SKA Science Update

- Data Challenges Update
- Science Meetings
- AOB
Science Data Challenge 2 results paper

- Describing the Challenge, the simulations, teams’ methods, results and analysis
- Submitted to MNRAS
- Over 100 challenge participants
- Over 50 worldwide institutions

**SKA Science Data Challenge 2: analysis and results**

P. Martley1,2, A. Bouwdel1, R. Brou1, J. N. H. S. Aditya1, S. Ascheid2, L. Agra2,2, A. Chakraborty4, X. Chen4, S. Choudhuri5, A. O. Clark1, J. Colón5, J. S. Collins3, D. Cornu1, L. Derriën2, M. Dell’Oro1, J. Ferretti1, B. Franço6, A. Galan4, J. Gérard2, F. Gevry2, H. Håkansson1, M. J. Hardcastle3, C. Henkel2, D. Herrero2, K. M. Hess1,2,12,22,23,25, J. Jogaka2,3, S. Jia7, R. J. Jeeck1,2, D. Kneer5, S. Klima5, D. Klein2,2, B. Lau4, L. La5, A. Mazarante4, J. Mioduszewski1, R. Monzal7,7, S. Nish7, M. Oxborrow5,5, M. Parra1,2, N. Patrício3,3, A. Pest1,5, F. Salomé1, S. Sánchez-Espejo1,2, M. Sargenson1,6,7,8, H. Semelin1, P. Serra5, A. K. Shov1, A. X. Shyu1,3,4, A. Sjoberg1,2, L. Smit1,8, A. Sone2, V. Stolyarov10,3, E. Tolley16, G. C. Torninc2, J. M. van der Hulst2, A. Vafaei Sadir2, L. Verde2, Montenegro2, T. Winter2, K. Yu1, L. Yu3, L. Zhang15, X. Zhang15, Y. Zheng, A. Albeni1, M. Ashdown2, C. Bon3, M. Brüggen2, J. Cannon2,4, R. Chen1, F. Costa1,3,3, J. Cowan3, F. Courbin1, J. Ding1, G. Feuerstein4, J. Freeman4,2, L. Guo2, C. Ghezal4, Q. Guo1, E. Gustafsson1,2, M. Jirsa1, M. G. Jones15, G. Józsa15, P. Kamphuis15, J. -P. Klein15, M. Lindqvist15, B. Liu3, Y. Liu1, Y. Mao1, A. Marchal1, J. Márquez1, A. Mescheryakov1, M. Oberl1, N. Ooerla1, M. Pardey-Pommerin1, W. Pet1, B. Peng1, J. Subbian1, A. Surgo1, J. Sturck15, C. Tasse1,15, A. Wang1, Y. Wang1, H. Xi1, X. Ying, H. Zhang1, J. Zhang1, M. Zhao1, S. Zuo1

Affiliations can be found at the references.

Accepted MNRAS. Received VYK is original from MRR

**ABSTRACT**

The Square Kilometre Array Observatory (SKA) will explore the radio sky to new depths in order to conduct transformative science. SKAO data products made available to astronomers will be correspondingly large and complex, requiring the application of advanced analysis techniques in order to extract key science findings. To this end, SKA is conducting a series of Science Data Challenges, each designed to familiarise the scientific community with SKA data and to drive the development of new analysis techniques. We present the results from Science Data Challenge 2 (SDC2), which invited participants to find and characterise 23245 (neutral) hydrogen (HI) sources in a simulated data product representing ≈ 5% of the SKA MDR spectral line observations from redshift 0.5 to 5.0. Through the generous support of eight international supercomputing facilities, participants were able to undertake the Challenge using dedicated computational resources. Alongside the main challenge, reproducibility awards were made in recognition of the pipelines which demonstrated Open Science best practices. The Challenge saw over 100 participants develop a range of new and existing techniques, in results which highlight the synergy of multidisciplinary and collaborative effort. The winning strategy, which combined predictions from two independent machine learning techniques to yield a 2% improvement in overall performance – surpassing the Open Science Challenge target for this method – is likely to be the combination of methods in a so-called ensemble approach.

**Keywords:** methods data analysis – radio data – galaxies – techniques imaging spectroscopy – galaxies: statistics – surveys – software: simulations

1. **INTRODUCTION**

© XIZ The Author
Science Data Challenge 2 results paper

High level findings:

- Complementary methods

Mix of new and existing techniques; machine learning and non-machine learning

SoFiA package very popular thanks to excellent documentation and ease of use

Analysis of biases and HI mass recovery with redshift

SKA Science Data Challenge 2: analysis and results


Abstract

The Square Kilometre Array Observatory (SKA) will explore the radio sky to new depths in order to conduct transformational science. SKA data products made available to astronomers will be correspondingly large and complex, requiring the application of advanced analysis techniques in order to extract key science findings. To this end, SKA-CAT is conducting a series of Science Data Challenges (SDCs), each designed to familiarise the scientific community with SKA data and to drive the development of new analysis techniques. We present the results from Science Data Challenge 2 (SDC2), which invited participants to find and characterize 250245 sources in a simulated data product representing a 2000 hr SKA MDR spectral line observation from redshifts 0.2 to 5.5. Through the generous support of eight international supercomputing facilities, participants were able to undertake a high-performance, distributed computational experiment. Against this main challenge, “reproducibility awards” were made in recognition of eight pipelines which demonstrated Open Science best practices. The Challenge saw over 100 participants develop a range of new and existing techniques, in results which highlighted the synergies of multi-disciplinary and collaborative effort. The winning strategy — which combined predictions from two independent machine learning techniques to yield a 24 per cent improvement in overall performance — underscores some of the most important lessons of such challenges.

Key words: methods data analysis — radio lines: galaxies — techniques: imaging spectroscopy — galaxies: statistics — surveys — software: simulations

1 Introduction

The Square Kilometre Array (SKA) project was born from an ambition to create a telescope sensitive enough to trace the formation of...
Science Data Challenge 2 results paper

- Expressing SDC2 scores in terms of source signal-to-noise values
- Meaningful measure of signal-to-noise
- Use SKA MID noise properties:
  - RMS noise remains ~constant when spatially smoothing up to ~70 arcsec FWHM
- Possible implications for source finding approaches
Science Data Challenge 2 results paper
Reproducibility awards

Reproducibility:

*Is the software:*
- Well-documented
- Easy to install
- Easy to use

Reusability:

*Does the software:*
- Use an open licence
- Have findable code
- Use code standards
- Use built-in tests

Find the award checklist at: sdc2.astronomers.skatelescope.org/sdc2-challenge
Science Data Challenge 2 results paper
Reproducibility awards

Reproducibility:
Is the software:
• Well-documented
• Easy to install
• Easy to use

Reusability:
Does the software:
• Use an open licence
• Have findable code
• Use code standards
• Use built-in tests

Results

<table>
<thead>
<tr>
<th>Team name</th>
<th>Reproducibility award</th>
<th>Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPFL</td>
<td>Bronze</td>
<td><a href="https://github.com/epfl-radio-astro/LiSA">https://github.com/epfl-radio-astro/LiSA</a></td>
</tr>
<tr>
<td>FORSKA-Sweden</td>
<td>Silver</td>
<td><a href="https://github.com/FraunhoferChalmersCentre/ska-sdc-2">https://github.com/FraunhoferChalmersCentre/ska-sdc-2</a></td>
</tr>
<tr>
<td>HI-FRIENDS</td>
<td>Gold</td>
<td><a href="https://github.com/HI-FRIENDS-SDC2/h-friends">https://github.com/HI-FRIENDS-SDC2/h-friends</a></td>
</tr>
<tr>
<td>NAOC-Tianlai</td>
<td>Bronze</td>
<td><a href="https://github.com/kfyu/SDC2-tianlai">https://github.com/kfyu/SDC2-tianlai</a></td>
</tr>
<tr>
<td>SHAO</td>
<td>Bronze</td>
<td><a href="https://github.com/astrosumit/SDC2-SHAO">https://github.com/astrosumit/SDC2-SHAO</a></td>
</tr>
<tr>
<td>Team SoFiA</td>
<td>Silver</td>
<td><a href="https://github.com/SoFiA-Admin/SDC2-SoFiA">https://github.com/SoFiA-Admin/SDC2-SoFiA</a></td>
</tr>
</tbody>
</table>

Award announcement to be featured in next edition of Contact
Tiered EoR Data Challenge

- **SDC3a Foregrounds**
- Foreground Subtraction + 21cm Power Spectrum Extraction (SWG contacts: Trott & Jelic)
- Target Participants: SWGs like CD/EoR, Cosmology, Continuum, etc.
  - Input Data: Calibrated Visibilities and High Fidelity Image
- Challenge will be based on:
  a) Ability to remove the point source + diffuse foregrounds from the data-set
  b) Ability to extract the cylindrical power spectrum
- Verification of the results from participants
  c) Comparison with the original input signal power spectrum
Tiered EoR Data Challenge

- **SDC3b Inference**

- Extraction of reionization parameters (SWG contacts: Mesinger & Mellema)

- Target Participants: SWGs like CD/EoR
  - Input Data: EoR PS + noise and residual foreground contamination

- Challenge will be based on:
  a) Ability to extract the IGM and source properties
  b) Verification of the results from participants
    - Comparison with the input ionisation history
Tiered EoR Data Challenge: Timeline

• SDC3a foregrounds: end of 2022, 6 months duration
• SDC3b inference: after SDC3 foregrounds, 6 months duration
  • Two independent datasets, different EoR model
  • Teams will be able to complete them individually
• SDC3 foregrounds results will be propagated to the SDC3 inference simulation by adding foreground residuals to the input EoR PS and/or filtering some modes
Purpose

As with our previous two data challenges (SDC1 and SDC2), our goal is to prepare the radio-astronomical community for the novel nature of the data expected from the Square Kilometre Array. Given the order-of-magnitude improvement in sensitivity, new analysis methods are required for both the challenging nature of resulting data, but also for the previously untouched science. Thus, realistic, synthetic datasets emulating the telescope’s capabilities will be disseminated to the community to test the suitability of existing methods and foster the development of new ones on these next-generation, scientific datasets. Ultimately, results of each of the competing teams’ approaches will be compared via a standard figure-of-merit, instigating a competitive nature to our challenges.
Science Data Challenge 3

Foregrounds

Our ‘Foregrounds’ challenge asks participants to remove obscuring sources of emission which prevent analysis of the underlying hydrogen-21cm signal from the Epoch of Reionisation. This foreground emission stems from both Galactic and extragalactic sources, both of which have previously observed, and unobserved components.

Given the lack of a model for the finer structure of Galactic emission at SKA-LOW frequencies, the removal of Galactic emission from the dataset represents a significant challenge. By similar reasoning, source confusion from previously unknown extragalactic sources, especially at the coarser resolution at metre-wavelengths, complicates the matter further.

From our synthetic datasets, participants are asked to extract the cylindrically-averaged power spectrum of the EoR signal, clean from foregrounds contamination.

To assess resulting submissions, our scoring (‘figure-of-merit’) algorithms will take resulting cylindrical power spectra, and return a score. Ancillary analytical data products can be assessed, however, the cylindrical power spectra is the primary focus of the challenge.

Our expected analysis will include:
-自動化による放射源の位置推定
-データセットからパワースペクトラムの推定
-フィギュア・オブ・メリットによる評価
EoR Data Challenge: Computational Facility Partners

- Why computational facility partners?
  - Store the dataset in multiple locations, where teams will be able to access
  - Provide computational resources to inspect and analyse the dataset without transferring
- How will it work?
  - Teams will state their computational needs as part of the SDC3 registration
  - The SDC team will collaborate with the facility partners to identify the best matches with teams
  - Teams will access the data through the chosen facility
  - The data will be made available at multiple facilities at the same time to ensure a fair challenge
  - Teams will be able to process the data there
- Which facilities for SDC3?
  - IRIS
  - INAF ICT facility
  - SPSRC
  - GENCI-IDRIS
  - EngageSKA - UCLCA
  - Swiss SRC
  - ChinaSRC
  - ASTRON/SURF
  - AUS SRC
  - JPSRC
**SDC3a Registrations**

- Registration started on 10\textsuperscript{th} October 2022
- To-date, there have been 18 registrations from 7 countries
- Expect more by end of registration today (15\textsuperscript{th} November 2022)
SDC3a Registrations

17 registrations (per 14 November)
SDC3a Registrations

Data Format Preference
- UVFITS: 13
- MS: 5

Data Type(s) Wanted
- Images: 15
- Both: 3

HPC Requested?
- Yes: 15
- No: 3
Science Meetings

- Joint ESO/SKAO Conference and Workshop was planned for week of 14 November 2022 “Coordinated Surveys of the Southern Sky”, in Garching: week of 27 February 2023
- Joint SKAO/ngVLA Science Conference week of 30 April 2023, in Vancouver
  - Web site in development, SOC formed
- EAS 2023, SKAO Lunch Session (1.5 hour) proposed
- IAU GA 2024 in Cape Town, several Letters of Intent for SKAO related Symposia have been submitted, including in EoR and HI areas, any news here?
Any Other Business

• New SWG mailing lists are now available (including core sub-lists) with same conventions throughout
  • e.g. swg-transients@skao.int, swg-col-core@skao.int, swg-vlbi@skao.int, swg-particles@skao.int
  • Old list names will still work
• Is there general SWG interest in central hosting of WG notes?
• News from SWG Chairs?
• ...
We recognise and acknowledge the Indigenous peoples and cultures that have traditionally lived on the lands on which our facilities are located.