Learning the Composition of Ultra High Energy Cosmic Rays

+

Resolution of (Heavy) Primaries in Ultra High Energy Cosmic Rays

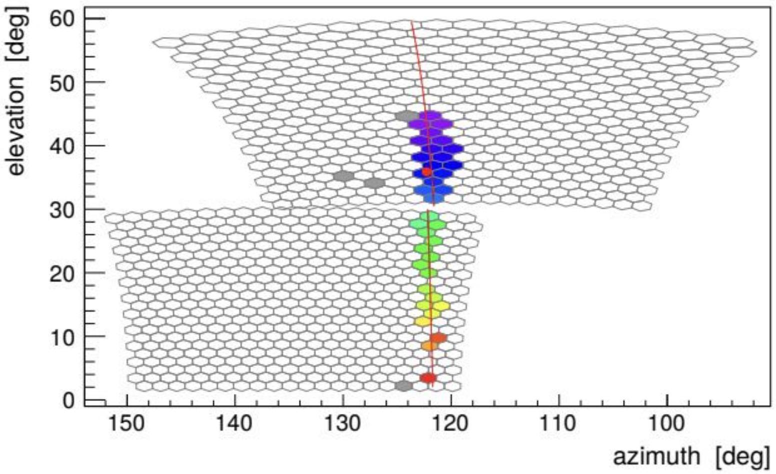
Blaž Bortolato, Jernej F. Kamenik, Michele Tammaro

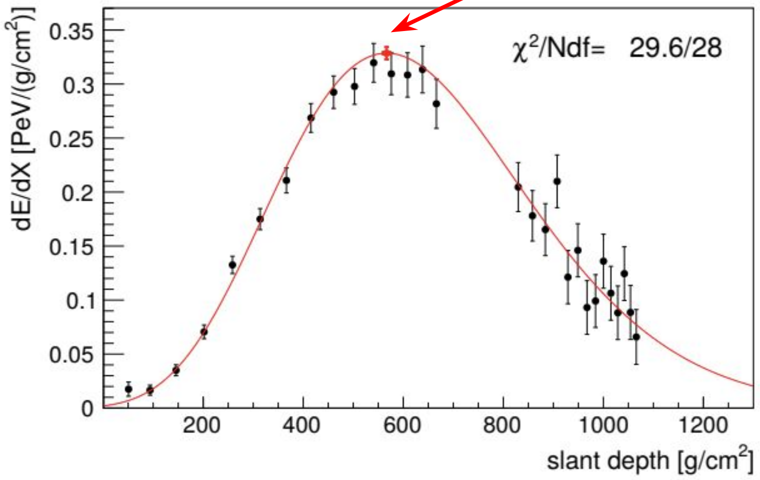
Outline

- Two papers – the second one, originally proposed by Denise, heavily relies on the first
- **Learning the Composition of Ultra High Energy Cosmic Rays** (2022)
  - New methodology for mass fractions fit
  - Some results, but more as a showcase
- **Resolution of (Heavy) Primaries in Ultra High Energy Cosmic Rays** (2024)
  - Further application of the methodology
  - Claims about possible biases in PAO’s results!

Background

- Pierre Auger Observatory has a fluorescence detector that *directly* measures X_{max}

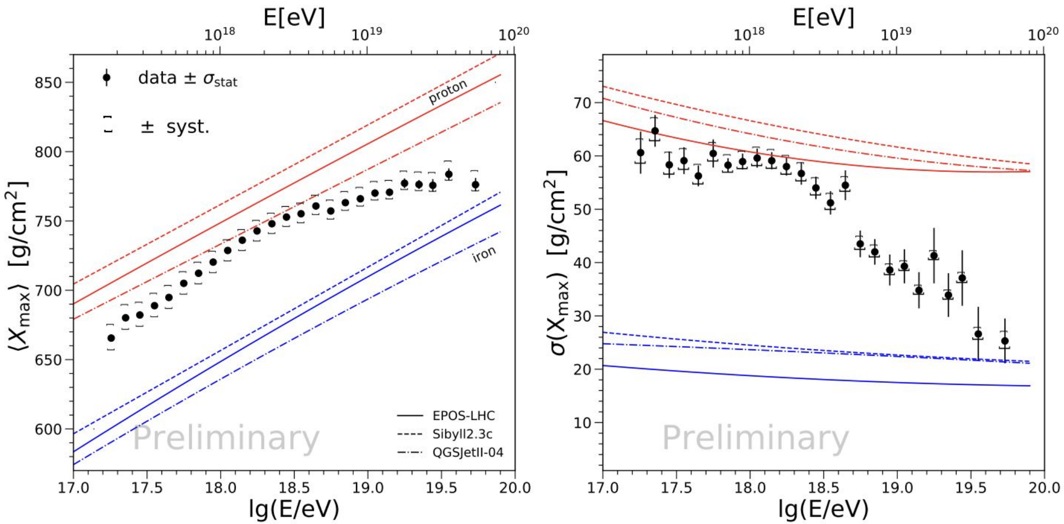




The Pierre Auger Collaboration, Aab, A., Abreu, P., et al. 2019 (arXiv), https://arxiv.org/abs/1909.09073

Background

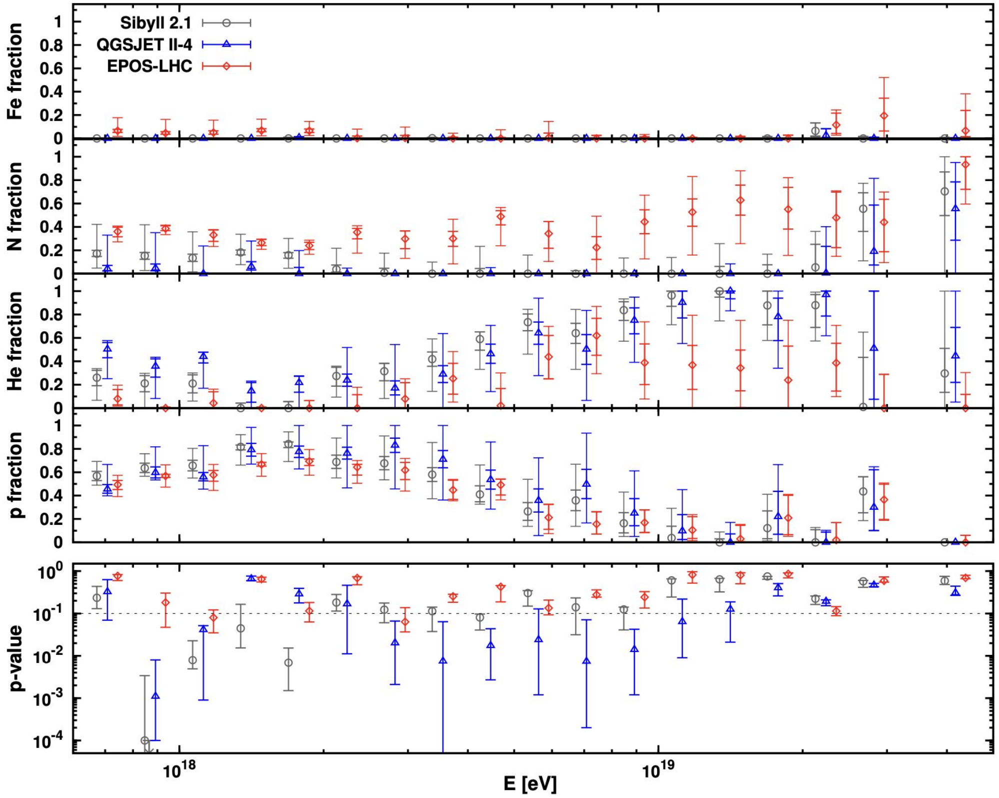
- X_{max} has a direct dependence on the primary particle (and on E, H, ...)



The Pierre Auger Collaboration, Aab, A., Abreu, P., et al. 2019 (arXiv), https://arxiv.org/abs/1909.09073

Background

- Inference of mass composition is possible from X_{max}!
- Within PAO, the analysis is performed by fitting a mixture of simulated X_{max} distribution to the measured one



Aab, A., Abreu, P., Aglietta, M., et al. 2014, Physical Review D, Vol. 90 (American Physical Society (APS)), http://dx.doi.org/10.1103/PhysRevD.90.122006

Paper 1: **Learning the Composition of Ultra High Energy Cosmic Rays**

Bortolato, B., Kamenik, J. F., & Tammaro, M. 2023, Physical Review D, Vol. 108 (American Physical Society (APS)), http://dx.doi.org/10.1103/PhysRevD.108.022004

Outline

- A novel approach to the problem by people outside of PAO collaboration
- Some key points:
  - **Kernel estimation** instead of binning for the X_{max} PDF
  - **Bootstrapping** procedure to evaluate uncertainties in the PDF
  - Generalized **central moments** decomposition for the X_{max} distribution
  - **All 26 primaries** (H - Fe) analysed (+ a discussion in the next paper!)
  - **Cumulative fractions** as a new way of reporting results

Data

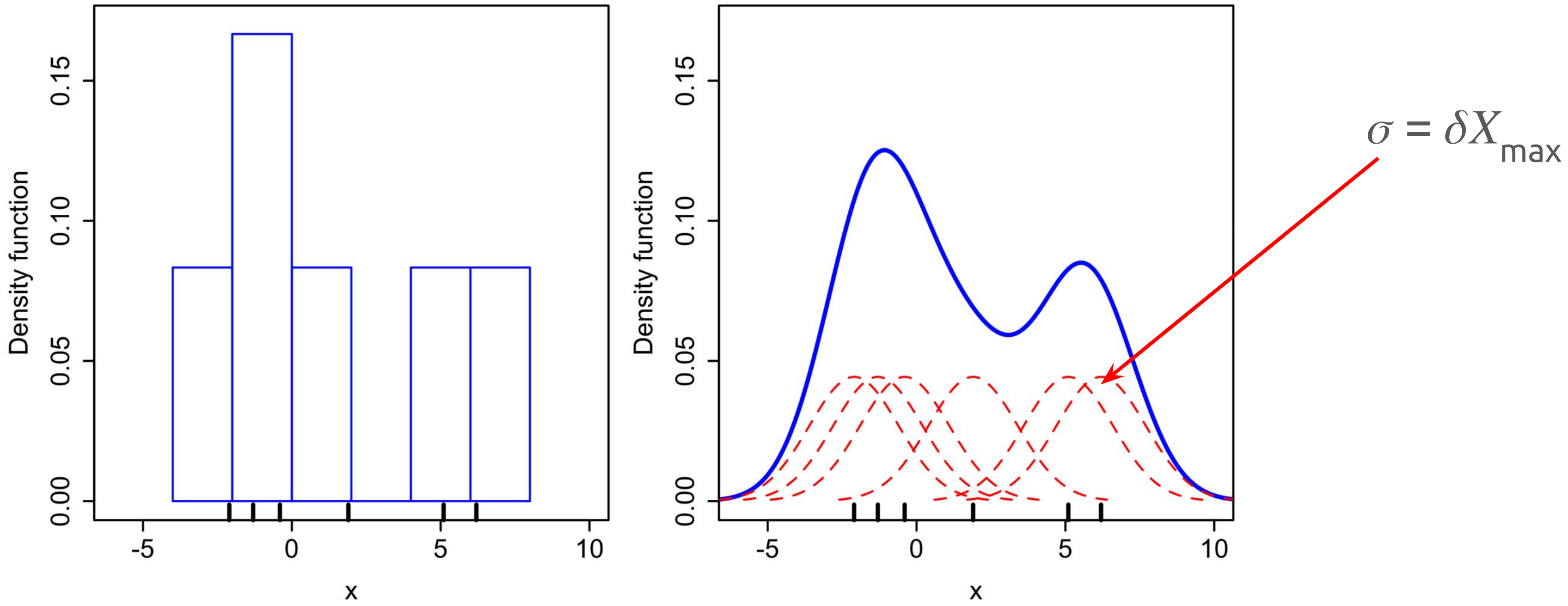
- Pierre Auger Open Data – “the public release of 10% of the Pierre Auger Observatory cosmic-ray data published in recent scientific papers and at International conferences”
- 3156 FD events available, sorted in 3 energy bins

$$E \leq 1\ \mathrm{EeV},\ 1 < E \leq 2\ \mathrm{EeV}\ \text{and}\ 2 < E \leq 5\ \mathrm{EeV}$$

- X_{max} distribution is estimated through a **kernel density** method with Gaussian kernel with width = reported reconstructed uncertainty δX_{max}

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

Sidenote: kernel density estimation (KDE)



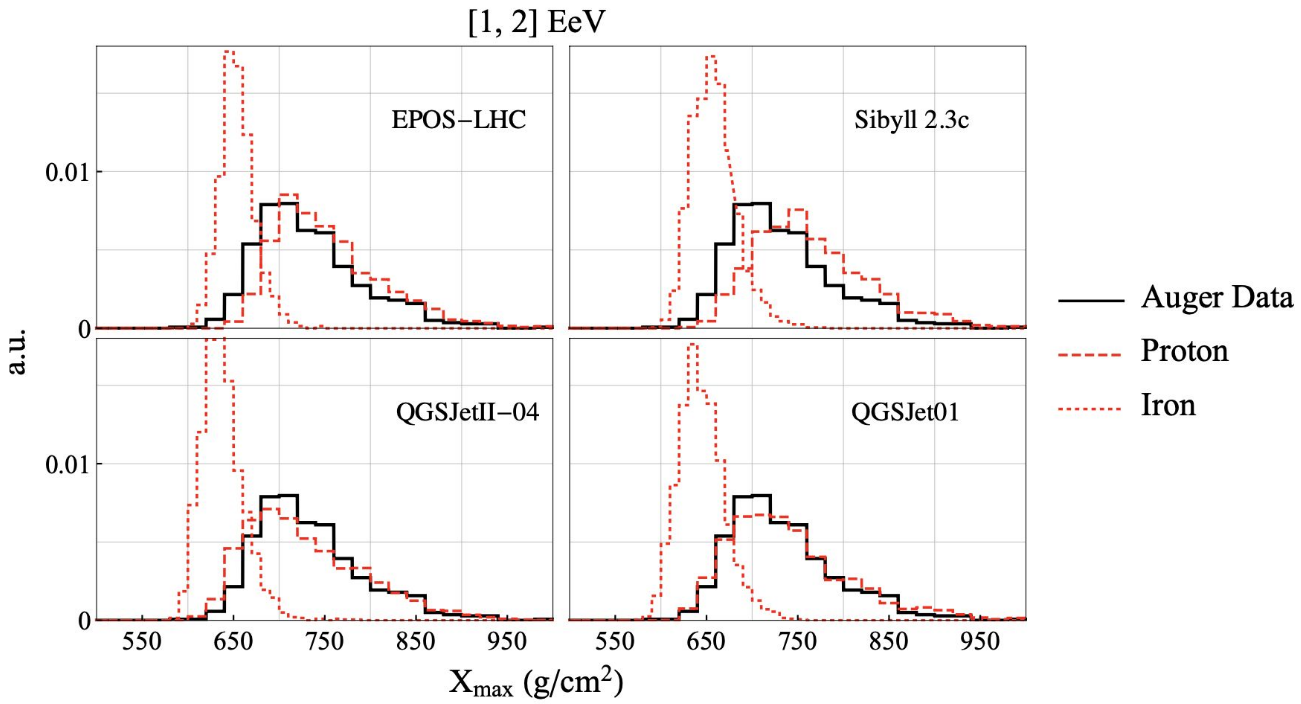
https://commons.wikimedia.org/wiki/File:Comparison_of_1D_histogram_and_KDE.png

Simulations

- CORSIKA7.7401 used to simulate showers
- All primaries with Z from 1 to 26 (most abundant stable isotopes)
- Flat energy distribution
- Hadronic models: QGSJET01, QGSJetII-04 , EPOS and Sibyll 2.3c
- 2000 simulations × 26 primaries × 3 energy bins × 4 hadronic models = 624000 simulated showers
- Again, they estimate X_{max} PDF with KDE (modified)

$$P_{\mathrm{sim}}(X_{\mathrm{max}} \mid S) = \frac{1}{\tilde{N}} \sum_j \int \mathrm{d}\tilde{X}\, \mathcal{N}\left(\tilde{X} \mid X^j_{\mathrm{max}}, \delta X^j_{\mathrm{max}}\right) \times R(X_{\mathrm{max}} - \tilde{X}) \times \epsilon(\tilde{X})$$

Simulations



Central moments decomposition

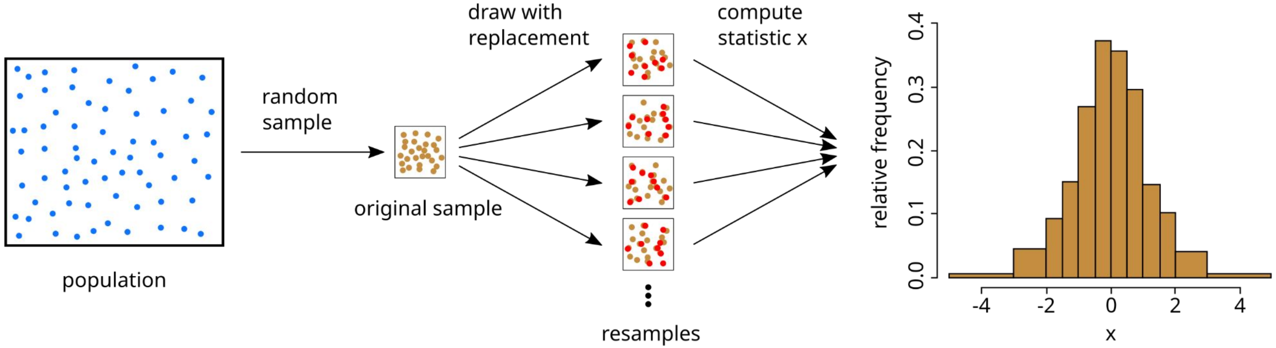
$$z_1 \equiv \langle X_{\max} \rangle = \frac{1}{N} \sum_{i=1}^{N} X_{\max,i} \,,$$

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

- With $n \to \infty$ the moments completely characterize the distribution
- With a finite n it's a dimensionality reduction technique
- Can be effectively estimated from data and simulations through their PDF

Evaluating uncertainties

- PDF approximation with KDE and the moments calculated from it are all *point estimates* – but how it would vary under repeated measurement/simulation?
- Bootstrapping is a generic name for a set of techniques to evaluate properties of point estimates

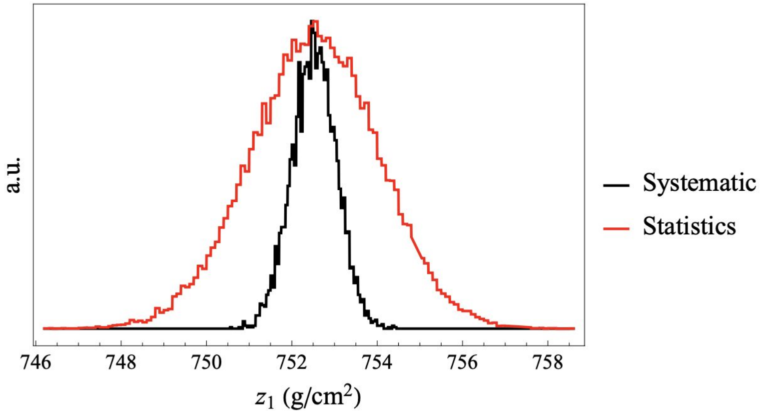


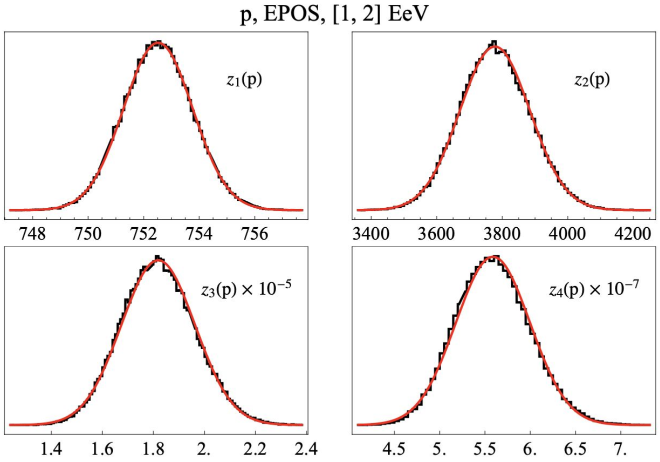
https://commons.wikimedia.org/wiki/File:Illustration_bootstrap.svg

Evaluating uncertainties

- **Systematic**: for every shower the X_{max} is resampled from it's contribution to the KDE
- **Statistical**: the whole dataset is resampled with replacement and the new KDE is produced

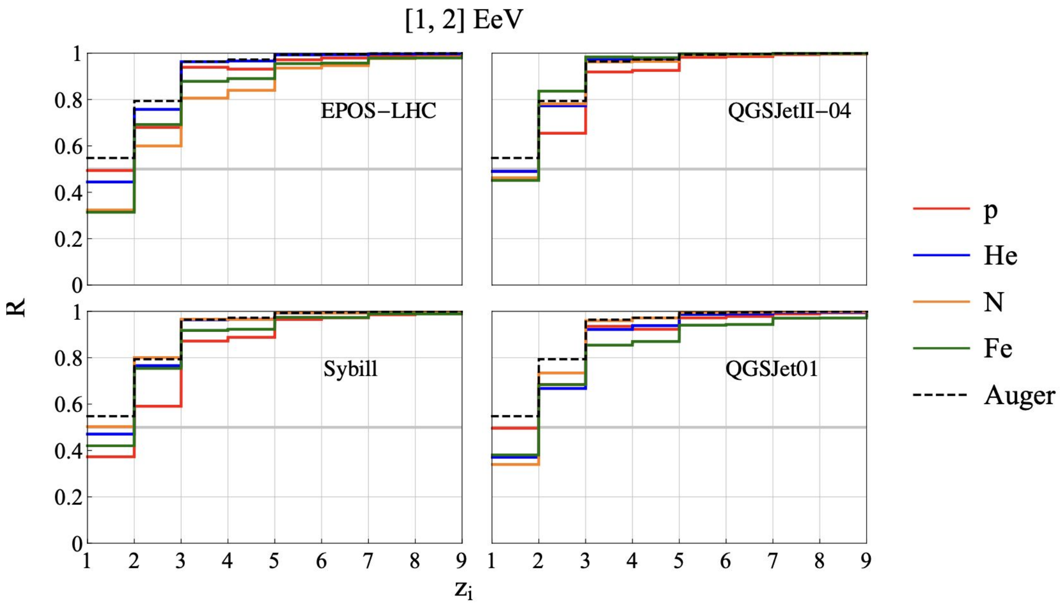
The result is a series of resampled PDF estimations → a sample of central moments → by fitting it with a multinomial normal distribution, the μ and Σ are obtained





How many moments is enough?

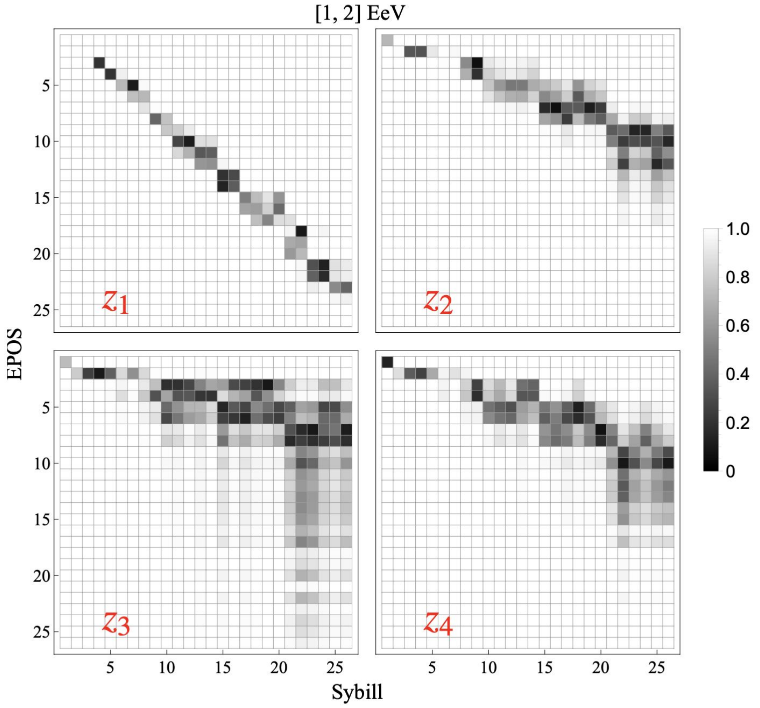
- Having performed the bootstrapping, we can look at the correlations between subsequent central moments
- Low correlation – the moments capture different distribution features, so it's worth going higher
- High correlation = we can drop the higher moments without too much info lost



3 first moments are selected based on this

How different are hadronic interaction models?

- Distributions can be compared with Hellinger distance
- Models (here EPOS and Sibyll) agree on the 1st moment (mean)
- Higher moments show relative bias – e.g. Z=10 according to EPOS is as spread out as Z=20-25 according to Sibyll



Inferring the composition

- From bootstrapping procedure the distribution of central moments vector is obtained

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

- This is true for both measurement and simulation (single primary)
- A mixture of simulations for a given fractions also leads after some transformations to, again, normal approximation

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

- Now, the fluctuations of both data and MC are encoded in the matrices Σ

Likelihood

- Introducing a vector of **unknown** true moments z as a nuisance parameter:

$$\tilde{\mathcal{L}}(z, w) = \mathcal{N}_n\left(z \mid \tilde{\mu}, \tilde{\Sigma}\right) \times \mathcal{N}_n\left(z \mid \mu_w, \Sigma_w\right)$$

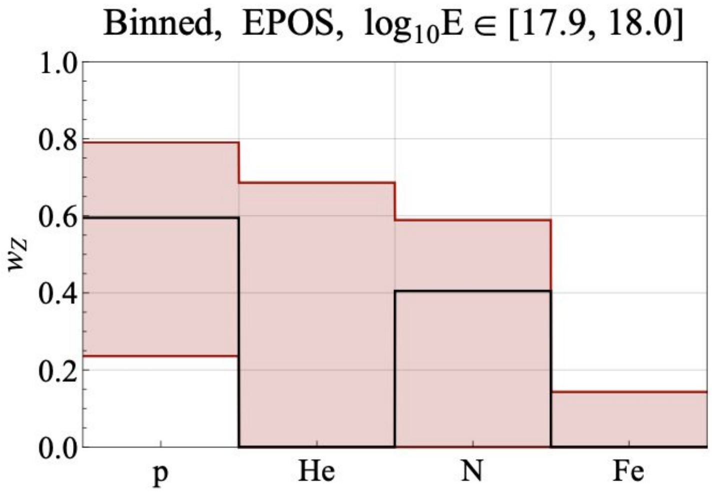
- … and in a Bayesian fashion, marginalizing over them right away

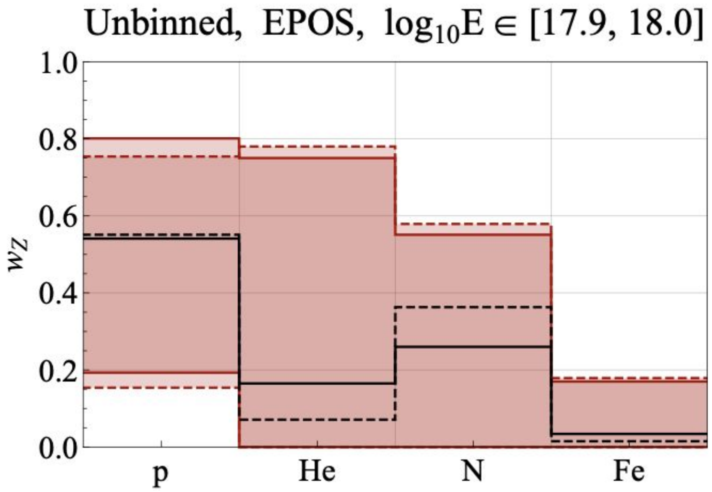
$$\mathcal{L}(w) = \int \tilde{\mathcal{L}}(z, w)\ \mathrm{d}^n z\,.$$

- For normal approximation, this integral is solved analytically
- The likelihood can be maximized to obtain MLE
- Or it can be plugged into a Bayesian MCMC procedure to produce a sample from the posterior and confidence intervals

Results – 4 primaries

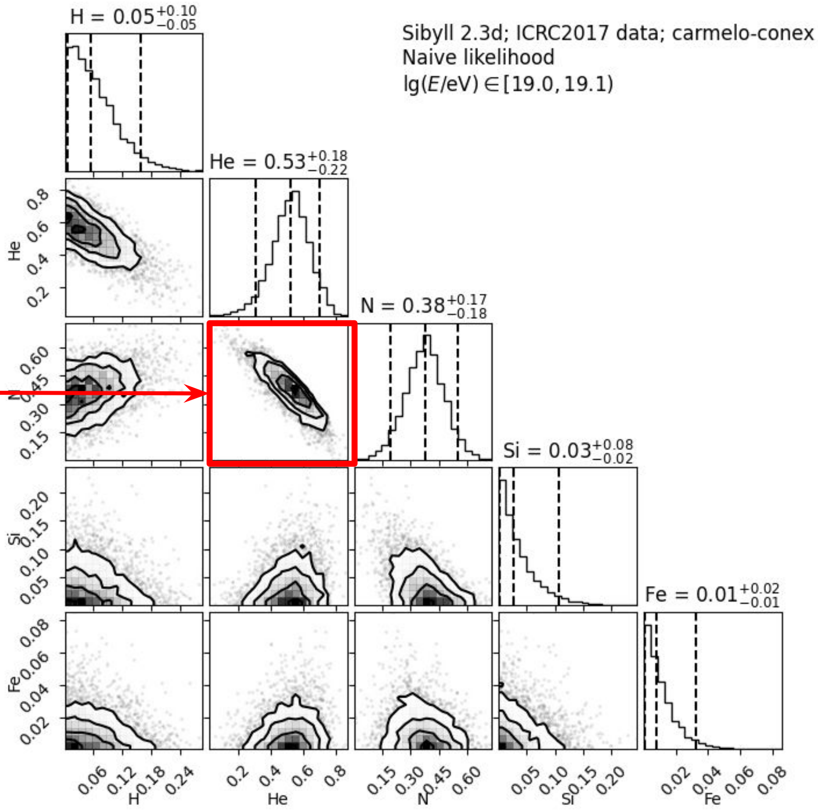
Comparison with the template fit modeled after 10.1103/PhysRevD.90.122006 (on the same PAOD data)





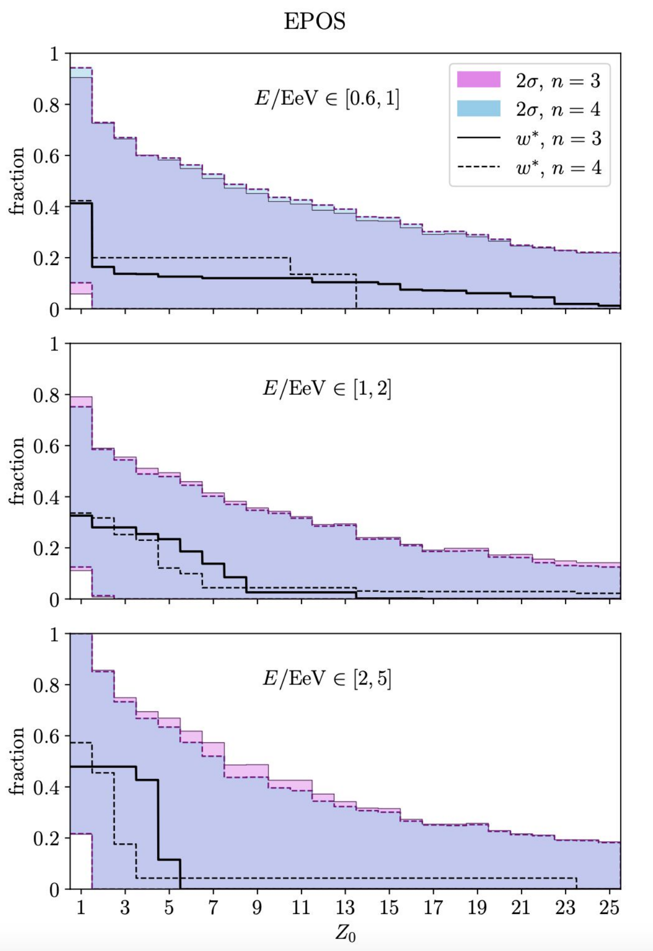
Inferring all the primaries

- There's no point in talking about each of the 26 individual primaries – the composition of close primaries are highly degenerate!
- Instead, we can talk about cumulative distribution: for every Z plot the fraction of elements heavier than Z
- In this representation, the degeneracies are not so important



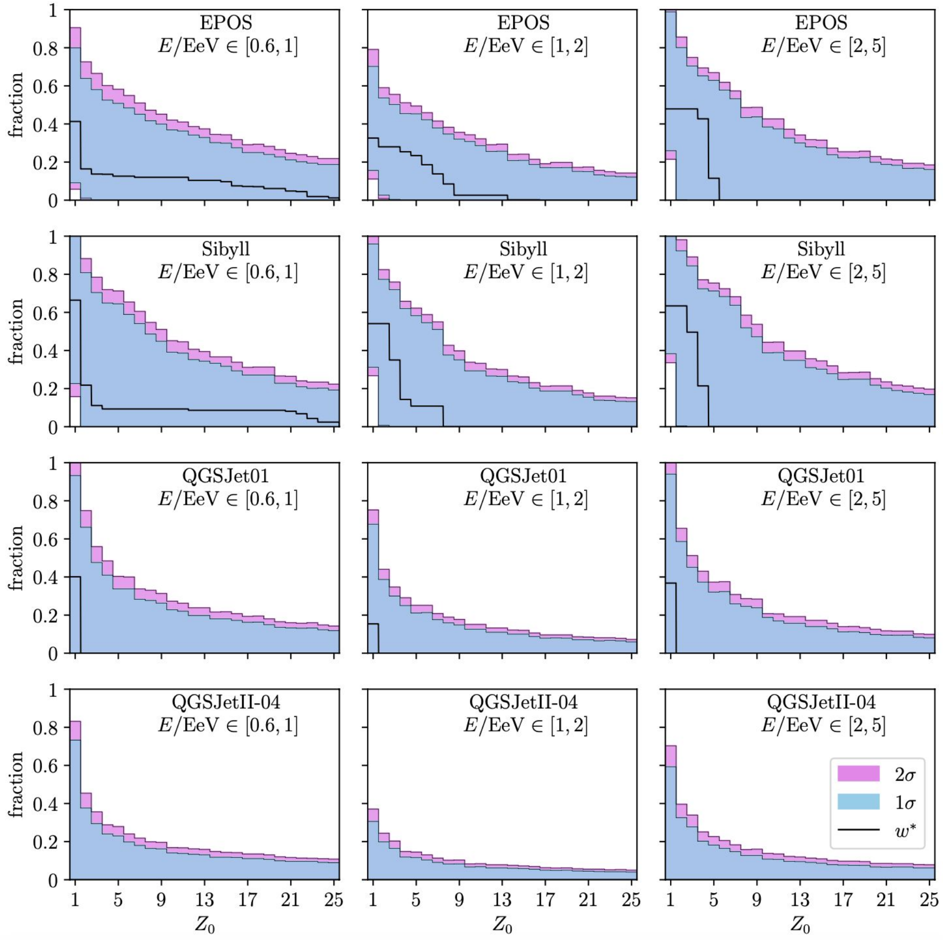
Results – all primaries

- The statements conveyed by this representation are weaker: e.g. we can only exclude that >90% of the showers are sourced by protons (at least 10% of the showers are sourced by elements heavier than protons)
- Still, it's clear that higher energy shower show heavier composition



Results – hadronic model comparison

- QGSJet is consistently disfavored
- EPOS and Sibyll fit the data similarly well, but give different best-fit compositions



Conclusions

- Central moments offer an intuitive low-dimensional representation of X_{max} distribution
- Bootstrapping procedure allows one to quantify the systematic and statistical uncertainties both in data and in MC
- Likelihood function, relying on the multinomial normal approximation, has a compact form and is computationally cheap
- Nested sampling allows including the full range of primaries into the analysis

From me

- Cumulative distribution is a good format for reporting the composition analysis results, mitigating the problems with conveying correlated fractions

Paper 2: **Resolution of (Heavy) Primaries in Ultra High Energy Cosmic Rays**

Bortolato, B., Kamenik, J. F., & Tammaro, M. 2024 (arXiv), https://arxiv.org/abs/2409.06841

Outline

- The paper extends the previous analysis in two ways:
  - What happens if we include the transiron elements into the analysis?
  - What happens if we include only a subset of primaries?
- Both questions challenge the common assumption that the CR composition is well-represented by just a few primaries, often H, He , C/N/O, Fe

Data, simulations, and methodology

- Data – the same as before, but only in $E \in [0.65, 1]$ EeV (934 events)
- Simulations – generally the same as before, but
  - Only EPOS model
  - 10000 showers per primary
  - H - Fe, plus Z = 27, 28, 29, 30, 34, 39, 40, 42, 43, 44, 47, 50, 57, 58, 64, 72, 81, 82, 91, 92, 94
- Uncertainties are evaluated with bootstrapping, as before
- Additionally, the case of (projected) larger statistics is considered by reducing uncertainties by 1/f (f – statistical multiplier)
- n=3 first central moments are used, determined to be sufficient before
- The inference procedure is the same as before

Distance in the space of primaries

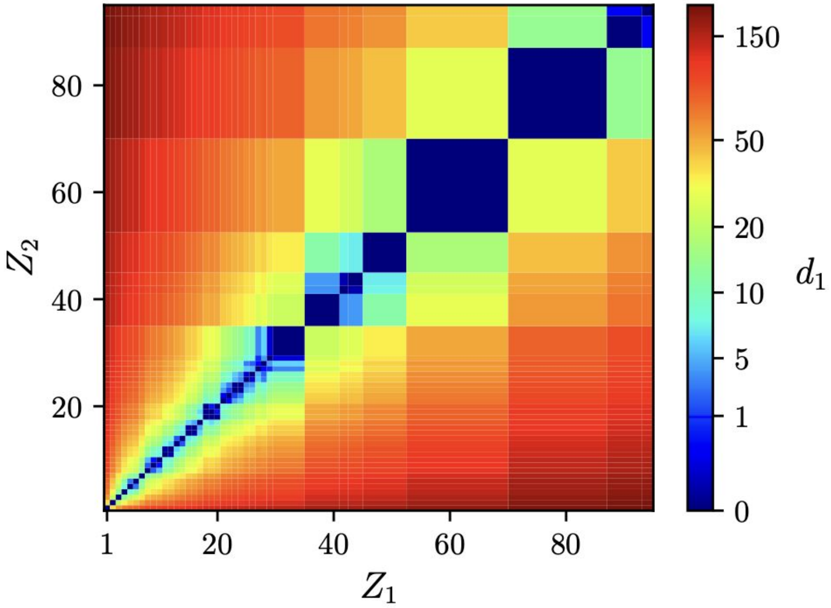
- In general, we would like to include in the fit only sufficiently distinguishable primaries
- To quantify the distinguishability, introduce the distance measure

$$d_n^2(Z_1, Z_2) \equiv (\mu_{Z_1} - \mu_{Z_2})^T (\Sigma_{Z_1} + \Sigma_{Z_2})^{-1} (\mu_{Z_1} - \mu_{Z_2})$$

- Here, μ and Σ are mean and covariance (n-vector and n by n matrix) of bootstrapped distribution of n first central moments of two primaries
- The measure has a straightforward interpretation of a χ^2 random variable
- For example, if for n=1 the distance between two primaries is 2.7, we can interpret this as a 90% confidence that they are distinguishable only by their first moment

Distance in the space of primaries

- This distance depends on the statistics – in the limit $N \to \infty$ all primaries would be distinguishable
- The authors use a conservative value for N equal to the measured dataset size
- Authors find that distances for $n>1$ are practically the same as for $n=1$, so they use it throughout



Choosing primaries wisely...

- The distance measure one to choose the primaries so that they are approximately equidistant from each other

d_0	N_L (N_H)	List of atomic numbers Z
16.6	4 (2)	1, 3, 10, 24, 52, 94
6.4	8 (5)	1, 2, 4, 6, 9, 13, 19, 27, 37, 50, 67, 89
4.0	12 (7)	1, 2, 3, 4, 5, 7, 9, 11, 14, 17, 21, 26, 31, 37, 44, 53, 63, 75, 89
2.8	16 (10)	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 15, 17, 20, 23, 26, 30, 34, 39, 44, 50, 57, 64, 72, 81, 91
2.0	20 (14)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 20, 22, 24, 26, 29, 32, 35, 38, 42, 46, 50, 54, 59, 64, 70, 76, 82, 89
1.3	24 (21)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 43, 46, 49, 52, 55, 58, 62, 66, 70, 74, 78, 83, 88, 93

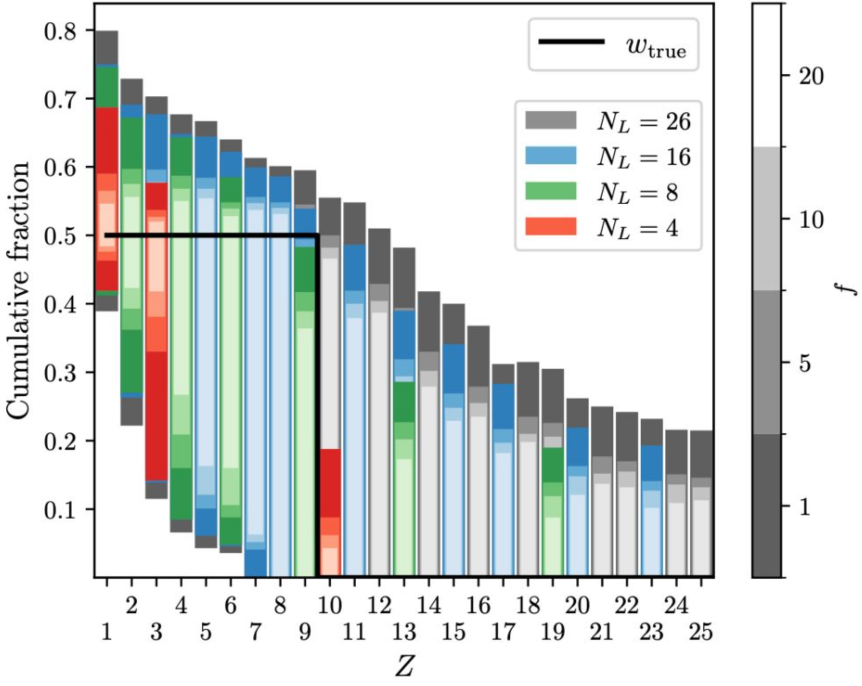
For reference, PAO analyses are:

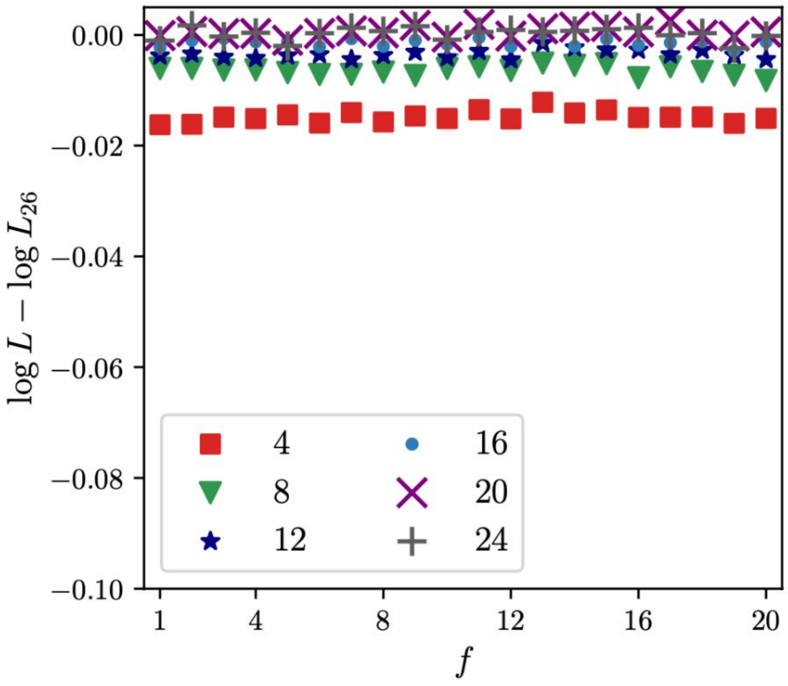
1 2 7 (14) 26

Inference with a subset of primaries

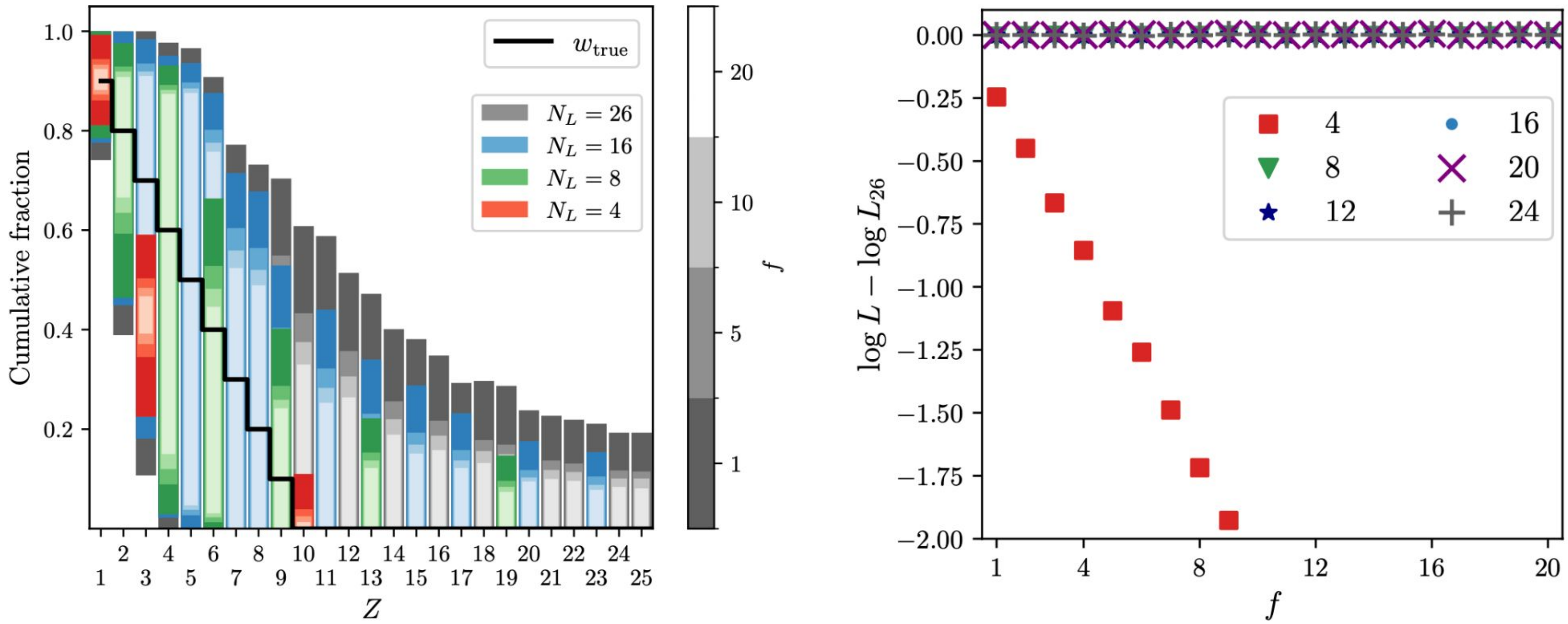
- To understand the influence of using the subset of primaries, the authors perform their analysis on mock data:
  - **Ex1**: 50% of protons (Z=1), 50% of neon nuclei (Z=10)
  - **Ex2**: 10% of each of Z=1, 2, ..., 10
- Analyses are performed with a different number of optimally chosen primaries N_L = 4, 8, 12, 16, 20, 24, 26
- Analyses are performed with different statistical multiplier f = 1, 5, 10, 20

Inference with a subset of primaries – **Ex1**

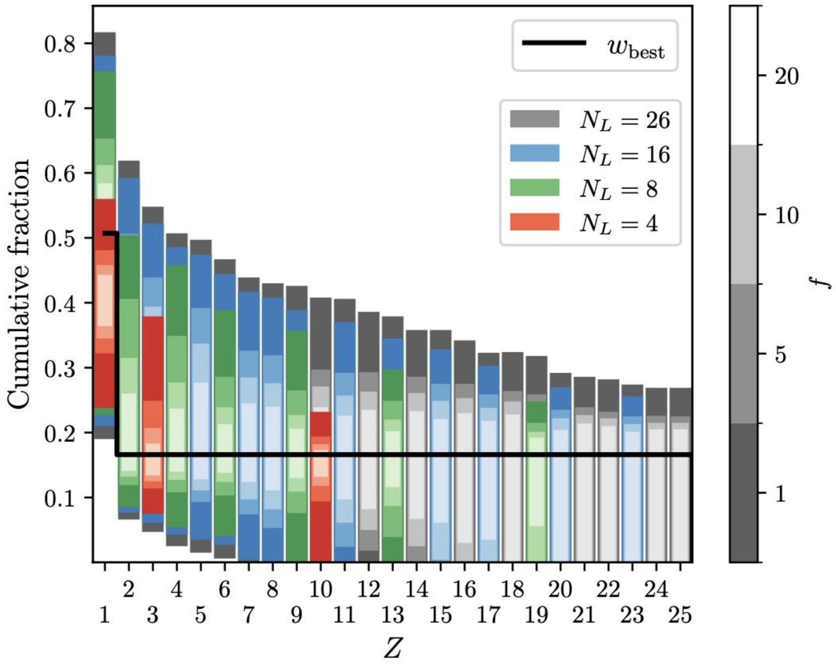




Inference with a subset of primaries – Ex2



Inference with a subset of primaries – PAOD fit

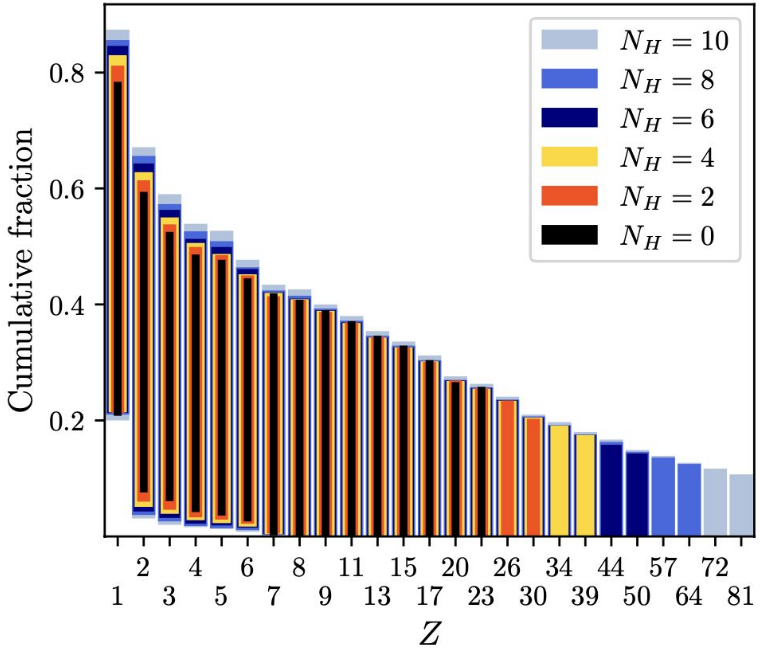


Inference with transiron primaries

- The inclusion of heavier nuclei follow the same equidistant prescription
- Confidence intervals are slightly enlarged for the light component
- Upper bound on transiron elements in the PAOD:

$$w(Z > 26, E \in [0.65, 1]\ \mathrm{EeV}) \leq 24\%,$$

$$w(Z > 26, E \in [1, 2]\ \mathrm{EeV}) \leq 18\%.$$



Conclusions

- A choice of primaries to be included in the composition fit imposes a strong prior on the results
- The strategy to select the most representative and distinguishable primaries is developed, based on the bootstrapped uncertainty and distribution similarity measure
- At least 16 primaries is needed to provide a diverse enough set
- When using lower numbers of primaries, biases arise (confidence intervals don’t count the true value), exacerbated with increasing statistics
- The method is trivially extendable to heavier primaries and yields upper limits for them

Additional slides

Notes from the point of view of Auger analysis

- Estimating PDF for measured X_{max}, the authors effectively smear it with detector resolution! And since the problem of the too narrow measured histograms is present in the PAO analysis, this can be a clue as to what we're doing wrong
- Bootstrapping procedure is taken by authors at face value, it would be interesting to validate its performance and/or compare with other methods