Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

haute école
neuchâtel berne jura
arc ingénierie
www.he-arc.ch

PRISME-21 dataset

Public Reference for Interferometric Simulations for Foreground Mitigation in the Epoch of Reionization 21-cm signal

SKACH Winter Meeting 2026 – Emmanuel de Salis

Team : Michele Bianco, Sambit. K. Giri, Rohit Sharma, Tianyue Chen, Shreyam Parth Krishna, Chris Finlay, Viraj Nistane, Philipp Denzel, Massimo De Santis, Hatem Ghorbel

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

haute école
neuchâtel berne jura
arc ingénierie
www.he-arc.ch

The team behind this dataset

- This dataset is the result of a collaboration to tackle the SKAO Science Data Challenge 3a (SDC3a) ‘Foregrounds’, by the SKACH team.
- The SKACH team is composed of experts from astrophysics and data science

  ➥ *Disclaimer : I am from the Data Science side.*

SDC3a Foregrounds

- The 'Foregrounds' challenge goal is to remove obscuring sources of emission which prevent analysis of the underlying hydrogen-21cm signal from the Epoch of Reionisation (EoR).
- Participants are asked to extract the cylindrically-averaged power spectrum of the EoR signal, clean from foregrounds contamination.
- To do so, we were expected to simulate our own dataset for the creation and training of our models.
- This resulted of a lot of work to be able to simulate coherent, rich and relevant data.
- After the challenge, various colleagues and researchers expressed their interest in this dataset, which lead to an effort to make it available to the community.

Creating data

A Challenge by itself

- Creating an artificial dataset raise multiple challenges, especially with data representing real physical elements:
  - From a physics perspective: the dataset should align with real data, with similar behavior. It should "feel" and "act" as real data, to make analysis based on this data credible.
  - From a Machine Learning (ML) perspective : the underlying method for generating the data should be complex and hidden enough so that the ML models, that relies on statistical patterns, cannot catch the generation method.
- Dataset was created in Python, using *21cmFast*, that allows simulating both the cosmic structure formation and ionization of the intergalactic medium. The total size of this dataset is around 7Tb.
- Full details about data generation can be found in Bianco et al. (2024)[1]

1 : Bianco, M., Giri, S. K., Sharma, R., Chen, T., Krishna, S. P., Finlay, C., Nistane, V., Denzel, P., De Santis, M. & Ghorbel, H. (2025). Deep learning approach for identification of Hii regions during reionization in 21-cm observations–III. image recovery. *Monthly Notices of the Royal Astronomical Society*, staf973.

Accessing the data

Also a challenge by itself

- Hosting 7Tb of data in open access and according to the FAIR guidelines rules out many open science platforms, which have a limit of 50Go per dataset usually.
- Fortunately, a long-term solution was found thanks to the CSCS.
- Dataset :
  - Hosted on : https://rgw.cscs.ch/ska:sdc3-simdata
  - Walkthrough access : https://github.com/Emmanuel-desalis/sdc3-simdata
  - DOI and citation : https://zenodo.org/records/18326619

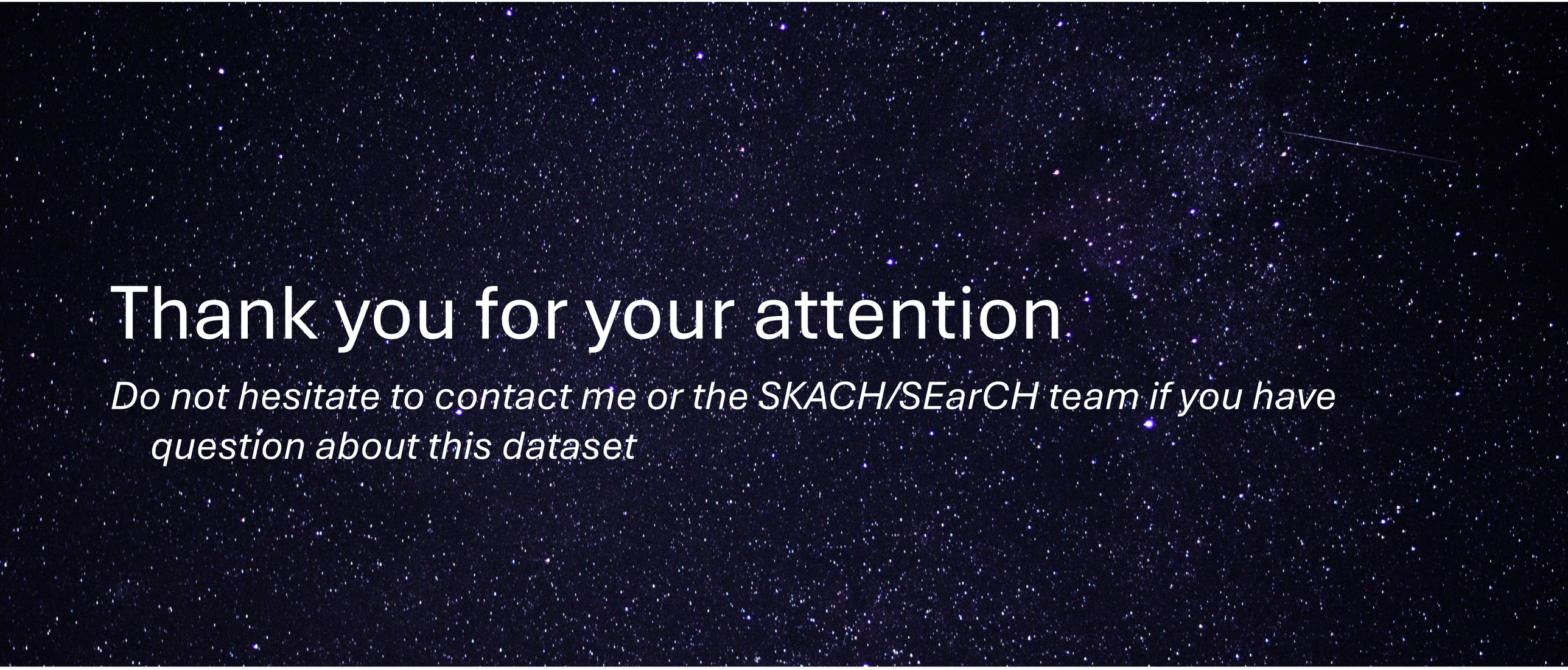
- ***No account needed to download the data***
- ***Only requirement : Python 3.9+***
- ***You can visualize the dataset, and download only subset of the data if you want***

Usage of the data

- Creating new foregrounds mitigation method → *I am currently doing this*
- Benchmarking existing foregrounds mitigation techniques.
- Assess your own foregrounds mitigation method on another dataset than the one you used to create it.
- But also :
  - Use it as a training dataset for your students
  - Use it as a starting point for your PhD research
  - *… and probably many other use-cases.*

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

haute école
neuchâtel berne jura
arc ingénierie
www.he-arc.ch



Thank you for your attention

Do not hesitate to contact me or the SKACH/SEarCH team if you have question about this dataset