# Learning the composition of Ultra-High Energy Cosmic Rays

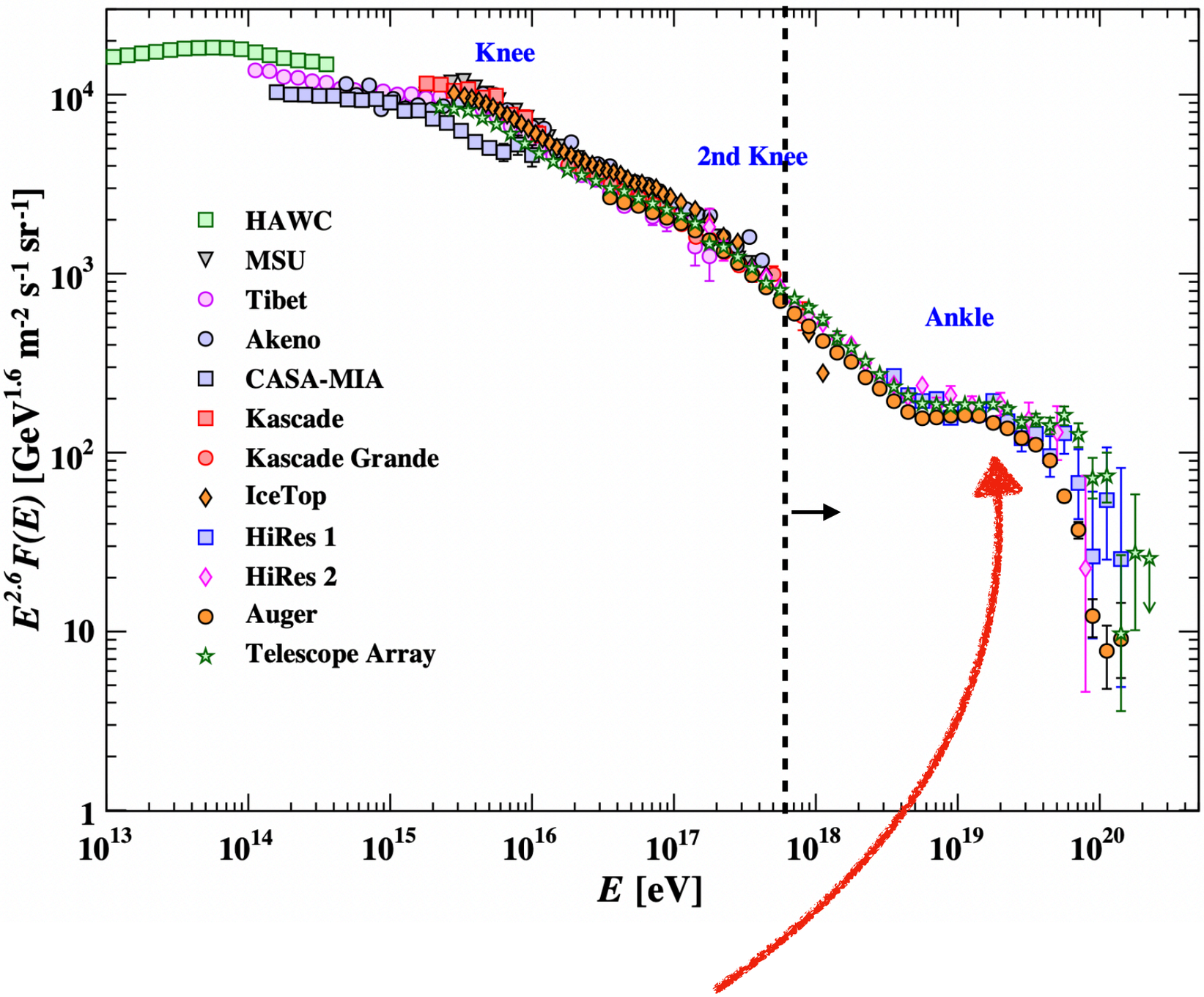
Michele Tammaro

@ Theory Group Day, 25/03/2024

with B. Bortolato and J.F. Kamenik (2212.04760, 2304.11197 and more to come…)


Istituto Nazionale di Fisica Nucleare
SEZIONE DI FIRENZE

# Ultra High-Energy Cosmic Rays (UHECR)




$$\phi \sim 10^{-2}\ \mathrm{km}^{-2}\ \mathrm{yr}^{-1}$$

Where?

Astrophysical sources of UHECR

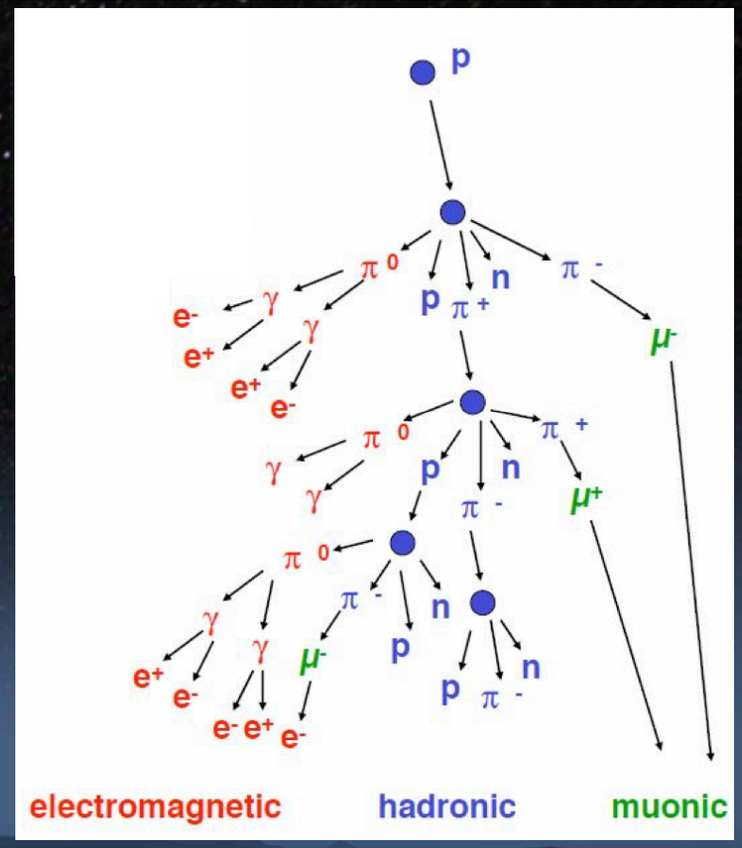
How?

Acceleration mechanisms

What?

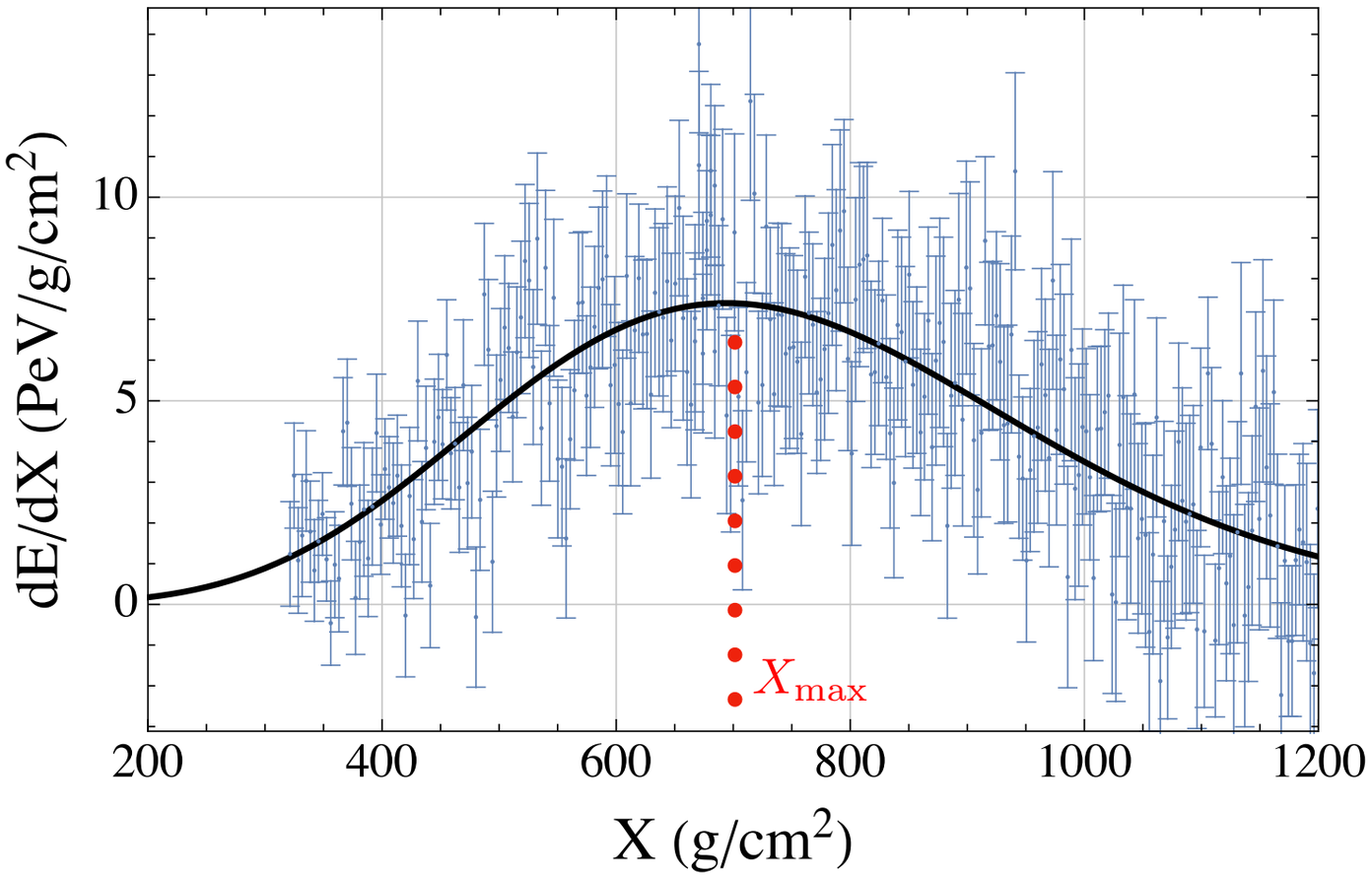
Mass composition

# Hybrid Showers



# Hybrid Showers

$$X_{\mathrm{ground}} \sim 1200\ \mathrm{g/cm^2}$$

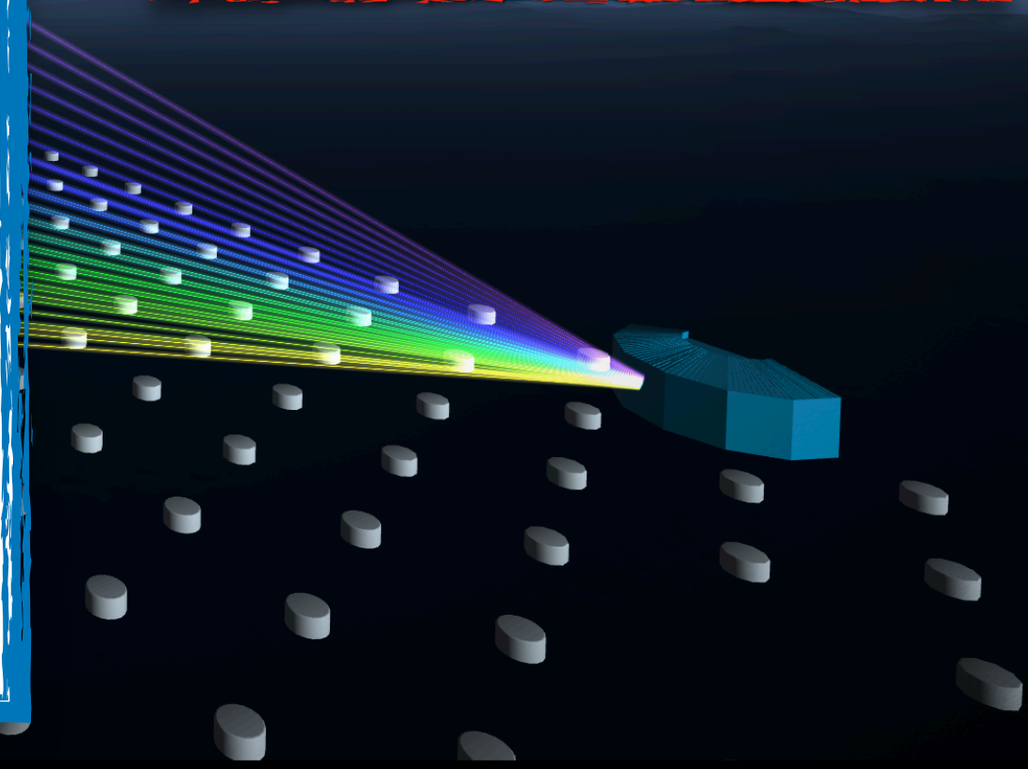



$$\langle X_{\mathrm{max}} \rangle \propto \log E - \log A$$

FD runtime:
~10% total run

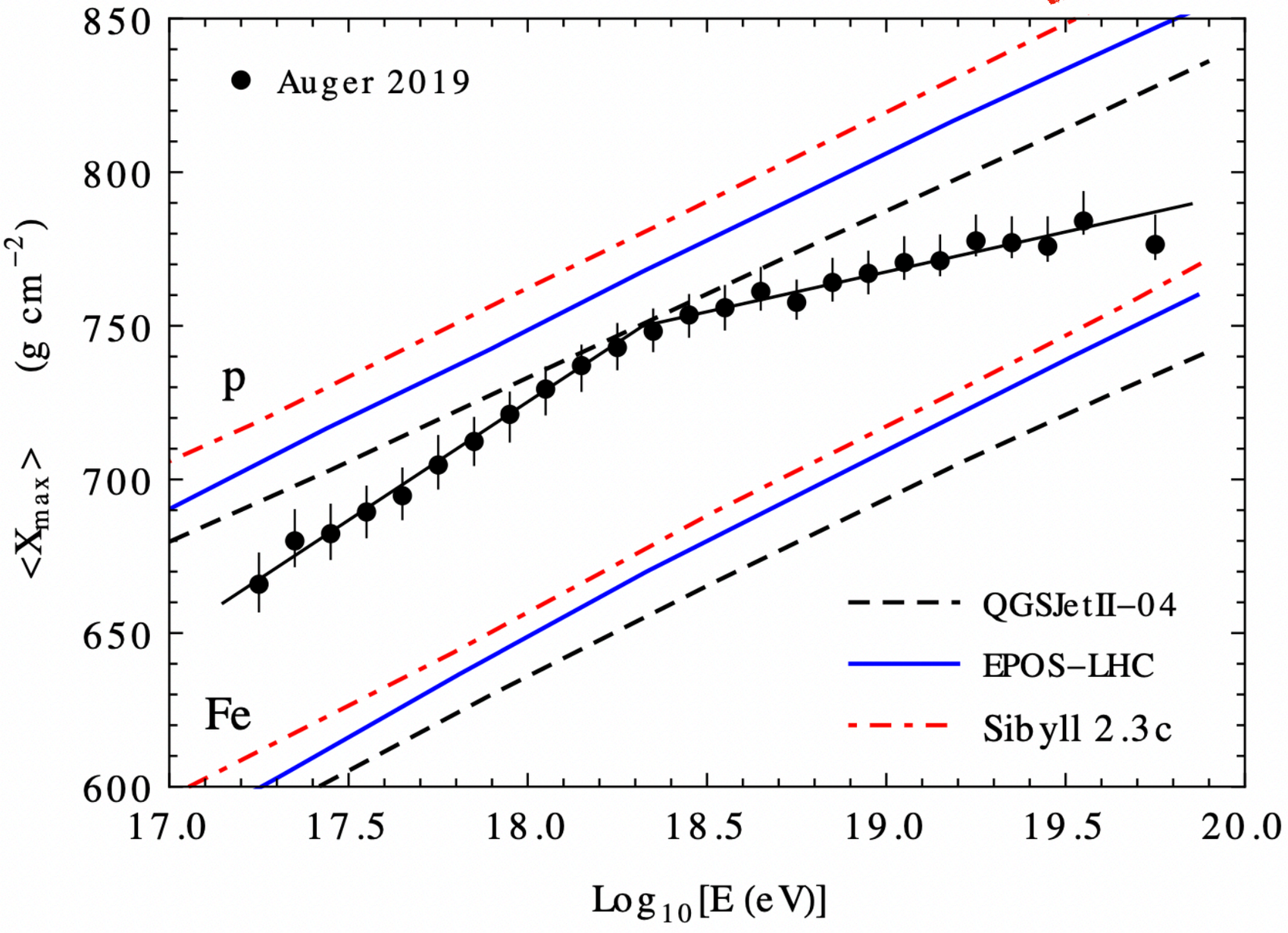
Public FD data:
~10% total data

Low statistics Alert!



Different models,
which one to use?

(Lipari: 2012.06861)




Simulations perfomed with
CORSIKA
(https://www.iap.kit.edu/corsika/)

Auger data seems to indicate heavier primaries at higher energies

# What has been done

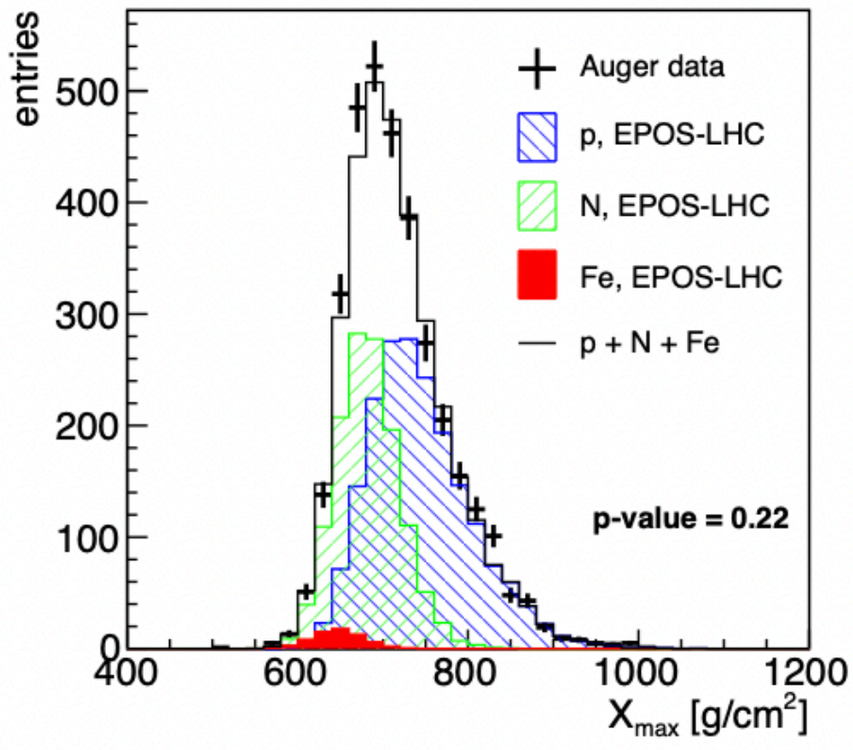
Fit $X_{\max}$ distribution with mixture

$$(p,\ He,\ N,\ Fe)$$

Bin and count

(Arsene, Sima: 2001.02667)
(Auger Coll.: 1409.5083)
(Lipari: 2012.06861)

$$p \sim 65\%, \quad N \sim 35\%$$




# What we do

$$w = (w_p,\ w_{He}, \ldots,\ w_{Fe}) \qquad \sum_i^{26} w_i = 1 \qquad 25 \text{ free parameters}$$

Decompose distribution in moments ⟶ Unbinned likelihood

Faster computation with Nested Sampling

# What has been done

Fit $X_{\max}$ distribution with mixture

$$(p,\ He,\ N,\ Fe)$$

Bin and count

(Arsene, Sima: 2001.02667)
(Auger Coll.: 1409.5083)
(Lipari: 2012.06861)

$$p \sim 65\%, \quad N \sim 35\%$$



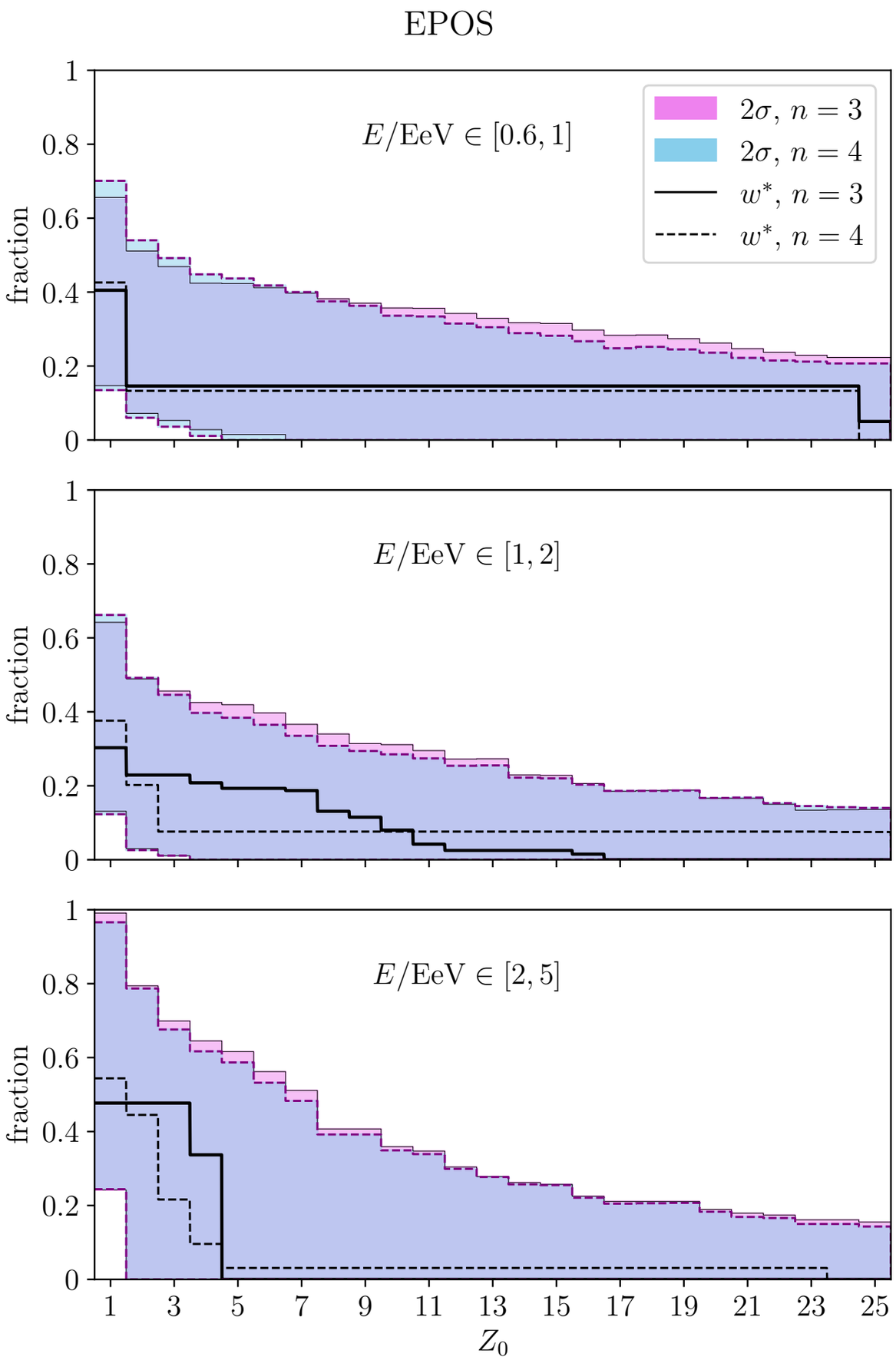

# What we do

Decompose distribution in moments ⟶ Unbinned likelihood

$$w = (w_p,\ w_{He}, \ldots,\ w_{Fe}) \qquad \sum_i^{26} w_i = 1 \qquad \text{25 free parameters}$$

Faster computation with Nested Sampling

# Full (cumulative) composition - EPOS




Fraction of elements heavier than $Z_0$

Can exclude 100% proton compositions

Results are unchanged increasing the number of features

# Including ground data

Hybrid

$$(x_i, y_i)_{i=1,\ldots,N}$$

FD

SD

Correlation $\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)}$

Non-Hybrid

$$(\hat{x}(\hat{y}_j), \hat{y}_j)_{j=1,\ldots,M}$$

How much information is added?

Assumption:
the underlying distribution of events
is the same for both sets

# Including ground data

Hybrid

$$(x_i, y_i)_{i=1,\ldots,N}$$

FD SD

Correlation $\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)}$

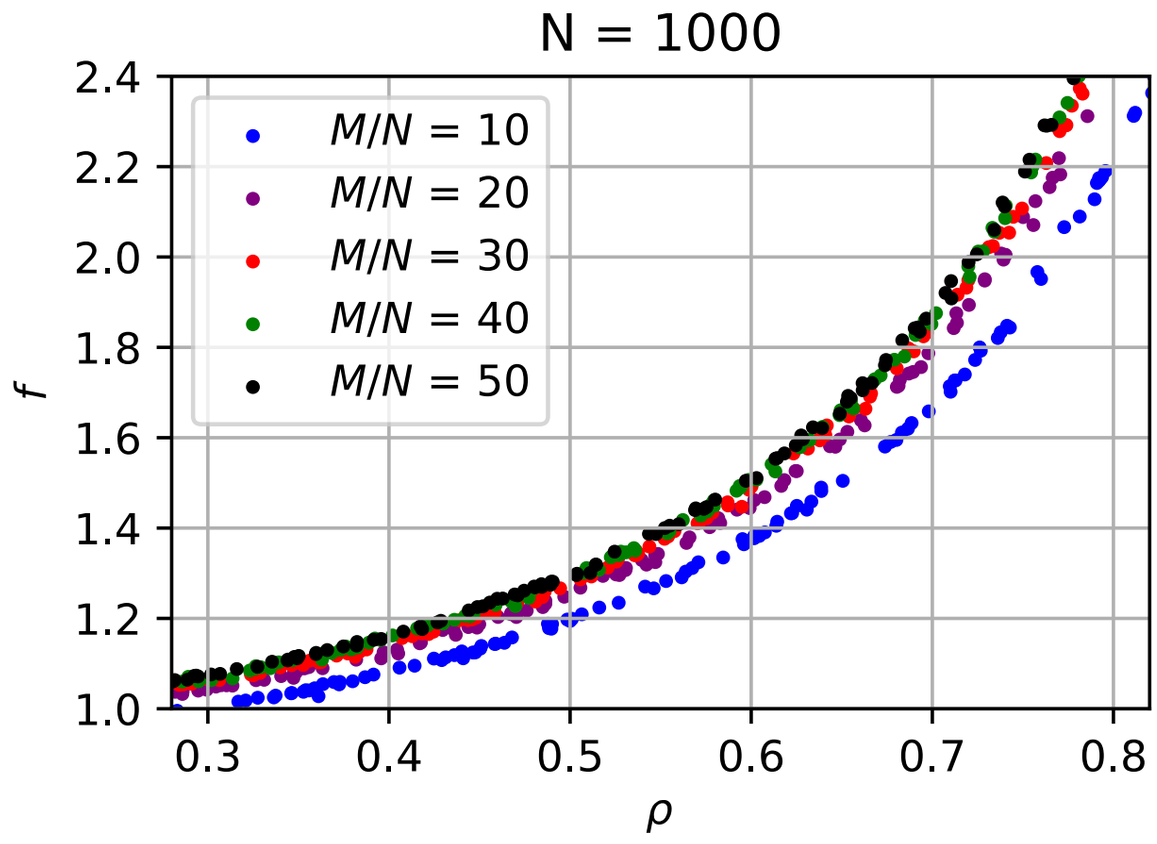
Non-Hybrid

$$(\hat{x}(\hat{y}_j), \hat{y}_j)_{j=1,\ldots,M}$$

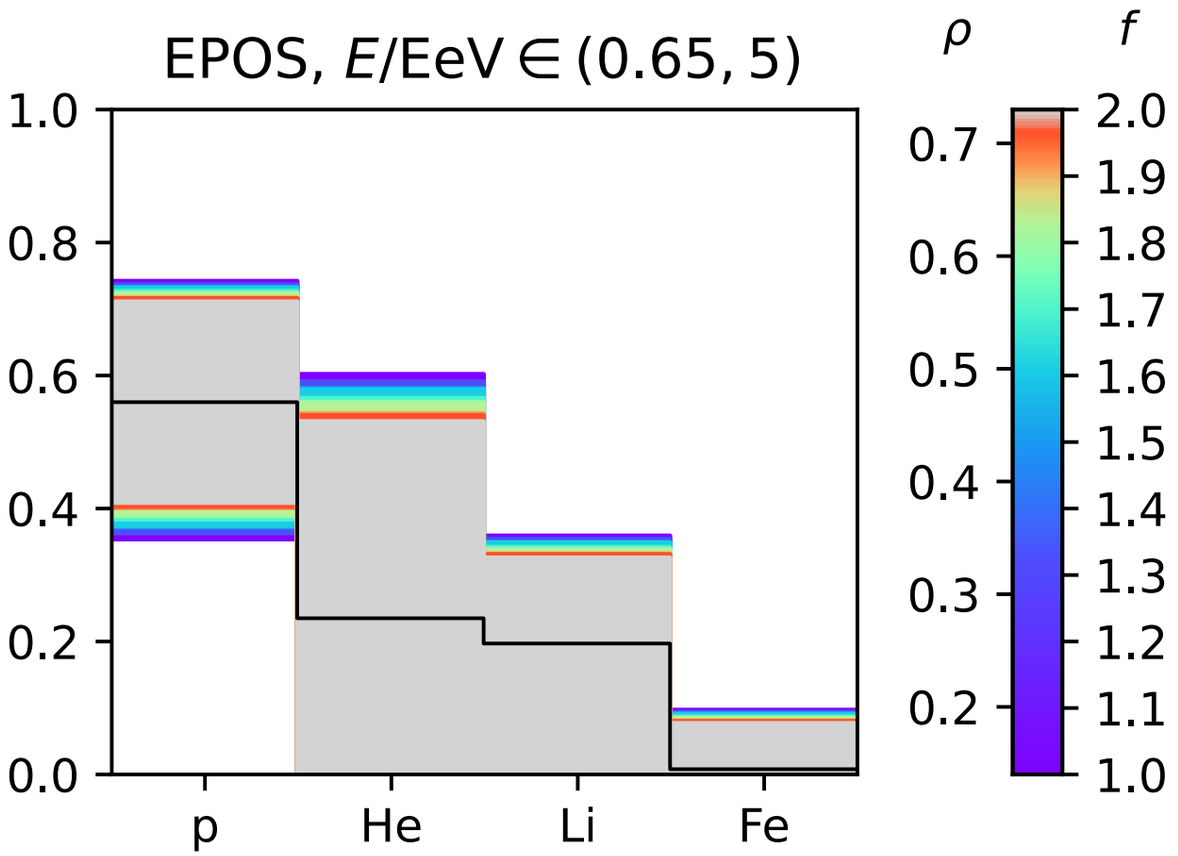
$$\mathcal{X}_l^{\mathrm{comb}} = \mathcal{X}_\ell \cup \mathcal{X}_\ell^{\mathrm{inf}} = \{(X,Y)\}_\ell$$

$$f \equiv \frac{\mathrm{Var}(\bar{x})}{\mathrm{Var}(\bar{X})}$$

Effective factor of included events



# Including ground data



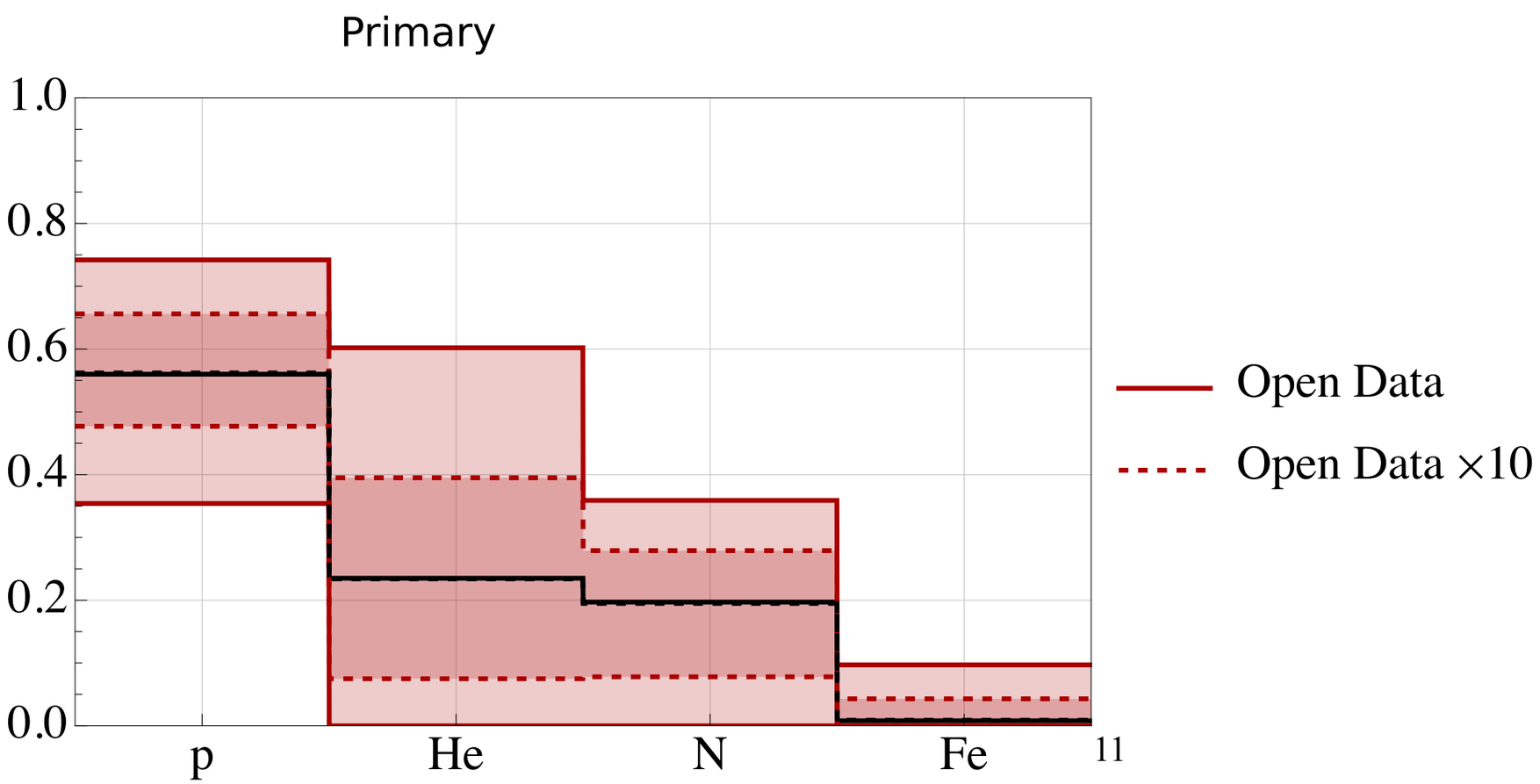

ρ~30% with (non)-linear fits

(Auger Coll.: 1710.07249)

ρ~70% with Deep Neural Network

(Auger Coll.: 2101.02946)

With ~70% correlation we can effectively double the statistical power



# Summary

- We build a simple framework for inference of UHECR mass composition

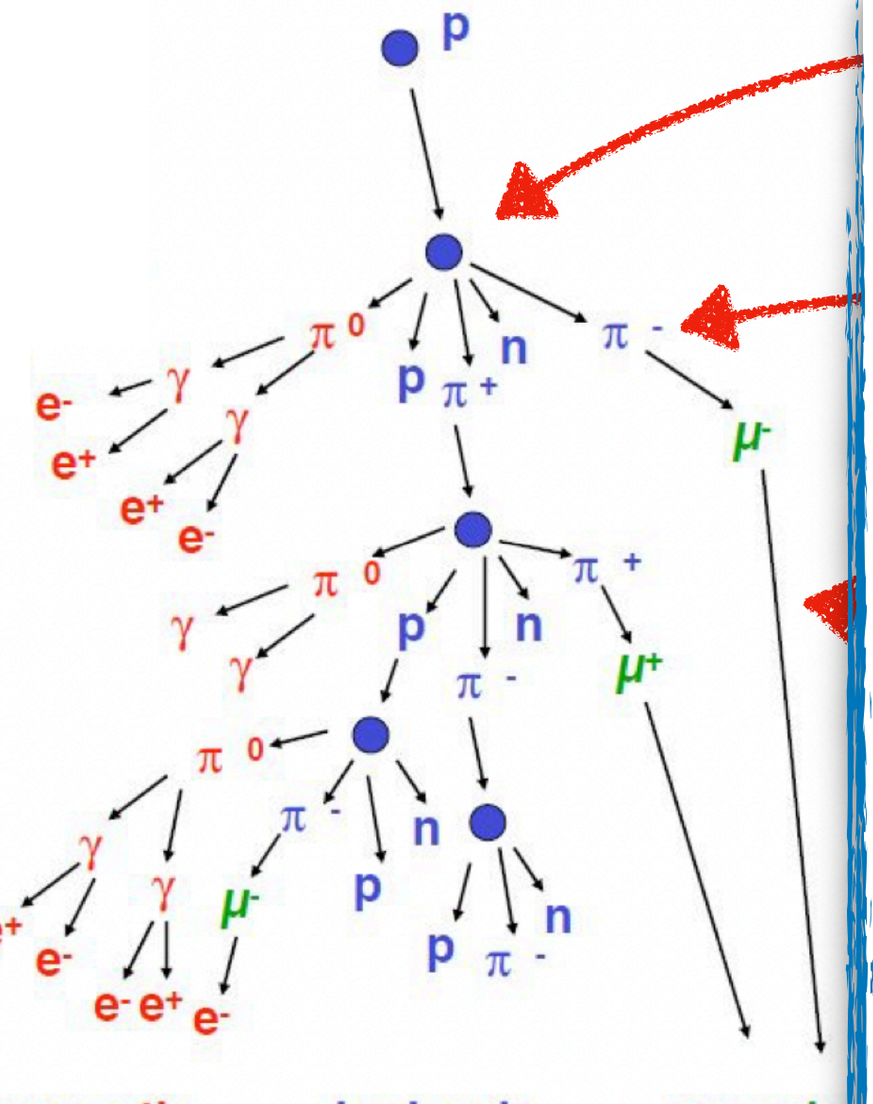
- Limited by very low statistics…

- …but we include ground data to increase statistical power

- Many ideas for improvements

# Backup slides

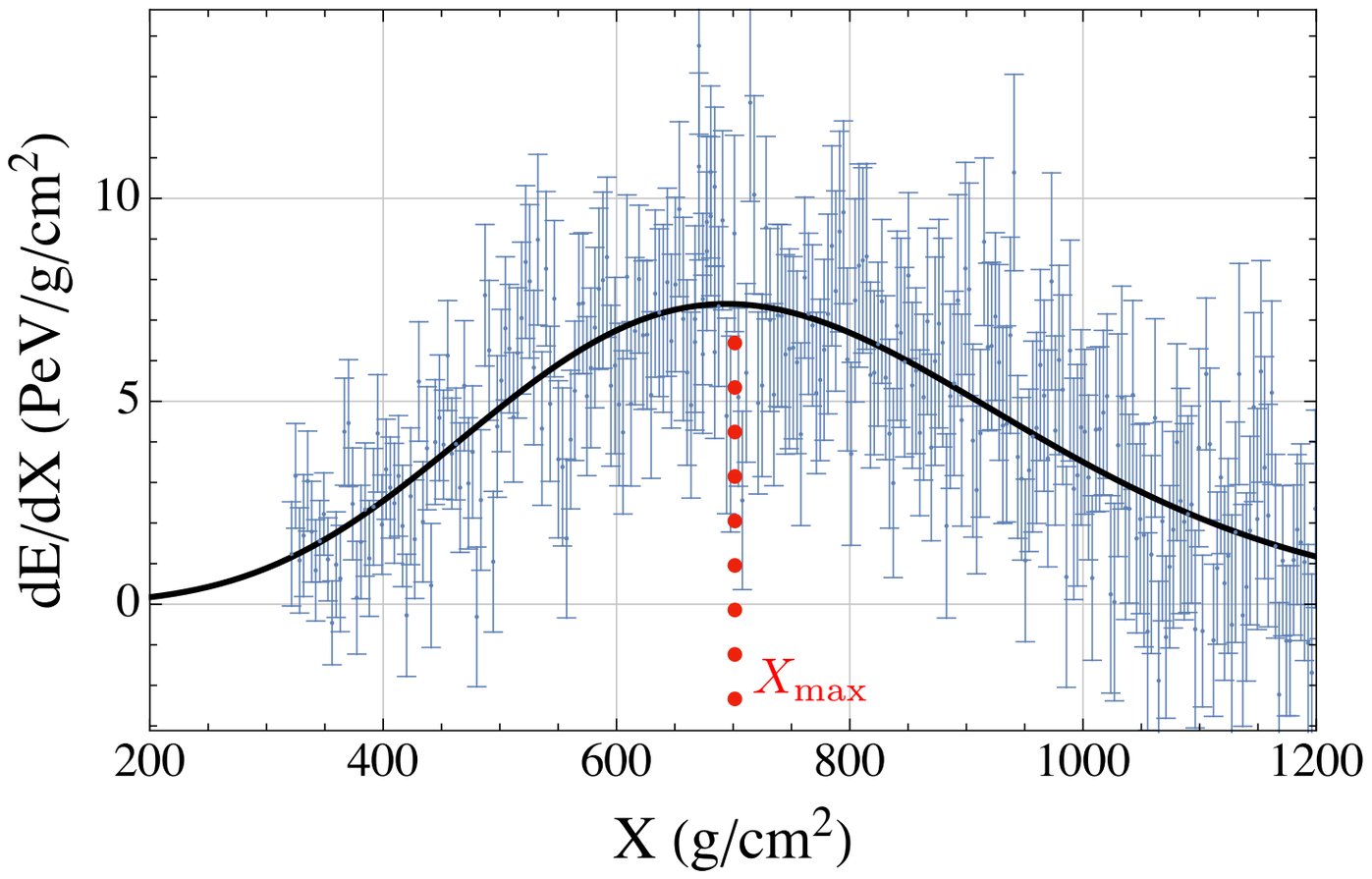
# Extensive Air Showers (EAS)







$$X_{\rm ground} \sim 1200\ {\rm g/cm^2}$$

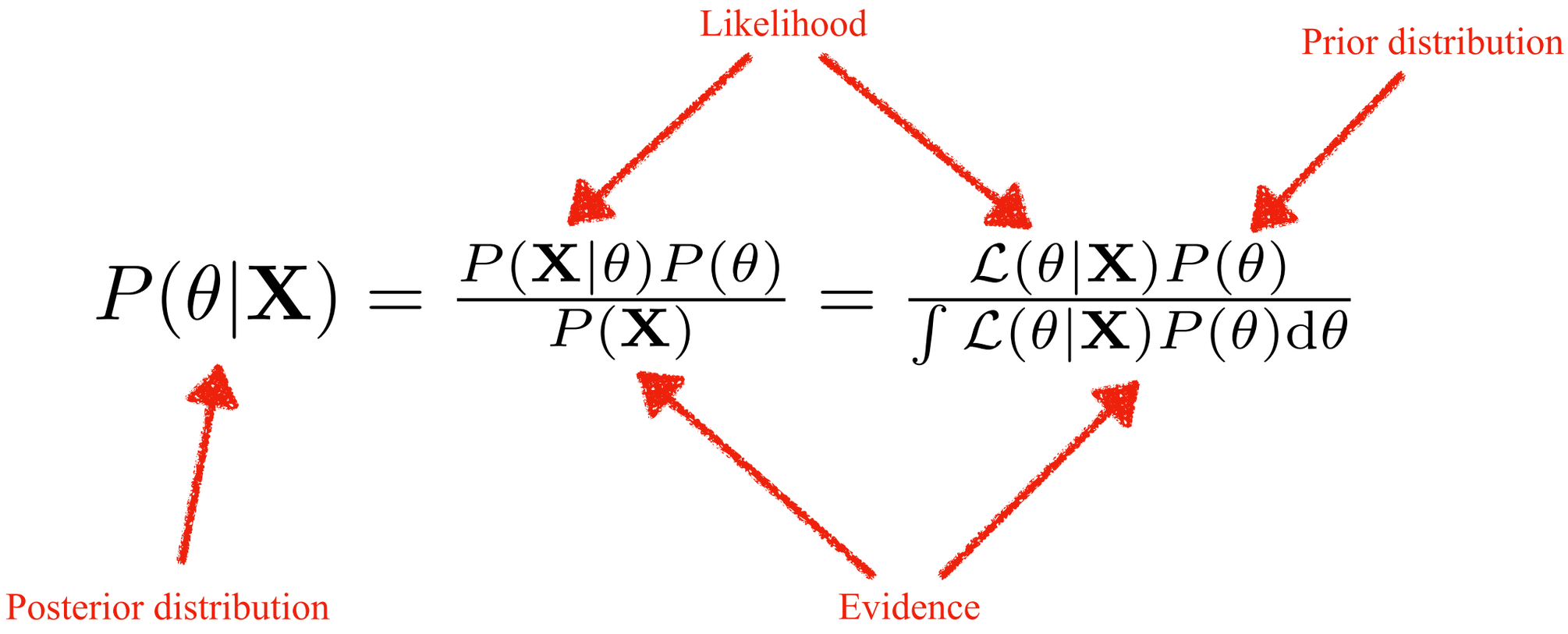



(Gaisser, Hillas, '77)

$$f_{GH}(X) = \left(\frac{{\rm d}E}{{\rm d}X}\right)_{\rm max} \left(\frac{X - X_0}{X_{\rm max} - X_0}\right)^{\frac{X_{\rm max} - X_0}{\lambda}} \exp\left(\frac{X_{\rm max} - X}{\lambda}\right)$$

# Interlude: Bayesian inference

$\mathbf{X} = x_1, \ldots, x_n$ data sample $\theta$ parameters, $x \sim p(x|\theta)$




Maximum Likelihood Estimate (MLE): select $\theta$ that maxime $\mathcal{L}$

Maximum a Posteriori Estimation (MPE) for Bayesian people

# Interlude: Bayesian inference

$\mathbf{X} = x_1, \ldots, x_n$ data sample $\quad$ $\theta$ parameters, $x \sim p(x|\theta)$

$\mathbf{X} \equiv X_{\max}$ $\quad$ $\theta \equiv w = (w_{\rm p}, w_{\rm He}, \ldots, w_{\rm Fe}) \quad \sum_i w_i = 1$

$\mathcal{L}(\theta|\mathbf{X})$ from simulations

Likelihood

Prior distribution

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} = \frac{\mathcal{L}(\theta|\mathbf{X})P(\theta)}{\int \mathcal{L}(\theta|\mathbf{X})P(\theta)\mathrm{d}\theta}$$

Posterior distribution

Evidence

Maximum Likelihood Estimate (MLE): select $\theta$ that maxime $\mathcal{L}$

Maximum a Posteriori Estimation (MPE) for Bayesian people

# Interlude: Bayesian inference

$\mathbf{X} = x_1, \ldots, x_n$ data sample $\quad$ $\theta$ parameters, $x \sim p(x|\theta)$

$\mathbf{X} \equiv X_{\max}$ $\quad$ $\theta \equiv w = (w_{\rm p}, w_{\rm He}, \ldots, w_{\rm Fe})$ $\quad \sum_i w_i = 1$

Likelihood

Prior distribution

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} = \frac{\mathcal{L}(\theta|\mathbf{X})P(\theta)}{\int \mathcal{L}(\theta|\mathbf{X})P(\theta)\mathrm{d}\theta}$$

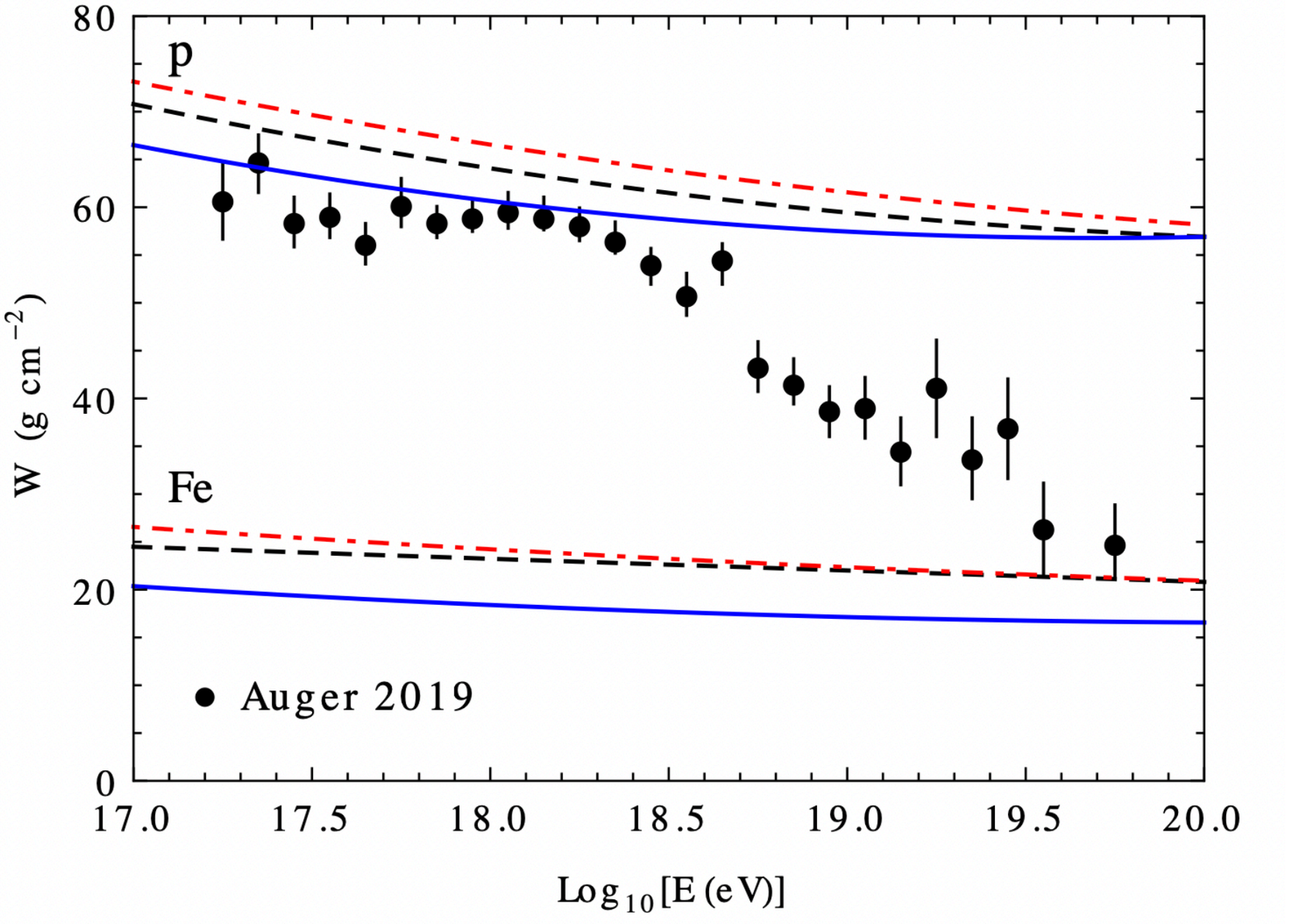
Posterior distribution

Evidence

Maximum Likelihood Estimate (MLE): select $\theta$ that maxime $\mathcal{L}$

Maximum a Posteriori Estimation (MPE) for Bayesian people

(Lipari: 2012.06861)




$$W = \left(\langle X_{\rm max}^2\rangle - \langle X_{\rm max}\rangle^2\right)^{1/2}$$

Simulations perfomed with CORSIKA (https://www.iap.kit.edu/corsika/)

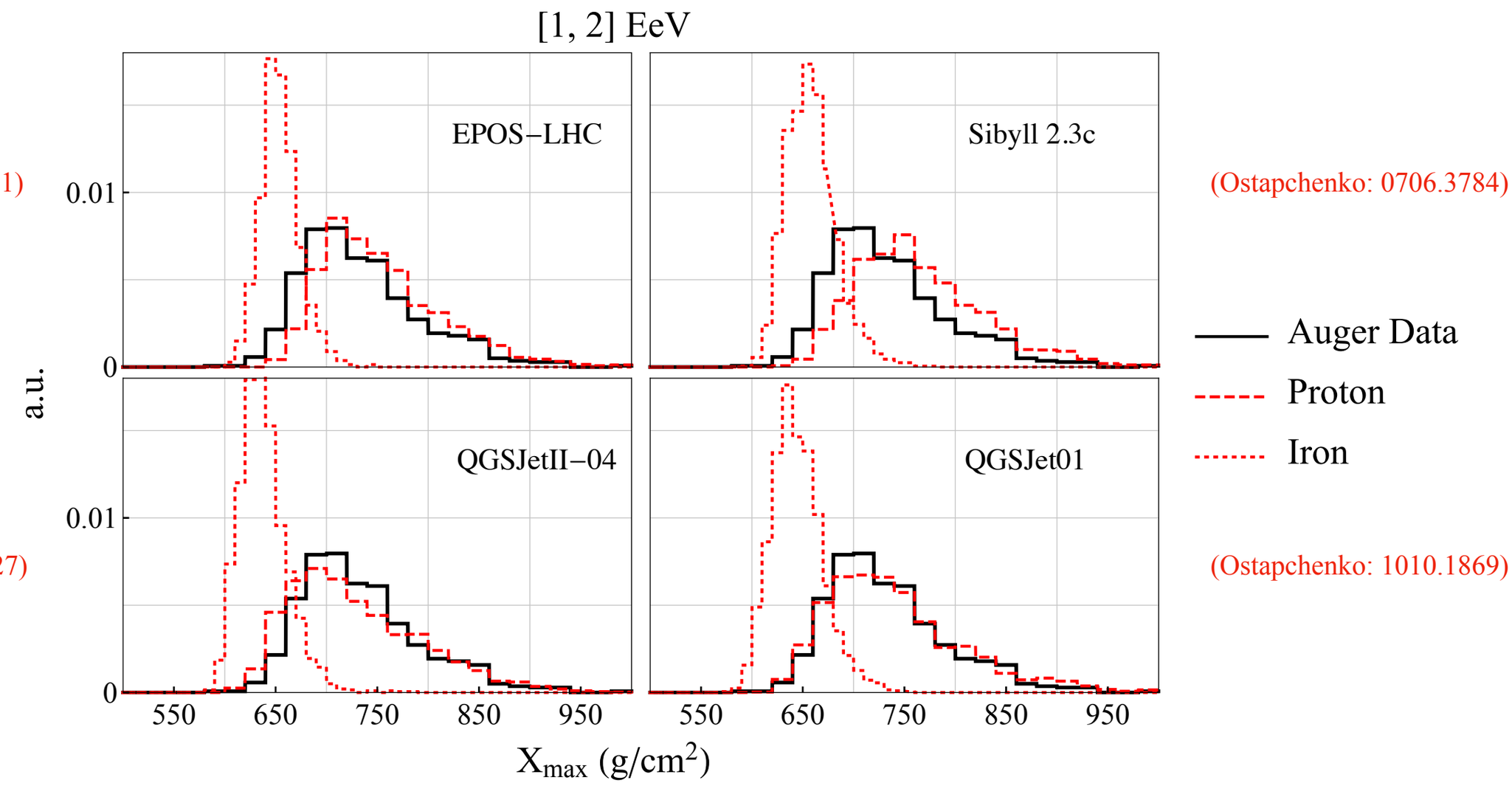
Single or two-element mixture cannot reproduce the data

(depending on the model…)

# The results will be model-dependent

(Pierog et al: 1306.0121)

(Riehn et al: 1709.07227)




(Ostapchenko: 0706.3784)

(Ostapchenko: 1010.1869)

# Data

Pierre Auger Open Data 2021

- $\sim 10\%$ of total data
- 1602 hybrid showers in $E \in [0.6, 60]$ EeV
- Split in three bins: $E \in [0.6, 1],\ [1, 2],\ [2, 5]$ EeV
- $\sim 500$ events per bin

$$P_{\rm Aug}(X_{\rm max}|E)$$

$$P_{\rm sim}(X_{\rm max}|Z, E)$$

Convolute all with detector effects

(Auger Coll.: 1409.5083)

Simulations with CORSIKA

- 4 hadronic models
- 26 primaries from $p$ to $Fe$
- 2000 shower per element/bin/model
- 624000 simulations

Bonus achievement: get complaints from both IJS and CERN clusters

# Probability Distribution Function (PDF)

Each data/simulation point is given as $(X_{\max}, \delta X_{\max})$

Data

$$P_{\mathrm{Aug}}(X_{\max}|E) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{N}\left(X_{\max} \mid X^{j}_{\max}, \delta X^{j}_{\max}\right)$$

Simulations

Set of inputs

Detector effects

$$P_{\mathrm{sim}}(X_{\max} \mid S=\{E,Z,H\}) = \frac{1}{\tilde{N}}\sum_{j}\int \mathrm{d}\tilde{X}\, \mathcal{N}\left(\tilde{X} \mid X^{j}_{\max}, \delta X^{j}_{\max}\right) \times R(X_{\max}-\tilde{X}) \times \epsilon(\tilde{X})$$

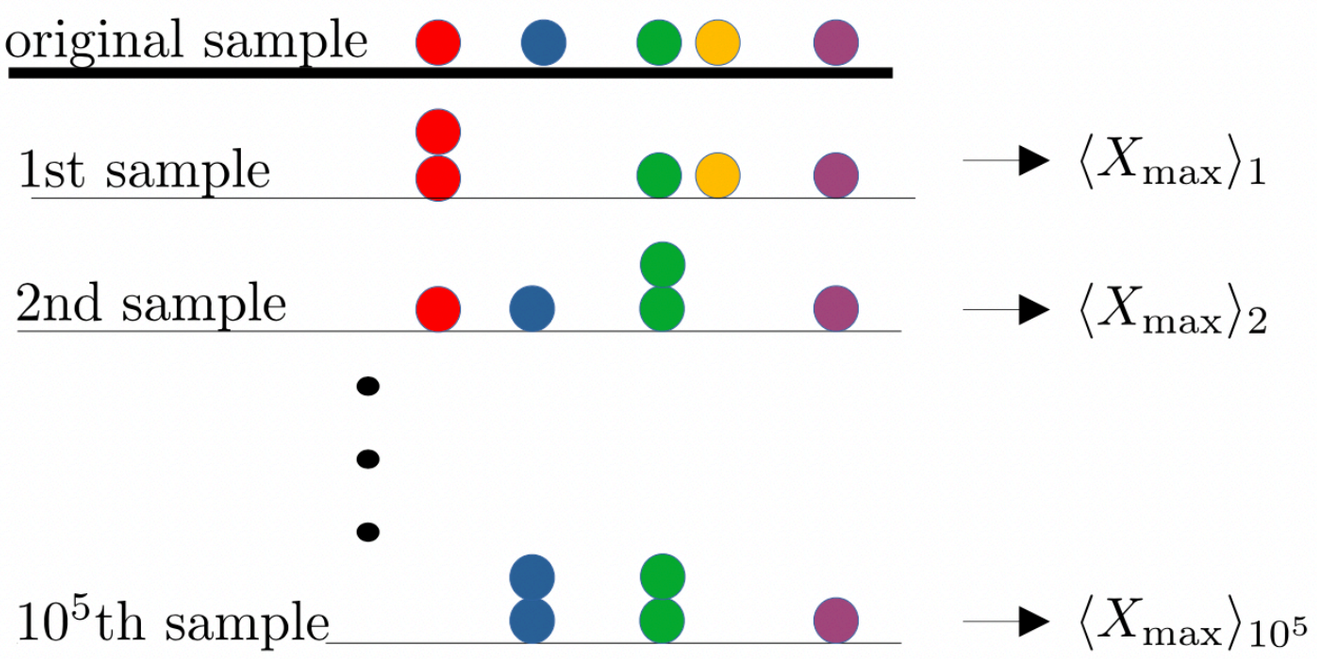
# Moments

$$\langle X^n_{\rm max}\rangle_Z = \frac{\int P(X_{\rm max} \mid Z)\ X^n_{\rm max}\ {\rm d}X_{\rm max}}{\int P(X_{\rm max} \mid Z)\ {\rm d}X_{\rm max}}$$

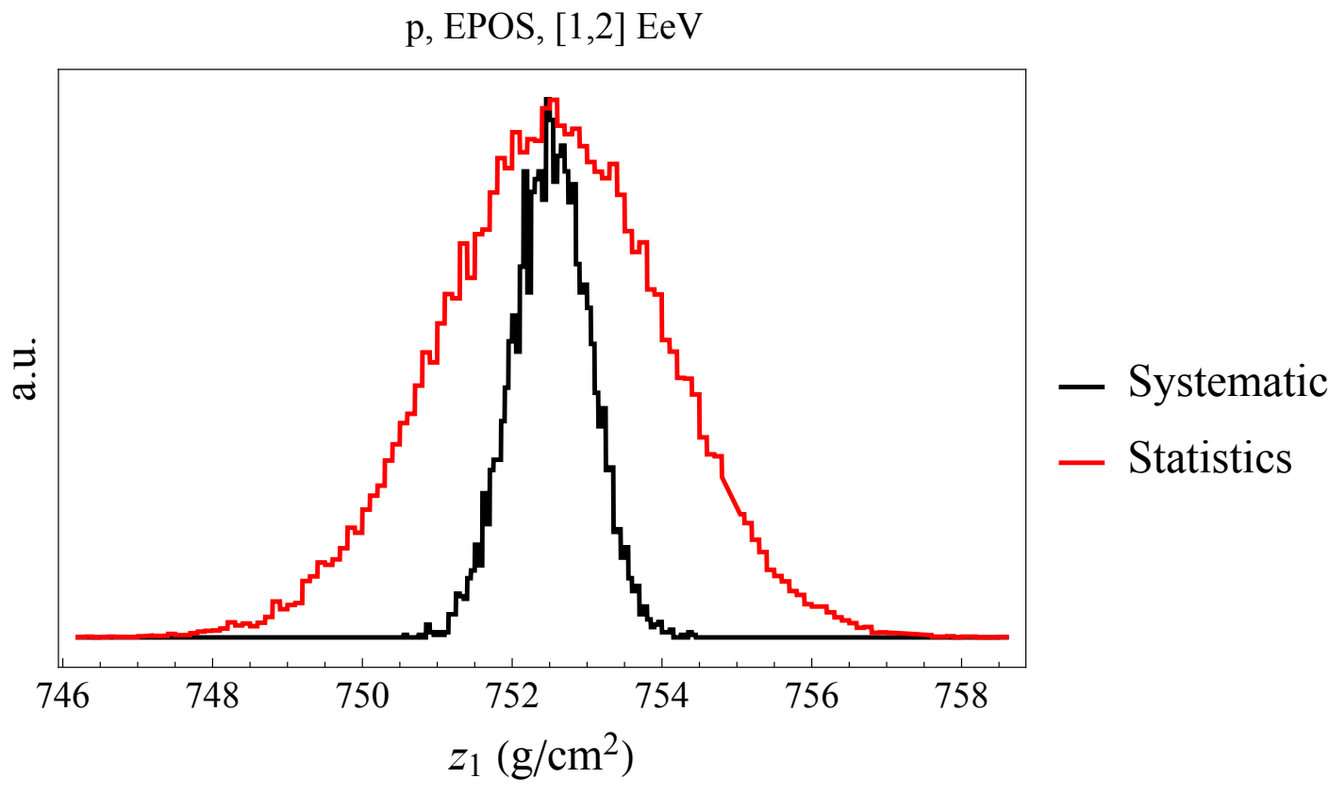
$$\langle X^n_{\rm max}\rangle(w) = \frac{\sum_Z \langle X^n_{\rm max}\rangle_Z\ w_Z}{\sum_Z\ w_Z}$$

All systematic uncertainties are included

**Bootstrapping**

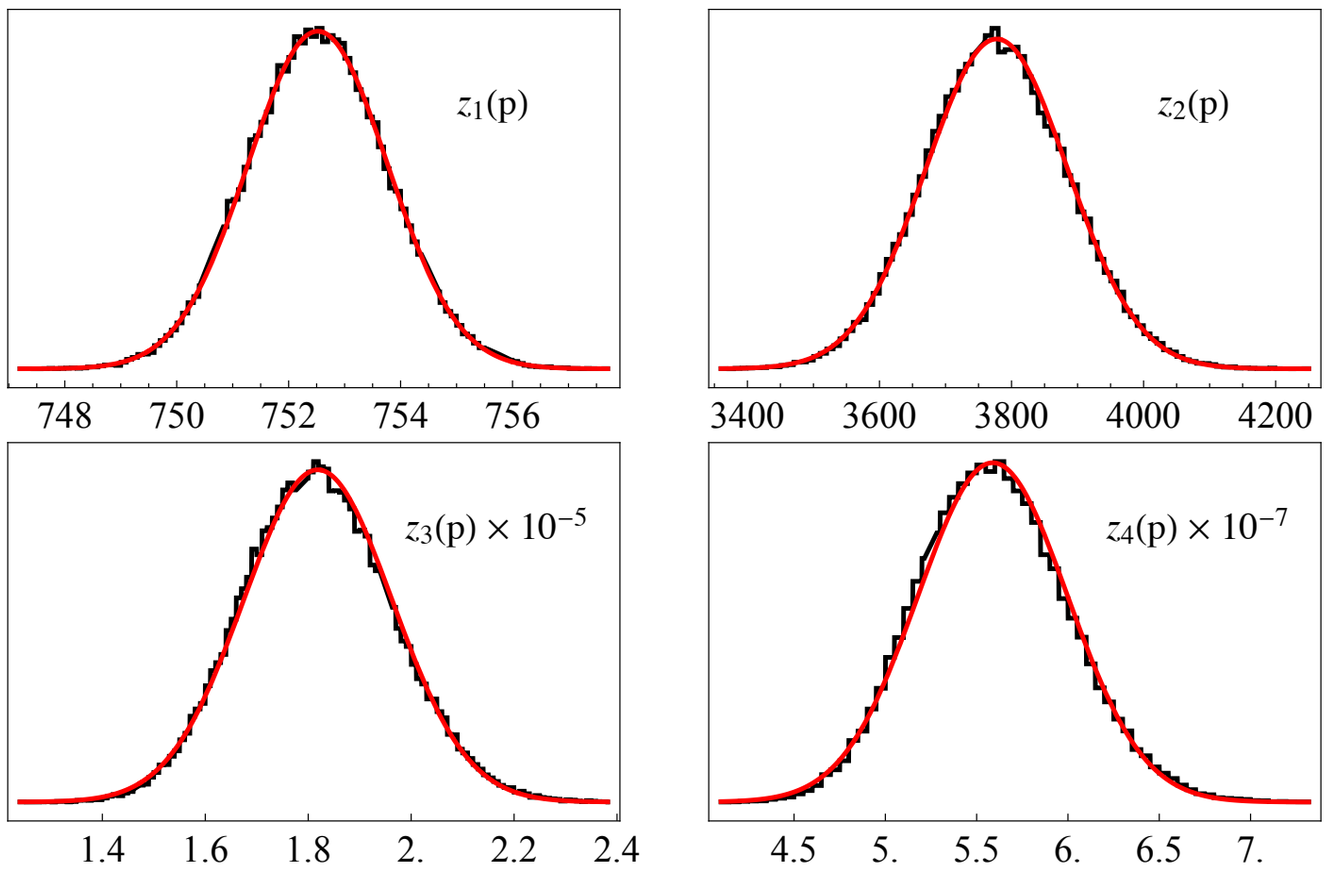



Include statistical uncertainties
(data and simulation)
by Bootstrapping



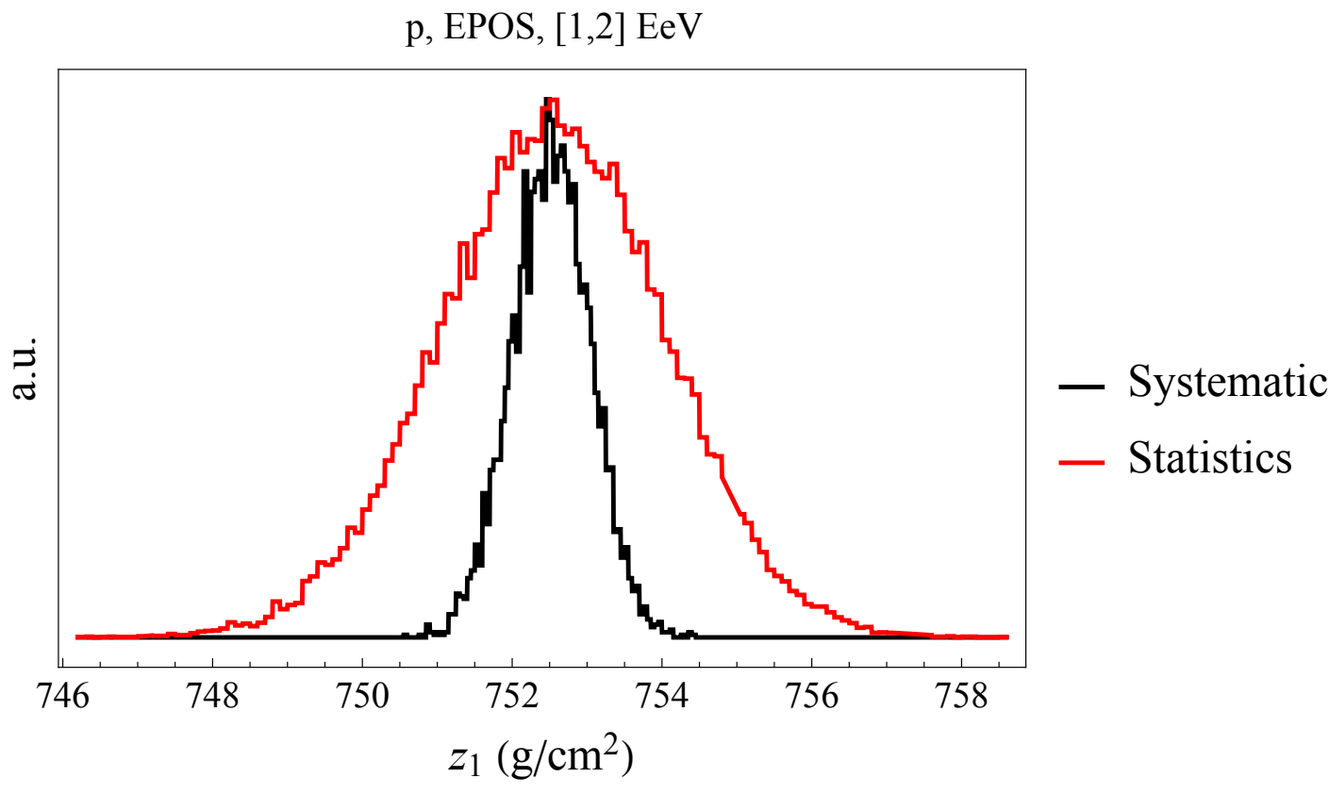

Simulations

$$z(w) \sim \mathcal{N}_n\Big(z \mid \mu(w),\ \Sigma(w)\Big)$$




Data

$$\tilde{z} \sim \mathcal{N}_n\left(z \mid \tilde{\mu}, \tilde{\Sigma}\right)$$

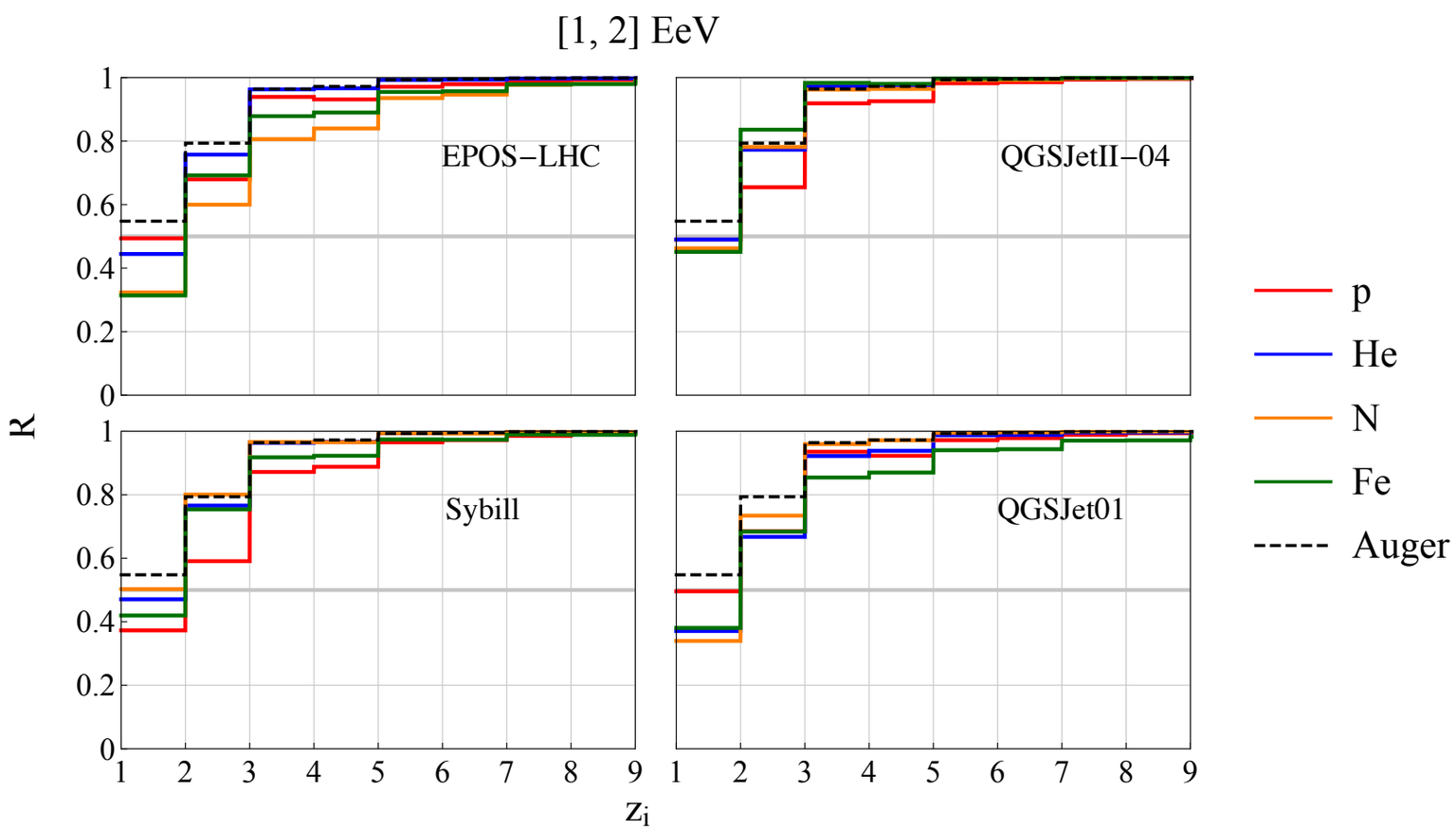



Simulations

$$z(w) \sim \mathcal{N}_n\Big(z \mid \mu(w),\ \Sigma(w)\Big)$$

Data

$$\tilde{z} \sim \mathcal{N}_n\Big(z \mid \tilde{\mu}, \tilde{\Sigma}\Big)$$




Higher moments are strongly correlated

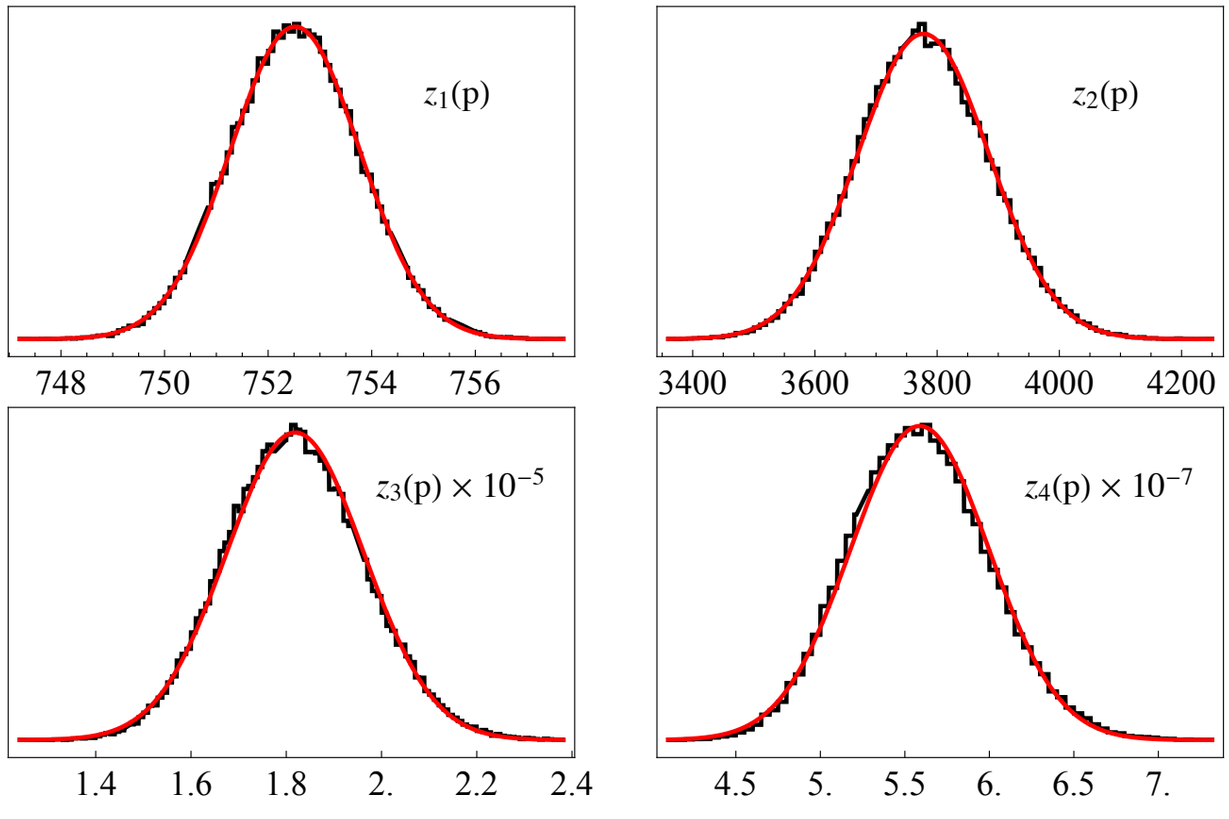
# Moments

$$\langle X_{\max}^n \rangle_Z = \frac{\int P(X_{\max} \mid Z)\ X_{\max}^n\ \mathrm{d}X_{\max}}{\int P(X_{\max} \mid Z)\ \mathrm{d}X_{\max}}$$

$$\langle X_{\max}^n \rangle(w) = \frac{\sum_Z \langle X_{\max}^n \rangle_Z\ w_Z}{\sum_Z\ w_Z}$$

All systematic uncertainties are included

p, EPOS, [1, 2] EeV




— Bootstrap
— Normal

Include statistical uncertainties (data and simulation) by Bootstrapping

Simulations

$$z(w) \sim \mathcal{N}_n\Big(z \mid \mu(w),\ \Sigma(w)\Big)$$

Data

$$\tilde{z} \sim \mathcal{N}_n\left(z \mid \tilde{\mu}, \tilde{\Sigma}\right)$$

# Nested Sampling

Evidence $$Z = \int \mathcal{L}(w)\ \mathrm{Dir}(w)\ \mathrm{d}^D w = \int_0^1 \mathcal{L}(X)\ \mathrm{d}X\ ,$$

- at step $k = 1$, sample $N_{live}$ points (compositions)
- select $w_1$ with lowest likelihood $L_1$; $w_1$ is a dead point
- at step $k > 1$, sample a new live point $w$ from prior, with constraint $\mathcal{L}(w) > L_{k-1}$
- Find the dead point $w_k$, with likelihood $L_k$
- calculate volume of prior region with $L_{k-1} < \mathcal{L}(w) \leq L_k$
- calculate evidence shift $\delta Z_k = L_k \delta X_k$

Can be estimated with Beta distributions

Output is a set of $w_k$ with weights $u_k = \delta Z_k / Z$

$$\mathrm{CL}(\mathcal{L}_0) = \sum_{(w_k, u_k)\ |\ \mathcal{L}(w) \geq \mathcal{L}_0} u_k$$

# Nested Sampling

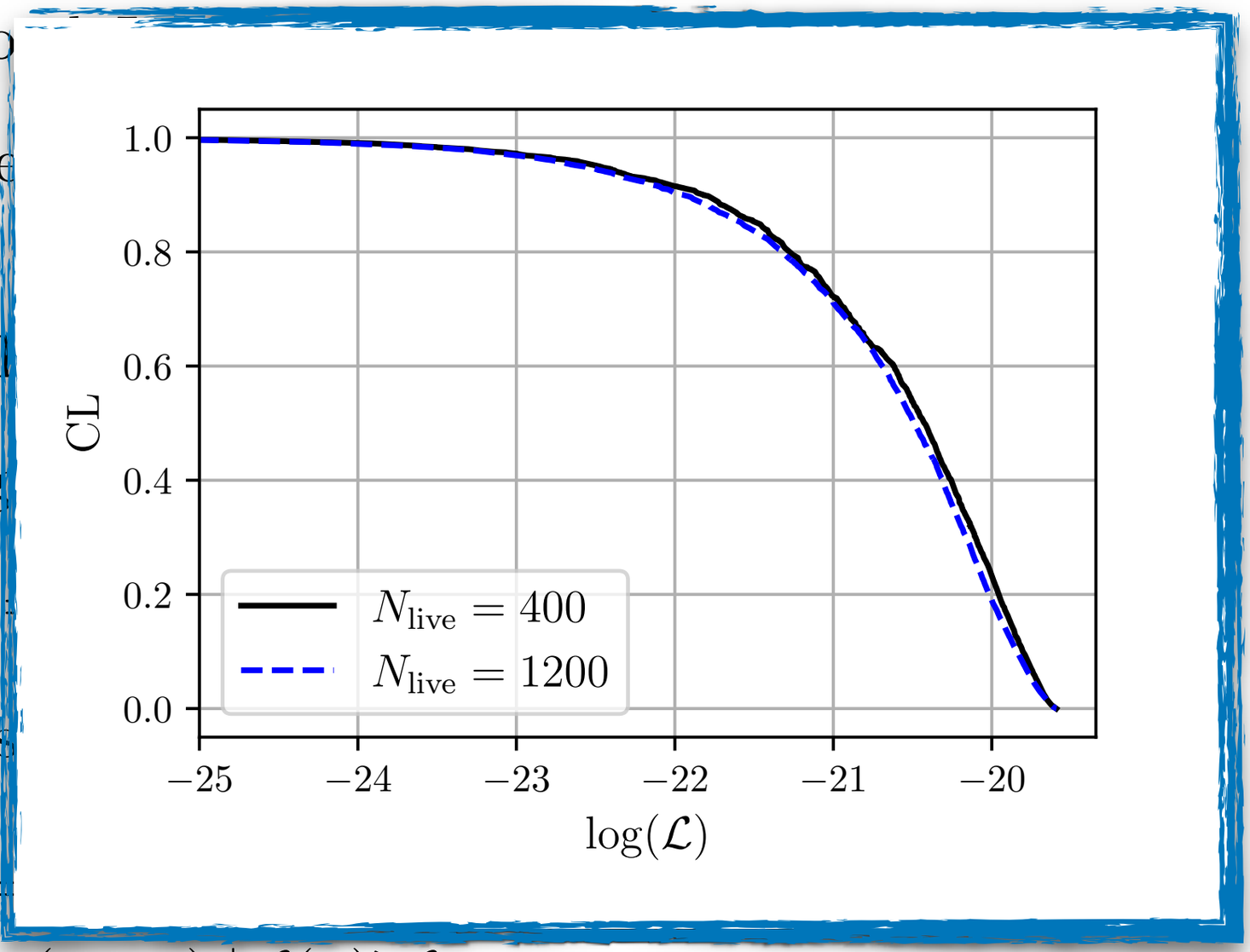
Evidence $$Z = \int \mathcal{L}(w)\ \mathrm{Dir}(w)\ \mathrm{d}^D w = \int_0^1 \mathcal{L}(X)\ \mathrm{d}X\ ,$$

- at step $k = 1$, sample $N_{live}$ points (compositions)
- select $w_1$ with lowest likeliho
- at step $k > 1$, sample a ne
  $\mathcal{L}(w) > L_{k-1}$
  - Find the dead point $w_k$, wit
  - calculate volume of prior reg
  - calculate evidence shift $\delta Z_k$

Output is a set of $w_k$ with weights

$$\mathrm{CL}(\mathcal{L}_0) =$$

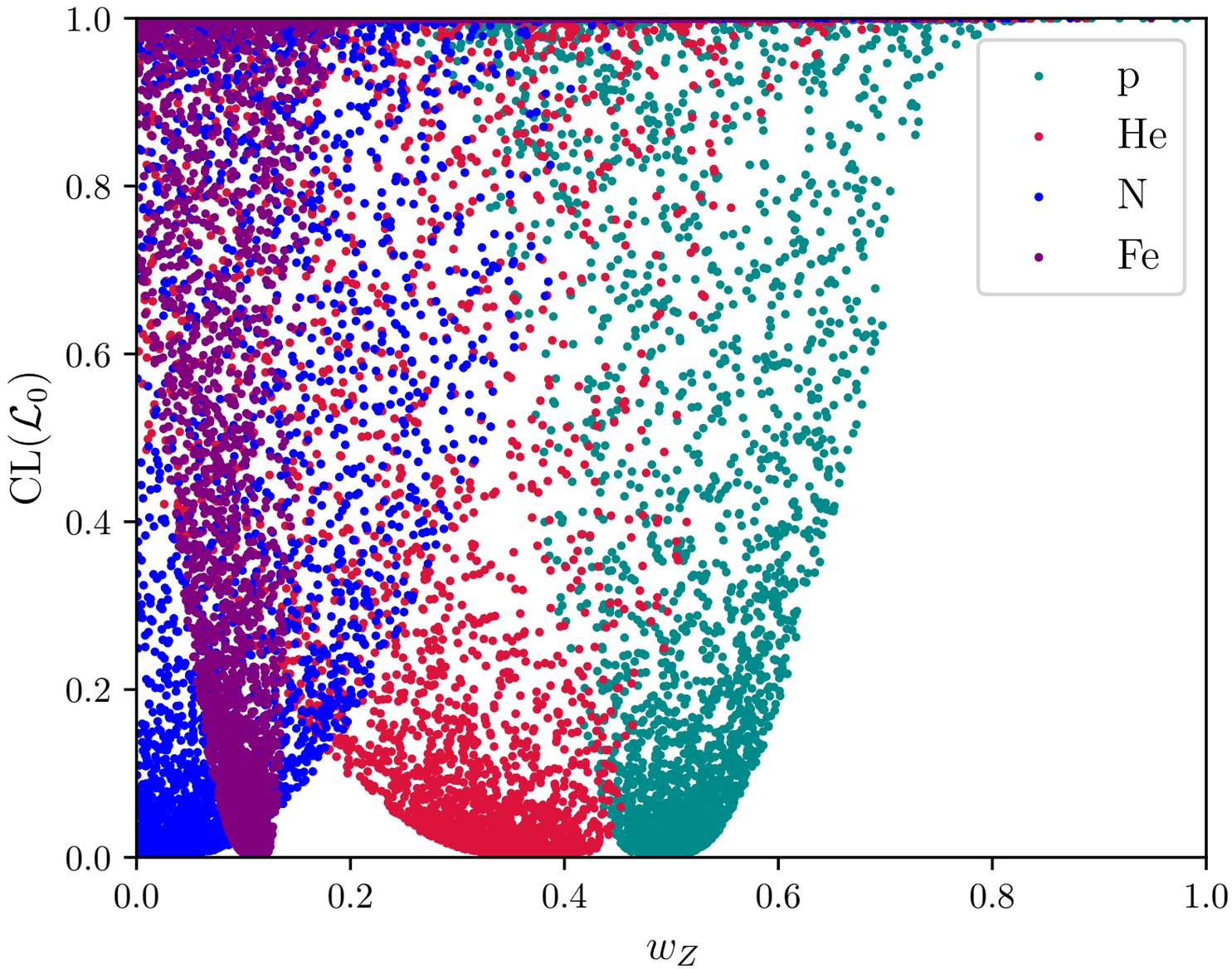
$(w_k, u_k) \mid \mathcal{L}(w) \geq \mathcal{L}_0$



Full likelihood form

$$\mathcal{L} = (2\pi)^{-\frac{D}{2}} \det\left(\Sigma_w + \tilde{\Sigma}\right)^{-\frac{1}{2}} \exp\Big[ -\frac{1}{2}\left(\mu_w^T \Sigma_w \mu_w + \tilde{\mu}^T \tilde{\Sigma} \tilde{\mu}\right) + \frac{1}{2}\left(\mu_w^T \Sigma_w^{-1} + \tilde{\mu}^T \tilde{\Sigma}^{-1}\right)\left(\Sigma_w^{-1} + \tilde{\Sigma}^{-1}\right)^{-1}\left(\Sigma_w \mu_w + \tilde{\Sigma}\tilde{\mu}\right)\Big]$$

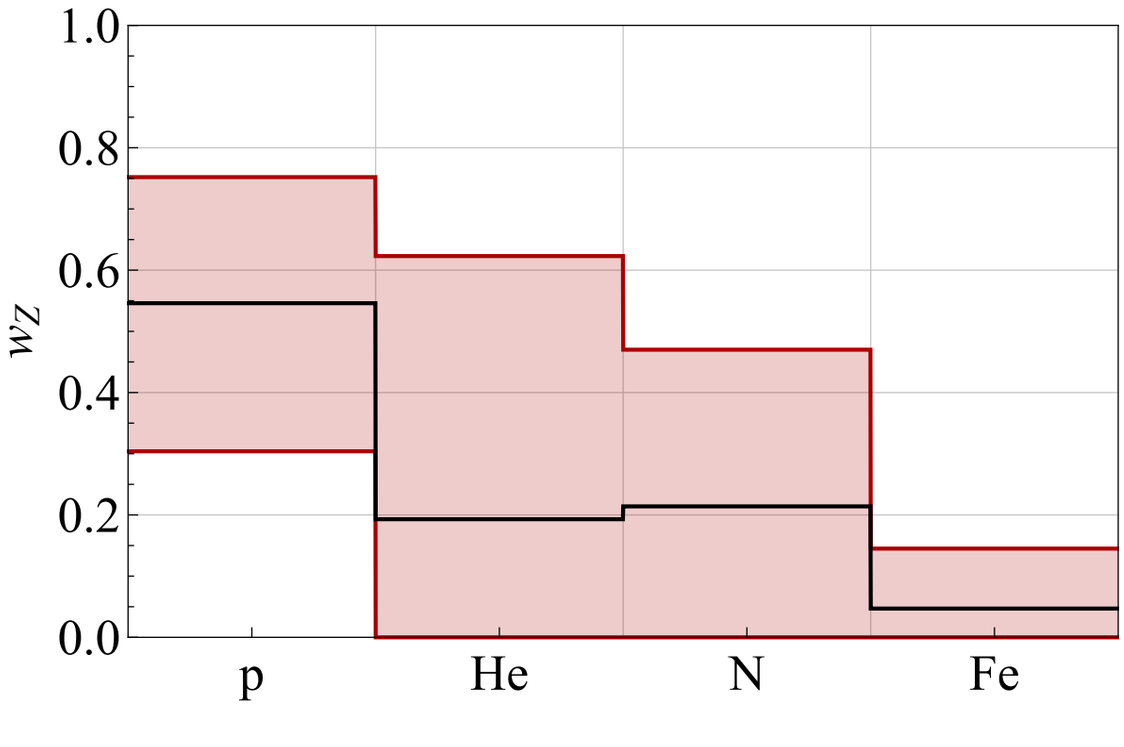
# 4 primary mixture

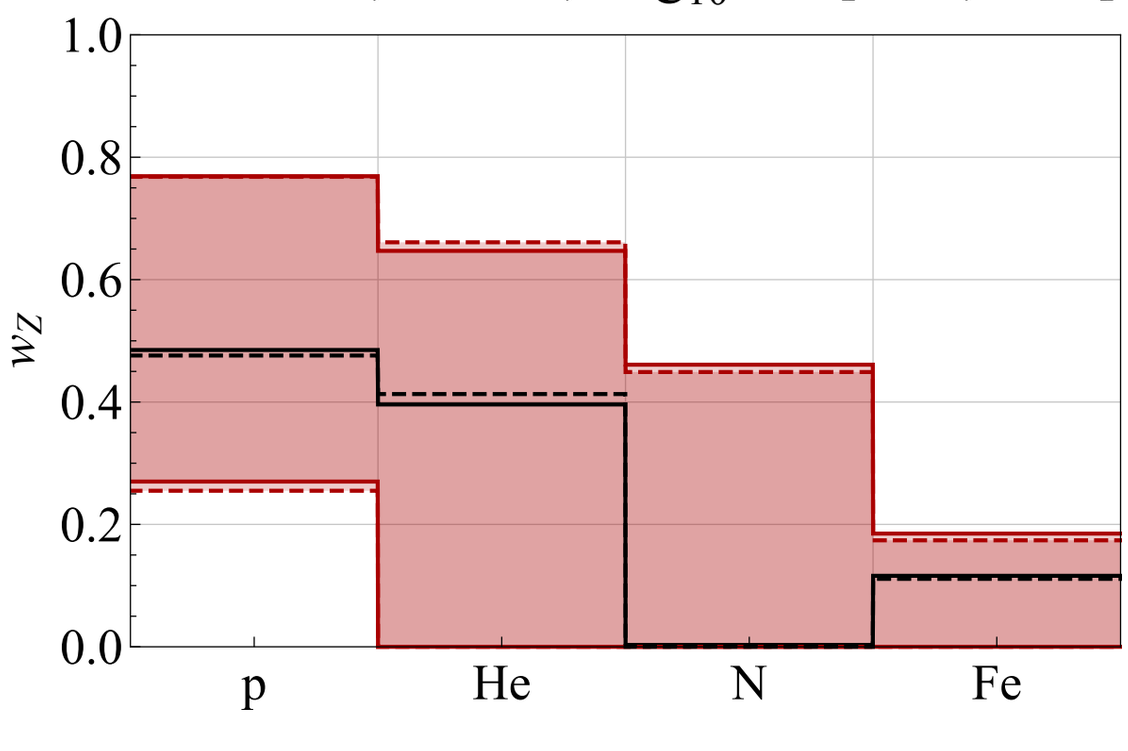



Projections of 4D log-likelihood

NS allows to efficiently sample the likelihood and find the Confidence Levels (CL)

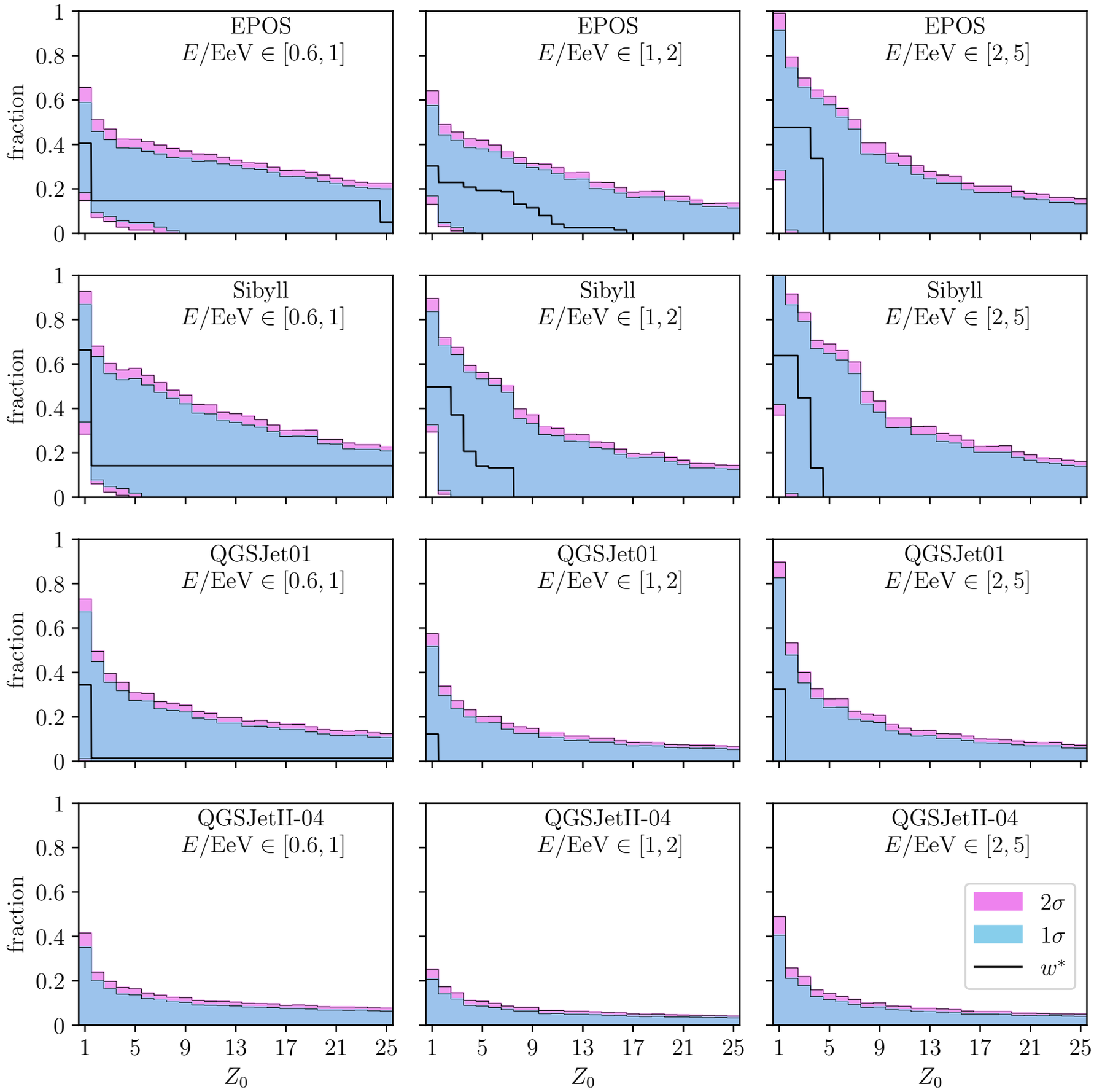
# 4 primary mixture




Consistent with results from 2001.02667

Results are unchanged increasing
the number of features



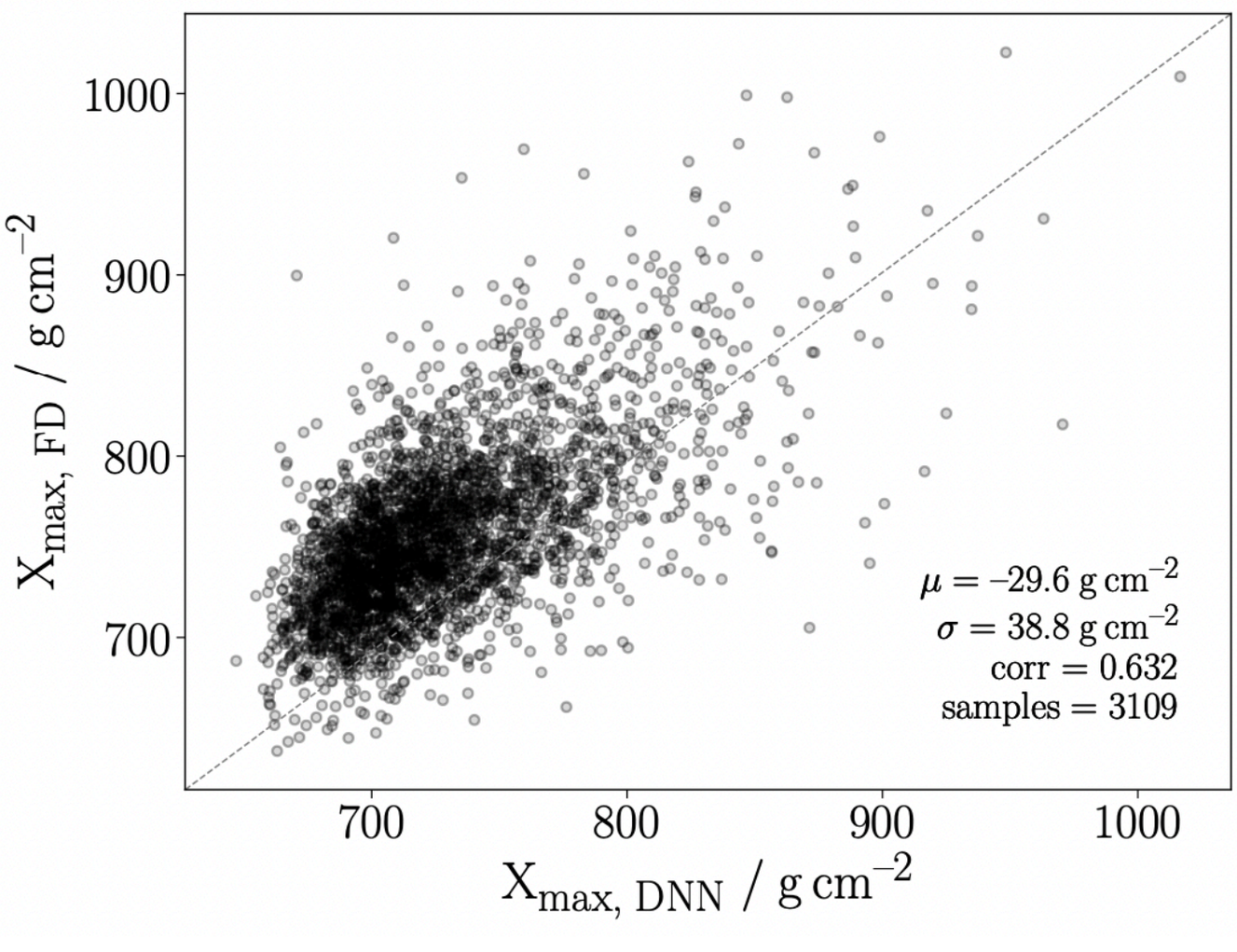

EPOS and Sibyll (LHC-based) exclude 100% proton and can give bounds on heavy elements

QGS models are consistent with 100% proton composition

# SD to FD

Build a map from Surface Detector data to Fluorescent Detector data ($X_{\max}$)

(Auger Coll.: 2101.02946)




Train Deep Neural Network (DNN) on simulated data

✓ Shows strong correlations

— As good as simulations (ground data sims are ~wrong)

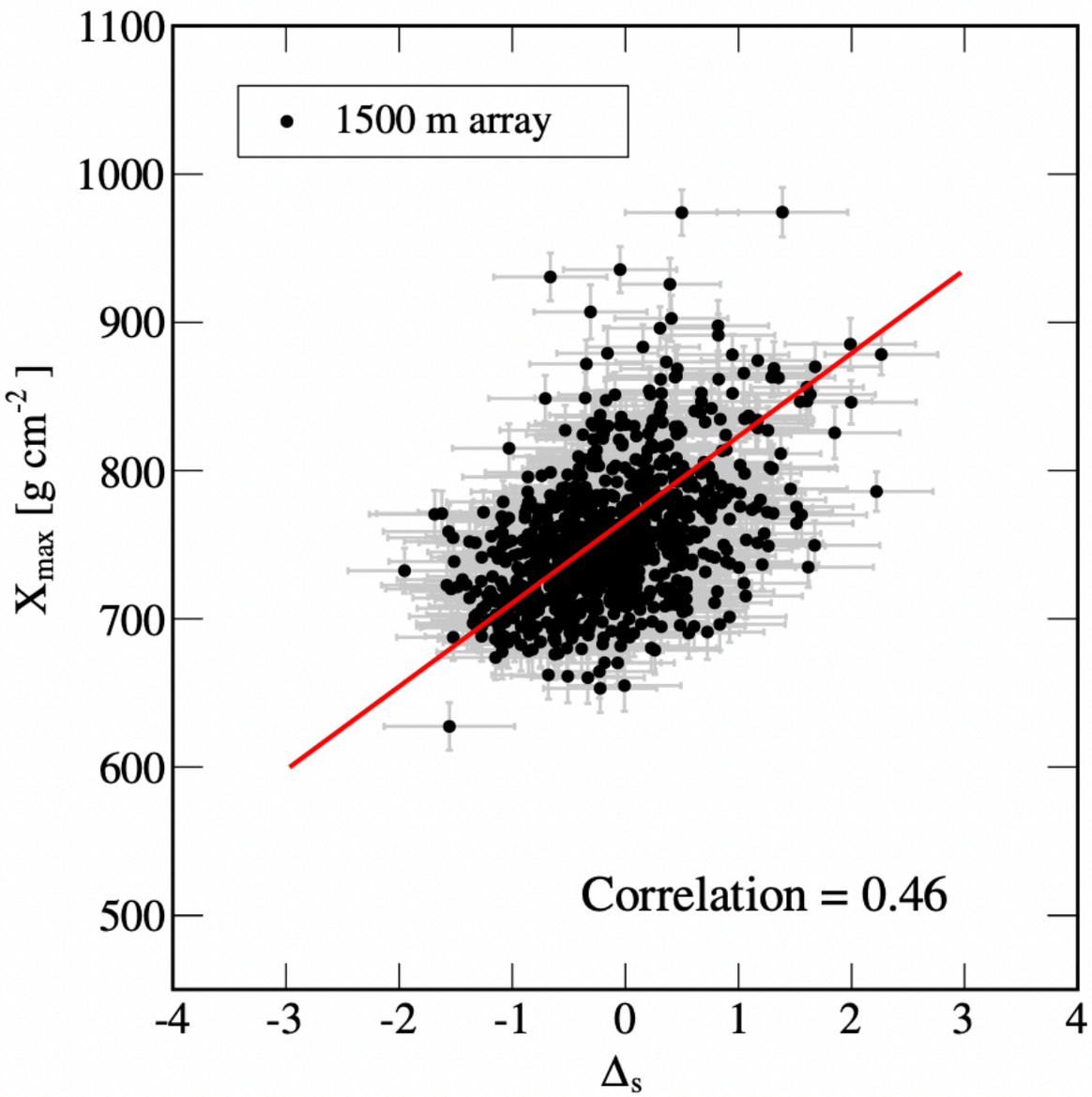
— Trained with 4 primaries

✗ We do not have access to this DNN

# SD to FD

Build a map from Surface Detector data to Fluorescent Detector data ($X_{\max}$)

(Auger Coll.: 1710.07249)




Build simple observable that correlates with FD

— Weaker correlation and large uncertainties

— Some quantities are given fits

✓ We can reproduce and use this

✓ We can (try to) improve on this