# Machine Learning Methods
# $\times$
# Ultra-High Energy Cosmic Rays

Blaž Bortolato

Wien, 14.12.2023

Institut
"Jožef Stefan"
Ljubljana, Slovenija

University of Ljubljana
Faculty of Mathematics and Physics

# Focus of the talk

→ **Infer** the composition of Ultra High-Energy Cosmic Rays

→ Solve **unfolding** problem in an unbiased way
(work in progress)

→ Remove bias in **classification** problems

→ Use results from **regression** in an unbiased way

B. Bortolato, J. F. Kamenik, and M. Tammaro, 2304.11197
B. Bortolato, J. F. Kamenik, and M. Tammaro, Phys. Rev. D 108, 022004 (2023), 2212.04760.

# Why Cosmic Rays?

**Ultra High-Energy Cosmic Rays (UHECR)**

→ atomic nuclei (H, He,..., Fe)

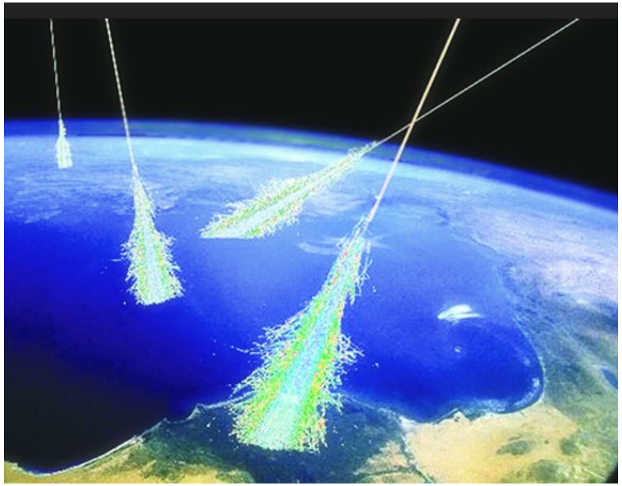
→ $E \geq 10^9$GeV (proton mass: 1GeV)

**Unknowns**:

→ acceleration mechanisms

→ sources

**This Talk** → composition (fractions of atomic nuclei)



Artist's impression of cosmic rays striking Earth (Simon Swordy/University of Chicago, NASA)

**Motivation**

→ galactic & inter-galactic magnetic fields, propagation

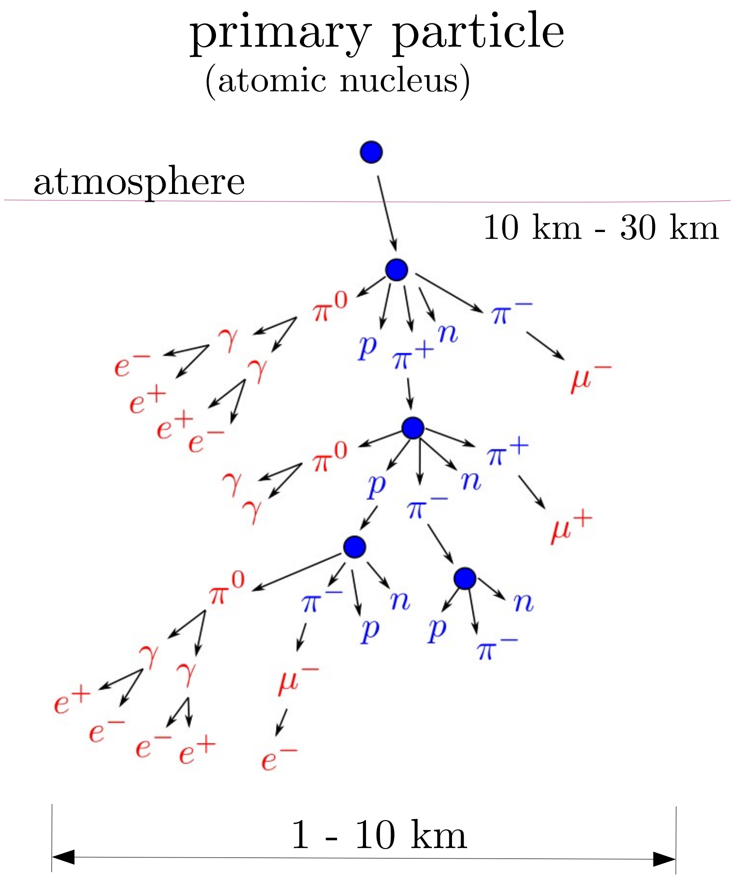
→ acceleration mechanisms (extreme events)

→ hadron interactions at ultra high energies

→ $10^5 \times - 10^7 \times$ the energy at LHC

→ **undiscovered particles**

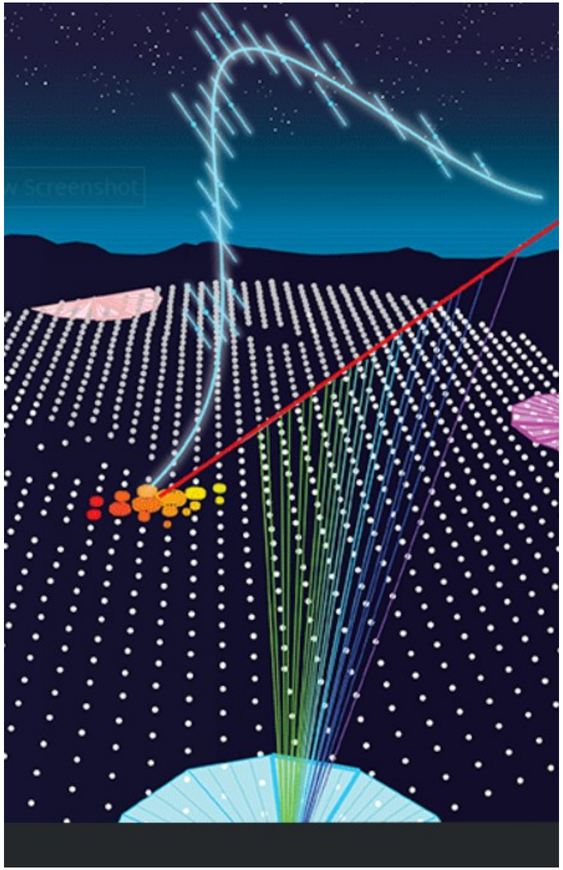
# Extensive Air Showers



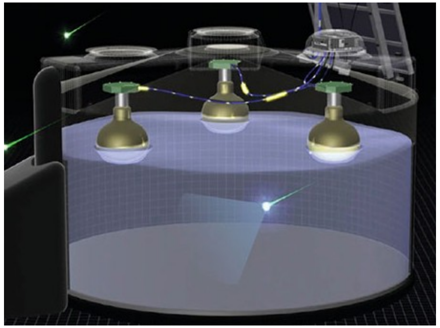

→ Most of the energy on the ground is carried by muons

→ Charged particles ionize nearby particles in the air, fluorescence light is emmited

# Pierre Auger Observatory

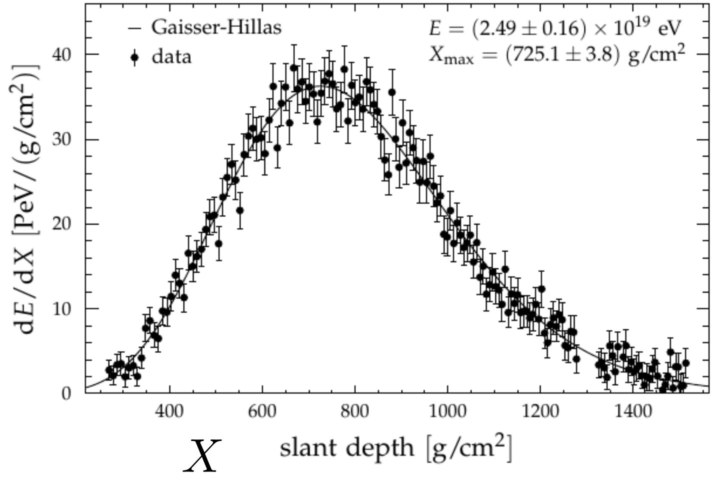


Surface Detectors (SD) are
**water Cherenkov detectors**



$$X = \int \rho(\vec{r})\ dr$$

Fluorescence detectors (FDs) operate only in clear, moonless nights




→ 1600 Surface detectors, 1.5 km apart, cover 3000 km$^2$

→ 4 × 6 Fluorescence detectors, 330–380 nm

Pierre Auger Collaboration, Alexander Aab (Siegen U.) et al. 1409.4809 [astro-ph.HE] Phys.Rev.D 90 (2014) 12, 122005

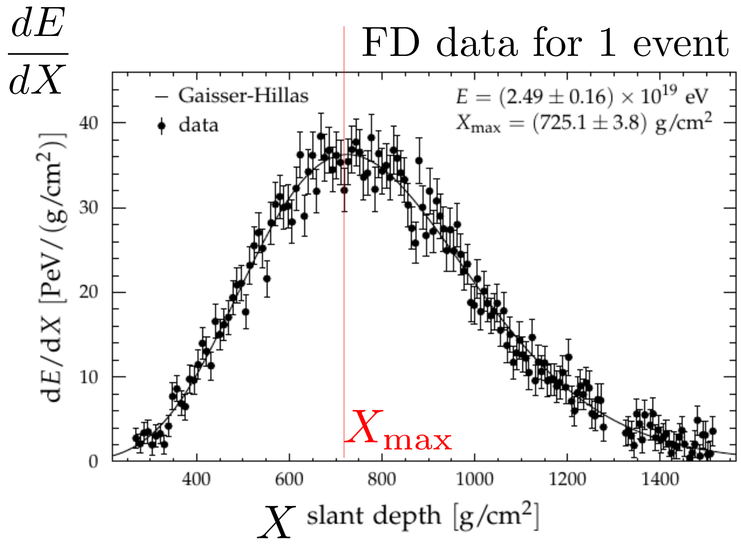
# Data

## Pierre Auger Observatory Open Data 2021

→ 10% of all observed events

∼ 1600 events with SD and FD data

∼ 22000 events with only SD data




**Composition is obtained by comparing simulations to measurements**

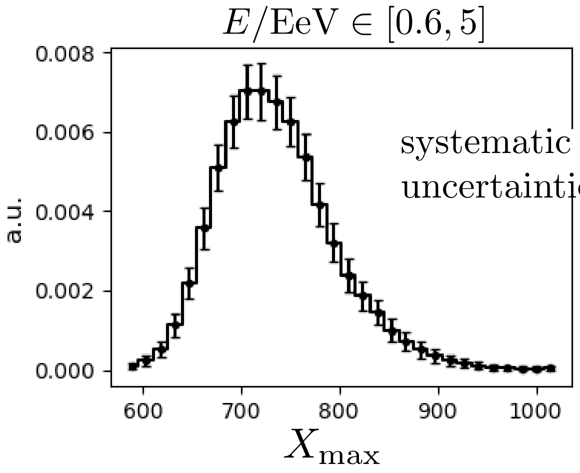
→ data from SD detectors cannot be reproduced by simulations

→ $X$ distribution can be well simulated

→ $X_{max}$ depends on mass of the primary particle

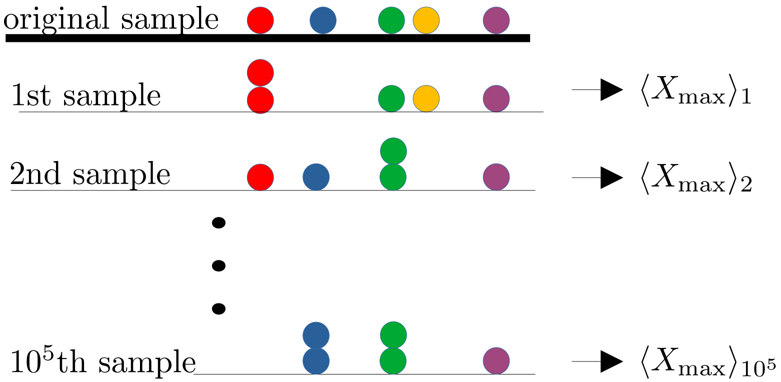
Measured values
(Pierre Auger Observatory)

# What we propose




systematic + statistical uncertainties

## Bootstrap method




Most of the information: $\mathbf{z} = (z_1, z_2, z_3, z_4)$

$$z_1 = \frac{1}{N} \sum_{i=1}^{N} X_{\mathrm{max},i} \qquad n = 1, 2, 3, 4$$

$$z_n = \frac{1}{N} \sum_{i=1}^{N} (X_{\mathrm{max},i} - z_1)^n$$

Distribution of observables

$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z} | \mu, \Sigma)$$

*This trick lowers computational costs, but it is not required

# What we propose

**Simulations**

→ 0.6 EeV - 5 EeV

→ 4 hadronic models

→ 26 primaries: p, He, Li,..., Fe

→ 6000 events/model/primary

$$P(\mathbf{z}|\mathbf{w}) = \mathcal{N}(\mathbf{z}|\mu(\mathbf{w}), \Sigma(\mathbf{w}))$$

$$\mathbf{w} = (w_p, ..., w_{Fe})$$

**Measurements**

$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu, \Sigma)$$

**Problem**

observed data

simulations

Distribution of compositions

$$P(\mathbf{z}) = \int P(\mathbf{z}|\mathbf{w})\ P(\mathbf{w})\ d\mathbf{z}^n$$

**Solution**

simulations

observed data

$$\log L(\mathbf{w}) = \int \log[P(\mathbf{z}|\mathbf{w})]\ P(\mathbf{z})\ d^n\mathbf{z}$$

Bayes theorem

$$P(\mathbf{w}) \propto L(\mathbf{w})\ \text{Prior}(\mathbf{w})$$

# What we propose

**Log-Likelihood**

simulations observed data

$$\log L(\mathbf{w}) = \int \log[P(\mathbf{z}|\mathbf{w})]\ P(\mathbf{z})\ d^m\mathbf{z}$$

$\rightarrow$ analytical solution if $P(\mathbf{z}|\mathbf{w})$ and $P(\mathbf{z})$ are normal

**Bayes theorem**

$$P(\mathbf{w}) \propto L(\mathbf{w})\ \mathrm{Prior}(\mathbf{w})$$

$\rightarrow$ 25 dimensional distribution

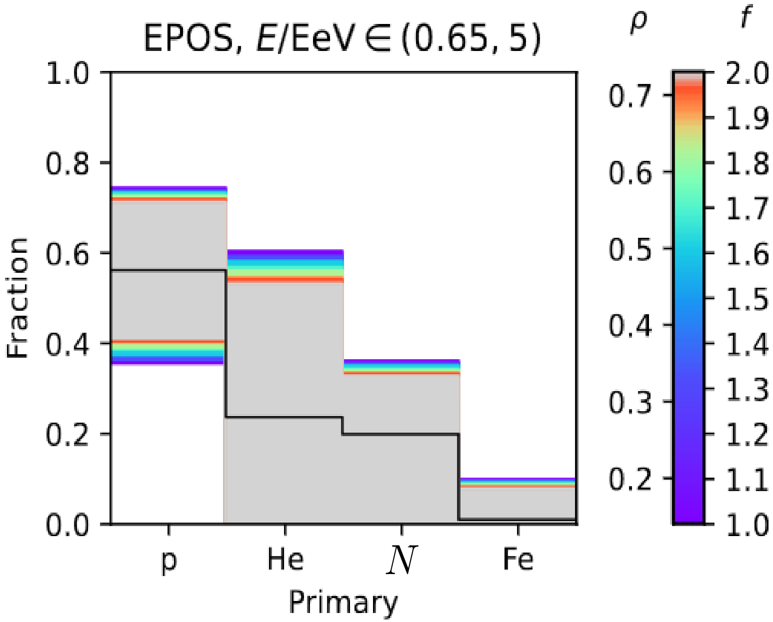
$\rightarrow$ sample from $P(\mathbf{w})$ with **Nested Sampling (NS)**

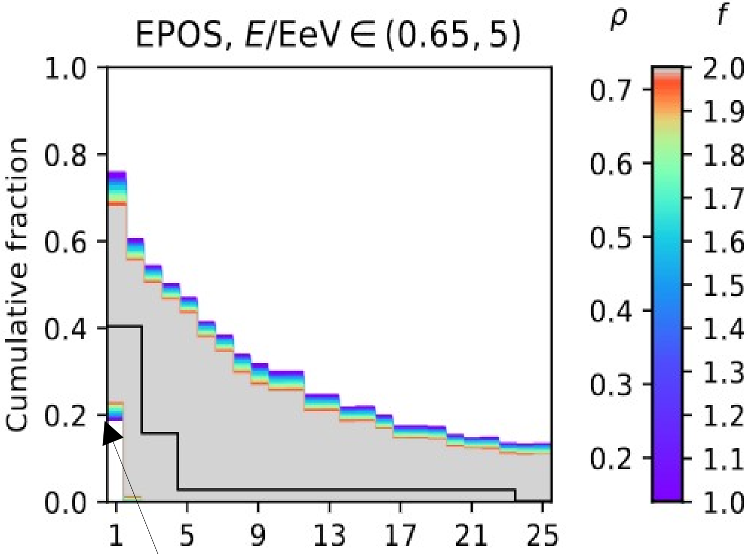
$\rightarrow$ NS estimates the volume of $P(\mathbf{w})$ as a function of $\log L$

**Most probable composition $\mathbf{w}^*$:**

$$\log L(\mathbf{w}^*) = \max$$

# Results




atomic number $Z$

fraction of elements with $Z > 1$

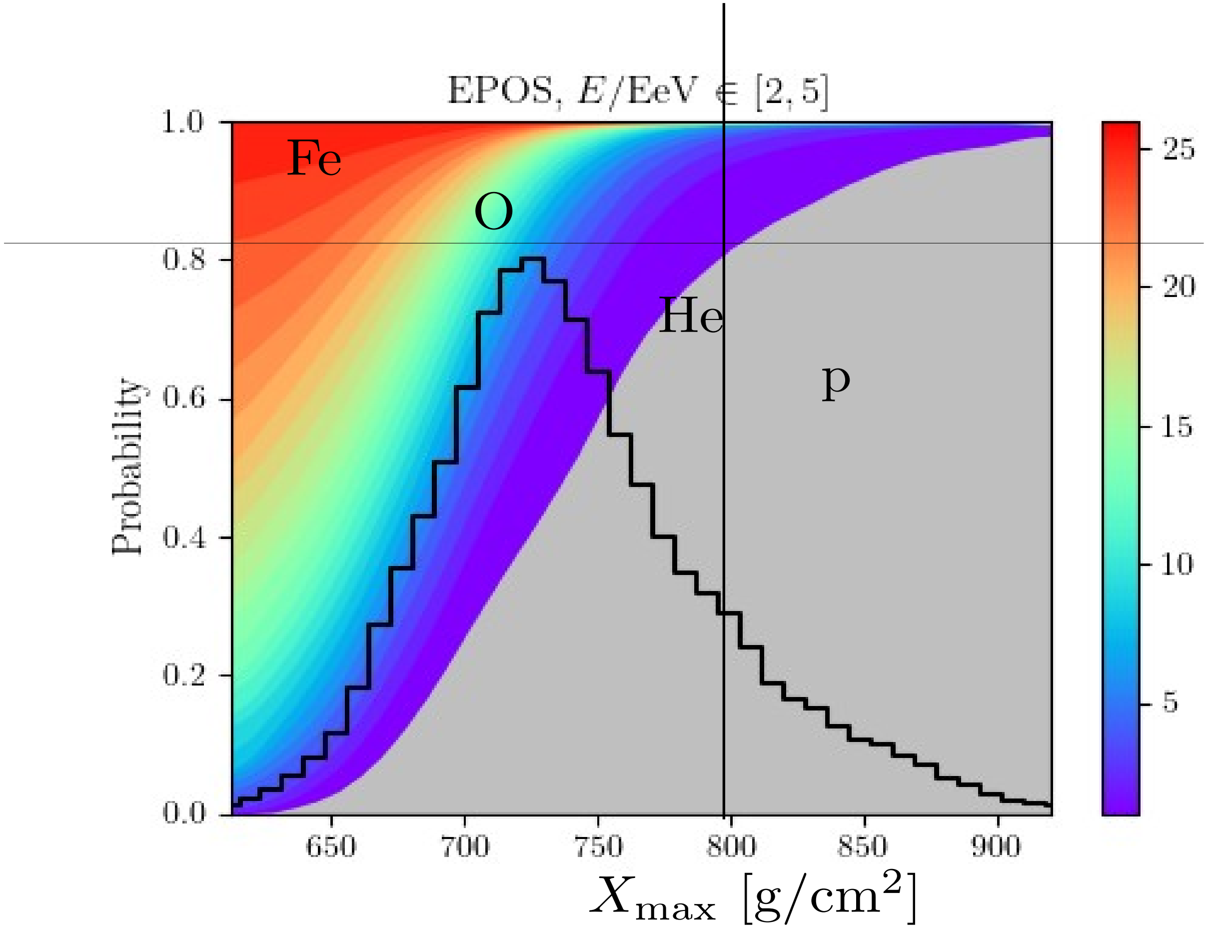
→ 95% CL

→ 1600 events

20% – 80% of primaries are not protons (95% CL)

B. Bortolato, J. F. Kamenik, and M. Tammaro, 2304.11197

# Results




**"Expected Classification"**

# Applications

**Unfolding problem**

$$P(\mathbf{y}) = \int P(\mathbf{y}|\mathbf{x})\ P(\mathbf{x})\ d\mathbf{x}^n$$

Unbiased solution

$$\log L(\mathbf{w}) = \int \log[P(\mathbf{z}|\mathbf{w})]\ P(\mathbf{z})\ d^n\mathbf{z}$$

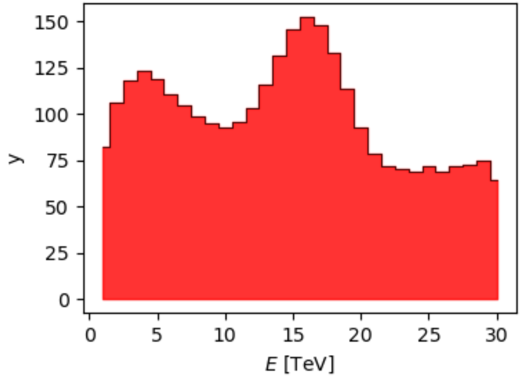
Bayes theorem

$$P(\mathbf{w}) \propto L(\mathbf{w})\ \mathrm{Prior}(\mathbf{w})$$

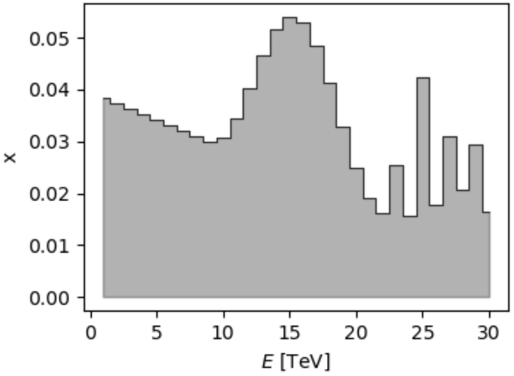
observed distribution

detector response

actual distribution




Unfolding



# Applications

**Removing bias in a classification problem**

$$P(\mathbf{y}) = \int P(\mathbf{y}|\mathbf{x})\ P(\mathbf{x})\ d\mathbf{x}^n$$

Unbiased solution

$$\log L(\mathbf{w}) = \int \log[P(\mathbf{z}|\mathbf{w})]\ P(\mathbf{z})\ d^n\mathbf{z}$$

Bayes theorem

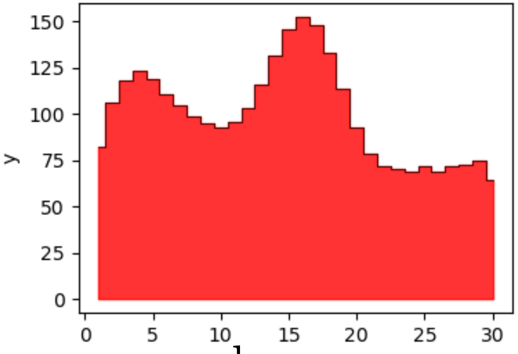
$$P(\mathbf{w}) \propto L(\mathbf{w})\ \text{Prior}(\mathbf{w})$$

predicted probabilities

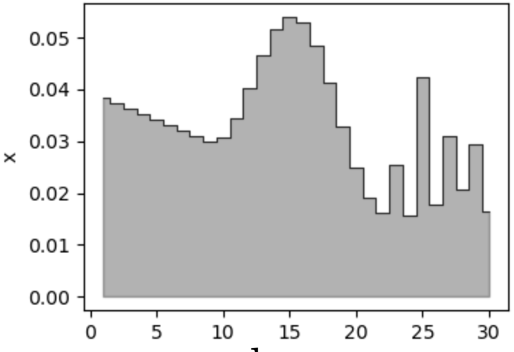
actual labels

classifier response (confusion matrix)

Unfolding




classes




classes

# Applications

**Removing bias in a classification problem (step by step)**

→ train a classifier (on balanced dataset)

→ evaluate the confusion matrix, normalize rows (bootstrap method)
$\{R_1, ..., R_N\}$ $\quad P(y|x) :$ x $\rightarrow \{\mathrm{R}_1 x, ..., R_N x\}$

→ predict distribution of classes on unlabelled dataset (bootstrap method)
$\{y_1, ..., y_N\} \rightarrow P(y)$

→ infer distribution $P(x)$ of actual distributions of classes $x$ $\quad \log L(\mathbf{x}) = \int \log[P(\mathbf{y}|\mathbf{x})]\ P(\mathbf{y})\ d^n\mathbf{y}$
$P(\mathbf{x}) \propto L(\mathbf{x})\ \mathrm{Prior}(\mathbf{x})$

→ reweight classifier outputs

**Classifier is treated as a black box →**
**predictions are statistically completely meaningful**

# Applications

If $P(x)$ is parametrized with $P(x|\theta)$,
optimal parameters are given by maximizing:

observed data

$$P(\mathbf{y}) = \int P(\mathbf{y}|\mathbf{x})\, P(\mathbf{x})\; d\mathbf{x}^n$$

detector/ model

$$\int \log\left\{\int P(\mathbf{y}|\mathbf{x})P(\mathbf{x}|\theta)\; d^n\mathbf{x}\right\}\; P(\mathbf{y})\; d^m\mathbf{y} = \max$$

detector/ model — observed data

→ NN

→ normalizing flows

→ ...

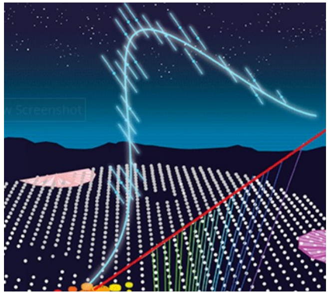
In practise...

$$\left\langle \log\left\{ \langle P(\mathbf{y}|\mathbf{x})\rangle_{\mathbf{x}\sim P(\mathbf{x}|\theta)} \right\} \right\rangle_{\mathbf{y}\sim P(\mathbf{y})} = \max$$

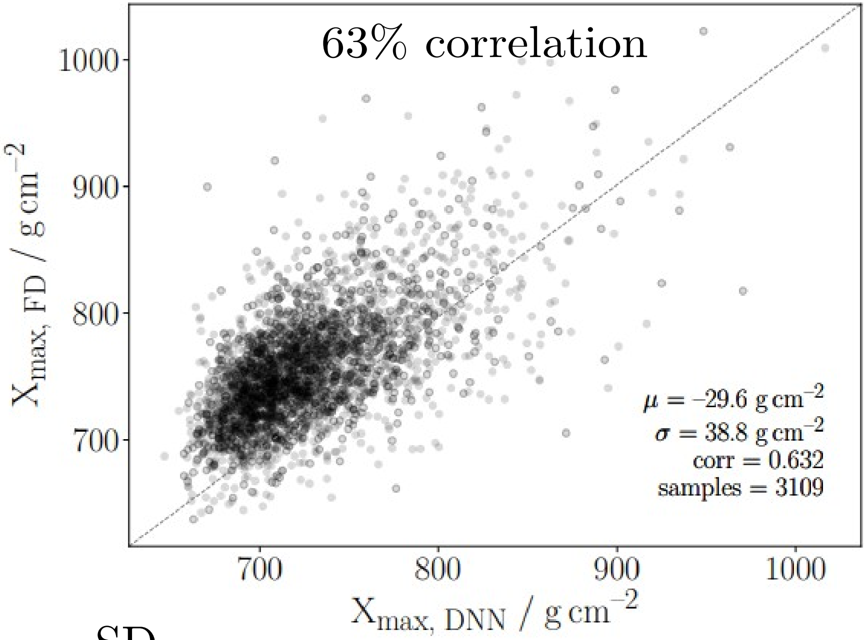
detector/ model — observed data

# Regression

Measured $X_{\max}$



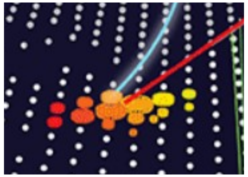
FD




Events with both SD and FD data

SD

Surface detectors data



$\rightarrow$ DNN $\rightarrow X_{\max,\mathrm{DNN}}$

Pierre Auger Collaboration, Alexander Aab (Nijmegen U., IMAPP) et al., 2101.02946 [astro-ph.IM], DOI: 10.1088/1748-0221/16/07/P07019

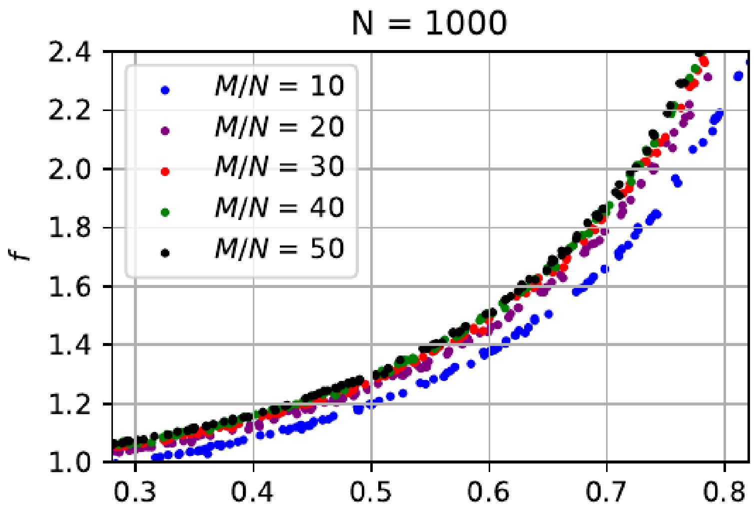
# Regression

**How do we use predicted $X_{\mathrm{max,DNN}}$?**

1) estimate $P(X_{\mathrm{max}}|X_{\mathrm{max,DNN}})$
from events with SD & FD

2) sample $X_{\mathrm{max,SD}}$ for events with only SD

bootstrap method




correlation between $X_{\mathrm{max,FD}}$ and $X_{\mathrm{max,DNN}}$

Measure of statistical power:

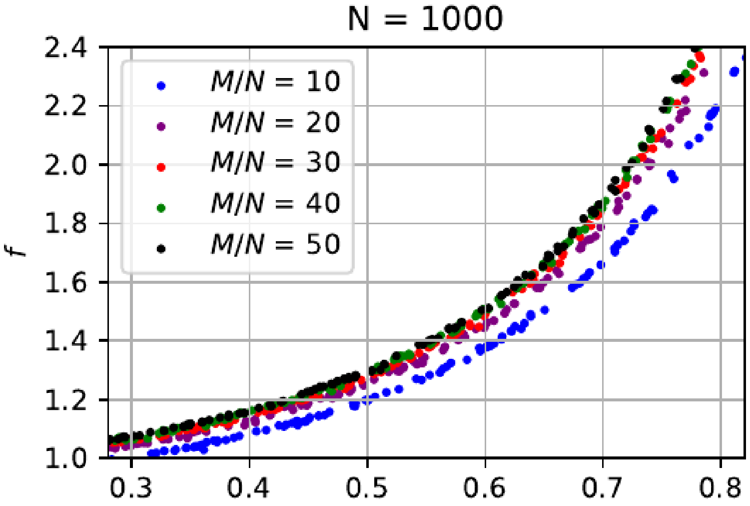
$$f = \frac{\mathrm{Var}(\overline{X^{\mathrm{FD}}_{\mathrm{max}}})}{\mathrm{Var}(\overline{X^{\mathrm{FD+SD}}_{\mathrm{max}}})}$$

variance of the mean of $\{X_{\mathrm{max,FD}}\}$

variance of the mean of $\{X_{\mathrm{max,FD}}\} \cup \{X_{\mathrm{max,SD}}\}$

# Regression

$$f = \frac{\mathrm{Var}(\overline{X_{\mathrm{max}}^{\mathrm{FD}}})}{\mathrm{Var}(\overline{X_{\mathrm{max}}^{\mathrm{FD+SD}}})}$$




correlation between $X_{\mathrm{max,FD}}$ and $X_{\mathrm{max,DNN}}$

Events with FD and SD data (∼ 200 events)

$X_{\mathrm{max}}$

Events with ONLY SD data (22000 events)

$X_{\mathrm{max}} \sim P(X_{\mathrm{max}} | X_{\mathrm{max,DNN}})$ (large uncertainties)

Effectively $> 200$ events with FD and SD

If correlation $|\rho(X_{\mathrm{max}}, X_{\mathrm{max,DNN}})| \sim 75\%$,
$2 \times 200$ events with FD and SD

# Summary

→ Progress on inferring the composition of UHECR

→ composition for 26 primaries with 95% confidence level

→ classification of primaries based on $X_{\max}$

→ Applications of the method

(work in progress)

→ unbiased solution to the unfolding problem

→ removing bias on trained classifiers

observed data — detector/model — Goal

$$P(\mathbf{y}) = \int P(\mathbf{y}|\mathbf{x})\, P(\mathbf{x})\, d\mathbf{x}^n$$

→ Statistical power on predictions of a regression model

# Thanks!


PIERRE
AUGER
OBSERVATORY
Event ID: 092415269700
Time: 30 Aug 2009 04:38:40
Energy: 36.57 EeV
Number of Stations: 21

# Cosmic Rays

**What are made of?**

$e^+$, $p^-$, $e^-$, $p^+$, stable atomic nuclei (H, He,..., Fe)

**Typical sources**

→ gamma-ray bursts

→ active galactic nuclei

→ supernovae

| | | | |
|---|---|---|---|
| **Source** | solar system | galactic sources | extra-galactic sources |
| **Flux** | | $\lesssim \frac{1 \text{ event}}{m^2 \cdot \text{ year}}$ | $\frac{1 \text{ event}}{(km)^2 \cdot \text{ year}}$ |
| **Energy** [GeV] | $\lesssim 10^3$ | $10^3$ - $10^8$ | $> 10^8$ |

**This talk**

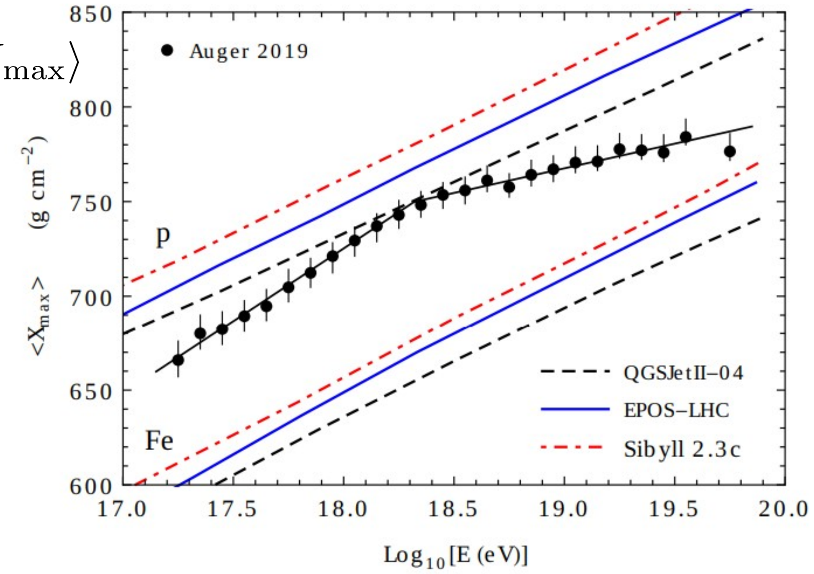
Ultra High-Energy Cosmic Rays (UHECR)

$E \sim 10^9$ GeV

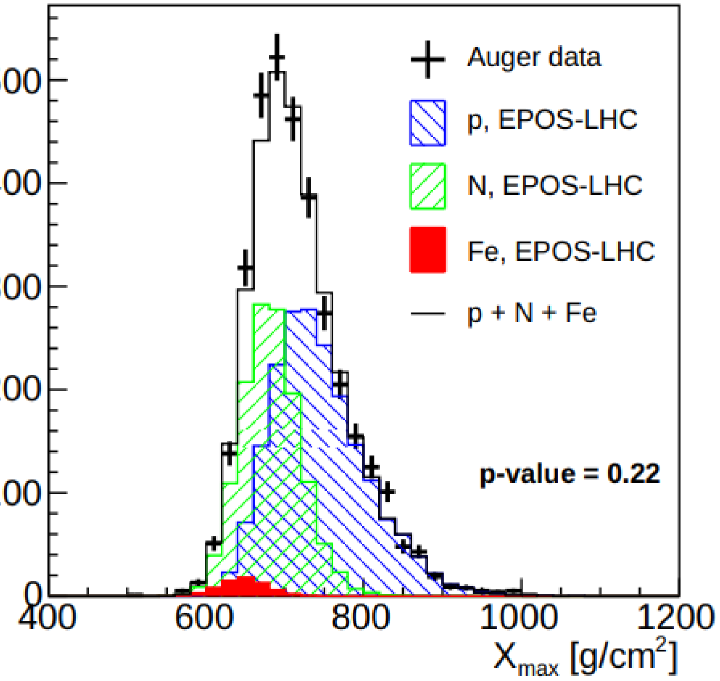
$E \sim 1$ EeV

# Composition from literature

$\langle X_{\max} \rangle$




$\log_{10}(E/\mathrm{eV}) \in [17.9, 18.0]$



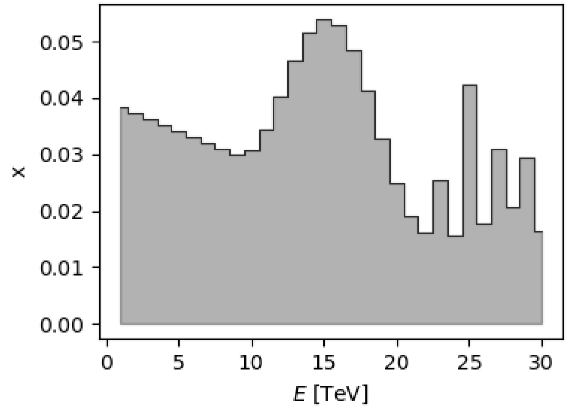

$w_p \sim 60\%$
$w_{He} \sim 0$
$w_N \sim 35\%$
$w_{Fe} \sim 5\%$

$\text{Likelihood}(w) \sim \prod_{bin}$ Poisson distribution
$w = \text{argmax}(\text{Likelihood})$


N. Arsene, O. Sima, 2001.02667 [astro-ph.HE], Eur.Phys.J.C 80 (2020).
P. Lipari, 2012.06861 [astro-ph.HE], Phys.Rev.D 103 (2021) 10, 103009.

# Unfolding problem 1/2

True distribution of events



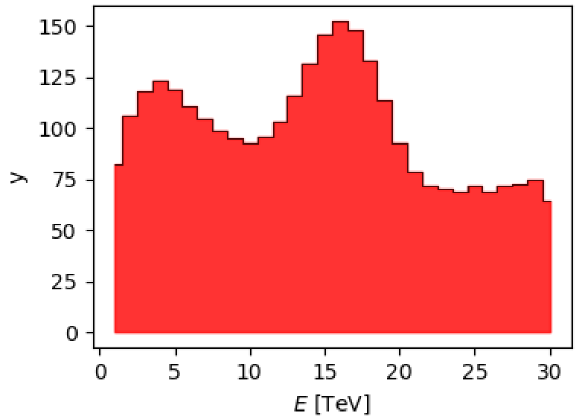

Detector measures

$$\mathbf{y} = (y_1, \ldots, y_m)$$

given original values:

$$\mathbf{x} = (x_1, \ldots, x_n)$$

Observed number of events



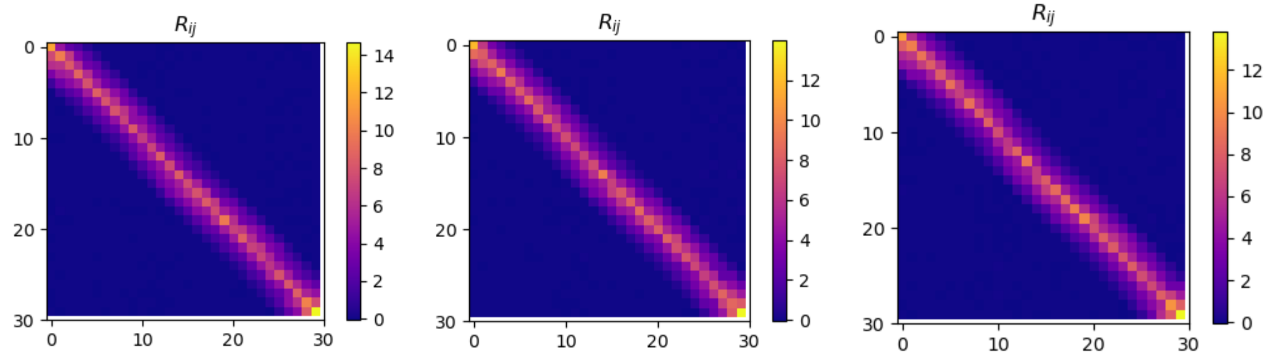

Assume detector response $P(\mathbf{y}|\mathbf{x})$ can be modeled as:

$$y_k = N_y \sum_j R_{kj} \, x_j$$

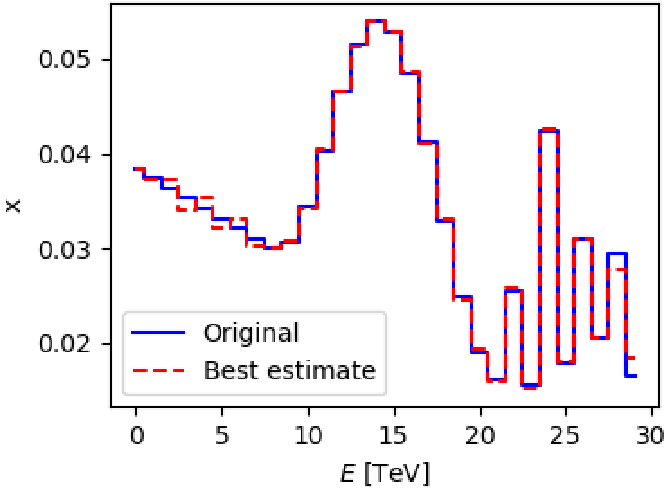
$$R_{kj} = \overline{R}_{kj} + \delta R_{kj} \, \epsilon$$

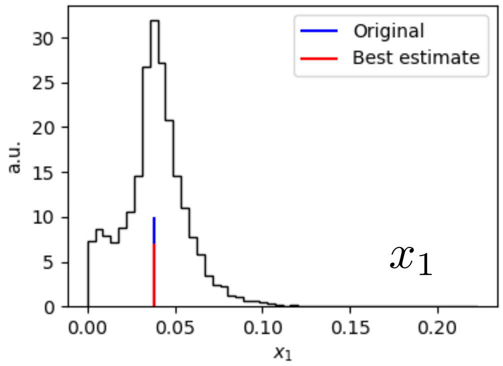
$$\epsilon \sim \mathcal{N}(0, 1)$$

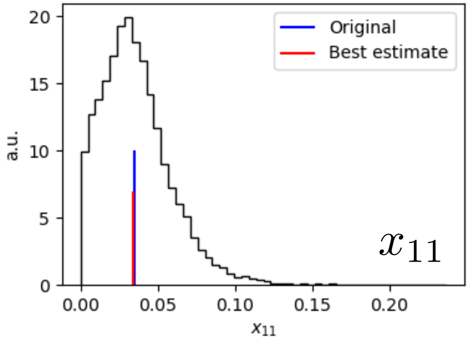
R - response matrix

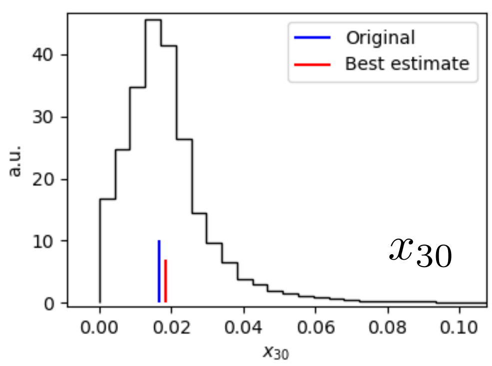


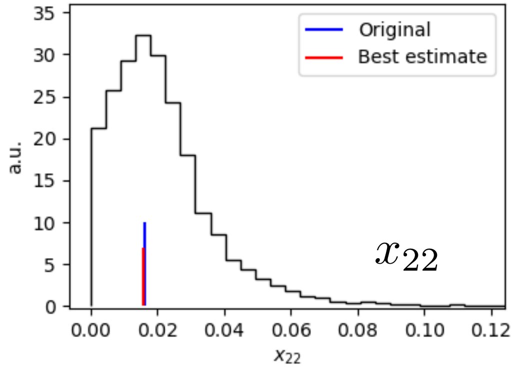
# Unfolding problem 2/2




$$x^{\text{best estimate}} = \text{argmax} \log L(x)$$

$$\log\left(\frac{L(\mathbf{x}^{\text{original}})}{L(\mathbf{x}^{\text{best estimate}})}\right) \approx 0.0047$$