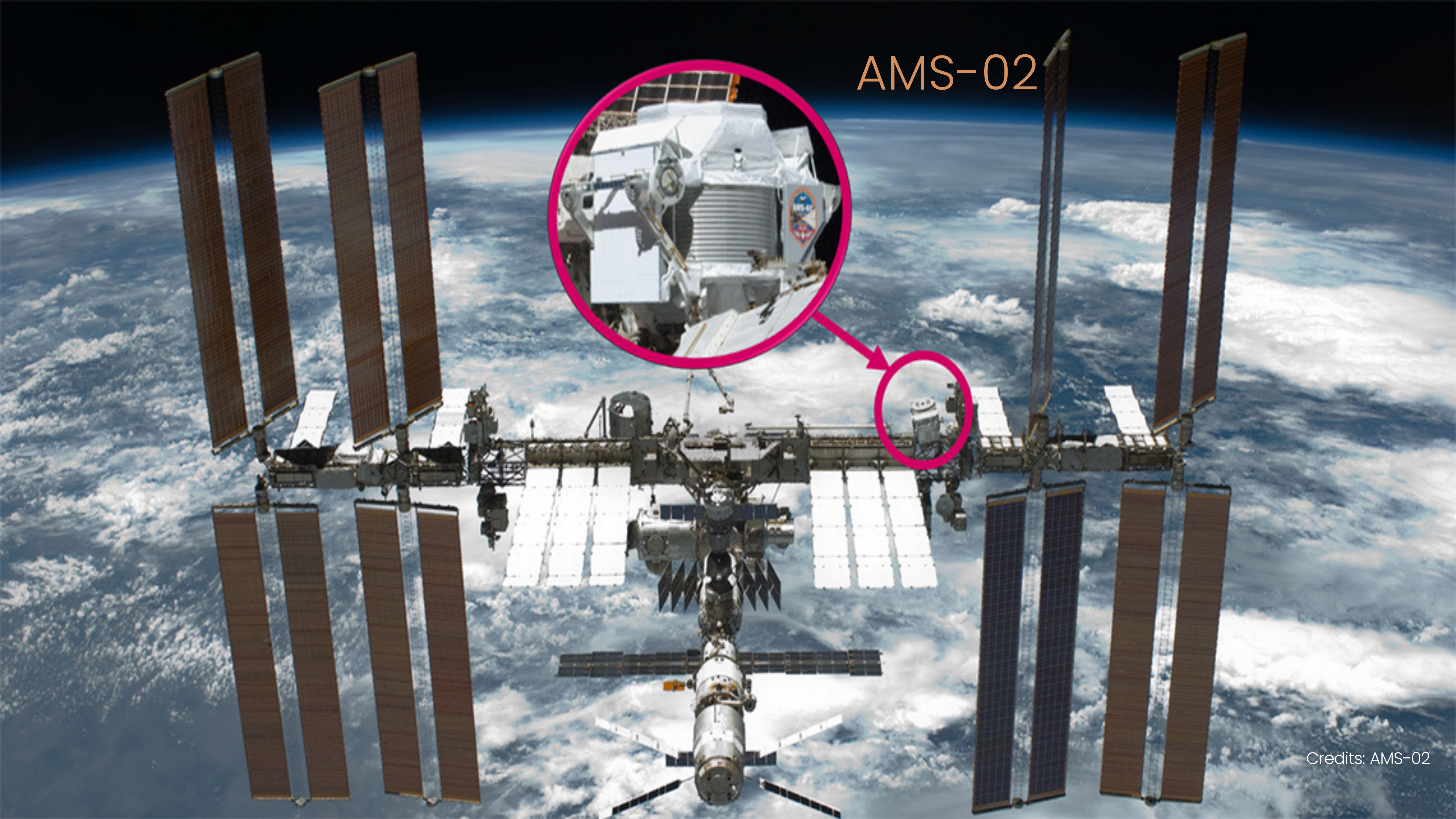
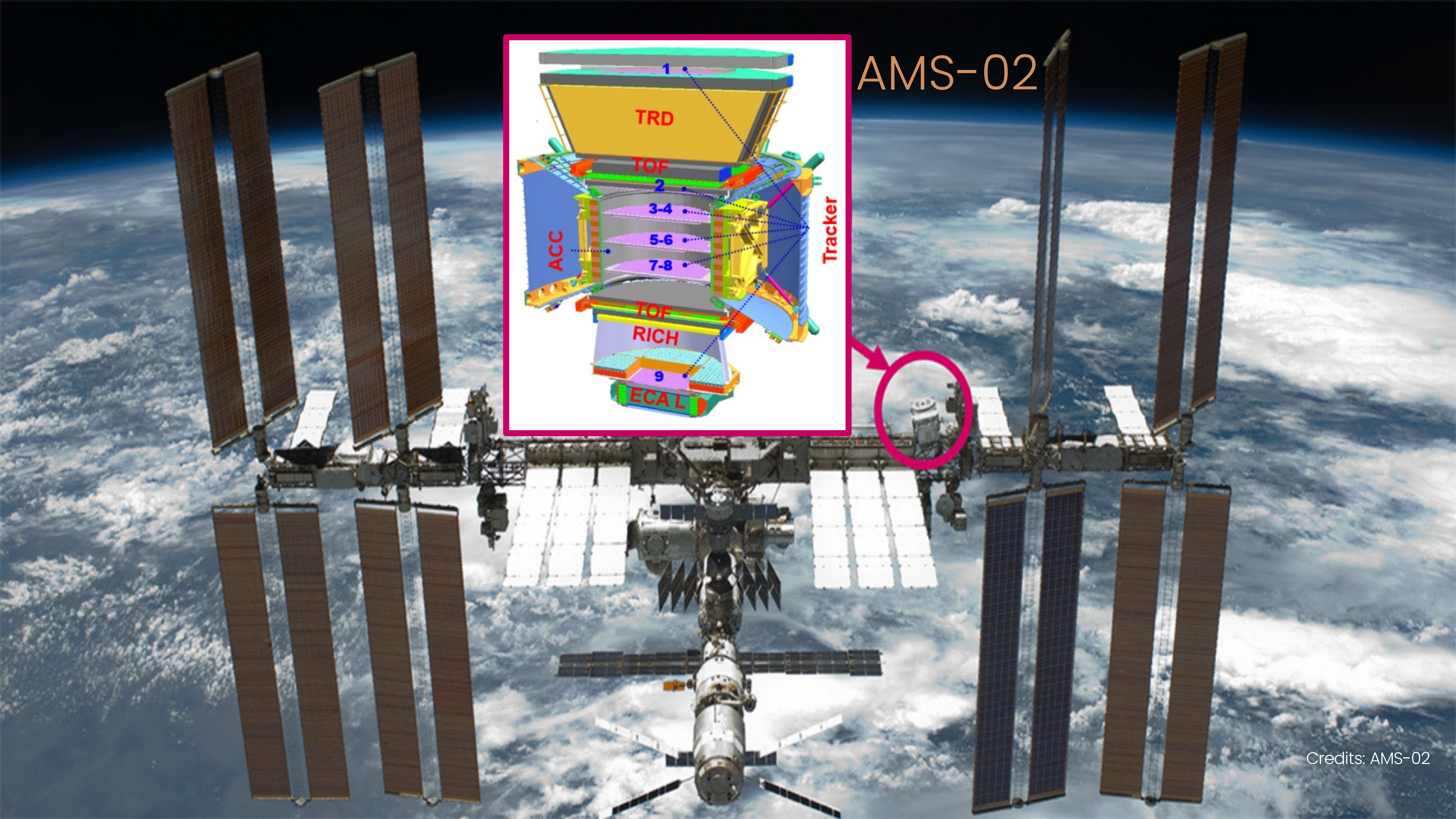
Feature selection techniques for CR isotope identification with the AMS-02 experiment in space.

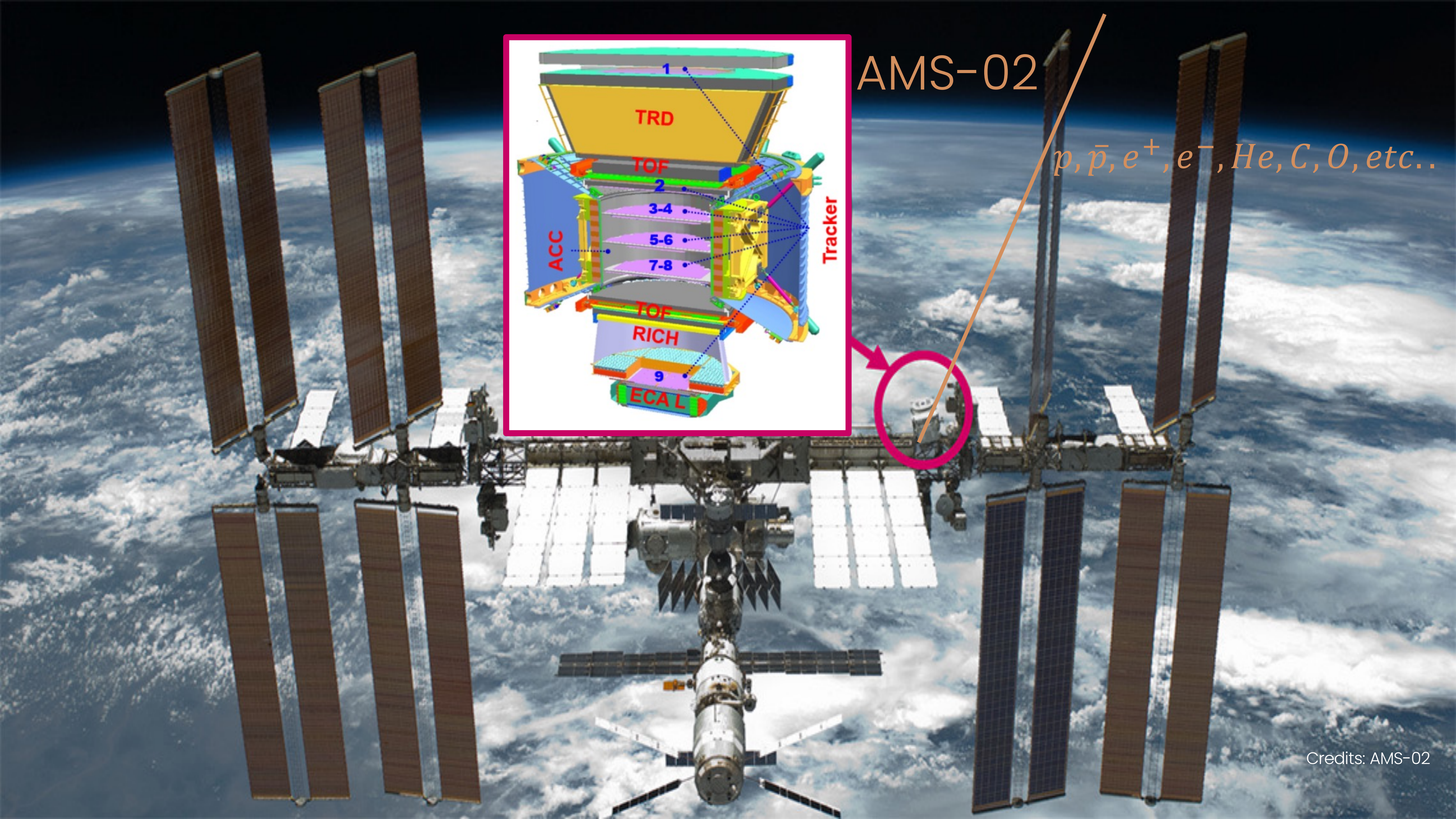
Marta Borchiellini

Kapteyn Astronomical Institute, RUG

How to improve isotope identification with AMS-02 using Machine Learning feture selection methods?


AMS-02
Credits: AMS-02


AMS-02
TRD
TOF
1
2
3-4
5-6
7-8
9
ACC
Tracker
TOF
RICH
ECAL
Credits: AMS-02


AMS-02
$p, \bar{p}, e^+, e^-, He, C, O, etc..$
TRD
TOF
1
2
3-4
5-6
7-8
9
ACC
Tracker
TOF
RICH
ECAL
Credits: AMS-02

RICH Background rejection

INPUT DATASET

- Events labelled on the basis of beta reconstruction
- 130 features linked to RICH

→

BUILDING THE ESTIMATOR

Training (BDT)
+
Evaluating the performance of the BDT

RICH Background rejection

INPUT DATASET

- Events labelled on the basis of beta reconstruction
- Variables linked to RICH

→

FEATURE SELECTION

ML Learning

vs

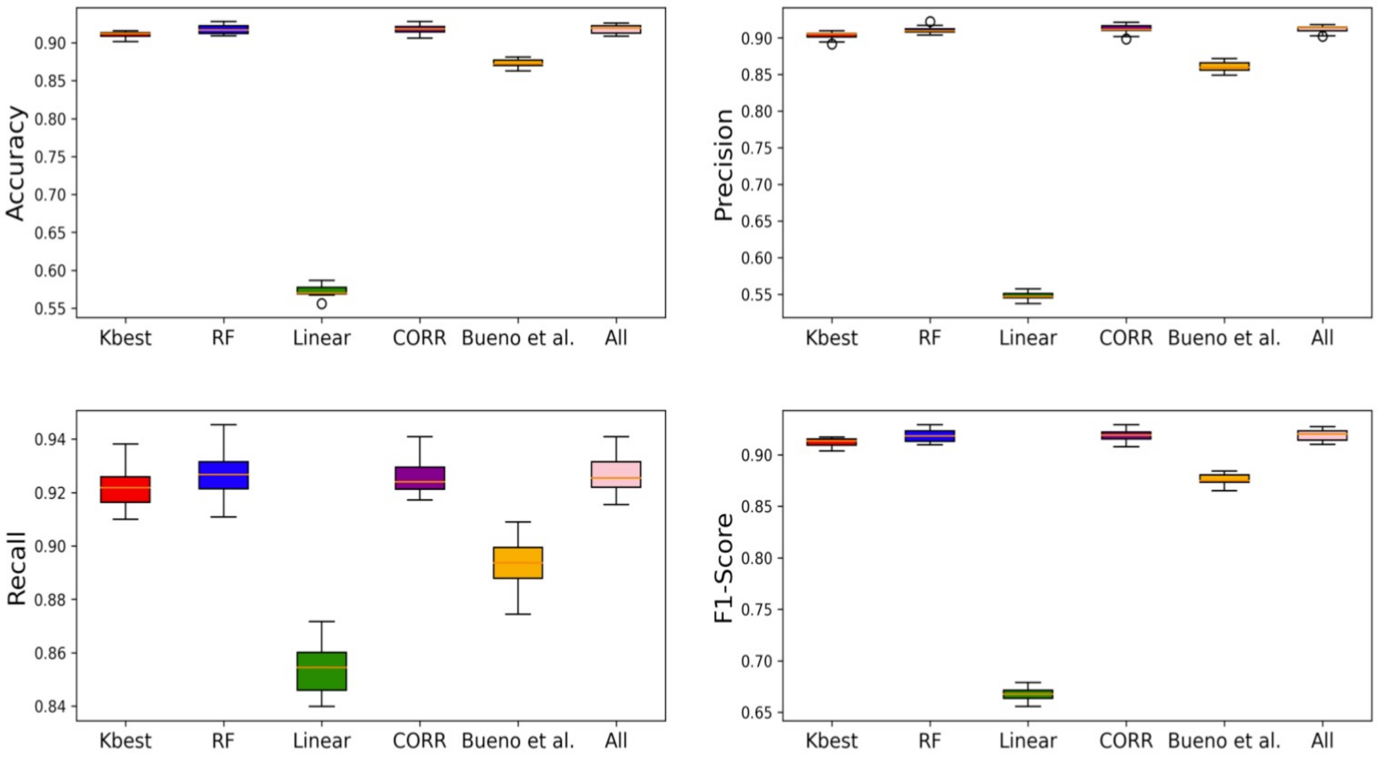
Physics driven

→

BUILDING THE ESTIMATOR

Training (BDT)
+
Evaluating the performance of the BDT

Results



Borchiellini et al., Particles 2024, 7(2), 417-434

- Almost every ML algorithm performs better than the physics-driven method (Bueno et al.)
- Random Forest allows for 90% background rejection and 92% signal efficiency
- With the Random Forest technique is possible to achieve similar outcomes compared to All while reducing overfitting risks and slightly decreasing training time

Thank you!

If you want to know more:

M. Borchiellini, L. Mano, F. Barão, M.Vecchi. 2024. "Feature Selection Techniques for CR Isotope Identification with the AMS-02 Experiment in Space», Particles 7, no. 2: 417-434. https://doi.org/10.3390/particles7020024

Or come to the poster session tomorrow!

interTwin

interTwin: An interdisciplinary Digital Twin Engine for Science

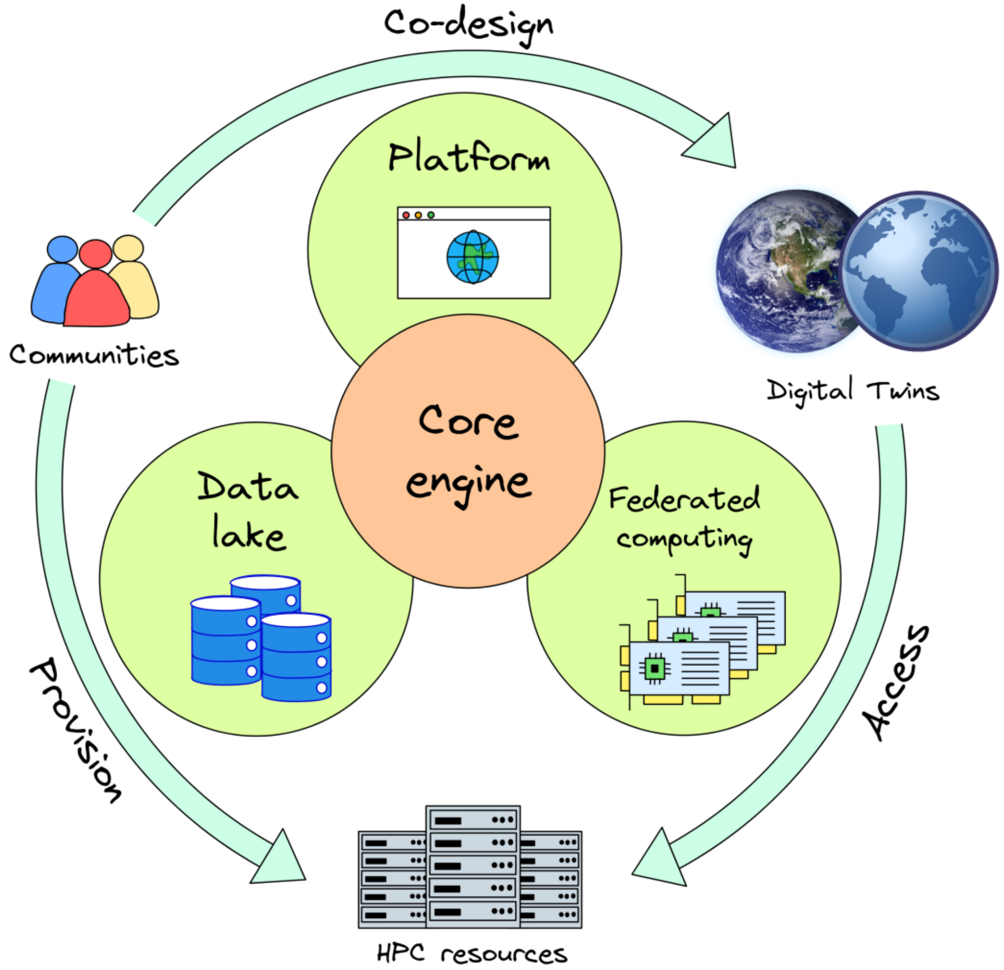
European AI for Fundamental Physics Conference 2024, Amsterdam, 30.04.2024

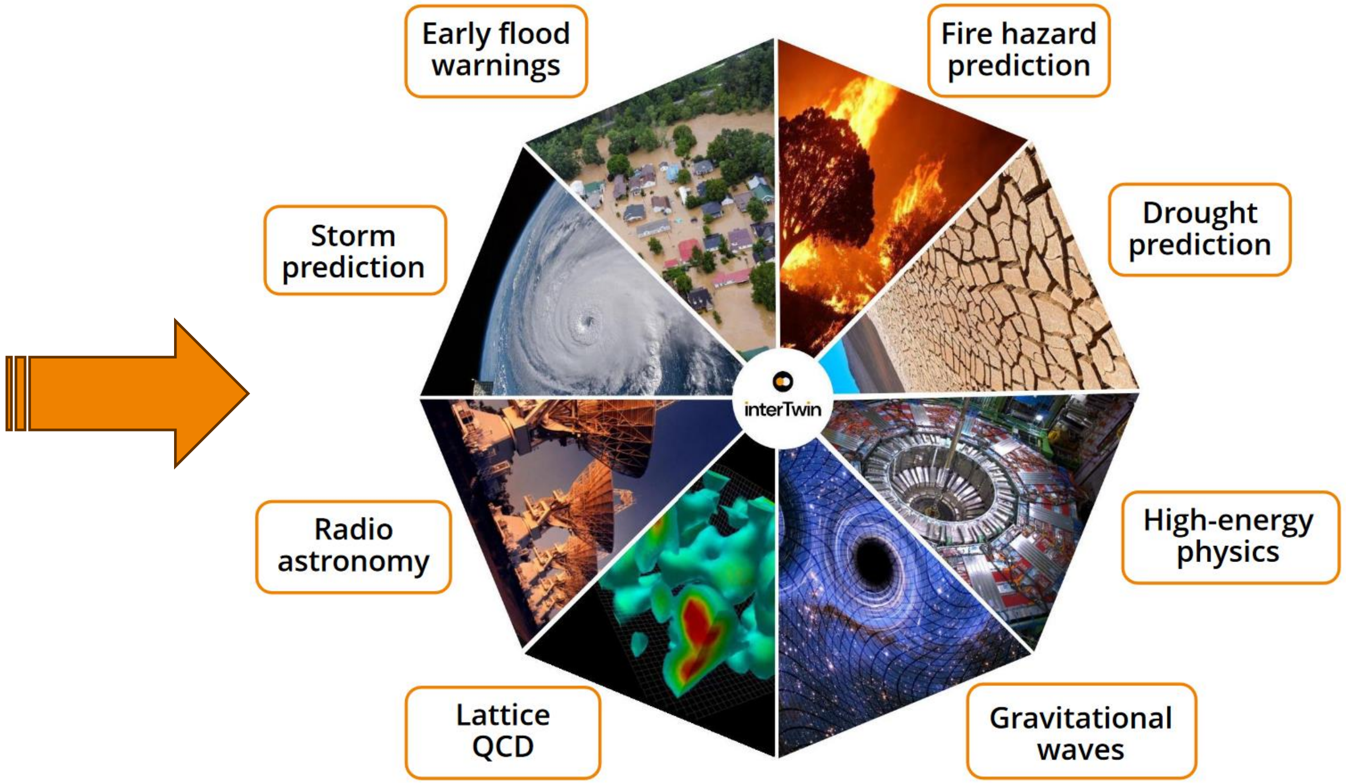
Kalliopi Tsolaki, Matteo Bunino, Alexander Zoechbauer, Ilaria Luise, Maria Girone, Sofia Vallecorsa, David Rousseau, Alberto Di Meglio, CERN-IT & CNRS/IN2P3 on behalf of interTwin consortium

Funded by the European Union

The interTwin project is funded by the European Union - Grant Agreement Number 101058386

interTwin project: digital twin engine for science

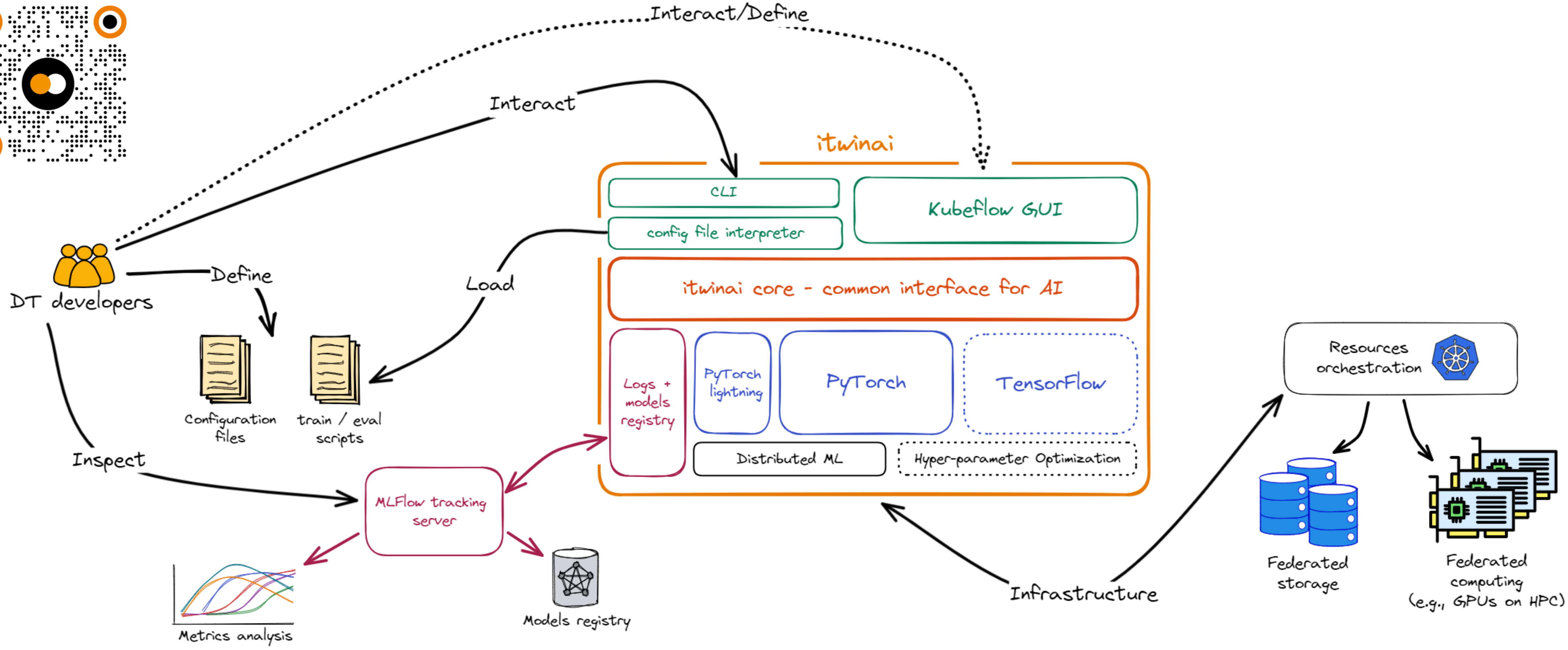




https://www.intertwin.eu/

itwinai: an interTwin module for AI workflows

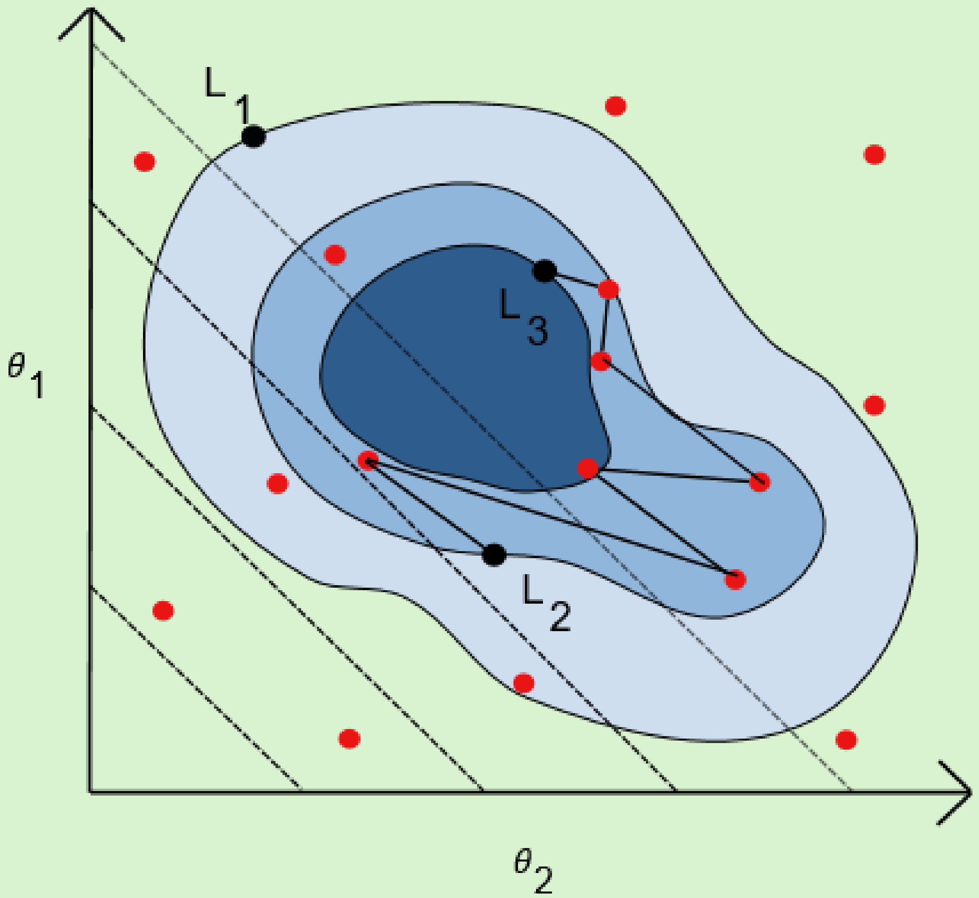
Reproducibility, Reusability, Framework-independent, Scalability, Access to Cloud and HPC

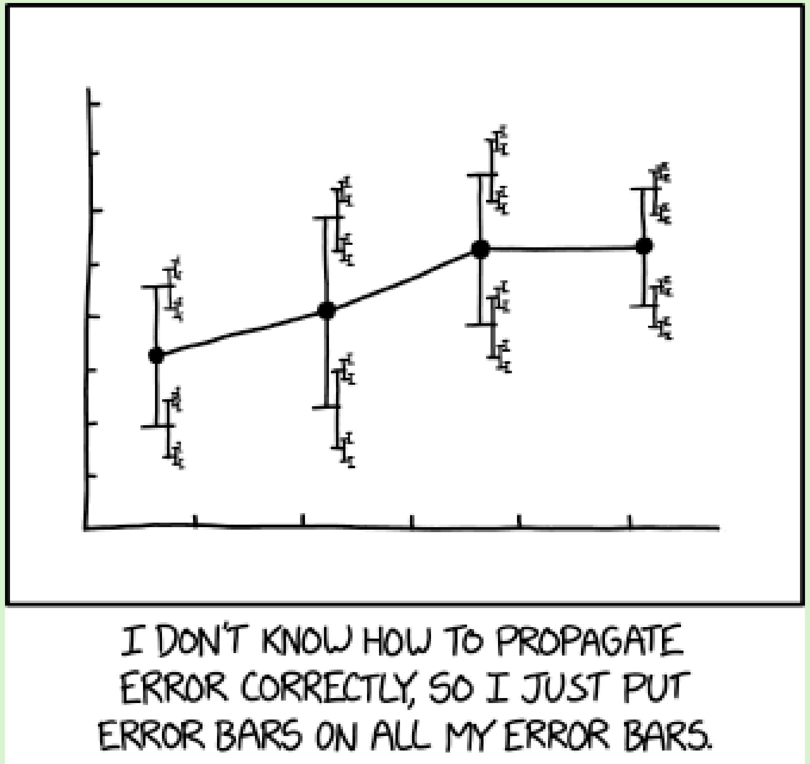


https://www.intertwin.eu/

Costless correction of nested sampling parameter estimation

Metha Prathaban



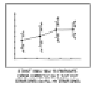


[https://xkcd.com/2110/]

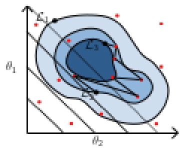
Phantom points in nested sampling parameter estimation

Metha Prathaban <myp23@cam.ac.uk>
Kavli Institute for Cosmology · Cavendish Laboratory · University of Cambridge

Nested sampling parameter estimation differs from evidence estimation, in that it incurs an additional source of error. This error affects estimates of parameter means and credible intervals in gravitational wave analyses and beyond, and yet, it is typically not accounted for in standard error estimation methods. We present two novel methods to quantify this error more accurately for any chain-based nested sampler, using the additional likelihood calls made at runtime in producing independent samples. Using injected signals of black hole binary coalescences as an example, we demonstrate how these extra points may be carefully utilised to estimate the true error correctly, and provide a way to check the accuracy of the resulting error bars.



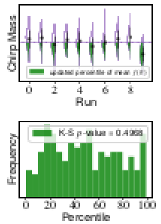
1. Nested sampling and phantom points



Nested sampling (NS) is a popular Bayesian inference tool for parameter estimation and model comparison. A set of live points is drawn from the prior and at each iteration, the live point with the lowest likelihood is deleted. A new point is drawn, with the constraint that its likelihood must be higher than that of the deleted point [1, 2]. In this way, a series of nested iso-likelihood contours are defined [3].

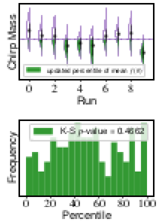
There are many way to generate a new live point with the hard likelihood constrain $\mathcal{L} > \mathcal{L}_i$. However, many NS implementations use a Markov-Chain based procedure, where new points are continually generated within the likelihood contour until we are satisfied that the new point is independent from the deleted point. This point is then assigned as the new live point, and the points generated in the chain between the deleted and new live point (red) are typically discarded. These '**phantom points**', though deemed too correlated to use in evidence estimation, can provide useful information about the parameter space, though this has been largely unexplored.

3. Likelihood binning method



- Bin phantom points by their likelihood values, such that each dead point is associated with a set of phantom points from the run which sit very close to the contour defined by it.
- We make the assumption that, though phantom points do not lie exactly on the dead point's iso-likelihood contour, they are still representative of the $f(\theta)$ values along the contour.
- For each dead point in equation 4, resample a new $f(\theta)$ value from the associated bin (which includes the original dead point itself) and a new ΔX.
- Repeat many times to obtain the error as the standard deviation of the resulting distribution of estimates.

Reconstructed runs method



- All phantom points are perfectly valid, except that they may be too correlated with their associated **dead** point to use both in the same run.
- We can therefore take the 1[st] phantom point in every chain in the run and combine these carefully to form an equally valid nested sampling run to the original.
- Repeat with other phantom points to reconstruct multiple valid runs from original.
- Combine the parameter estimates from each of these reconstructed runs, as well as the original run, and compute the corresponding new error bar from this.

zenodo repository with the source code and pdf for this poster.

References

[1] W. J. Handley, M. P. Hobson and A. N. Lasenby. polychord: [illegible]. Monthly Notices of the Royal Astronomical Society, [illegible]
[2] W. J. Handley, M. P. Hobson, and A. N. Lasenby. [illegible] nested sampling for cosmology. Monthly Notices of the Royal Astronomical Society: Letters, [illegible]
[3] John Skilling. Nested sampling for general Bayesian computation. Bayesian Analysis, [illegible]
[4] [illegible]
[5] [illegible]
[6] [illegible]
[7] [illegible]

[illegible]

Errors in evidence vs parameter estimation

Evidence estimation

From Bayes' theorem,

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \quad (1)$$

Parameter estimation

The expected value of $f(\theta)$ is [4]

$$E[f(\theta)] = \int f(\theta)\frac{\mathcal{L}(\theta)\pi(\theta)}{\mathcal{Z}}d\theta \quad (2)$$

Changing the integration variable to the fractional prior volume within an iso-likelihood contour, X, and approximating this as a sum over the dead points [3,5]:

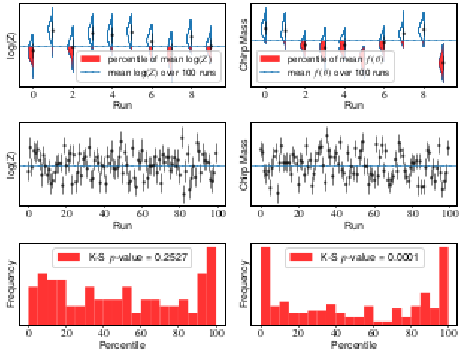
$$\int_0^1 \mathcal{L}(X)dX \approx \sum_{i \in \text{dead points}} \mathcal{L}_i \Delta X_i. \quad (3)$$

By construction, all points along a given contour have the same likelihood value; using the likelihood value of a single dead point, $\mathcal{L}_i$, as a proxy for $\mathcal{L}(X)$ is an exact substitution. The dominant error is the unknown volumes of the fractional prior volume 'shells' between contours, ΔX.

$$\frac{1}{\mathcal{Z}}\int \tilde{f}(X)\mathcal{L}(X)dX \approx \frac{1}{\mathcal{Z}}\sum_i \tilde{f}(X_i)\mathcal{L}_i\Delta X_i. \quad (4)$$

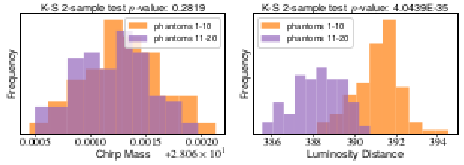
We are required to use $f(\theta_i)$ as a proxy for $\tilde{f}(X_i)$. This is **not** an exact substituion, and here this becomes the dominant error. Consider the example of estimating $\theta_1+\theta_2$ from the figure to the left; contours of constant parameter values are shown in dashed lines.

For a binary black hole, we can apply the 'simulated weights' method [6, 7], suggested by Skilling, to estimate the errors on the evidence and chirp mass per run. If this is sufficient, the percentiles of the 'true evidence' and 'true chirp mass' (estimated over 100 runs) should be uniformly distributed.



Verifying accuracy of error bars

For certain parameters, the chain length of the sampler may not be long enough to accumulate sufficient uncorrelated phantom points to use these methods. We can check from a single run whether we have the correct error bars by splitting the phantom points in two halves, and applying either method separately to the two sets of points to check for convergent results using the K-S 2-sample test. The default chain length in PolyChord [1,2] is long enough for these methods to work well on the chirp mass, but not the luminosity distance:



We can apply something similar at runtime to reduce the sampling time by calculating the optimal chain length.

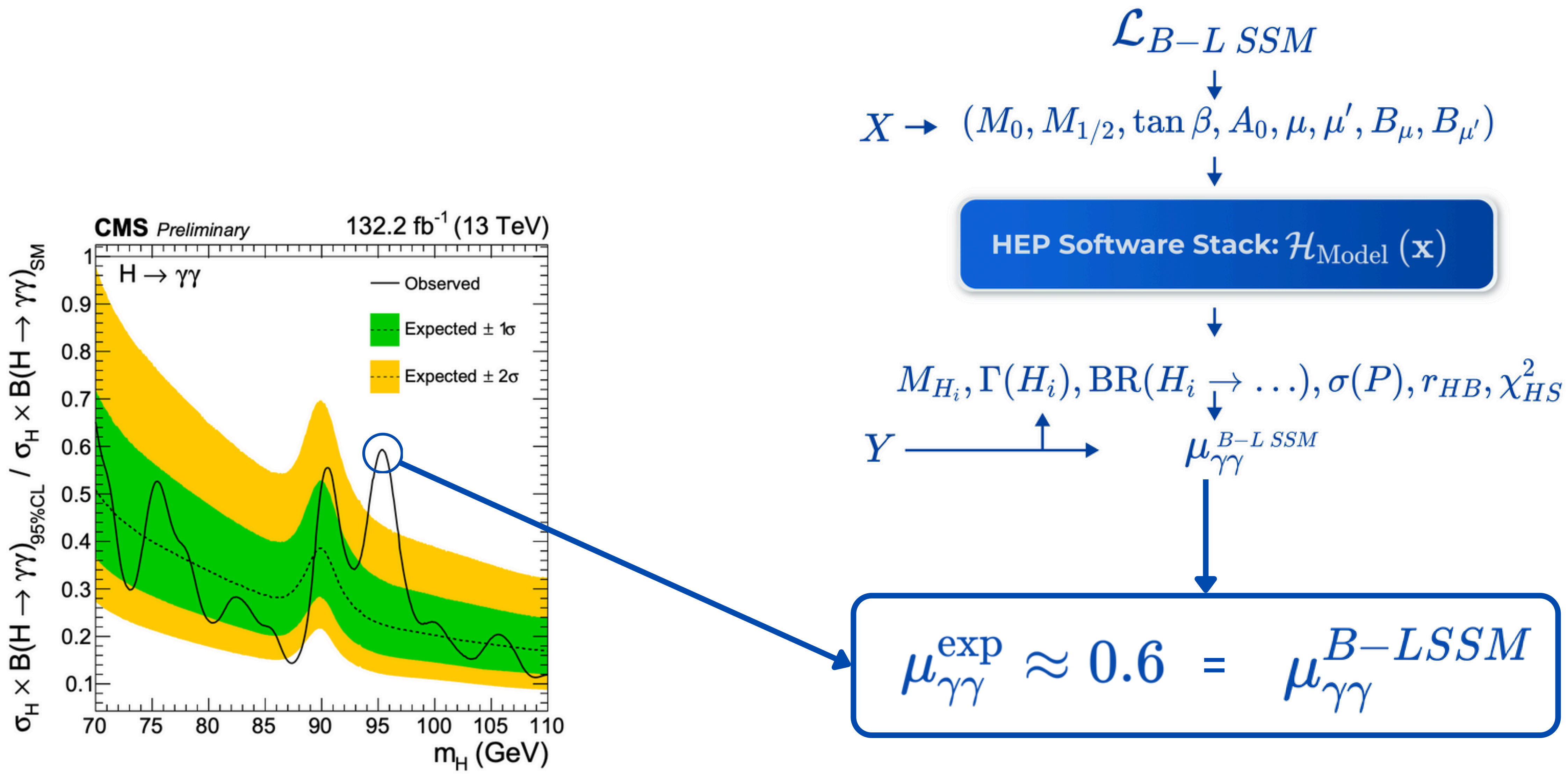
Efficient Parameter Space Exploration in BSM Theories
with Batched Multi-Objective Constraint Active Search

Physics and Astronomy
Electronics and Computer Science
University of Southampton

Mauricio A. Diaz, *Giorgio Cerro, Stefano Moretti, Srinandan Dasmahapatra.*

Several hints of new physics exist, and more are emerging:

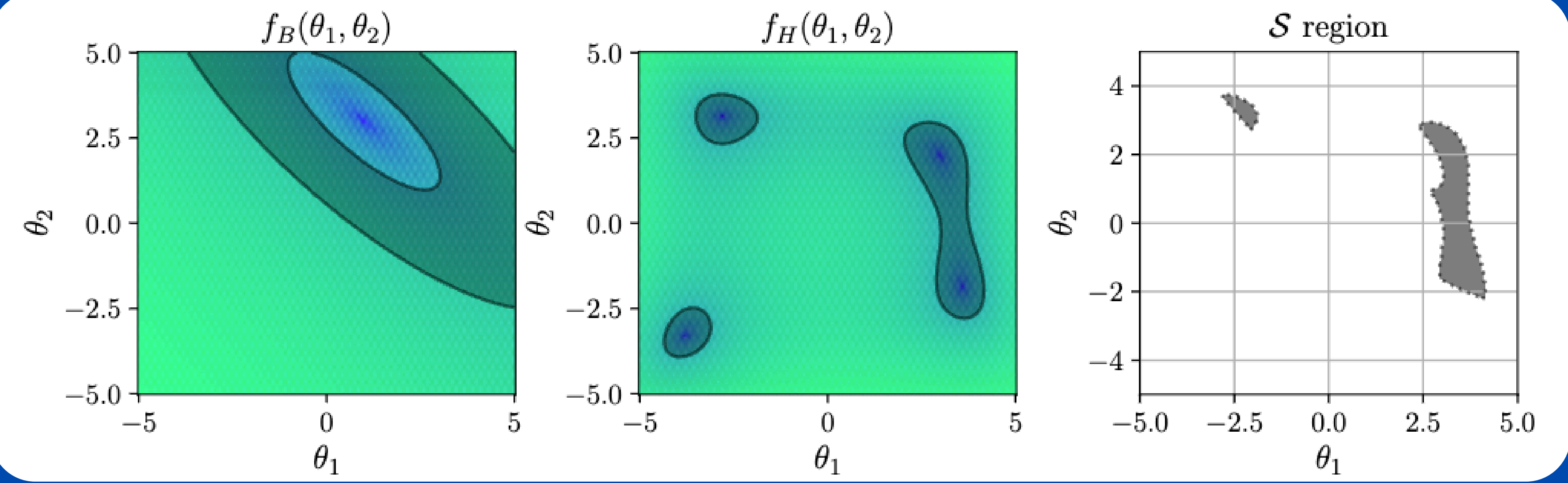
- **Neutral Scalars**
- Flavour anomalies
- Neutrino masses
- Dark matter



PARAMETER SPACE SCANS

Parameter scan methods aim to identify a set of points that belong to a rare category defined by constraints

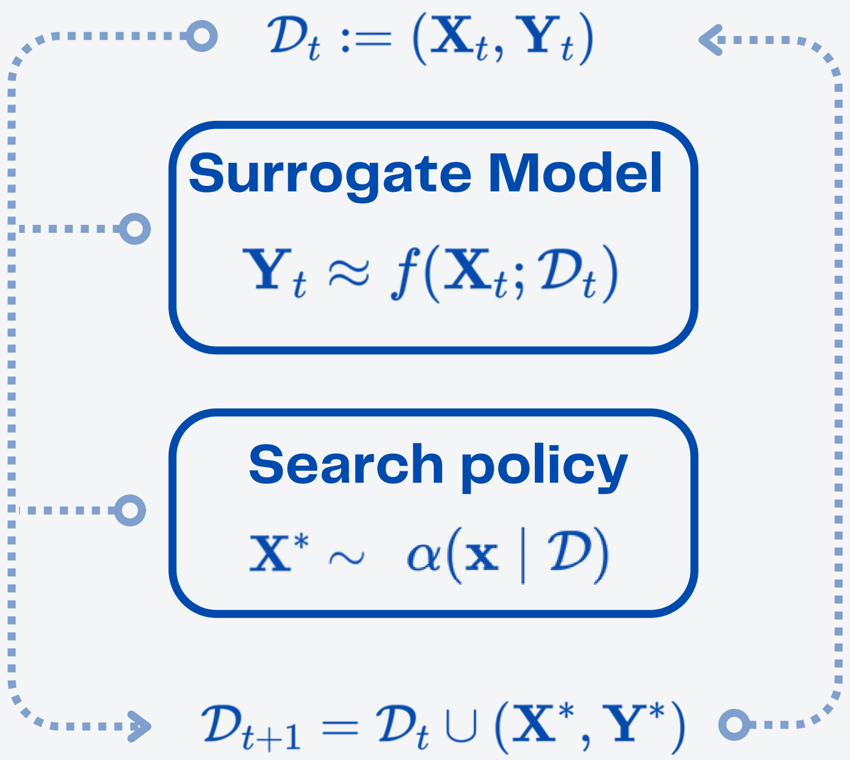
$$\mathcal{S} = \{\mathbf{x} \mid \mathbf{y} = \mathcal{H}_{\text{Model}}(\mathbf{x}) \wedge y_i \in \tau_i\}$$



- **Full $\mathcal{S}$ characterisation**
- **Diverse and dense filling**

- $\mathcal{S}$ region might be sparse and disconected
- $\mathcal{H}_{\text{Model}}(\mathbf{x})$ is expensive to evaluate

Mauricio A. Diaz

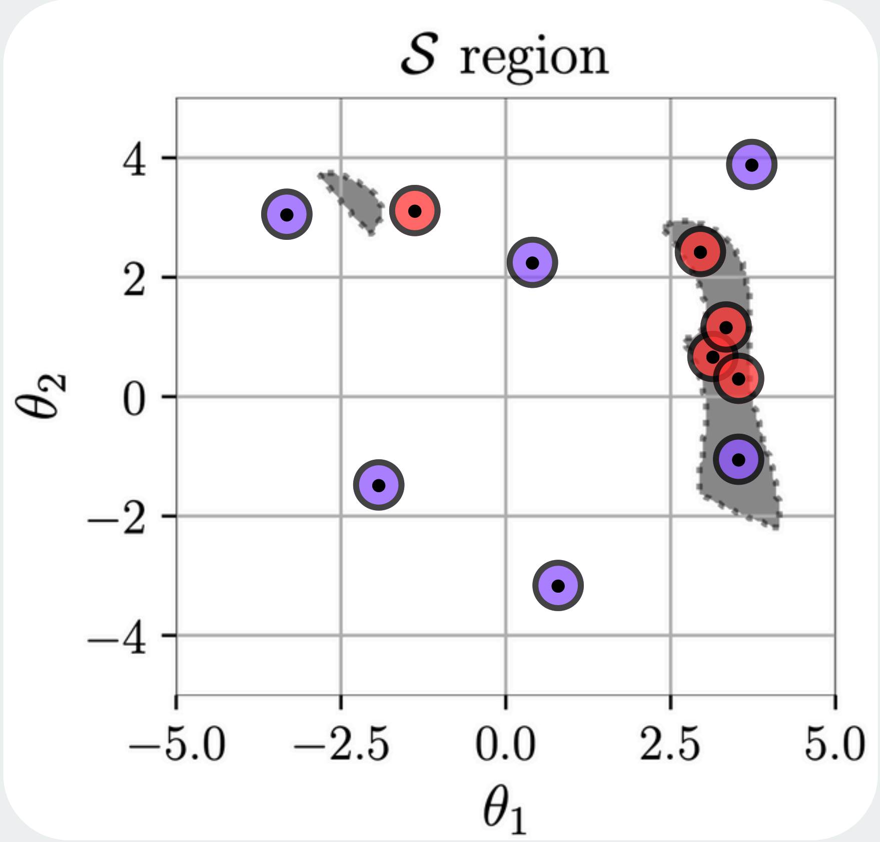


Active Search Formulation

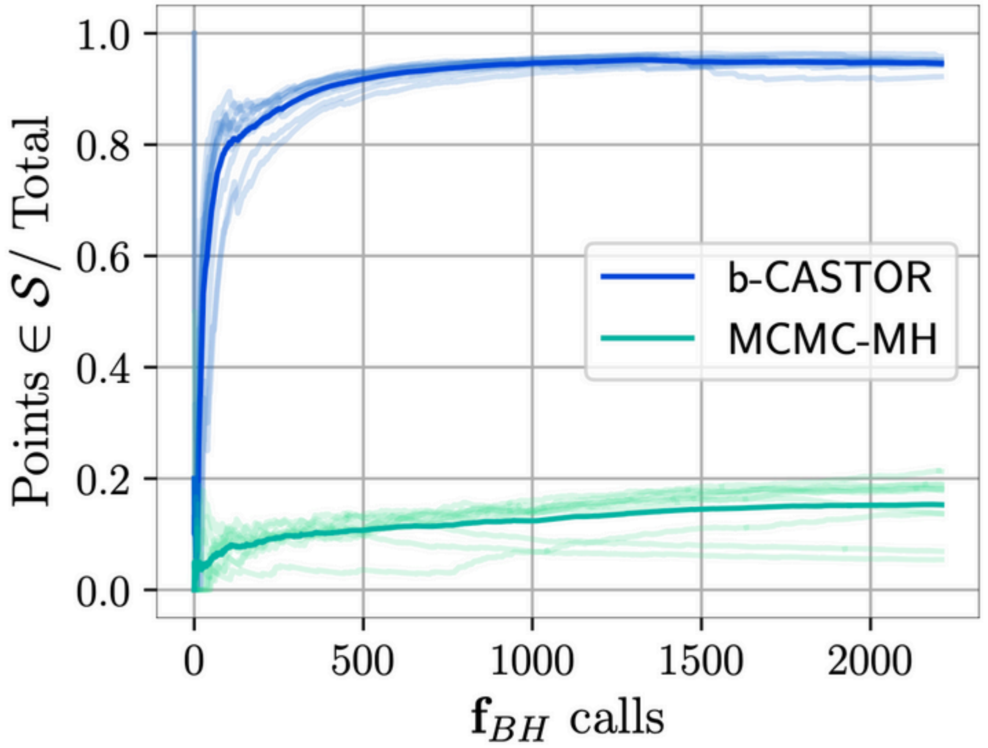
We introduce

b-CASTOR

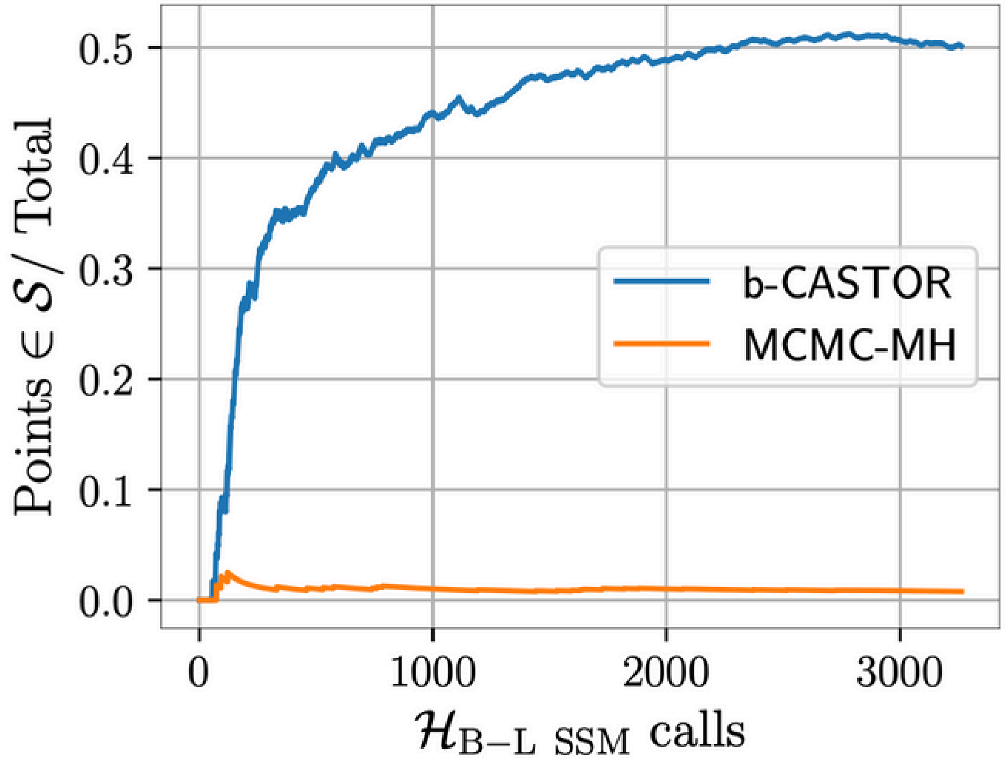
batched **C**onstraint **A**ctive **S**earch[1] with **T**PE **O**ptimisation and **R**ank based sampling



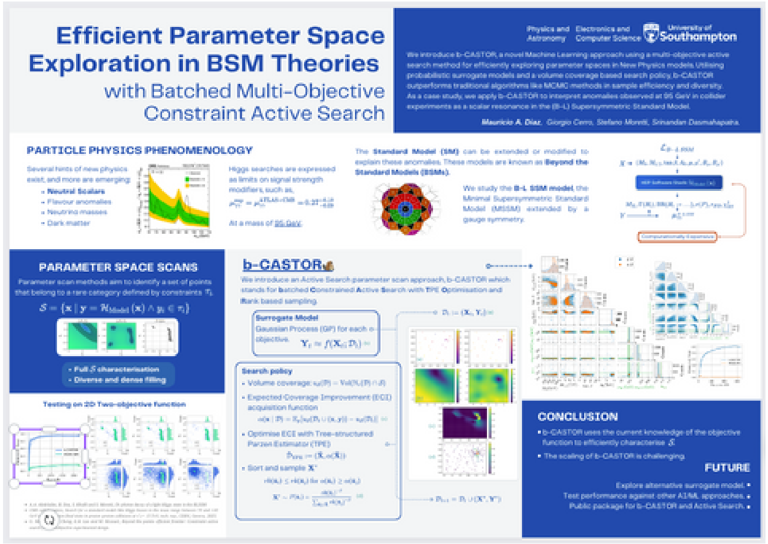
Test Function: 2D Two Objectives



B – L SSM study: 8D Five-Objectives



A 27



Mauricio A. Diaz

1. *G. Malkomes, B. Cheng, E.H. Lee and M. Mccourt, Beyond the pareto efficient frontier: Constraint active search for multiobjective experimental design*