What slow down cosmic ray analysis and what can we do about them?

Several Practical Considerations in Cosmic Ray Analysis & Machine Learning

Xinhua Bai

South Dakota School of Mines and Technology

Acknowledgement: U.S. National Science Foundation-EPSCoR (RII Track-2 FEC, award #2019597)

Outline

1. Discovering and understanding surprises in MC and data.
2. Fluctuations and training sample size
3. Phase space
4. Uncertainties
5. Using data of all kinds
6. Data, data transparency and knowledge transparency
7. Accuracy limit

More about questions rather than solutions

1. Discovering and understanding surprises

To what extent can we claim the NN selection and training is sufficient?

Surprises in:

- Human ignorance
  - Hardware
  - Software
  - Calibration
- Human errors
  - Hardware
  - Software
  - Calibration
- Physics

Examples:

- Shower core:
  - Mis-reconstruction of those landing on the edge of the surface array
  - Double core events
- Shower size:
  - Trigger and event builder
  - Saturation in detectors
- Ground particle signals in surface array – often hard to simulate precisely:
  - Fluctuations in different stages of the detection associated with different physics processes
- Physics: may beyond reliable simulation or physics we know

Note: Some example simulation plots are available from presenter per request.

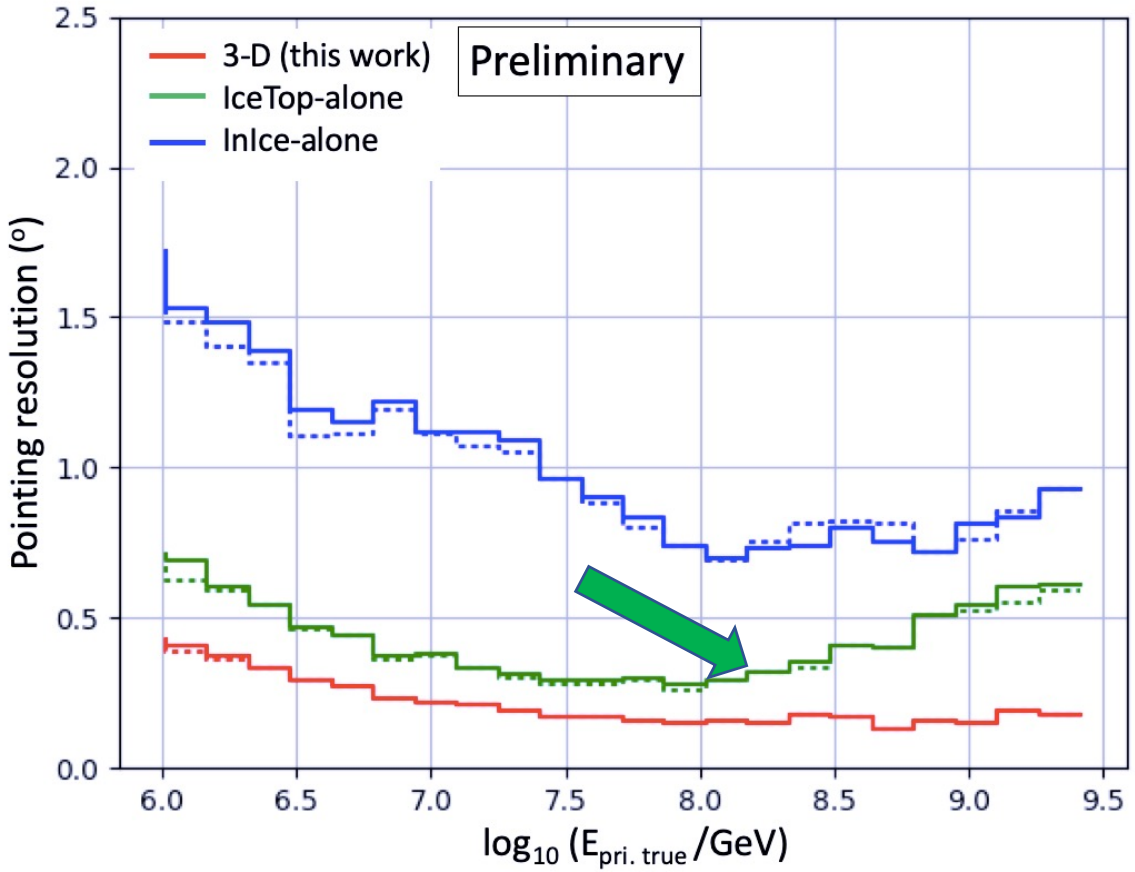
2. Fluctuations and training sample size

An old long-standing problem: The deterioration of angular resolution at higher energies.

Example: IceTop official/released reconstruction:

$$\Delta t(R) = C_2 R^2 + C_1(1 - exp(-\frac{R^2}{2\sigma^2}))$$

$$\sigma_t(R) = a + bR^2$$

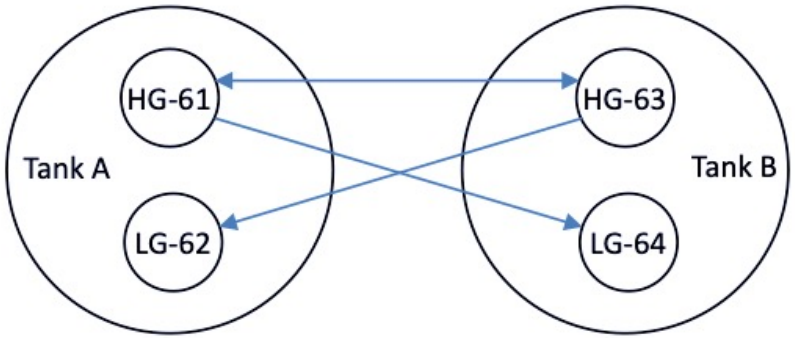


PoS ICRC2019 (2020) 244

2. Fluctuations and training sample size – cnt.

One solution: Flexible curvature ⊕ “proper” timing fluctuation ***for each EAS***

Figure: NIM A700:188,2013



$$\sigma_{ti} = C\frac{\sqrt{\Sigma_{j=1}^{2}\left(t_{ij} - \left(\frac{t_{i1}+t_{i2}}{2}\right)\right)^2}}{\left(\Sigma_{j=1}^{2} Q_{ij}\right)^{a}} + b$$

Events reconstructed successfully:
Laputop: 186490
Flexible curvature & old fluc. 186491
Flexible curvature & this fluc. 186481

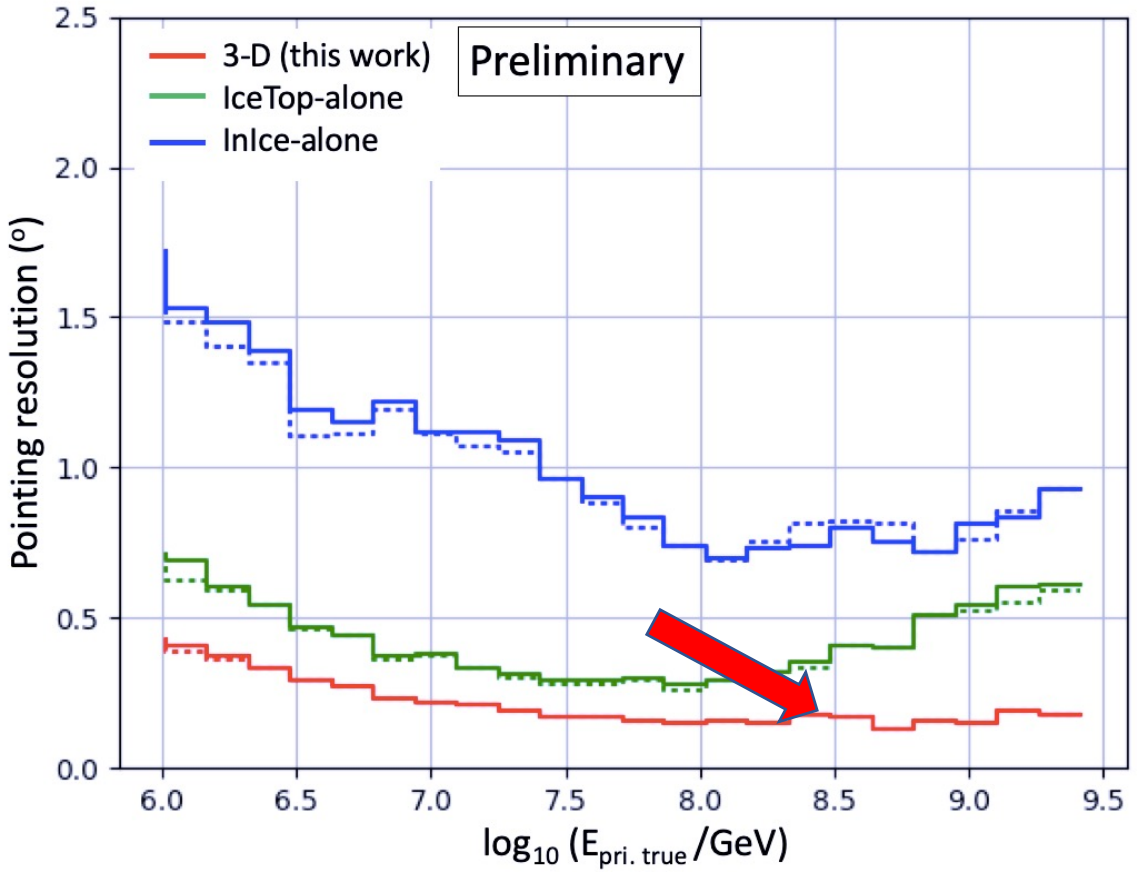


Figure: PoS ICRC2019 (2020) 244

2. Fluctuations and training sample size – cnt.

What's the optimal size of the training sample?

Different impacts related with:

- True/physical correlations and fluctuations in training sample.
- Information in individual showers **vs** Information in large data/training samples
- How they are processed by different architectures, parameters, algorithms in ML

Some key factors:

- Fluctuations in training samples:
  - Showers of different primary particles
  - Showers with different primary energies
  - Showers at different zenith angles
  - Timing of detected particles → direction
  - Size of measured pulses → core & energy
- They are different and have different impact on CR reconstruction accuracy
- They make the separation between "correlations" and "causality" uneasy.

Could the unsupervised ML be the ultimate solution?

3. Phase space

Is it possible to cover sufficient phase space from available/limited phase space with ML? How? To what accuracy?

- Primary energy (Figure)
- Primary composition
- Zenith angle
- Interaction models

➢ more hybrid detections

➢ time dependent variables (ex. snow accumulation, seasonal variations of atmosphere, noises, etc.)

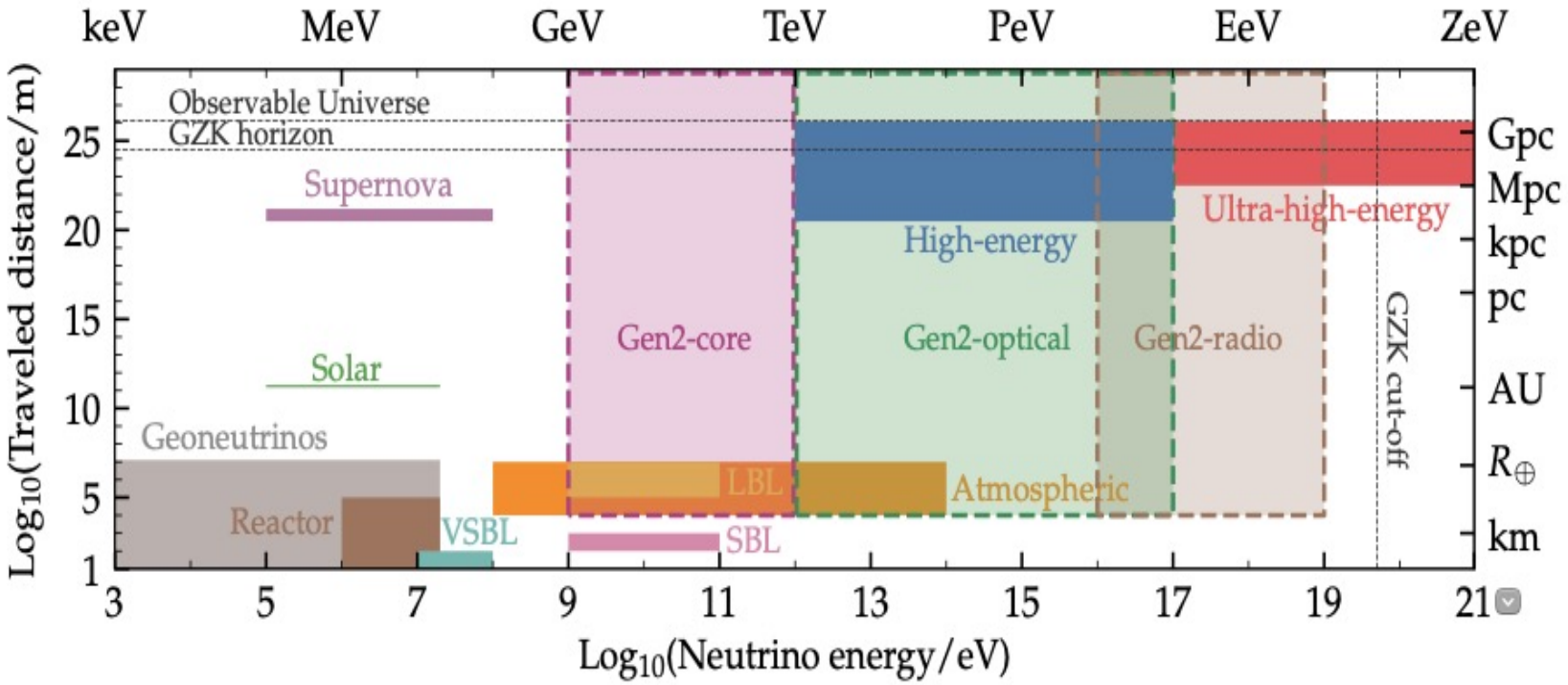


Figure: J. Phys. G 48 (2021) 6, 060501

4. Using data of all kinds

Is this possible to cover sufficient phase space by using existing data of all kinds to save time and Machine Learning somehow?

1) Variations from sub datasets and/or data covering sub phase space
   - How? How much can we gain?
2) Bridging MC data produced for different purposes and/or by different projects
   - How? How much can we gain?

5. Uncertainties

Quantitative uncertainties help us:
(1) understand the cause of weaknesses in experiments and modeling
(2) find better solutions or develop better techniques.

Statistical uncertainty ⊕ Systematic uncertainties of all kinds ⊕

Training uncertainty ⊕

ML-intrinsic uncertainty (architectures, parameters, algorithms, etc.) ⊕ …

→ Uncertainty in ML-outcomes

Fundamental questions:

- Is it necessary to distinguish errors of all kinds/from different sources?
  - Yes. How to measure and report them?
  - No. How to use ML to target the weaknesses and improve them?

6. Data transparency and knowledge transparency

How much we can achieve true ***knowledge transparency*** by using advanced data techniques?

- **Reality**: Data > Transparent data >> Transparent knowledge

  To benefit young generations at all levels everywhere,

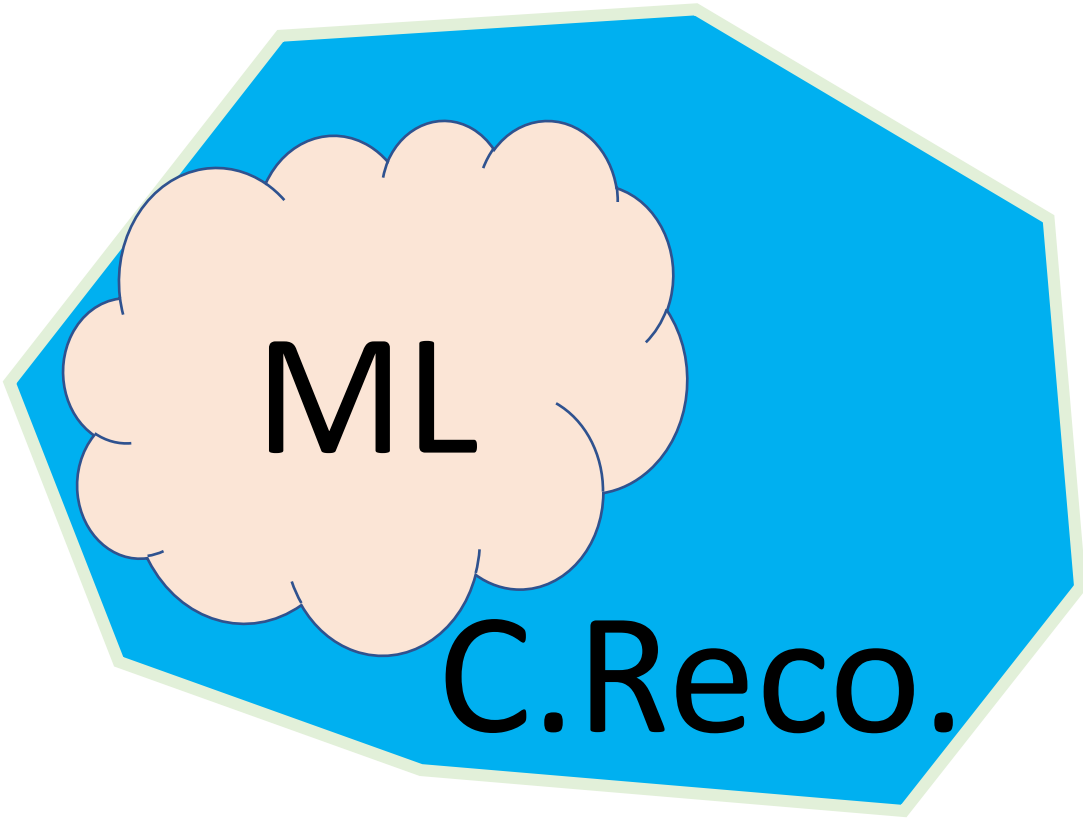
- **We need**: Data ≈ Transparent data ≈ Transparent knowledge

- Biggest difficulty: Most current data are not produced / organized / published for the purpose of knowledge transparency for learners.
- I don't see an easy way of doing this. But it may become a major demand in the future.

7. Accuracy limit

Where are the accuracy limits and what are the causes of them in ML?

- We have a good understanding and description of the accuracy limits in conventional/analytical reconstruction and analysis
  - Statistics, error analysis theory, …
- How ML and conventional reconstruction ***compensate*** each other and ***improve*** CR analysis accuracy depends on:
  - how well we know where the ML accuracy (theoretical) limits are.
  - how to form a “common ground” for users given the limitations in ML? **(next slide)**



One that works: More practice and sharing – This workshop👏

7. Accuracy limit – cnt.

What do major limitations of machine learning mean for CR analysis?

Peter Voss, 2016:

1) **Every narrow application requires special training**
2) **Need a lot handmade structured training data**
3) It is usually necessary to supervise learning: training data must be marked
4) Requires lengthy offline/batch training
5) Don't learn incrementally or interactively in real time
6) **Poor transfer learning, module reusability and integration**
7) **The system is opaque**, making them difficult to debug
8) **"Long tail" cannot be reviewed or guaranteed performance**

Monte Carlo ≠ Reality

Your results, my results ≠ Physics results unless there exists a consensus about:

(i) choices in utilizing NN,
(ii) confidence level of ML outcomes,
(iii) …

7. Accuracy limit – cnt.

What do major limitations of machine learning mean in CR analysis?

Peter Voss, 2016 – cnt.:

9) **They encode correlations, not causal or ontological relationships**
10) Do not encode entities or spatial relationships between entities
11) Handle only very narrow aspects of natural language
12) **Not suitable for advanced, symbolic reasoning or planning**

My biggest concern: Better prediction (with a large amount of parameters) ≠ Better (human) understanding

Summary

1. There are plenty topics to study, some are fundamental.
2. There are some details to consider in order to make good us of the powerful tool ML. Some are tricky to answer: ex. fluctuations, uncertainties.
3. ML has limitations that also matter to physicists.
4. Conventional reconstruction/analysis should still be emphasized.
5. Comparison between results from ML and conventional reconstruction/analysis is valuable.
   - Especially for problems in which both **correlation and causality** matter.
6. To make knowledge more transparent should be a goal in the era of Big Data.
7. Pay attention to new development in ML.