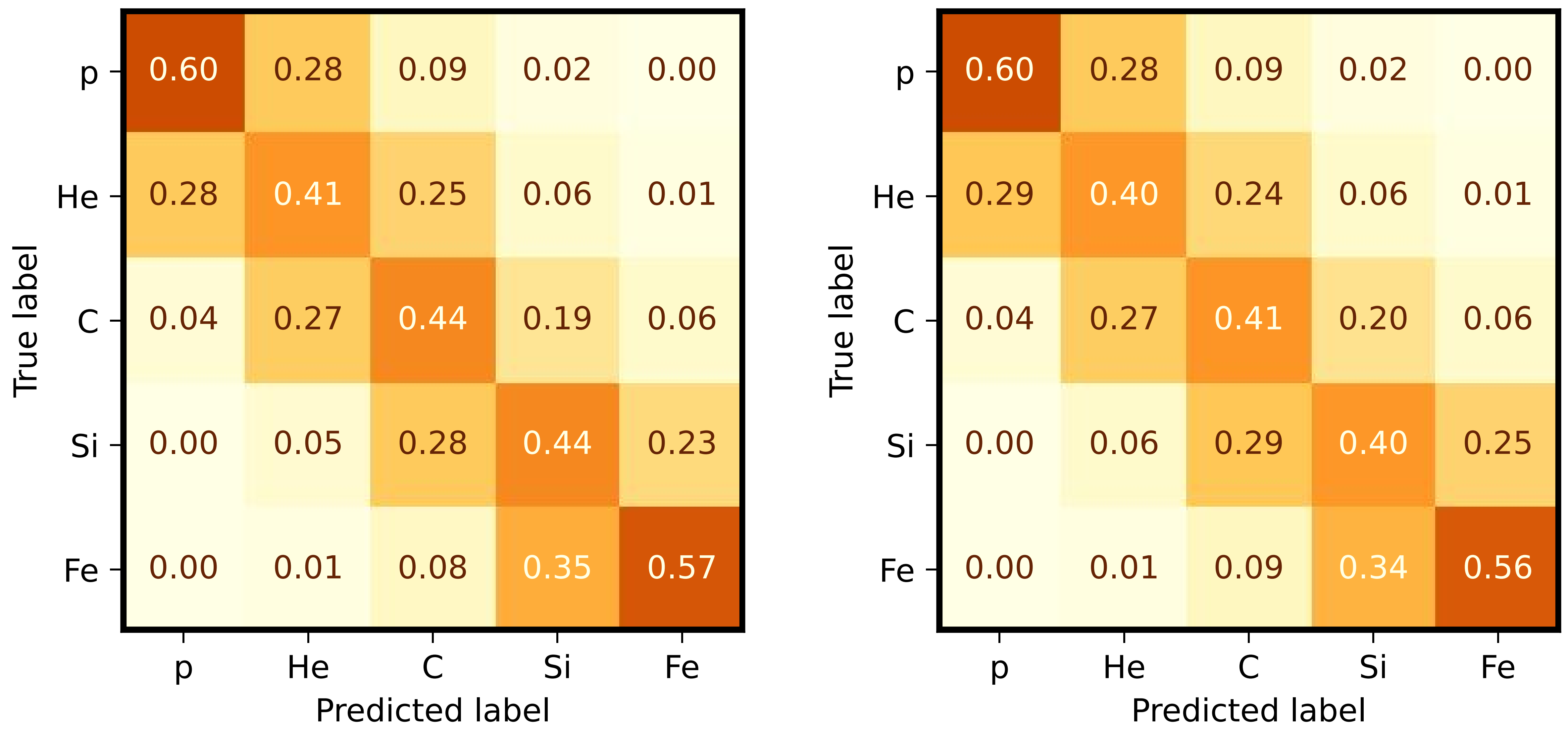


Dependence of the ratio of the predicted flux to the true flux on the zenith angle θ for different energy ranges.
Top for a default CNN, bottom for a CNN that does not use θ

Missing detectors



Example of spoiled Monte Carlo event (dashed area shows detectors not working)



True label \ Predicted label	p	He	C	Si	Fe
p	0.60	0.28	0.09	0.02	0.00
He	0.28	0.41	0.25	0.06	0.01
C	0.04	0.27	0.44	0.19	0.06
Si	0.00	0.05	0.28	0.44	0.23
Fe	0.00	0.01	0.08	0.35	0.57

True label \ Predicted label	p	He	C	Si	Fe
p	0.60	0.28	0.09	0.02	0.00
He	0.29	0.40	0.24	0.06	0.01
C	0.04	0.27	0.41	0.20	0.06
Si	0.00	0.06	0.29	0.40	0.25
Fe	0.00	0.01	0.09	0.34	0.56

Confusion matrices for CNNs trained on e/γ, μ energy releases, before (left) and after (right) "spoiling" the dataset