Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **523,** 1967–1993 (2023)
Advance Access publication 2023 May 31

https://doi.org/10.1093/mnras/stad1375

SKA Science Data Challenge 2: analysis and results

P. Hartley,[1★] A. Bonaldi,[1,2] R. Braun,[1] J. N. H. S. Aditya,[3] S. Aicardi,[4] L. Alegre,[1,5] A. Chakraborty,[6] X. Chen,[7] S. Choudhuri,[8,9] A. O. Clarke,[1] J. Coles,[10] J. S. Collinson,[1] D. Cornu,[11] L. Darriba,[12] M. Delli Veneri,[13] J. Forbrich,[14] B. Fraga,[15] A. Galan,[16] J. Garrido,[12] F. Gubanov,[17] H. Håkansson,[18] M. J. Hardcastle,[14] C. Heneka,[19] D. Herranz,[20] K. M. Hess,[12,21,22] M. Jagannath,[23] S. Jaiswal,[3] R. J. Jurek,[24] D. Korber,[16] S. Kitaeff,[25] D. Kleiner,[26] B. Lao,[3] X. Lu,[11] A. Mazumder,[6] J. Moldón,[12] R. Mondal,[27] S. Ni,[28] M. Önnheim,[18] M. Parra,[12] N. Patra,[6,29] A. Peel,[16] P. Salomé,[11] S. Sánchez-Expósito,[12] M. Sargent,[16,30,31] B. Semelin,[11] P. Serra,[26] A. K. Shaw,[32] A. X. Shen,[33,34] A. Sjöberg,[18] L. Smith,[10] A. Soroka,[17] V. Stolyarov,[10,35] E. Tolley,[16] M. C. Toribio,[36] J. M. van der Hulst,[22] A. Vafaei Sadr,[37] L. Verdes-Montenegro,[12] T. Westmeier,[25] K. Yu,[7] L. Yu,[38] L. Zhang,[39,40] X. Zhang,[28] Y. Zhang,[3] A. Alberdi,[12] M. Ashdown,[10] C.R. Bom,[15] M. Brüggen,[19] J. Cannon,[41] R. Chen,[38] F. Combes,[11,42] J. Conway,[36] F. Courbin,[16] J. Ding,[39] G. Fourestey,[16] J. Freundlich,[43] L. Gao,[28] C. Gheller,[26] Q. Guo,[7] E. Gustavsson,[18] M. Jirstrand,[18] M. G. Jones,[44] G. Józsa,[45] P. Kamphuis,[46] J.-P. Kneib,[16] M. Lindqvist,[36] B. Liu,[38] Y. Liu,[7] Y. Mao,[47] A. Marchal,[48] I. Márquez,[12] A. Meshcheryakov,[49] M. Olberg,[36] N. Oozeer,[45] M. Pandey-Pommier,[50] W. Pei,[7] B. Peng,[38] J. Sabater,[5] A. Sorgho,[12] J. L.Starck,[16] C. Tasse,[51,52] A. Wang,[3] Y. Wang,[7] H. Xi,[38] X. Yang,[3] H. Zhang,[39] J. Zhang,[28] M. Zhao[28] and S. Zuo[47]

Affiliations are listed at the end of the paper

Accepted 2023 April 21. Received 2023 March 30; in original form 2022 October 18

ABSTRACT

The Square Kilometre Array Observatory (SKAO) will explore the radio sky to new depths in order to conduct transformational science. SKAO data products made available to astronomers will be correspondingly large and complex, requiring the application of advanced analysis techniques to extract key science findings. To this end, SKAO is conducting a series of Science Data Challenges, each designed to familiarize the scientific community with SKAO 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 233 245 neutral hydrogen (H I) sources in a simulated data product representing a 2000 h SKA-Mid spectral line observation from redshifts 0.25–0.5. 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 those pipelines which demonstrated Open Science best practice. The Challenge saw over 100 participants develop a range of new and existing techniques, with results that highlight the strengths of multidisciplinary and collaborative effort. The winning strategy – which combined predictions from two independent machine learning techniques to yield a 20 per cent improvement in overall performance – underscores one of the main Challenge outcomes: that of method complementarity. It is likely that the combination of methods in a so-called ensemble approach will be key to exploiting very large astronomical data sets.

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

1 INTRODUCTION

The Square Kilometre Array (SKA) project was born from an ambition to create a telescope sensitive enough to trace the formation of the earliest galaxies. Observing this era via the very weak emission from neutral hydrogen atoms will be possible only by using a collecting area of unprecedented size: large enough not only to provide a window onto *Cosmic Dawn* but – thanks to its increase in sensitivity over current instruments – also to explore new frontiers in galaxy evolution and cosmology, cosmic magnetism, the laws of gravity, extraterrestrial life and – in the strong tradition of radio astronomy (Wilkinson et al. 2004) – the unknown (see the SKA

[★] E-mail: philippahartley@hotmail.com

Published by Oxford University Press on behalf of Royal Astronomical Society

Science Book, Braun et al. 2015 for a comprehensive description of the full SKA science case).

First light at the SKA Observatory (SKAO) will mark a paradigm shift not only in the way we see the Universe but also in how we undertake scientific investigation. In order to perform such sensitive observations and extract scientific findings, huge amounts of data will need to be captured, transported, processed, stored, shared, and analysed. Innovations developed in order to enable the SKAO data journey will drive forward data technologies across software, hardware, and logistics. In a truly global collaborative effort, preparations, are underway under the guidance of the SKA Regional Centre Steering Committee to build the required data infrastructure and prepare the community for access to it (Chrysostomou et al. 2020). Alongside operational planning, scientific planning – undertaken by the SKAO Science Working Groups – is underway in order to maximize the exploitation of future SKAO data sets. The SKA model of data delivery will provide science users with data in the form of science-ready image and non-image SKAO products, with calibration and pre-processing having been performed by the Observatory within the Science Data Processor (SDP) and at the SKA Regional Centres (SRCs). While this model reduces by many orders of magnitude the burden of data volume on science teams, the size and complexity of the final data products remains unprecedented (Scaife 2020).

The primary goal of the SKAO Science Data Challenge (SDC) series is defined thus:

(i) To support future observers to prepare for SKAO data.

This goal is achieved via the following objectives:

To familiarize the astronomy community with the nature of SKAO data products.

To drive forward the development of data analysis techniques.

The first objective allows participants not only to gain familiarity with the size and complexity of SKAO data, but also with the provision of data products in science-ready form. It is achieved through the distribution of publicly available[1] real or simulated data sets designed to represent as closely as possible future SKAO data. A successful Challenge will see engagement and participation representing a broad range of geography and expertise, and a step forward by participants in the understanding and skills involved in analysing SKA-like data. The second objective is achieved through the application of new or existing methods in order to extract findings from the data. Standardized cross-comparisons of methods, which would require a strict set of running conditions and constraints on participants, are not performed. Instead, the focus is on inclusion, training, and the generation of ideas. A successful Challenge will see the application of diverse ideas and methods to the problem, and an understanding of the ability of respective methods to produce useful findings.

The SKAO is committed to Open Science values and the FAIR data principles (Wilkinson et al. 2016; Katz et al. 2021) of Findability, Accessibility, Interoperability, and Reproducibility. Accordingly, we aim to ensure equal accessibility to the Challenges for all participants. In the latest Challenge, teams were able to access the ∼1 TB Challenge data set and computational resources at one of eight partner supercomputing facilities, at which each could deploy their own analysis pipelines (Section 2.2). This model also served as a test bed for a number of future SRC technologies. Throughout the Challenge, a strong emphasis was placed on the reproducibility and reusability of software solutions. All teams taking part in the Challenge were eligible to receive a reproducibility prize, awarded against a set of pre-defined criteria. We thus identified two secondary goals for this Challenge:

(i) To test SKA Regional Centre prototyping.
(ii) To encourage Open Science best practice.

Science Data Challenge 1 (SDC1; Bonaldi et al. 2020) saw participating teams find and characterize sources in simulated SKA-Mid continuum images, with results that demonstrate the complementarity of methods, the challenge of finding sources in crowded fields, and the importance of careful image partitioning. Domain knowledge proved important not only in the design of pipelines but in the application of correct unit conversions specific to radio astronomy. SDCs benefit from additional domain reference material to support participants who do not have a radio astronomy background.

Science Data Challenge 2[2] (SDC2) involved a simulated spectral line observation designed to represent the SKAO view of neutral hydrogen (H I) emission up to $z = 0.5$, again inviting participants to attempt source finding and characterization within a very large data product. Resulting from the 'spin-flip' of an electron in a neutral hydrogen atom, 21-cm spectral line emission and absorption traces the distribution of H I across the history of the Universe. This cold gas exists in and around galaxies, fueling star-formation via ongoing infall from the cosmic web. Observations of individual H I sources can reveal the interactions between galaxies and the surrounding intergalactic medium (IGM; Popping et al. 2015), can probe stellar feedback processes within the interstellar medium (ISM; de Blok et al. 2015), and can allow us to study the impact of active galactic nuclei (AGN) on the large-scale gas distribution in galaxies (Morganti, Sadler & Curran 2015). H I dynamics also provide a measurement of the dark matter (DM) content of galaxies (Power et al. 2015). Deep H I surveys are therefore crucial for our understanding of galaxy formation and evolution over cosmic time (Power, Baugh & Lacey 2010; Blyth et al. 2015; Meyer et al. 2017; Dodson et al. 2022).

The faintness of H I emission has until recently limited survey depths to up to $z \sim 0.25$ [see Sancisi et al. (2008), van der Hulst & de Blok (2013), and Koribalski et al. (2020) for reviews of the results so far]. H I emission has now been imaged in a starburst galaxy at $z \sim 0.376$ (Fernández et al. 2016) using the Very Large Array within the COSMOS H I Large Extragalactic Survey (CHILES), and signals observed using the *Giant Metrewave Radio Telescope* (*GMRT*) have been stacked in order to make a successful measurement of the cosmic H I mass density at $0.2 < z < 0.4$ (Bera et al. 2019) and to detect the H I 21-cm signal from 2841 galaxies at average redshift $z \sim 1.3$ (Chowdhury et al. 2021). The MeerKAT telescope – a precursor to the SKAO – has now launched the Looking At the Distant Universe with the MeerKAT Array (LADUMA) survey (Blyth et al. 2016), which will image H I emission in the Chandra Deep Field-South out to $z \sim 1$. The SKA-Mid telescope will survey to depths of $z \sim 1$ in emission and $z \sim 3$ in absorption across a wider field. Comparing both surveys over 2000 h of observation, an SKA-Mid survey is likely to increase by 0.8 dex the number of detected galaxies, probing a cosmic volume $V_c \approx 185$ Mpc^3 versus $V_c \approx 74$ Mpc^3 and significantly reducing the sensitivity of the results to cosmic variance. The size of resulting data sets necessitates the use of automated source finding methods; several software tools are currently available for H I source detection and characterization (Flöer & Winkel 2012; Jurek 2012;

[1] https://astronomers.skatelescope.org/ska-data-challenges/

[2] https://sdc2.astronomers.skatelescope.org/

Whiting 2012; Westerlund & Harris 2014; Teeninga et al. 2015; Serra et al. 2015a; Westmeier et al. 2021) and a comparative study based on WSRT data has recently been performed (Barkai et al. 2023).

In this paper we report on the outcome of SDC2. The structure of the paper is as follows: in Section 2 we define the Challenge; in Section 3 we describe the simulation of the SDC2 data sets; in Section 4 we present the methods used by participating teams to complete the Challenge; in Section 5 we describe the scoring procedure; in Sections 6 and 7 we present the Challenge results and analysis, before setting out our conclusions in Section 8.

2 THE CHALLENGE

In this section we present an overview of the Challenge delivery and the data product supplied to Challenge teams, followed by the definition of the Challenge undertaken.

2.1 Challenge overview

Participating teams were invited to access a 913 GB data set hosted on dedicated facilities provided by the SDC2 computational resource partners (Section 2.2). The data set, 5851 × 5851 × 6668 pixels in size, simulates an H I imaging datacube representative of future deep SKA-Mid spectral line observations, with the following specifications:

(i) 20 square degrees field of view.
(ii) 7 arcsec beam size, sampled with 2.8 × 2.8 arcsec pixels.
(iii) 950–1150 MHz bandwidth, sampled with a 30 kHz resolution. This corresponds to rest frame velocity widths 7.8 and 9.5 km s^{-1} at the upper and lower limits, respectively, of the redshift interval $z = 0.235–0.495$.
(iv) Noise consistent with a 2000-h total observation, in the range 26–31 μJy per channel.
(v) Systematics including imperfect continuum subtraction, simulated RFI flagging and excess noise due to RFI.

The H I datacube was accompanied by a radio continuum datacube covering the same field of view at the same spatial resolution, with a 950–1400 MHz frequency range at a 50 MHz frequency resolution.

Challenge teams were invited to use analysis methods that were any combination of purpose-built and bespoke to existing and publicly available. Together with the full-size Challenge data set, two smaller data sets were made available for development purposes. Generated using the same procedure as the full-size data set but with a different statistical realization, the ‘development’ and ‘large development’ data sets were provided along with truth catalogues listing H I source property values. A further, ‘evaluation’, data set was provided without a truth catalogue, in order to allow teams to validate their methods in a blind way prior to application to the full data set. The evaluation data set was also used by teams to gain access to the full-size datacube hosted at an SDC2 partner facility. Access was granted upon submission of a source catalogue based on the evaluation data set and matching a required format. The development and evaluation data sets were made available for download prior to and during the Challenge.

The Challenge description, its rules, its scoring method, and a description of the data simulations were provided on the Challenge website before and during the Challenge. A dedicated online discussion forum was used throughout the Challenge to provide information to participants, to answer questions about the Challenge and to facilitate participant interaction. Definitions of conventions and units applicable to the challenge were circulated to participants before and during the Challenge.

2.2 Supercomputing partner facilities

The following eight supercomputing centres formed an international platform on which the full Challenge data set was hosted and processed:

AusSRC and Pawsey – Perth, Australia, aussrc.org
China SRC-proto – Shanghai, China, An et al. (2022)
CSCS – Lugano, Switzerland, www.cscs.ch
ENGAGE SKA-UCLCA – Aveiro and Coimbra, Portugal, www.engageska-portugal.pt; www.uc.pt
GENCI-IDRIS – Orsay, France, www.genci.fr
IAA-CSIC – Granada, Spain, Garrido et al. (2021)
INAF – Rome, Italy, www.inaf.it
IRIS (STFC) – UK, www.iris.ac.uk

Collectively, the Challenge facilities provided 15 million CPU hours of processing and 15 TB of RAM to participating teams.

2.3 The challenge definition

The Challenge results were scored on the full-size data set, on which teams undertook:

Source finding, defined as the determination of the location in RA (degrees), Dec (degrees), and central frequency (Hz) of the dynamical centre of each source.

Source characterization, defined as the recovery of the following properties:

(i) Integrated line flux (Jy Hz): the total flux density S integrated over the signal $\int S \mathrm{d}_\nu$.
(ii) H I size (arcsec): the H I major axis diameter at 1 $M_\odot\ pc^{-2}$.
(iii) Line width (km s^{-1}): the observed line width at 20 per cent of its peak.
(iv) Position angle (degrees): the angle of the major axis of the receding side of the galaxy, measured anticlockwise from North.
(v) Inclination angle (degrees): the angle between line-of-sight and a line normal to the plane of the galaxy.

Catalogues listing measured properties were submitted via a dedicated scoring service (see Section 5.1), which compared each submission with the catalogue of truth values and returned a score. For the duration of the Challenge, scores could be updated at any time; the outcome of the Challenge was based on the highest scores submitted by each team. The Challenge opened on 1 February 2021 and closed on 31 July 2021.

2.4 Reproducibility awards

Alongside the main challenge, teams were eligible for ‘reproducibility awards’, which were granted to all teams whose processing pipelines demonstrated best practice in the provision of reproducible methods and Open Science. An essential part of the scientific method, reproducibility leads to better, more efficient science. Open Science generalizes the principle of reproducibility, allowing previous work to be built upon for the future. Reproducibility awards ran in parallel and independently from the SDC2 score, and there was no cap on the number of teams to whom the awards were given.

3 THE SIMULATIONS

Simulation of the H I datacubes involved three steps: source catalogue generation, sky model creation, and telescope simulation. All codes used to generate the data set are publicly available.[3]

3.1 Source catalogues

To produce a catalogue of sources with both continuum and H I properties we used the Tiered Radio Continuum Simulation (TRECS; Bonaldi et al. 2019) as updated by Bonaldi et al. 2023 Initial catalogues of H I emission sources were generated by sampling from an H I mass function derived from the ALFALFA survey results (Jones et al. 2018),

$$\phi(M_{\rm HI}, z) = \ln(10)\,\phi_* \left(\frac{M_{\rm HI}}{M_*(z)}\right)^{\alpha+1} e^{-\frac{M_{\rm HI}}{M_*(z)}}, \tag{1}$$

where the knee mass, $M_* = 8.71 \times 10^9$ M$_\odot$, marks the exponential decline from a shallow power law parameterized by $\alpha = -1.25$, and $\phi_* = 4.5 \times 10^{-3}$ Mpc^{-3} dex^{-1} is a normalization constant. A mild redshift dependence was applied by using $\log(M_*(z)) = \log(M_*) + 0.075z$.

Conversion from H I mass in units solar mass to integrated line flux F followed the relation from Duffy et al. (2012):

$$M_{\rm HI} = 49.8\ F\ {D_{\rm L}}^2, \tag{2}$$

where luminosity distance, $D_{\rm L}$, is measured in Mpc and is obtained via the source redshift. A lower integrated flux limit of 1 Jy Hz was made, such that a fully face-on and unresolved source at this limit would produce a peak flux density approximately equal to the noise r.m.s. The catalogue also included a position angle θ drawn from a uniform distribution between 0–360 degrees, and an inclination angle i from the probability distribution function $f(i) = \sin(i)$.

Catalogues of radio-continuum sources – star-forming galaxies (SFGs) and AGN – were then generated using the Tiered Continuum Radio Extragalactic Continuum Simulation (T-RECS; Bonaldi et al. 2019) for the frequency interval 950–1400 MHz. A flux density limit of 2×10^{-7} Jy at 1150 MHz was applied, corresponding to k-corrected radio luminosities $L_{1150\,\rm MHz} = 1.58 \times 10^{19}$ W Hz^{-1} and $L_{1150\,\rm MHz} = 8.59 \times 10^{19}$ W Hz^{-1} at the lower and upper redshift limits, respectively, for a source with spectral index $\alpha = -0.7$. Continuum T-RECS catalogue properties included DM mass, star-formation rate and redshift.

The H I catalogue and the portion of the radio continuum catalogue covering the same redshift interval were then further processed to identify those that would constitute a counterpart, i.e. be hosted by the same galaxy (see Bonaldi et al. 2023 for more details).

In order to generate source positions in RA (x), Dec (y), and redshift (z) and to provide a realistic clustering signal, the galaxies were associated with DM haloes from the P-Millennium simulation (Baugh et al. 2019). Both the mass and environment of host DM haloes were considered; galaxies were associated with available DM haloes having the closest mass in the same redshift interval, and preferential selection of DM haloes with local density lower than 50 objects per cubic Mpc was made for H I-containing sources. The redshift of each source was converted to obtain the observed frequency (ν) at its dynamical centre.

3.2 Sky model

The sky model was generated using the PYTHON scripting language, making use of the ASTROPY, SCIPY, and SCIKIT-IMAGE libraries for image and cube generation, and using a modified version of FITSIO for writing to file.

3.2.1 H I emission datacube

H I sources were injected into the field using an atlas of high quality H I source observations. The atlas, containing 55 sources in total, was collated using samples available from the WSRT Hydrogen Accretion in LOcal GAlaxieS (HALOGAS) survey (Fraternali et al. 2002; Oosterloo, Fraternali & Sancisi 2007; Heald et al. 2011) – available online – and the THINGS survey (Walter et al. 2008), made available after the application of multiscale beam deconvolution. The preparation of atlas sources involved the following steps:

(i) Measurement of H I major axis diameter at a surface density of 1 M$_\odot$ pc^{-2}, made after converting source flux to mass per pixel.

(ii) Masking of all pixels with surface density less than 1 M$_\odot$ pc^{-2}, in order to produce a positive definite noiseless model.

(iii) Rotation, using published source position angles, to a common position angle of 0 degrees.

(iv) Preliminary spatial resampling, such that the physical pixel size of the resampled data would be no lower than required for the lowest redshift simulated sources. A smoothing filter was applied prior to resampling, in order to prevent aliasing.

(v) Preliminary velocity resampling after application of a smoothing filter.

Though modestly sized, the atlas sample of real H I galaxies represented considerable morphological diversity, containing examples of Hubble stages 2–10. The parameter space representing catalogue sources was not completely covered. Physical properties of the atlas sample covered the SFR range 0.004–6.05 M$_\odot$ y^{-1}, the H I mass range 1.20×10^7 to 1.41×10^{10} M$_\odot$, and the H I major axis diameter 2.29–102.23 kpc. Catalogue sources covered the SFR range 0.0039–251 M$_\odot$ y^{-1} (median 0.97), H I masses $M_{\rm H\,I} = 6.99 \times 10^7$ M$_\odot$ and 4.08×10^8 M$_\odot$ at the lower and upper limits of the simulated redshift range, respectively (with median 1.14×10^9 and maximum 1.08×10^{11}), and H I diameters $S = 4.78$–270 kpc (median 24.7).

For each source from the simulation catalogue, a source from the prepared atlas of sources was chosen from those nearby in normalized H I mass-inclination angle parameter space. Once matched with a catalogue source, atlas sources underwent transformations in size in the spatial cube dimensions x and y and in velocity dimension V in order to obtain the H I size S, minor axis size b, and line width w_{20}. An appropriate smoothing filter was applied prior to all scalings, in order to avoid aliasing effects. Transformation scalings were determined using the catalogue source properties of H I mass, inclination angle, and redshift, and making use of the following relations:

$$S = 0.51 \log M_{\rm HI} - 3.32, \tag{3}$$

from Broeils & Rhee (1997), in order to determine spatial scalings for mass;

$$V_{\rm rot}^2 = \frac{{\rm G}M_{\rm dyn}}{r}, \tag{4}$$

where $V_{\rm rot}$ is the rest frame rotational velocity at radius r and $M_{\rm dyn}$ is the dynamical mass and is set using $M_{\rm dyn}/M_{\rm H\,I} = 10$, in order to

[3] https://github.com/PhilippaHartley/SKAO-SDC2

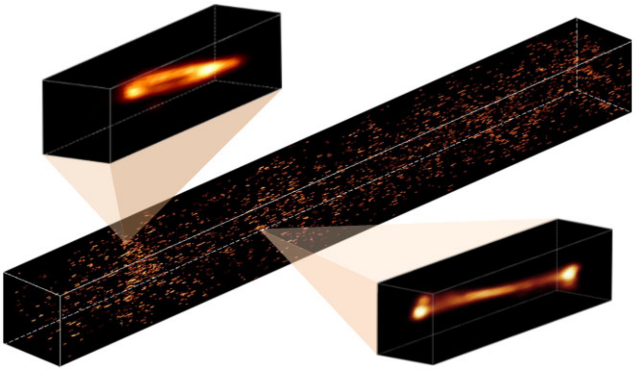

Figure 1. 3D view of the 'development' H I emission datacube, containing 2683 H I sources. The cube uses 1286 × 1286 × 6668 pixels to represent a 1 square degree field of view across the full Challenge frequency range 0.95–1.15 GHz (redshift 0.235–0.495). A log scaling has been applied to image pixel values. The two shorter axes represent the spatial dimensions and the longer axis the frequency dimension.

determine frequency scalings for H I mass;

$$\cos^2(i) = \frac{(b/S)^2 - \alpha^2}{(1-\alpha^2)}, \tag{5}$$

where $\alpha = 0.2$, in order to determine spatial scalings for inclination;

$$V_{\rm rad} = V_{\rm rot}\ \sin(i), \tag{6}$$

where $V_{\rm rad}$ is the rest frame radial velocity, and

$$w_{20} = \sqrt{(V_T^2 + 2V_{\rm rad}^2)}, \tag{7}$$

where V_T is the contribution to line width from turbulence, in order to determine velocity scalings for inclination. While a best fit to ALFALFA data finds a value $V_T = 90$ km s^{-1}, a lower value, $V_T = 40$ km s^{-1}, is chosen in order to avoid excessive scaling between peaks in velocity. Spatial scalings for redshift were determined by calculating the angular diameter distance $D_{\rm A}$, assuming a standard flat cosmology with $\Omega_{\rm m} = 0.31$ and $\rm H_0 = 67.8$ km s^{-1} Mpc^{-1} (Planck Collaboration 2016).

Finally, each transformed object was rotated to its catalogued position angle, convolved with a circular Gaussian of 7 arcsec FWHM and scaled according to total integrated H I flux, before being placed in the full H I emission field at its designated position in RA, Dec, and central frequency (Fig. 1).

3.2.2 Continuum emission datacube

The treatment of continuum counterparts of H I objects was dependent on the full width at half-maximum (FWHM) continuum size. An empty datacube with spatial resolution matching the H I datacube and an initial frequency sampling of 50 MHz was first generated. Each counterpart was then injected into the simulated field as either:

(i) an extended source, for those objects with a continuum size greater than 3 pixels;

(ii) a compact source, for those objects with a continuum size smaller than 3 pixels.

All compact sources were modelled as unresolved, and added as Gaussians of the same size as the synthesized beam. Images of all extended sources were generated according to their morphological parameters and then added as 'postage stamps' to an image of the full field, after applying a Gaussian convolving kernel corresponding to the beam.

The morphological model for the extended SFGs is an exponential Sersic profile, projected into an ellipsoid with a given axis ratio and position angle. The AGN population comprises steep-spectrum AGN, exhibiting the typical double-lobes of FRI and FRII sources, and flat-spectrum AGN, exhibiting a compact core component together with a single lobe viewed end-on. Within both classes of AGN all sources are treated as the same object type viewed from a different angle. For the steep-spectrum AGN we used the Double Radio-sources Associated with Galactic Nucleus (DRAGNs) library of real, high-resolution AGN images (Leahy, Bridle & Strom 2013), scaled in total intensity and size, and randomly rotated and reflected, to generate the postage stamps. All flat-spectrum AGN were added as a pair of Gaussian components: one unresolved and with a given 'core fraction' of the total flux density, and one with a specified larger size.

The continuum catalogues accompanying the Challenge data sets report the continuum size of objects as the Largest Angular Size and the exponential scale length of the disk for AGN and SFG populations, respectively.

3.2.3 Net emission and absorption cube

The H I emission cube described in Section 3.2.2 was further processed to introduce absorption features and the effect of imperfect continuum subtraction. H I absorption occurs if a radio continuum source is at a higher redshift along the same line of sight as an H I source. The intensity of the effect depends on both the brightness temperature of the continuum source and the H I opacity $\tau\Delta V$ of the H I source. Absorption features were introduced on the pixels of the H I model cube only if a background continuum source was present having at least a brightness temperature $T_{\rm min} = 100$ K. This corresponds to a flux density of $S_{\rm min} = 7.35 \times 10^{-4} T_{\rm min} \Delta\phi^2/\lambda^2$, with $\Delta\phi$ the beam size in arcsec and λ the observing wavelength in cm, yielding $S_{\rm min}$ in Jy per beam.

The absorption signature, $S_{\rm HIA}(\nu)$, was calculated as:

$$S_{\rm HIA}(\nu) = S_{\rm C}[1 - e^{(-\tau\Delta V/{\rm d}V)}], \tag{8}$$

where $S_{\rm C}$ is the continuum model flux density at this frequency and dV is the actual channel sampling in units of km s^{-1}. When observed with 100 pc or better physical resolution, the apparent H I column density $N_{\rm H\,I}$, can be related to an associated H I opacity (Braun 2012), as

$$N_{\rm HI} = N_0 e^{-\tau\Delta V} + N_\infty(1 - e^{-\tau\Delta V}), \tag{9}$$

where $N_0 = 1.25 \times 10^{20}$ cm^{-2}, $N_\infty = 7.5x10^{21}$ cm^{-2} and a nominal $\Delta V = 15$ km^{-1} provide a good description of the best observational data in hand. In turn, the hydrogen column density, $N_{\rm H\,I}$, associated with every pixel in the H I model cube can be obtained with

$$N_{\rm HI} = 49.8\, S_{\rm L}(\nu)\, \Delta\nu\, {\rm M}_\odot (1+z)^4/(N_{\rm p}\, m_{\rm H}\, \Delta\theta^2\, C_{\rm M}^2), \tag{10}$$

where $S_{\rm L}$ is the H I brightness in the pixel in Jy per beam, $\Delta\nu$ the channel spacing in Hz, $\rm M_\odot$ a solar mass, z the redshift of the H I 21-cm line that applies to this pixel, $N_{\rm p}$ the number of pixels per spatial beam, $m_{\rm H}$ the hydrogen atom mass, $\Delta\theta$ the spatial pixel size in radians, and $C_{\rm M}$ a Mpc expressed in cm. The preceding constant in the equation follows the flux density to H I mass conversion of Duffy et al. (2012).

In the current case, the physical resolution is too coarse – some 10 kpc per pixel – to resolve the individual cold atomic clouds that give rise to significant H I absorption opacity. The apparent column densities per pixel have therefore been subjected to an arbitrary power law rescaling designed to render a plausible amount of observable absorption signatures. We used

$$N'_{\mathrm{HI}} = 10^{19+[\log 10(N_{\mathrm{HI}})-19]\beta}, \quad (11)$$

if $N_{\mathrm{H\,I}} > 10^{19}$, with power law index $\beta = 1.9$. This is followed by a hyperbolic tangent asymptotic filtering,

$$N''_{\mathrm{HI}} = N_\infty[\mathrm{e}^{2N'_{\mathrm{HI}}/N_\infty} - 1]/[\mathrm{e}^{2N'_{\mathrm{HI}}/N_\infty} + 1], \quad (12)$$

in order to avoid numerical problems when solving for the opacity.

In order to simulate imperfect continuum emission subtraction within the final H I datacube, a noise cube representing gain calibration errors was produced. We first interpolated the simulated continuum sky model, $S_{\mathrm{C}}(\nu)$, to a frequency sampling of 10 MHz, before producing for each channel a two dimensional image of uncorrelated noise to represent an r.m.s. gain calibration error of $\sigma = 1 \times 10^{-3}$ and with spatial sampling 515 × 515 arcsec. The spatial and frequency samplings were chosen in order to represent the residual bandpass calibration errors that might result from the typical spectral standing wave pattern of an SKA dish at these frequencies, together with the angular scale over which direction dependent gain differences might be apparent.

The coarsely sampled noise field was then interpolated up to the 2.8 × 2.8 arcsec sampling of the sky model and a deliberately imperfect version of the continuum sky model, $S_{\mathrm{NC}}(\nu)$, was constructed by multiplying each pixel in the perfect model by (1 + N), where N is the value of the corresponding pixel in the noise cube. Finally, both the perfect and imperfect continuum models were downsampled to the final simulation frequency interval of 30 kHz. The net continuum-subtracted H I emission and absorption cube, $S(\nu)$) is finally calculated from the sum

$$S(\nu) = S_{\mathrm{L}}(\nu) + S_{\mathrm{C}}(\nu) - S_{\mathrm{NC}}(\nu) - S_{\mathrm{HIA}}(\nu). \quad (13)$$

3.3 Telescope simulation

The simulation of telescope sampling effects has been implemented by using PYTHON to script tasks from the MIRIAD package (Sault, Teuben & Wright 1995). Multiprocessing parallelization is exploited by applying the procedure over multiple frequency channels simultaneously.

3.3.1 *Calculation of effective PSF and noise level*

The synthesized telescope beam was based on a nominal 8-h duration tracking observation of the complete SKA-Mid configuration. A 1-min time sampling interval was used in order to make beam calculations sufficiently realistic while limiting computational costs. The thermal noise level was based on nominal system performance (Braun et al. 2019) for an effective on-sky integration time of 2000-h distributed uniformly over the 20 deg^2 survey field. The effective integration time per unit area of the survey field increases towards lower frequencies in proportion to wavelength squared. This is due to the variation in the primary beam size in conjunction with an assumed survey sampling pattern that is fine enough to provide a uniform noise level even at the highest frequency channel. The nominal r.m.s. noise level, σ_N therefore declines linearly with frequency between 950 and 1150 MHz.

Observations of the South Celestial Pole (Experiment ID 20190424–0024) using MeerKAT, which is located on the future SKA-Mid site and will constitute part of the SKA-Mid array, have been used to obtain a real world total power spectrum. With this power spectrum we can estimate the system noise temperature floor of the MeerKAT receiver system as a function of frequency, in addition to an estimate of any excess average power due to radio frequency interference (RFI). The ratio of excess RFI to system noise temperature, γ_{RFI}, was used to scale the nominal noise in each frequency channel and to determine the degree of simulated RFI flagging to apply to the nominal visibility sampling. Flagging was applied to all baselines from a minimum $\mathrm{B}_{\mathrm{min}} = 0$ up to a maximum B_{max} according, in units of wavelength, to

$$B_{\mathrm{max}} = 71 \times 10^{(\gamma_{\mathrm{RFI}}-1)^{1/3}}, \quad (14)$$

which produced maximum baseline lengths ranging from under 15 m to around 10 km across the relevant range of observing frequencies. The duration of RFI flagging, ΔHA, was determined, in hours, from

$$\Delta\mathrm{HA} = \begin{cases} 0, & \text{if } \gamma_{\mathrm{RFI}} < \gamma_{\mathrm{min}} \\ 8\,(\gamma_{\mathrm{RFI}} - \gamma_{\mathrm{min}})/(\gamma_{\mathrm{max}} - \gamma_{\mathrm{min}}), & \text{if } \gamma_{\mathrm{min}} > \gamma_{\mathrm{RFI}} > \gamma_{\mathrm{max}} \\ 8, & \text{if } \gamma_{\mathrm{RFI}} > \gamma_{\mathrm{max}} \end{cases}$$

where $\gamma_{\mathrm{min}} = 1.1$ and $\gamma_{\mathrm{max}} = 2$, are used to define the ranges of RFI ratios over which flagging is absent, intermittent, or continuous. Intermittent flagging intervals were placed randomly within the nominal HA = −4h to + 4h tracking window.

After application of flagging to the nominal visibility sampling, the synthesized beam and corresponding 'dirty' noise image were generated for each frequency channel. During imaging, a super-uniform visibility weighting algorithm was employed that makes use of a 64 × 64 pixel FWHM Gaussian convolution of the gridded natural visibilities in order to estimate the local density of visibility sampling. The super-uniform re-weighting was followed by a Gaussian tapering of the visibilities to achieve the final target dirty PSF properties, namely the most Gaussian possible dirty beam central lobe with 7 × 7 arcsec FWHM. The effective PSF is then modified to account for the fact that the survey area will be built up via the linear combination of multiple, finely spaced, telescope pointings on the sky. The effective PSF in this case was formed from the product of the calculated dirty PSF with a model of the telescope primary beam at this frequency, as documented in Braun et al. (2019). The dirty noise image for each channel was then rescaled to have an r.m.s. fluctuation level, σ_i, corresponding to the nominal sensitivity level of the channel degraded by its RFI noise ratio,

$$\sigma_i = \sigma_N \gamma_{\mathrm{RFI}}. \quad (15)$$

3.3.2 *Simulated sampling and deconvolution*

The H I net absorption and emission datacube (Section 3.2.3) was subjected to simulated deconvolution and residual degradation by the relevant synthesized dirty beam. Any signal, both positive and negative, in excess of three times the local noise level, $3\sigma_i$, was extracted as a 'clean' image with the threshold signal retained to form a residual sky image. The residual sky image was subjected to a linear deconvolution (via FFT division) with a 7 × 7 arcsec Gaussian, truncated at 10 per cent of the peak and then convolved with the dirty beam. The final data product cube was formed by summing for each channel the dirty residuals, the previously extracted clean feature image and the dirty noise image.

3.4 Limitations of the simulated data products

While significant effort has been expended to make a realistic data product for the Challenge analysis, there are many limitations to the degree of realism that could be achieved. Some of the most apparent are outlined below.

(i) Telescope sampling limitations, arising from the adoption of image plane sky model convolution to approximate the actual imaging process. This forms the most significant limitation to the simulations, but is necessitated by the fact that working instead in the visibility plane would require processing of data sets 7.4 PB in size, far exceeding current capabilities.

(ii) Realism of the noise properties: systematic effects such as residual RFI, bandpass ripples, residual continuum sidelobes, and deconvolution artifacts were not included in the simulation. Additionally, the properties of the errors that have been included feature mostly Gaussian, uncorrelated noise, which may not represent the complexity of those those found in real interferometric data.

(iii) H I emission model limitations, arising from the limited number of real H I observations used to generate simulated H I subcubes.

(iv) Catalogue limitations, arising from the independent generation of H I and continuum catalogues.

(v) H I absorption model limitations, due to very coarse sampling used to assess physical properties along the line of sight in order to introduce H I absorption signatures. Further, the relatively low resolution of the simulated observation results in a low apparent brightness temperature of continuum sources (< 100 K), such that the occurrence of absorption signatures has been restricted only to those continuum sources that exceed this brightness limit.

(vi) Continuum emission model limitations, arising from the use of simple models to describe SFGs and flat-spectrum AGN sources, and from the limited number of real images used to generate steep spectrum sources.

(vii) An assumption of negligible H I self-opacity which, although widely adopted in the current literature, is unlikely to be the case in reality (see e.g. Braun 2012).

(viii) The overall translation of truth catalogue inputs to simulated source morphologies: the Challenge scoring definition measures the recovery of truth catalogue inputs, while teams themselves measure properties from a simulated realization of those inputs. This could introduce a degeneracy in the evaluation of method performance.

The limitations listed above would in turn place limits on how well teams' performances on this data set would transfer to real data.

4 METHODS

Participating teams made use of a range of methods to tackle the problem, first making use of the smaller development data set and truth catalogue in order to investigate techniques. 12 teams made a successful submission entry using the full Challenge data set. The methods employed by each of those finalist teams are presented below and are summarised in Table1.

4.1 Coin

C Heneka, M Delli Veneri, A Soroka, F Gubanov, A Meshcheryakov, B Fraga, CR Bom, M Brüggen

During the Challenge the Coin team tested several modern ML algorithms from scratch alongside the development our own wavelet-based 'classical' baseline detection algorithm. For all approaches we first flagged the first 324 channels in order to remove residual RFI, as measured by the per-channel signal mean and variance. We considered the following ML architectures for object detection: 2D/3D U-Nets, R-CNN and an inception-style network that mimics filtering with wavelets. The to-date best-performing architecture was a comparably shallow segmentation U-Net that translated the 2D U-Net in Ronneberger, Fischer & Brox (2015a) to 3D. It was trained on 3D cubic patches taken from the development cube, each containing a source and with no preprocessing applied. We mitigated High (> 90 per cent) rates of false positives to moderate levels (~ 50 per cent; see Fig. 2) by imposing interconnectivity and size cuts on the potential sources and discarding continuum-bright areas. We obtained a roughly constant ~50:50 ratio between true and false positives for 0.25 deg^2 cutouts across the development cube and the full Challenge cube. Our 'classical' baseline performed an alternative detection procedure, first using Gaussian filtering in the frequency dimension followed by wavelet filtering and thresholding. Interscale connectivity (Scherzer 2010) and reconstruction were performed on the denoised and segmented output. This pipeline detected < 10 per cent true positives for the Challenge data release: an order of magnitude higher false positive rate than the ML-based pipeline.

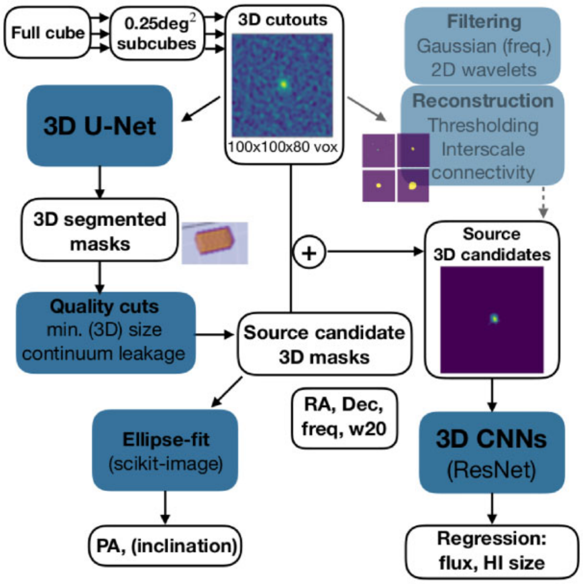

Figure 2. Data processing pipeline used by the Coin team.

Source positions (RA, Dec, central frequency, line width) were directly inferred from the obtained segmentation maps via the `regionprops` function of the SCIKIT-IMAGE PYTHON package (van der Walt et al. 2014). Source properties (flux, size) were derived through a series of ResNet convolutional neural networks (CNNs; He et al. 2016) applied to the source candidate 3D cutouts. The position angle was measured using the SCIKIT-IMAGE package to fit ellipses to sources masks; inclination could not be fitted for most objects.

We conclude that further cleaning and denoising and the application of techniques from the 'classical' baseline, such as wavelet filtering, is needed to improve on our machine learning (ML) pipeline method. Alternatively, further steps that include classification and a more curated training set could be desirable. Lessons learned in these 'from-scratch' developments can give valuable insights into the performance and application of said algorithms, such as the suitability of 3D U-Nets for segmentation of tomographic H I data and the need for additional cleaning algorithms jointly with networks or multistep procedures, such as a classification step, when faced with low S/N data.

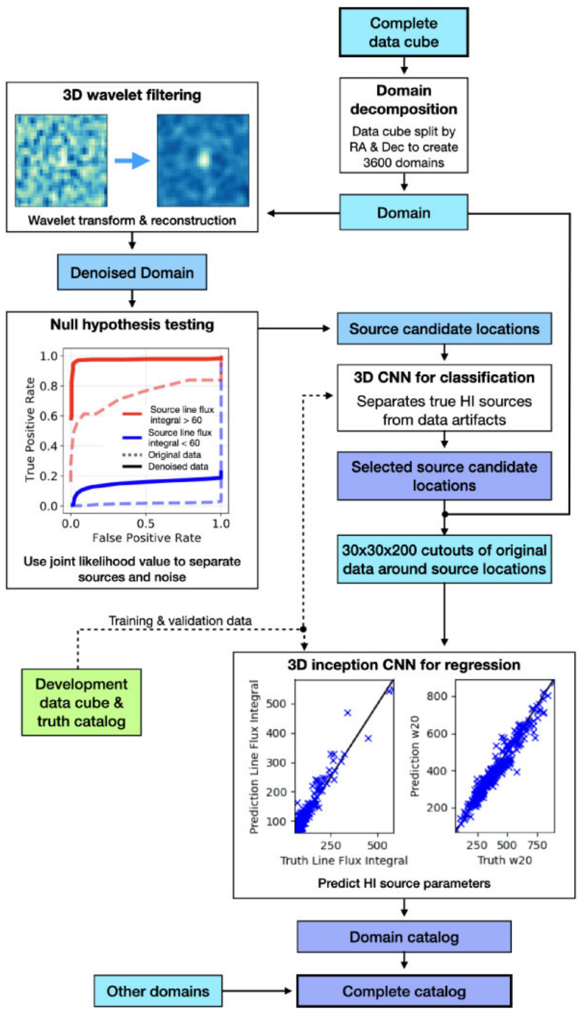
Complete data cube
Domain decomposition
Data cube split by RA & Dec to create 3600 domains
3D wavelet filtering
Wavelet transform & reconstruction
Domain
Denoised Domain
Null hypothesis testing
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
Source line flux integral > 60
Source line flux integral < 60
Original data
Denoised data
Use joint likelihood value to separate sources and noise
Source candidate locations
3D CNN for classification
Separates true HI sources from data artifacts
Selected source candidate locations
30x30x200 cutouts of original data around source locations
Training & validation data
Development data cube & truth catalog
3D inception CNN for regression
Prediction Line Flux Integral
Truth Line Flux Integral
100
200
300
400
500
250
500
Prediction w20
Truth w20
200
400
600
800
250
500
750
Predict HI source parameters
Domain catalog
Other domains
Complete catalog

Figure 3. Data processing pipeline used by the EPFL team.

4.2 EPFL

E Tolley, D Korber, A Peel, A Galan, M Sargent, G Fourestey, C Gheller, J-P Kneib, F Courbin, J-L Starck

The EPFL team used a variety of techniques developed specifically for the Challenge and which have been collected into the LiSA library (Tolley et al. 2022) publicly available on GITHUB.[4] The pipeline (Fig. 3) first decomposed the Challenge data cube into overlapping domains by dividing along RA and Dec. Each domain was then analysed by a separate node on the computing system. A pre-processing step used 3D wavelet filtering to denoise each domain: decomposition in the 2D spatial dimensions used the Isotropic Undecimated Wavelet Transform (Starck, Fadili & Murtagh 2007), while the decimated 9/7 wavelet transform (Vonesch, Blu & Unser 2007) was applied to the 1D frequency axis. A joint likelihood model was then calculated from the residual noise and used to identify H I source candidates through null hypothesis testing in a sliding window along the frequency axis. Pixels with a likelihood score below a certain threshold (i.e. unlikely to be noise) were grouped into islands. The size and arrangement of these islands were used to reject data artefacts. Ultimately the location of the pixel with the highest significance was kept as an H I source candidate location.

A classifier CNN was used to further distinguish true H I sources from the set of candidates. The final H I source locations were then used to extract data from the original, non-denoised domain to be passed to an Inception CNN which calculated the source parameters. The Inception CNN used multiple modules to examine data features at different scales. Finally, the H I source locations and features for each domain were concatenated to create the full catalogue. Both CNNs were trained on the development data set using extensive data augmentation.

4.3 FORSKA-Sweden

H Håkansson, A Sjöberg, MC Toribio, M Önnheim, M Olberg, E Gustavsson, M Lindqvist, M Jirstrand, J Conway

The FORSKA-Sweden team performed source detection using a U-Net (Ronneberger, Fischer & Brox 2015b) CNN with a ResNet (He et al. 2016) encoder. Our methods are presented more in detail in Häkansson et al. (2023), and all related code is published on GITHUB.[5] A training set was generated from the lower 80 per cent of the development cube, split along the x-axis, by applying a binary mask to all pixels within range of a source defined by a cylinder using source properties (major axis, minor axis, line width) from the truth catalogue. Batches of 128 cubes of size 32 × 32 × 32 pixels were sampled from the training area. Half of these cubes contained pixels assigned to a source in the target mask, which caused galaxy pixels to be over-represented in a training batch compared to the full development cube. This over-representation made training more efficient. The remaining 20 per cent of the development cube was used for frequent validation and tuning of model hyperparameters.

We used the soft Dice loss as the objective function (Milletari, Navab & Ahmadi 2016; Khvedchenya 2019). The initial weights of the model, pretrained from ImageNet, were provided by the PYTORCH-based SEGMENTATION MODELS package (Yakubovskiy 2020). Each 2D $k \times k$-filter of the pretrained model was converted to a 3D filter with a procedure similar to Yang et al. (2021). We aligned two dimensions to the spatial plane, and repeated the same 2D filter for k frequencies, which resulted in a $k \times k \times k$ filter. The Adam optimizer (Kingma & Ba 2014) with an initial learning rate of 10^{-3} was used for training the model. The trained CNN was applied to the raw Challenge data cube to produce a binary segmentation mask assigning each pixel either to a galaxy or not (Fig. 4).

The *merging* and *mask dilation* modules from SOFIA 1.3.2 (Serra et al. 2015a) were employed to post-process the mask and extract coherent segments into a list of separated sources. The last step of the pipeline was to compute the characterization properties for each extracted source. Some source properties were estimated in the aforementioned SOFIA modules, while others had to be computed outside in our code. The most recent weights obtained from CNN training and a fixed set of hyperparameters from the post-processing step were used to compute a score intended to mimic the scoring of the Challenge. The best model from training was then used as a basis for hyperparameter tuning, again using the mimicked scoring.

4.4 HI FRIENDS

J Moldón, L Darriba, L Verdes-Montenegro, D Kleiner, S Sánchez, M Parra, J Garrido, A Alberdi, JM Cannon, Michael G Jones, G Józsa, P Kamphuis, I Márquez, M Pandey-Pommier, J Sabater, A Sorgho

[4] https://github.com/epfl-radio-astro/LiSA

[5] https://github.com/FraunhoferChalmersCentre/ska-sdc-2/tree/cb3d34ebd944f3332de661cfb8fd7d3403cf9a45

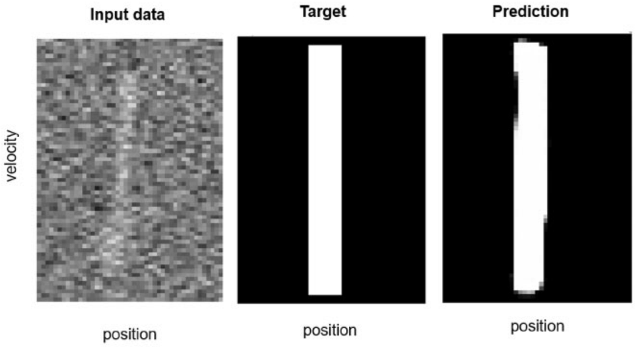

Figure 4. Cross-section images of input data, target and prediction with velocity and one positional dimension for one of the sources in the cube by team FORSKA-Sweden. The position axis is aligned with the major axis of the source.

The HI-FRIENDS team implemented a workflow (Moldon et al. 2021a) based on a combination of SOFIA-2 (Westmeier et al. 2021) and PYTHON scripts to process the data cube. The workflow, which is publicly available in GITHUB,[6] is managed by the workflow engine SNAKEMAKE (Mölder et al. 2021), which orchestrates the execution of a series of steps (called rules) and parallelizes the data analysis jobs. SNAKEMAKE also manages the installation of the software dependencies of each rule in isolated environments using conda (Anaconda 2020). Each rule executes a single program, script, shell command or JUPYTER notebook. With this methodology, each step can be developed, tested, and executed independently from the others, facilitating modularization and reproducibility of the workflow.

First, the cube is divided into smaller subcubes using the SPECTRAL-CUBE library. Adjacent subcubes include an overlap of 40 pixels (112 arcsec) in order to avoid splitting large galaxies. In the second rule, source detection and characterization is performed on each subcube using SOFIA-2 (Westmeier et al. 2021). We optimized the SOFIA-2 input parameters based on visual inspection of plots of the statistical quality of the fit and of some individual sources. In particular, we found that the parameters `scfind.threshold`, `reliability.fmin`, and `reliability.threshold` were key to optimizing our solution. We found that using the spectral noise scaling in SOFIA-2 dealt well with the effects of RFI-contaminated channels and we did not include any flagging step.

The third rule converts the SOFIA-2 output catalogues to new catalogues containing the relevant SDC2 source parameters in the correct physical units. We computed the inclination of the sources based on the ratio of minor to major axis of the ellipse fitted to each galaxy, including a correction factor dependent on the intrinsic axial ratio distribution from a sample of galaxies, as described in Staveley-Smith, Davies & Kinman (1992). The next two rules produce a concatenated catalogue for the whole cube: we concatenate the individual catalogues into a main, unfiltered catalogue containing all the measured sources, and then we remove the duplicates coming from the overlapping regions between subcubes using the r.m.s. as a quality parameter to discern the best fit. Because the cube was simulated based on real sources from catalogues in the literature we further filtered the detected sources to eliminate outliers using a known correlation between derived physical properties of each galaxy. In particular, we used the correlation in fig. 1 in Wang et al. (2016) that relates the H I size and H I mass of nearby galaxies. Several plots are produced during the workflow execution, and a final visualization rule generates a JUPYTER notebook with a summary of the most relevant plots.

Our workflow aims to follow FAIR principles (Wilkinson et al. 2016; Katz et al. 2021) to be as open and reproducible as possible. To make it findable, we uploaded the code for the general workflow to Zenodo (Moldon et al. 2021b) and WorkflowHub (Moldon et al. 2021c), which includes metadata and globally unique and persistent identifiers. To make the code accessible, we made derived products and containers available on GITHUB and Zenodo as open source. To make it interoperable, our workflow can be easily deployed on different platforms with dependencies either automatically installed (e.g. in a virtual machine instance in myBinder (Project Jupyter et al. 2018) or executed through singularity, podman, or docker containers. Finally, to make it reusable we used an open license, we included workflow documentation[7] that contains information for developers; the workflow is modularized as SNAKEMAKE rules. We included detailed provenance of all dependencies and we followed The Linux Foundation Core Infrastructure Initiative (CII) Best Practices.[8] Therefore, the workflow can be used to process other data cubes and should be easy to adapt to include new methodologies or adjust the parameters as needed.

4.5 HIRAXers

A Vafaei Sadr, N Oozeer

The HIRAXers team used a multilevel deep learning approach to address the Challenge. The approach extends to 3D a method applied to a similar, 2D, challenge (Vafaei Sadr et al. 2019) and uses multiple levels of supervision. Prior to source finding, a pre-processing step is used to detect regions of interest. Motivated by the recent progress in image-to-image translation techniques, one can utilize prior knowledge about source shapes to magnify signals, effectively suppressing background noise in a manner similar to image cleaning. We investigated two pre-processing approaches to reconstruct a 'clean' image. For both approaches we used a training set generated by using 2D spatial slices of the development data set to produce a source map containing masks and probability values. The output of the trained model can then be interpreted as a probability map.

Our first preprocessing approach used 2D slices in frequency as greyscale images. The model learns to retrieve information employing only transverse information. For the second approach, we extended the inputs into 3D to benefit from longitudinal patterns by adding different frequencies as convolutional channels, thus forming a multichannel image. We used a 128 × 128 sliding window to manage memory consumption, a mean squared error loss function, and a decaying learning rate. We used the standard image processor in TENSORFLOW (Abadi et al. 2015) for minimal data augmentation, with ranges of one degree for rotation and one per cent for zoom range, in addition to horizontal and vertical flips.

We developed our pipeline to examine the following architectures: V-Net (Milletari, Navab & Ahmadi 2016); Attention U-Net (Oktay et al. 2018); R2U-Net (Alom et al. 2018); U^{2}net (Qin et al. 2020); UNet3 + (Huang et al. 2020); TransUNet (Chen et al. 2021); and ResUNet-a (Diakogiannis et al. 2020). One can find most of the implementations in the KERAS-UNET-COLLECTION (Sha 2021) package. The learning rate was initiated at 1×10^{-3} with a 0.95 decay per 10 epochs using the Adam optimizer. Our results using the

[6] https://github.com/HI-FRIENDS-SDC2/hi-friends

[7] https://hi-friends-sdc2.readthedocs.io/en/latest/

[8] https://bestpractices.coreinfrastructure.org/en/projects/5138

development data set found that the U^{2}net architecture achieved the best performance. U^{2}net employs residual U-blocks in a ‘U-shaped’ architecture. It applies the deep-supervision technique to supervise training at all scales by downgrading the output.

In the second step of our method we trained a model to find and characterize the objects. To find the objects, we applied a peak finder algorithm to the 3D output of U^{2}net. A peak is simply the pixel that is larger than all its 27 neighbours. The ‘found’ catalogue was then passed into a modified eight-layer HighRes3DNet (Li et al. 2017) as a regressor for characterization before generating the final catalogue.

4.6 JLRAT

L Yu, B Liu, H Xi, R Chen, B Peng

The JLRAT team first divided the whole data set into small cubes of size 320 × 320 × 160 (RA, Dec, frequency) before applying to each cube a CNN containing a fully convolutional layer and a softmax layer. The CNN used 1D spectra from the cube as inputs and produced a masked output of candidate spectral signals. Using the inner product, we computed the correlation in the space domain between each candidate spectrum and known spectra from the SDC2 development cube. The result provided us with a set of 3D cubes, each containing a predicted galaxy with approximate position and size, and accurate line width. A 2D Gaussian function was used to fit the moment zero map with an intensity cutoff at 1 $\mathrm{M}_\odot\ \mathrm{pc}^{-2}$. The fit produced an ellipse with central position (RA, Dec), major axis and position angle, and the inclination of the galaxy. The flux integral was obtained by integrating the spectra within the ellipse in both space and frequency.

4.7 MINERVA

D Cornu, B Semelin, X Lu, S Aicardi, P Salomé, A Marchal, J Freundlich, F Combes, C Tasse

The MINERVA team developed two pipelines in parallel. The final catalogue merges the results from the two pipelines.

4.7.1 YOLO-CIANNA

The YOLO-CIANNA pipeline implemented a highly customized version of a YOLO (You Only Look Once; Redmon et al. 2015; Redmon & Farhadi 2016, 2018) network, which is a regression-based object detector and classifier with a CNN architecture. Our YOLO implementation is part of our general-purpose CNN framework, CIANNA[9] (Convolutional Interactive Artificial Neural Networks by/for Astrophysicists).

The definition of the training sample was of major importance to get good results. Most of the sources in the large development data set are impossible for the network to detect, and tagging them as positive detections would lead to a poorly trained model. For YOLO we used a combination of criteria to define a training set: (i) the CHADHOC classical detection algorithm (see Section 4.7.2); (ii) a volume brightness threshold; (iii) a local signal-to-noise ratio (SNR) estimation. Our refined training set contains around ∼1500 ‘true’ objects, with 10 per cent set aside for validation. All inputs were augmented using position and frequency offsets and flips. Our retained network architecture for this challenge operates on sub-volumes of 48 × 48 × 192 (RA, Dec, Frequency) pixels. The network was trained by selecting either a sub-volume that contains at least one true source or a random empty field, in order to learn to exclude all types of noise aggregation and artefacts.

The network maps each sub-volume to a 6 × 6 × 12 grid, where each element corresponds to a region of 8 × 8 × 16 pixels inside the input sub-volume. We chose to have the network predict a single possible detection box per grid element, producing the following parameters: x, y, z the bounding-box central position in the grid element; w, h, d the bounding-box dimension. We modified the YOLO loss function to allow us to predict the required H I flux, size, line width, position angle, and inclination in a single network forward for each possible box. The retained network architecture is made of 21 3D-convolutional layers, which alternate several ‘large’ filters (usually 3 × 3 × 5) that extract morphological properties and fewer ‘smaller’ filters (usually 1 × 1 × 3) that force a higher degree feature space while preserving a manageable number of weights to optimize. Some of the layers include a higher stride value in order to progressively reduce the dimensions down to the 6 × 6 × 12 grid. The last few layers include dropout for regularization and error estimation. In total the network has of the order of 2.3×10^6 parameters. When applying on the full datacube, predicted boxes are filtered using an ‘objectness’ score threshold to maximize the SDC2 metric.

Despite the fact that YOLO networks are known for their computational performance, our retained architecture still requires up to 36 h of training on a single RTX 3090 GPU using FP16/FP32 Tensor Core mixed precision training. The trained network has an inference speed of 76 sub-volumes per second using a V100 GPU on Jean-Zay/IDRIS, but due to necessary partial overlap and RAM limitations, it still requires up to 20 GPU hours to process the full ∼1 TB data cube.

4.7.2 CHADHOC

The Convolutional Hybrid Ad-Hoc pipeline (CHADHOC) has been developed specifically for SDC2. It is composed of three steps: a traditional detection algorithm, a CNN for identifying true sources among the detections, and a set of CNNs for source parameter estimation.

For detection, we first smooth the signal cube by a 600 kHz width along the frequency dimension and convert to an SNR on a per channel basis. Pixels below a fixed SNR of ∼2.2 are filtered out, and the remaining pixels are aggregated into detected sources using a simple friend-of-friend linking process with a linking length of 2 pixels. The position of each detection is computed by averaging the positions of the aggregated pixels. A catalogue of detections is then produced, ordered according to the summed source SNR values. When applied to the full Challenge data set, we divide the cube into 25 chunks and produce one catalogue for each chunk.

The selection step is performed with a CNN. A training sample is built by cross-matching with the truth catalogue the 10^5 brightest detections in the development cube, thus assigning a True/False label to each detection. Unsmoothed signal-to-noise cutouts of 38 × 38 × 100 pixels around the position of each detection are the inputs for the network. The learning set is augmented by flipping in all three dimensions, and one third of the detections are set aside as a test set. The comparatively light network is made of five 3D convolutional layers, containing 8, 16, 32, 32, and 8 filters, and three dense layers, containing 96, 32, and 2 neurons. Batch normalization, dropouts, and pooling layers are inserted between almost every convolutional and dense layer. In total the network has of the order of 10^5 parameters. The training is performed on a single Tesla V100 GPU in at most a few hours, reaching best performances after a few tens of epochs.

[9] https://github.com/Deyht/CIANNA

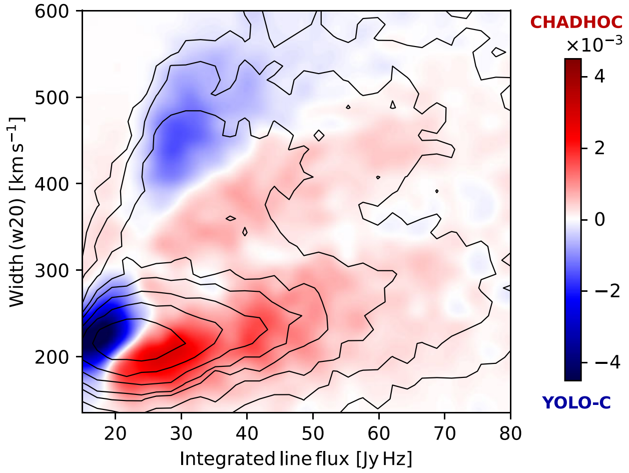

Figure 5. Team Minerva: Difference in the number of sources found between CHADHOC and YOLO-C catalogues in a flux against line width parameter space. The color encodes the difference in the local number of sources as a proportion of the total merged catalogue size (32652 predicted sources). The contours are the local number of sources averaged between the two catalogues with values: 6, 14, 30, 50, 64, 92, 128, 192. The density heatmap is computed on a 30 × 30 grid and plotted with interpolation.

The model produces a number between 0 (False) and 1 (True) for each detection. The threshold where the source is labelled as True is a parameter that must be tuned to maximize the metric defined by the SDC2. This optimization is performed independently of the training.

A distinct CNN has been developed to predict each of the source properties and includes a correction to the source position computed during the detection step. The architecture is similar to the one of the selection CNN, with small variations: for example, no dropout is used between convolutional layers for predicting the line flux. Cutouts around the ∼1300 brightest sources in the truth catalogue of the development cube are augmented by flipping and used to build the learning and tests sets. The networks are trained for at most a few hundred epochs in a few to 20 min each on a Tesla V100 GPU. Training for longer results in overfitting and a drop in accuracy.

Many details impact the final performance of the pipeline. Among them, the centering of the sources in the cutouts. Translational invariance is not trained into the networks. This is dictated by the nature of the detection process and is possibly the main limitation of the pipeline: the selection CNN will never be asked about sources that have not been detected by the traditional algorithm.

4.7.3 Merging the catalogues

If we visualize the catalogues produced by YOLO and CHADHOC in the sources parameter space (Fig. 5), we find that they occupy slightly different regions. For example, CHADHOC tends to find a (slightly) larger number of typical sources compared to YOLO, but misses more low-brightness sources because of the hard SNR threshold applied during the detection step. Thus, merging the catalogues yields a better catalogue.

Since both pipelines provide a confidence level for each source to be true, we can adjust the thresholds after cross-matching the two catalogues. In case of a cross-match we lower the required confidence level while when no cross-match is found we increase the required threshold. The different thresholds must be tuned to maximize purity and completeness. Finally, the errors on the source properties are at least partially uncorrelated between the two pipelines. Thus averaging the predicted values also improves the resulting catalogue properties.

4.8 NAOC-Tianlai

K Yu, Q Guo, W Pei, Y Liu, Y Wang, X Chen, X Zhang, S Ni, J Zhang, L Gao, M Zhao, L Zhang, H Zhang, X Wang, J Ding, S Zuo, Y Mao

After testing several methods, the NAOC-Tianlai team used the SOFIA-2 software to process of the SDC2 data sets. We optimized the SOFIA-2 input parameters by first performing a grid search in parameter space before refining the result using an MCMC simulation. We are currently developing a dedicated cosmological simulation on which to test our methods. However, during the Challenge time frame we mainly used the development and large development data sets to perform the optimization. The optimized parameters were then used for the processing of the full Challenge data set.

Due to the memory constraints and the consideration of avoiding excessive division along the spectral axis, the data sets were split into subcubes of size ∼330 × 330 × 3340 pixels for processing. Adjacent subcubes had an overlap of 10 or 20 pixels along each axis to ensure that H I galaxies on the border region were not missed. The full Challenge data set was therefore divided into 18 × 18 × 2 subcubes when processing.

Our main parameter selection procedure is as follows:

(i) We set a list of values to be searched for each parameter of interest, such as: `replacement`, `threshold` in the *scfind* module; `minSizeZ`, `radiusZ` in the *linker* module; and `minSNR`, `threshold`, `scaleKernel` in the *reliability* module. We then processed in parallel the development data set with the different combinations of parameters values.

(ii) Next, we selected the optimal parameter combination by comparing the output catalogues from the previous step with the development data set truth catalogue. To choose the optimal parameters, thresholds were applied to the *total detection number*, to the *match rate* (true detection/total detection), and to the final *score*.

(iii) To make the found optimal parameter combination more robust, different subcubes were processed following the procedure given above, and the combination that performed well on all subcubes was selected.

For reference, our trial produced the following optimized parameter settings: `scaleNoise.windowXY/Z = 55` for normalizing the noise across the whole datacube; `kernelsXY = [0, 3, 7]`, `kernelsZ = [0, 3, 7, 15, 21, 45]`, `threshold = 4.0`, `replacement = 1.0` in the *scfind* module for the S + C finder in SOFIA-2; `radiusXY/Z = 2`, `minSizeXY = 5`, `minSizeZ = 20` in the *linker* module for merging the masked pixels detected by the finder; and `threshold = 0.5`, `scaleKernel = 0.3`, `minSNR = 2.0` in the *reliability* module for reliability calculation and filtering. In our processing, each parameter combination instance took ∼5 min with one CPU thread to process one subcube.

Finally, we applied the optimal parameter combination to the processing of all subcubes from the Challenge data set, and merged the results.

4.9 SHAO

S Jaiswal, B Lao, JNHS Aditya, Y Zhang, A Wang, X Yang

The SHAO team developed a fully-automated pipeline in PYTHON to work on the Challenge data set. Our method involved the following steps: (1) We first sliced the datacube into individual frequency channel images and used SEXTRACTOR (Bertin & Arnouts 1996) to perform source finding on each image. We used a 2.5σ detection threshold (for ~99 per cent detection confidence) and minimum detection area of 2 pixels. (2) We cross-matched the sources found in consecutive channel images using the software TOPCAT (Taylor 2005) with a search radius of 1 pixel = 2.8 arcsec. (3) For each source detected in at least two consecutive channel images we estimated the range of channels for each source, adding 1 extra channel on both sides. (4) We extracted a subcube across the channel range obtained in the previous step, using a spatial size of 12 pixels around each identified source. (5) We made a moment-0 map for each extracted source using its subcube, after first masking negative flux densities. (6) We used SEXTRACTOR on the moment-0 map of each extracted H I source to estimate the source RA and Dec coordinates, major axis, minor axis, position angle, and integrated flux. Inclination angle was estimated using the relations given by Hubble (1926) and Holmberg (1946). (7) We constructed a global H I profile for each source by estimating the flux densities within a box of size 6 pixels around the source position in every channel of its subcube. (8) We finally fit a single Gaussian model to estimate the central frequency of H I emission and line width at 20 per cent of the peak.

The score obtained by this method is not very satisfactory. However, our investigations gave us confidence in dealing with a large H I cube and making the pipeline for the analysis. We will try to improve our pipeline by optimizing the input parameters and implementing different algorithms in the future. The use of ML techniques could be a good choice for such data sets.

4.10 Spardha

AK Shaw, NN Patra, A Chakraborty, R Mondal, S Choudhuri, A Mazumder, M Jagannath

The SPARDHA team developed a PYTHON-based pipeline which starts by dividing the 1 TB Challenge data set into several small cubelets. We performed source finding using an MPI-based implementation to run parallel instances of SOFIA-2 on each cubelet. We tuned the parameters of SOFIA-2 to maximize the number of detected sources. A total of 118 cubelets were analysed, which were categorized into two groups, namely: (1) Normal cubelets and (2) Overlapping cubelets. The whole datacube was first divided into consecutive blocks of equal dimensions to create Normal cubelets (Fig. 6). Overlapping cubelets were then centred at the common boundaries of Normal cubelets in order to detect sources that fall at their common boundaries.

In order to avoid source duplication, buffer regions were defined around the faces of each cubelet (see Fig. 6, top row). We always accepted any source whose centre was detected within the cubelet but not in the buffer zone (see Fig. 6, bottom row). We conservatively set the width of buffer zones based on the physically motivated values of the spatial and frequency extent of typical galaxies scaled at the desired redshifts. We chose the maximum extent of the galaxy on the sky plane to be ~80 kpc (Wang et al. 2016), corresponding to ~10 pixels in the nearest frequency channel. The buffer region was set to be twice this extent, i.e. 20 pixels. Overlapping regions were therefore 4 × 20 = 80 pixels wide. Along the frequency direction, galaxies can have a line-width extent of ~500 km s^{-1}, which corresponds to ~72 channels. The widths of the buffer regions and Overlapping regions along the frequency axis were therefore 144 and 288 channels, respectively. The acceptance regions of the cubelets (normal and overlapping) were such that they spanned the whole data cube contiguously when arranged accordingly. Although this approach increased the computation slightly due to analysing some regions of the data more than once, it ensured that there was no common source present in the list. Analysing cubelets was the most time-consuming part in our pipeline. We analysed 118 cubelets on 472 cores in parallel in around 15 min.

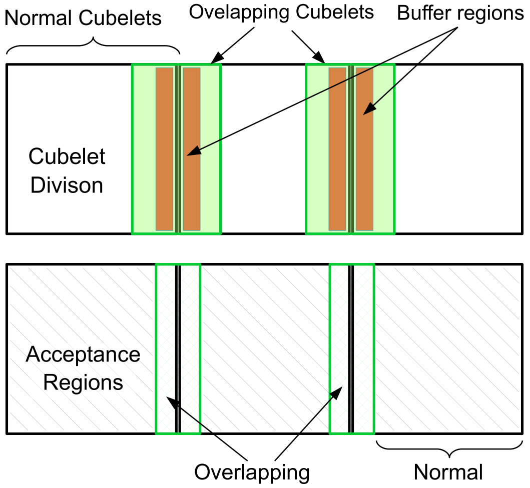

Figure 6. Team Spardha: The 2D projection along one axis of the schematic division of the data into *Normal* and *Overlapping* cubelets (top row), and the corresponding Acceptance regions (black hashing; bottom row). Normal cubelets are illustrated by black outlined boxes. Overlapping cubelets are centred at the common boundaries of Normal cubelets and are illustrated by green boxes. Orange regions (top row) represent buffer zones.

We used physical equations to convert the SOFIA-2 catalogue into the SDC-prescribed units and to discard bad detections such as those sources having *NaN* values in the columns or those with negative flux values. In the final stage we put limits on the line width, discarding detections with unusual values. Motivated by physical models and observations of galaxies, we conservatively accepted the sources having $w_{20} \in [60,\ 500]$ km s^{-1} (McGaugh et al. 2000). We finally arranged the catalogue in descending order of the flux values. Based on tests using the development datacube, for which the exact source properties are known, we chose the top 35 per cent of total sources to generate the final catalogue for submission.

4.11 Starmech

MJ Hardcastle, J Forbrich, L Smith, V Stolyarov, M Ashdown, J Coles

The Starmech tackled the Challenge from the point of view of dealing with the Challenge data set within the constraints of the resources provided to us (a single node with 30 cores and 124 GB RAM, 800 GB root volume, and 1 TB additional data volume). Some computational constraints will be a feature of future working in the field when computing resources are provided as part of shared SKA Regional Centres.

We considered existing source finding tools: PYBDSF (Mohan & Rafferty 2015), a continuum source finder, and SOFIA and SOFIA-2, two generations of a 3D source finder already optimized for H I (Westmeier et al. 2021). While PYBDSF readily generated a catalogue of the continuum sources and could be run on many slices in frequency, slicing and averaging with fixed frequency steps does not give good results since emission lines have a variety of possible widths in frequency space. Instead we focused on the two publicly available 3D source finders. Our tests showed that SOFIA-2's memory footprint is much lower than that of SOFIA for a given data cube and its speed significantly higher, so it became our algorithm of choice.

In order to work with the available RAM, we needed to slice the full Challenge datacube either in frequency or spatially. We chose to slice spatially because this allows SOFIA-2 to operate as expected in frequency space; essentially the approach is to break the sky down into smaller angular regions, run SOFIA-2 on each one in series, and then join and de-duplicate the resulting catalogue. Whether done in parallel (as in the MPI implementation SOFIA-X; Westmeier et al. 2021), or in series as we describe here, some approach like this will always be necessary for large enough H I series in the SKA era since the full data set sizes will exceed any feasible RAM in a single node for the foreseeable future.

Our implementation was a simple PYTHON wrapper around SOFIA-2. The code calculates the number of regions into which the input data cube needs to be divided such that each individual sub-cube can fit into the available RAM. Assuming a tiling of $n \times n$, it then tiles the cube with n^2 overlapping rectangular spatial regions. We define a guard region width g in pixels: each region passed to SOFIA overlaps the adjacent one, unless on an edge, by $2g$ pixels. Looping over the sub-cubes, SOFIA-2 is run on each one to produce n^2 overlapping catalogues in total. For our final submission we used SOFIA-2 default parameters with an `scfind.threshold` of 4.5σ, $g = 20$ pixels, a spatial offset threshold for de-duplication of 1 pixel, and a frequency threshold of 1 MHz. g was chosen to be larger than the typical size in pixels of any real source. We verified that there were no significant differences, using these parameters, between the reassembled catalogue for a smaller test cube and the catalogue directly generated by running SOFIA-2 on the same cube, using TOPCAT for simple catalogue visualization and cross-matching. Due to time constraints, we did not move on to the next obvious step of optimizing the parameters used for SOFIA-2 based on further runs on the test and development data sets.

We removed source duplication arising from overlapping regions by considering catalogues from adjacent sub-cubes pairwise. We firstly discarded all catalogue entries with pixel position more than g pixels from the edge of a sub-cube; these should already be present in another catalogue. The remaining overlap region, $2g$ pixels in width, height, or both, was cross-matched in position and sources whose position and frequency differ by less than user-defined threshold values were considered duplicates and discarded from one of the two catalogues. Finally the resulting n^2 de-duplicated catalogues were merged and catalogue values converted according to units specified by the submission format.

We would like to have explored the utility of dimensional compression of the data as part of the source finding, for example by using moment maps in an attempt to eliminate noise and better pinpoint source detection algorithms. *A priori*, this would have been of rather technical interest since any resulting bias on source detection would need to be considered. However, in this way, it may have been possible to identify candidate sources to then characterize based on observable parameters such as size and linewidth, in a first step as point sources vs resolved sources, and including flags for potential overlap in projection or velocity.

4.12 Team SOFIA

KM Hess, RJ Jurek, S Kitaeff, P Serra, AX Shen, JM van der Hulst, T Westmeier

Team SOFIA made use of the Source Finding Application (SOFIA; Serra et al. 2015a; Westmeier et al. 2021) to tackle the Challenge.

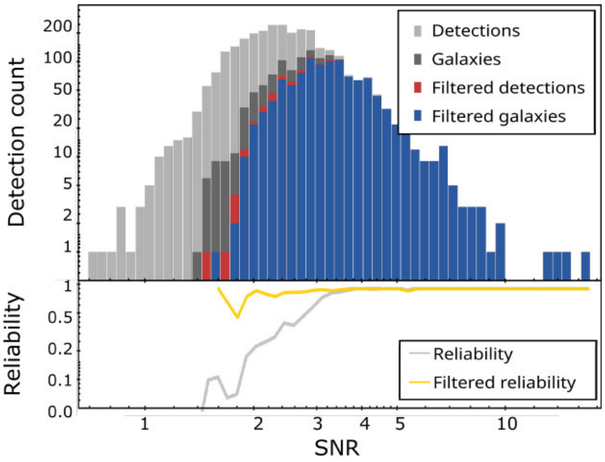

Figure 7. Team SOFIA: Histogram of total detections (light-grey), real galaxies (dark-grey), detections after filtering (red), and real galaxies after filtering (blue) as a function of integrated SNR from a SOFIA run on the development cube (top panel). The reliability of the original and filtered catalogue is shown as the grey and orange curve, respectively (bottom panel). Parameter space filtering significantly boosts SOFIA's reliability at low SNR. Note that we measure SNR within the actual SOFIA source mask, and the resulting values can not be directly compared with the optimized SNR defined in Section 6.2.

Development version 2.3.1 of the software, dated 2021 July 22,[10] was used in the final run submitted to the scoring service. To minimize processing time, 80 instances of SOFIA were run in parallel, each operating on a smaller region (≈11.8 GB) of the full cube. The processing time for an individual instance was just under 25 min, increasing to slightly more than 2 h when all 80 instances were launched at once due to overhead from simultaneous file access. The resulting output catalogues were merged and any duplicate detections in areas of overlap between adjacent regions discarded.

We ran SOFIA with with the following options: after flagging of bright continuum sources >7 mJy followed by noise normalization in each spectral channel, the S + C finder was run with a detection threshold of 3.8 times the noise level, spatial filter sizes of 0, 3, and 6 pixels and spectral filter sizes of 0, 3, 7, 15, and 31 channels. We adopted a linking radius of 2 and a minimum size requirement of 3 pixels/channels. Lastly, reliability filtering was enabled with a reliability threshold of 0.1, an SNR threshold of 1.5 and a kernel scale factor of 0.3.

Based on tests using the development cube, we improved the reliability of the resulting source catalogue from SOFIA by removing all detections with $n_{\mathrm{pix}} < 700$, $s < -0.00135 \times (n_{\mathrm{pix}} - 942)$ or $f > 0.18 \times \mathrm{SNR} + 0.17$, where n_{pix} is the number of pixels within the 3D source mask, s is the skewness of the flux density values within the mask, f is the filling factor of the source mask within its rectangular 3D bounding box, and SNR is the integrated SNR of the detection. Detection counts for the original and filtered catalogue from the development cube are shown in Fig. 7 as a function of SNR. Our final detection rate peaks at SNR ≈ 3, with a reliability of close to 1 down to SNR ≈ 2. The filtered catalogue from the full cube contains almost 25 000 detections, about 23 500 of which are real, implying a global reliability of 94.2 per cent.

[10] https://github.com/⟨0:sc ⟩Sofia⟨/0:sc⟩-Admin/⟨0:sc ⟩Sofia⟨/0:sc⟩-2/tree/11ff5fb01a8e3061a79d47b1ec3d353c429adf33

It should be emphasized that our strategy of first creating a low-reliability catalogue with SOFIA and then removing false positives through additional cuts in parameter space is based on development cube tests and was adopted to maximize our score. This strategy may not work well for real astronomical surveys which are likely to have different requirements for the balance between completeness and reliability than the one mandated by the scoring algorithm.

Lastly, the source parameters measured by SOFIA were converted to the requested physical parameters. As the calculation of disc size and inclination required spatial deconvolution of the source, we adopted a constant disc size of 8.5 arcsec and an inclination of 57.3 degrees for all spatially unresolved detections. In addition, statistical noise bias corrections were derived from the development cube and applied to SOFIA's raw measurement of integrated flux, line width and H I disc size.

5 SCORING

A live scoring service was provided for the duration of the Challenge. The service allowed teams to self-score catalogue submissions while keeping the truth catalogue hidden, and automatically updated a live leaderboard each time a team achieved an improved score. All participating teams were provided with credentials with which the scoring service could be accessed over the internet using a simple, pip-installable command line client. Participants used this client to upload submissions to the service, after which it was evaluated by a scoring algorithm against the truth catalogue. Once the score had been calculated, it could be retrieved from the scoring service using the client. Teams were limited to a maximum submission rate of 30 submissions per 24-h period.

5.1 Scoring procedure

The scoring algorithm[11] is written in PYTHON and makes use of the PANDAS and ASTROPY libraries. Scoring is performed by comparing submitted catalogues with a truth catalogue, each containing the same source properties. The first step of the scoring is to perform a positional cross-match between the true and the submitted catalogues. Matched sources from the submitted catalogue are then assigned scores according to the combined accuracy of all their measured properties. Finally, the scores of all matched sources are summed and the number of false detections subtracted, to give the overall Challenge score.

5.1.1 Source cross-match

Cross-matching is performed using the SCIKIT nearest neighbours classifier with the `kd_tree` algorithm, which uses a tree-based data structure for computational efficiency (Bentley 1975). The cross-match procedure considers the position of a source in the 3D cube, identified by RA, Dec, and central frequency. Each coordinate set is first converted to a physical position space via the source angular diameter distance. All submitted sources with positions within which a truth catalogue source is in range are then recorded as matches. For each submitted source, this range in the spatial and frequency dimensions is determined by the beam-convolved submitted H I size and the line width, respectively. Detections that do not have a truth source within this range are recorded as false positives. Matched detections are further filtered by considering the range of the matched truth sources. Detections which lie outside the beam-convolved H I size and the line width of the matched truth source are at this stage also rejected and recorded as false positives.

It is possible that the cross-match returns multiple submitted sources per true source. In that case, all matches are retained and scored individually. The reasoning behind this choice is that components of H I sources, especially in the velocity field, could be correctly identified but interpreted as separate sources. If that were the case, classifying them as false positives would be too much of a penalty. All submitted sources matched to the same true source are inversely weighted by the number of multiple matches during the scoring step. It is also possible for more than one truth source to be matched with a single submitted source. In these cases, only the match between the submitted source and truth source which yields the lowest multiparameter error (equation 16) is retained. This procedure ensures that matches in crowded regions take into account the resemblance of a truth source to a submitted source, in addition to its position.

A final step is performed to compare the multidimensional error with a threshold value, above which any nominally matched submitted sources are discarded and counted as false positives. The multiparameter error D is calculated using the Euclidean distance between truth and submitted sources in normalized parameter space,

$$D = (D^2_{\rm pos} + D^2_{\rm freq} + D^2_{\rm HI\,size} + D^2_{\rm line\,width} + D^2_{\rm flux})^{\frac{1}{2}}, \tag{16}$$

where the errors on parameters of spatial position, central frequency, line width and integrated line flux have been normalized following the definitions in Table 2. The error on H I size is at this stage normalized by the beam-convolved true H I size in order not to lead to the preferential rejection of unresolved sources. The multidimensional error threshold is set at 5, i.e. the sum in quadrature of unit normalized error values.

5.1.2 Accuracy of sources properties

For all detections that have been identified as a match, properties are compared with the truth catalogue and a score is assigned per property and per source. The following properties are considered for accuracy: sky position (RA, Dec), H I size, integrated line flux, central frequency, position angle, inclination angle, and line width. Each attribute j of a submitted source i contributes a maximum weighted score w_i^j of 1/7, so that the maximum weighted score w_i for a single matched source is 1,

$$w_i = \sum_{j=1}^{7} w_i^j. \tag{17}$$

The weighted score of each property of a source is determined by

$$w_i^j = \frac{1}{7} \min\left\{1, \frac{{\rm thr}_j}{{\rm err}_i^j}\right\}, \tag{18}$$

where ${\rm err}_i^j$ is the error on the attribute and ${\rm thr}_j$ is a threshold applied to that attribute for all sources. Errors calculated in this step are detailed in Table 2, along with corresponding threshold values, which have been chosen using the distribution of errors obtained during tests on the Challenge data products using the SOFIA source finder. Finally, the weighted scores of submitted sources are averaged over any duplicate matches with unique truth sources.

5.1.3 Final score per submission

The final score is determined by subtracting the number of false positives $N_{\rm f}$ from the summed weighted scores w_i of all $N_{\rm m}$ unique

[11] https://pypi.org/project/ska-sdc/

Table 1. The main features of the methods applied by each team to SDC2 are summarized for ease of reference.

Team name	Pre-processing	Detection	False-positive rejection	Characterization	Additional notes
Coin	RFI flagging	3D U-Net CNN Interscale connectivity	Size cuts Continuum rejection	ResNet CNNs Ellipse-fitting	Several CNNs tested
EPFL	Wavelet filtering	Joint likelihood	Size cut Classifier CNN	Inception CNN	Data augmentation
FORSKA-Sweden	–	3D U-Net CNN	SOFIA	SOFIA Modelling: check	–
HI-FRIENDS	SOFIA: Continuum flagging Noise normalization	SOFIA	SOFIA Additional parameter cuts	SOFIA Ellipse fitting	–
HIRAXers	U^2 net	Peak-finding	–	HighRes3DNet	Data augmentation
JLRAT	–	CNN Cross-correlation	–	Gaussian-fitting	Spectral inputs to CNN
MINERVA*	–| Smoothing | SNR mask	YOLO CNN| Friend-of-friend	–| CNN	YOLO CNN| CNNs	Training data refinement Data augmentation
NAOC-Tianlai	SOFIA: Continuum flagging Noise normalization	SOFIA	Parameter tuning	SOFIA	Gridsearch, MCMC
SHAO	–	SEXTRACTOR TOPCAT	–	SEXTRACTOR Gaussian fitting	–
Spardha	SOFIA: Continuum flagging Noise normalization	SOFIA	SOFIA Additional parameter cuts	SOFIA	Partitioning buffer zones
Starmech	SOFIA: Continuum flagging Noise normalization	SOFIA	SOFIA	SOFIA	TOPCAT for verification
Team SOFIA	SOFIA: Continuum flagging Noise normalization	SOFIA	SOFIA Additional parameter cuts	SOFIA	Noise bias corrections

The methodology is divided into pre-processing, source finding, false-positive rejection, and source characterization steps. The asterisk denotes the step taken by team MINERVA to combine the results of two independent methods, demarcated here by the pipe symbol, to form a final catalogue.

Table 2. Definitions of errors and threshold values for the properties of sources.

Property	Error term	Threshold
RA and Dec, x, y	$D_{\rm pos} = \frac{(x-x')^2+(y-y')^2}{\hat{S}'}$	0.3
H I size, S	$D_{\rm HI\,size} = \frac{\lvert S-S'\rvert}{\hat{S}'}$	0.3
Integrated line flux, F	$D_{\rm flux} = \frac{\lvert F-F'\rvert}{F'}$	0.1
Central frequency, ν	$D_{\rm freq} = \frac{\lvert \nu-\nu'\rvert}{w'_{20,\,{\rm Hz}}}$	0.3
Position angle, θ	$D_{\rm PA} = \lvert\theta-\theta'\rvert$	10
Inclination angle, i	$D_{\rm incl} = \lvert i-i'\rvert$	10
Line width, w_{20}	$D_{\rm line\,width} = \frac{\lvert w_{20}-w'_{20}\rvert}{w'_{20}}$	0.3

Prime denotes the attributes of the truth catalogue, x, y are the pixel coordinates corresponding to RA, Dec, ν is the central frequency, S is the H I major axis diameter and$\hat{S}$ is the beam-convolved major axis diameter, f is the source integrated line flux, θ is the position angle, i is the inclination angle, and w_{20} is the H I line width. Calculations of position angles take into account potential angle degeneracies by defining the angle difference as a point on the unit circle and taking the two-argument arctangent of the coordinates of that point: $|\theta - \theta'| = {\rm atan2}[\sin(\theta - \theta'), \cos(\theta - \theta')]$.

matched sources:

$$\text{final score} = \sum_{i}^{N_{\rm m}} w_i - N_{\rm f}. \tag{19}$$

False positives are linearly penalized in order to preserve equal weighting between characterization performance and the ability to remove false detections.

5.2 Reproducibility awards

Participating teams were encouraged to consider early on in the Challenge the overall architecture and design of their software pipelines. At the Challenge close, teams were invited to share pipeline solutions. Reproducibility awards were then granted in acknowledgement of those teams whose pipelines demonstrated best practice in the provision of reproducible results and reusable methods. Pipelines were evaluated using a checklist developed in partnership with the Software Sustainability Institute (SSI)[12] (Crouch et al. 2013), which was provided to teams for the purposes of self-assessment during the Challenge. The checklist[13] considered the following criteria:

Reproducibility of the solution. Can the software pipeline be re-run easily to produce the same results? Is it:

[12] https://www.software.ac.uk/

[13] https://sdc2.astronomers.skatelescope.org/sdc2-challenge/reproducibility-awards

Table 3. SDC2 finalist teams' scores are reported, rounded to the nearest integer.

Team name	Score	N_d	N_m	R	C	A
MINERVA	23 254	32 652	30 841	0.945	0.132	0.81
FORSKA-Sweden	22 489	33 294	31 507	0.946	0.135	0.77
Team SOFIA	16 822	24 923	23 486	0.942	0.101	0.78
NAOC-Tianlai	14 416	29 151	26 020	0.893	0.112	0.67
HI-FRIENDS	13 903	21 903	20 828	0.951	0.089	0.72
EPFL	8515	19 116	16 742	0.876	0.072	0.65
Spardha	5615	18 000	13 513	0.751	0.058	0.75
Starmech	2096	27 799	17 560	0.632	0.075	0.70
JLRAT	1080	2100	1918	0.913	0.008	0.66
Coin	−2	29	17	0.586	0.000	0.60
HIRAXers	−2	2	0	0.000	0.000	–
SHAO	−471	471	0	0.000	0.000	–

Also reported are the number of detections N_d and matches N_m (Section 5.1.1), and the overall reliability (R; equation 20) and completeness (C; equation 21) of each method. Finally, the source characterization accuracy (A; equation 22) reports the percentage accuracy of source property measurement averaged over all properties for all sources matched per team.

(i) Well-documented
(ii) Easy to install
(iii) Easy to use

Reusability of the pipeline. Can the code be reused easily by other people to develop new projects? Does it:

(i) Have an open licence
(ii) Have easily accessible source code
(iii) Adhere to coding standards
(iv) Utilize tests

All parts of the software pipeline developed by each team were evaluated, including packages that the teams have written and code that interacts with third party packages, but not including any third party packages themselves.

6 RESULTS AND ANALYSIS

In this Section we first present the overall Challenge results before reporting on the determination of source signal-to-noise values. We then analyse the results from source finding and characterization perspectives and present the results of the reproducibility awards.

6.1 Challenge results

The final scores of all teams who submitted a catalogue based on the full Challenge data set are reported in Table 3. Each team's number of detections, N_d – composed of matches, N_m, and false positives, N_f – are also listed, along with the number of matches, the overall reliability, R, and completeness, C, calculated as follows:

$$R = \frac{N_\mathrm{m}}{N_\mathrm{d}} = \frac{N_\mathrm{m}}{N_\mathrm{m} + N_\mathrm{f}}; \tag{20}$$

$$C = \frac{N_\mathrm{m}}{N_\mathrm{t}}, \tag{21}$$

where N_t is the number of sources in the truth catalogue. The overall characterization accuracy of each team's method, A, is defined as the accuracy of source property measurement according to Section 5.1.2, averaged over all properties for all matches per team:

$$A = \frac{\sum_i^{N_\mathrm{m}} w_i}{N_\mathrm{m}}. \tag{22}$$

We note that the scoring algorithm (Section 5), designed to penalize false detections, can result in a teams' highest scoring submission containing a significantly less complete catalogue than other submissions made by the same team if the number of false positives is high. This is the case for teams Coin, HIRAXers, and SHAO. With each teams' agreement therefore we have used the team's submission with the highest completeness for the following analysis, while leaving the leaderboard scores unchanged. This allows us more robustly to investigate the characterization performance of these teams' methods.

6.1.1 *Conventions and units*

Several conventions and conversions are used during the characterization of H I spectral line data which, without clear and unambiguous specification, can lead to inconsistencies between catalogues and between physical and measured properties. Room for error arose due to potential alternative position angle definitions and to the need to shift the rest frequency into the frame of the source. Where teams' catalogues have followed alternative conventions or incorrect conversions, catalogue corrections have been applied after the close of the Challenge leaderboard. While teams' scores are affected slightly, leaderboard positions do not change. The Challenge organizing team used the dedicated discussion forum (Section 2.1) to resolve misunderstandings in the rules and conventions as they arose. Future SKAO Science Data Challenges will benefit from additional instructions and examples where ambiguity or unfamiliarity can be anticipated. The reporting of observed rather than derived parameters would also reduce measurement inconsistencies.

6.2 Signal-to-noise

The appropriate definition and calculation of source signal-to-noise values is important in order to gain an understanding of the absolute performance of teams' methods and to transfer insights gained from SDC2 to other data sets. While the value of peak signal-to-noise is easy to define, it fails to capture any information about source extent. Alternatively, the integrated signal-to-noise can be evaluated for a chosen mask across the source. The total error contribution from the mask pixels can be calculated using the usual rules of correlated error propagation. However, due to the smoothing effect of beam sampling, the amount of true signal contained within a finite mask cannot be determined. Further, the application of smoothing kernels – routinely used in signal processing problems to boost signal with respect to noise – results in modification to the signal-to-noise properties of a given source. For the purpose of this analysis therefore we use a signal-to-noise definition based on the peak signal of a smoothed source. The definition adopted for this paper is intended to provide the most helpful insight into SDC2 results, but is not necessarily the best choice for other data sets.

A given signal in the presence of additive white Gaussian noise can be maximized with respect to the noise by applying a smoothing filter matched to the signal. In this case, the matched filter optimizes the trade-off between noise-suppression and signal-suppression. In the case of an SKA-observed spatial noise field, logarithmic spacing of the array configurations results in a relatively uniform sensitivity, in units of Jy per beam, across a wide range of angular scales (Braun et al. 2019). This property is evident upon Gaussian smoothing of the SDC2 simulated spatial noise field, which sees a slight reduction in beam-normalized r.m.s. noise to an approximately

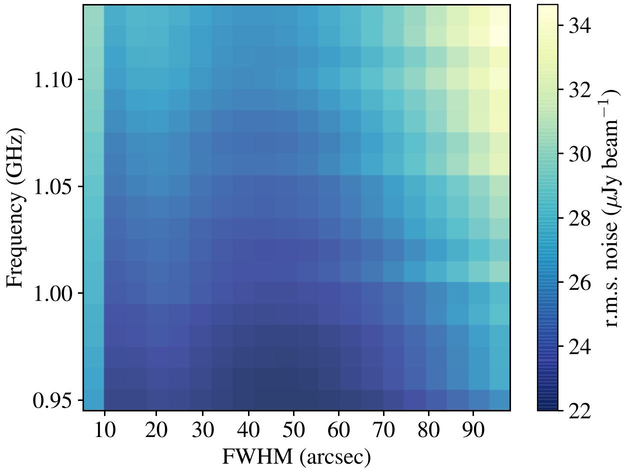

Figure 8. The r.m.s noise of the 2000h SKA-Mid 20 square degree noise field is plotted as a function of frequency and of smoothing. The smoothing FWHM presented is the result of adding in quadrature the 7 arcsec beam FWHM and the FWHM of a Gaussian smoothing filter applied to the field.

constant level between angular ranges ∼10 and 80 arcsec FWHM (see Fig. 8, which presents r.m.s. noise as a function of total spatial smoothing and frequency for a simulated 2000-h SKA-Mid observation of a 20 square degree field). The signal-to-noise of a source observed using the SKA can therefore be maximized in the spatial dimensions simply by applying a sufficiently large Gaussian smoothing kernel, provided that the source itself is no larger in spatial extent than the angular range of uniform sensitivity. Fig. 9 presents the effect on signal-to-noise of smoothing an SKA-observed Gaussian source using a range of Gaussian smoothing kernels.

For each SDC2 source, an SNR lue was obtained by first selecting the minimum r.m.s noise value, $\sigma_{\mathrm{rms},\,\nu}$, achieved by smoothing the SKA noise field at the source central frequency, ν, with a Gaussian smoothing kernel. The total smoothing scale is obtained by adding in quadrature the FWHM of the corresponding smoothing kernel to the SKA beam FWHM. Making the assumption that the spatial extent of the source is smaller than the total smoothing scale, such that the integrated source flux density per channel i would equal the peak value of the smoothed source per channel, the source pixel values were integrated over spatial dimensions to produce a spectral profile, $S(i)$. A tophat filter was then applied to the source spectral profile,

$$S'(i) = \frac{1}{k}\sum_{u=0}^{k-1} S(i-u), \tag{23}$$

and to a 1D white Gaussian noise field $N(i)$ with standard deviation equal to $\sigma_{\mathrm{rms},\,\nu}$,

$$N'(i) = \frac{1}{k}\sum_{u=0}^{k-1} N(i-u). \tag{24}$$

The size of the tophat filter, k, was chosen to equal the number of channels in the spectral profile with values greater than 10 per cent of the maximum value. The final SNR value,

$$\mathrm{SNR} = \frac{S'_{\mathrm{max}}}{\sigma'_{\mathrm{rms}}}, \tag{25}$$

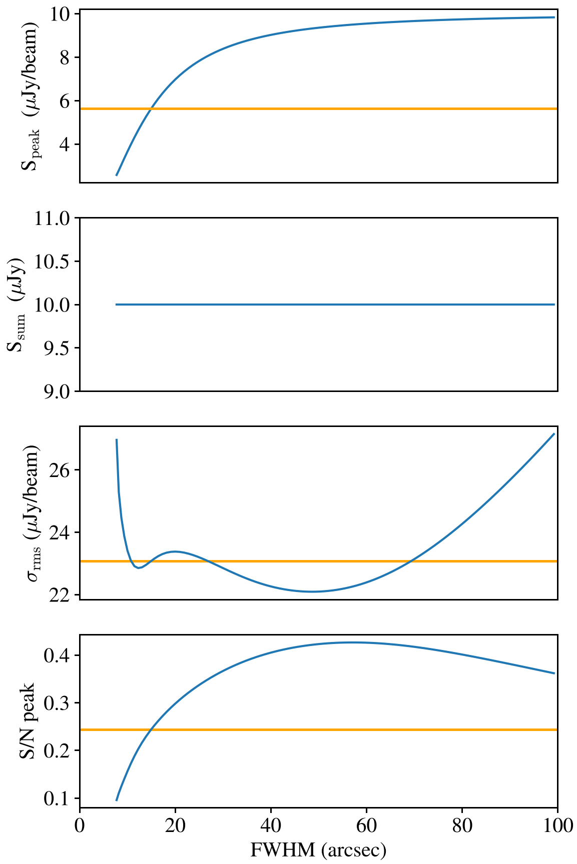

Figure 9. A simulated circular Gaussian source of FWHM 14 arcsec is convolved with a circular Gaussian 'beam' of 7 arcsec FWHM and used to illustrate signal-to-noise characteristics of the SKA-Mid field as a function of smoothing. A series of Gaussian smoothing filters is applied both to the beam-convolved source and to a simulated noise field representing 2000 h of Band 2 SKA-Mid observations of a 20 square degree field. The beam FWHM and smoothing FWHM are added in quadrature to obtain the total smoothing FWHM, which is represented by the abscissa. From top, in blue: peak smoothed source flux density; total source flux density; r.m.s. noise of the noise field; peak SNR obtained using the peak smoothed source flux density and the r.m.s noise. The orange horizontal line represents the values obtained by applying instead a filter matched to the source.

was calculated using the maximum value of the filtered spectral profile,

$$S'_{\mathrm{max}} = \max\{S'(i)\}, \tag{26}$$

and the r.m.s. value of the filtered Gaussian noise,

$$\sigma'_{\mathrm{rms}} = \sqrt{\langle N'(i)\rangle}. \tag{27}$$

Fig. 10 presents binned SNR values of all sources in the full SDC2 truth catalogue.

6.3 Source finding

Fig. 11 presents for each team the number of final matches and false positives, binned according to integrated line flux along with all sources from the truth catalogue. When considering matches, truth catalogue line flux values are used; when considering false positives, the lack of corresponding truth values necessitates the use of submitted line flux values. Fig. 12 presents reliability and completeness values as a function of integrated line flux, where

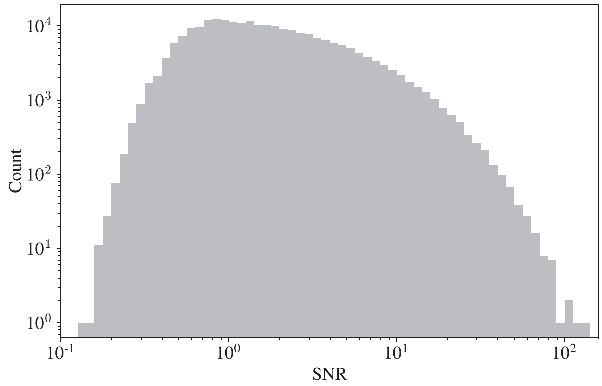

Figure 10. Truth catalogue sources are binned according to SNR values (see Section 6.2 for a description of the signal-to-noise calculation).

submitted values are again used in the calculation of reliability due to the absence of corresponding truth values for false positives. Fig. 12 also presents completeness as a function of SNR values.

6.4 Source characterization

In order to investigate the performance of teams' methods in the recovery of source properties, several relationships were investigated. Fig. 13 presents error terms (Table 2) calculated without using absolute values and plotted as a function of true property value and of SNR for flux, of true property value for size and line width measurements, and as a function of true size, for position and inclination angle measurements. Fig. 14 presents overall source characterization accuracy as a function of SNR. Characterization accuracy is determined according to Section 5.1.2, averaged over all properties except position in RA, Dec, and central frequency, for all matches per team in the given SNR interval.

Fig. 15 compares H I mass distributions constructed using teams' matched sources with the function constructed by taking the input redshift-dependent H I mass function $\phi(M_{\rm HI})$ (equation 1) and multiplying by the sky volume covered by a given redshift interval. True H I masses generated during our simulation (Section 3) were used to obtain for each team an H I mass distribution, $N_{\rm m}(M'_{\rm HI})$, by counting matched sources that fall within a logarithmic bin centred on true mass value $M'_{\rm HI}$.

A second H I mass distribution, $N_{\rm m}(M_{\rm H\,I})$, was constructed using submitted property values, F and ν, of teams' detections, which were converted to mass according to equation (2) (Duffy et al. 2012). The same conversion was applied to the full truth catalogue to produce the complete H I mass distribution, $N^C_{\rm m}(M'_{\rm HI})$, which was used to verify consistency between the input mass function and simulated observables.

Submitted and true values of teams' matches and detections, respectively, were used to plot the residual,

$$\Delta N_{\rm m}(M_{\rm HI}) = N_{\rm m}(M_{\rm HI}) - N_{\rm m}(M'_{\rm HI}), \tag{28}$$

after applying a second order spline interpolation to both distributions.

For each team, the H I mass distribution derived from true mass values, $N_{\rm m}(M'_{\rm HI})$, was interpolated and compared with the input H I mass distribution, $N_{\rm m}(M_{\rm H\,I})$, in order to identify the H I mass above which at least 50 per cent of truth catalogue sources are recovered (Table 4). Fig. 16 presents this mass for the top eight scoring teams as a function of redshift and compared with the H I mass function 'knee' mass (equation 1).

6.5 Reproducibility awards

Six teams submitted entries for the SDC2 reproducibility awards. Each pipeline was evaluated by an expert panel against the pre-defined award criteria (Section 5.2). Table 5 reports the awards granted to each team.

7 DISCUSSION

Challenge teams employed a variety of methods to tackle the simulated SKA-Mid H I data set. In this section we discuss the findings in terms of individual and collective method capabilities.

7.1 Source finding and characterization

The overall results (Table 3) show a wide range of performance both within and between methods. Reference to Table 1 indicates that strategies for false positive rejection are important. Further, the use of a refined training data set, as employed by team MINERVA, may be crucial.

While reliability and completeness (Fig. 12) generally show an increase with increasing flux and SNR, several teams show a drop-off at the brighter flux end. This is partly explained by a low number of sources, resulting in statistical noise. Reliability, in addition, will be particularly affected by the presence of brighter artefacts arising from imperfect continuum subtraction. Unreliability could in turn lead to a lower level of completeness in the corresponding flux bin, if source-finding methods themselves become correspondingly uncertain. For the top two scoring teams, a completeness of at least 50 per cent is achieved down to a limit of SNR∼5 and an integrated flux limit of ∼20 Jy Hz.

The analysis of individual source property recovery (Fig. 13) finds that of all properties, position angle is the most difficult to recover, with a standard deviation on the errors often covering most of the position angle range. This is understandable considering the large fraction of partially unresolved sources, and some teams are able to recover position angle well for resolved source sizes. Inclination angle, which gives rise to the radial velocity for a given rotational velocity (equation 6), and can therefore be approximated by making use of line width, flux, and size measurements, does not suffer the same problem. Source characterization could be improved by choosing a suitably high detection threshold. For example, analysis of characterization accuracy as as function of SNR (Fig. 14) finds a clear trend. The winning team, MINERVA, dominates across most of the SNR range, maintaining an average accuracy above 0.8 from SNR ∼10 to 60, and remaining around 10 per cent higher than the next team from SNR ∼3 to 60. At the very highest SNR, however, Team SOFIA achieved the greatest averaged accuracy, while the MINERVA performance falls slightly.

7.1.1 Noise biases

The analysis of integrated line flux measurements finds in general a positive excess at lower values. This demonstrates the problem of so-called 'flux boosting' as a result of increasing number counts in the presence of local noise fluctuations (Hogg & Turner 1998). In terms of SNR, flux boosting becomes apparent at SNR∼7 but remains minimal for the top three scoring teams, which see a flux boosting effect of ∼40 per cent at SNR = 3. Similar noise biases may be apparent in the measurement of H I size and line width, where there is a general tendency to overestimate smaller sizes and underestimate larger sizes. Some teams used the SDC2 development data set to

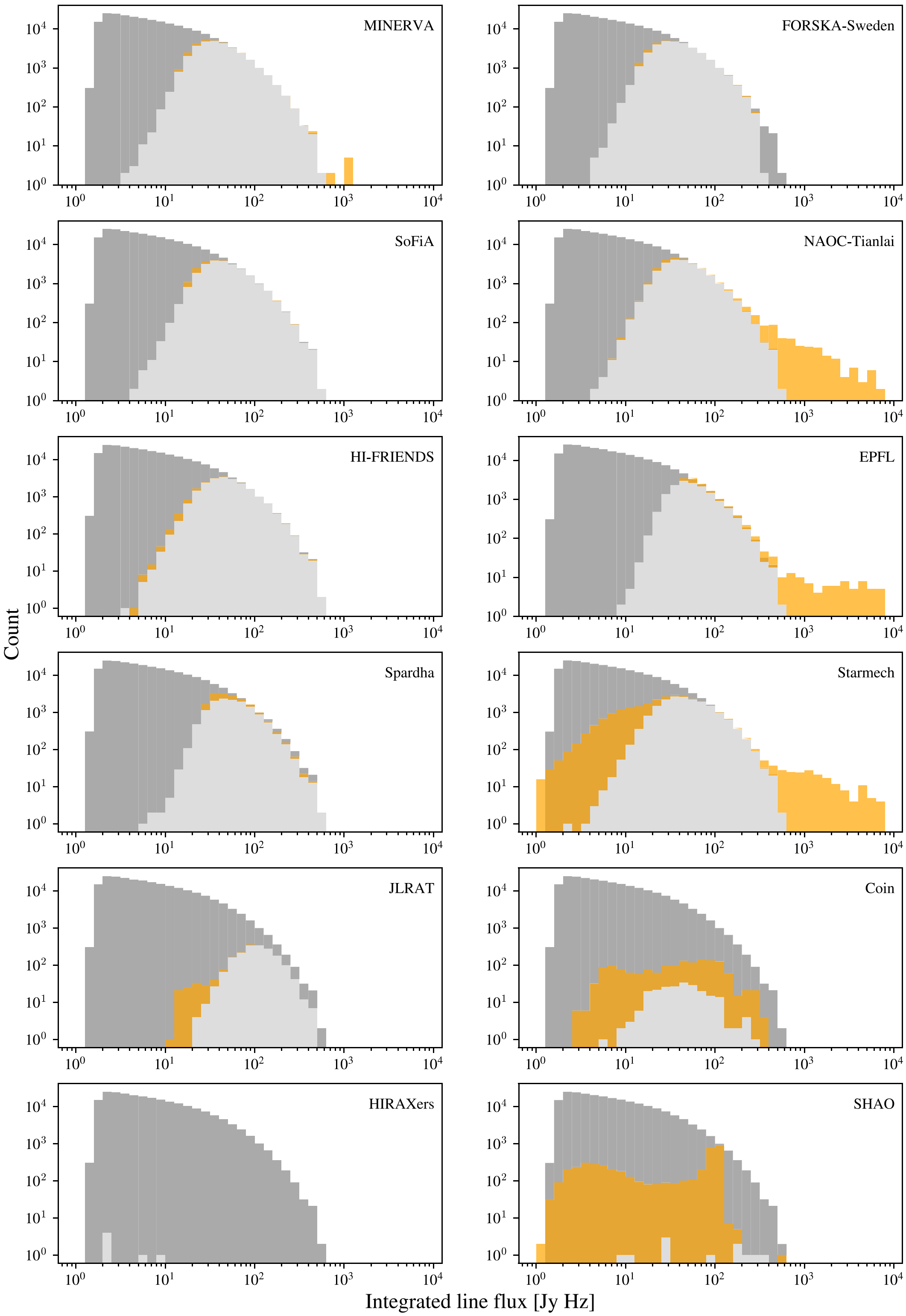

Figure 11. Sources in the full Challenge data set binned according to integrated line flux value. For each team, all sources in the full truth catalogue (dark grey) are overplotted by the true values of matches (light grey) and by the submitted values of false detections (yellow).

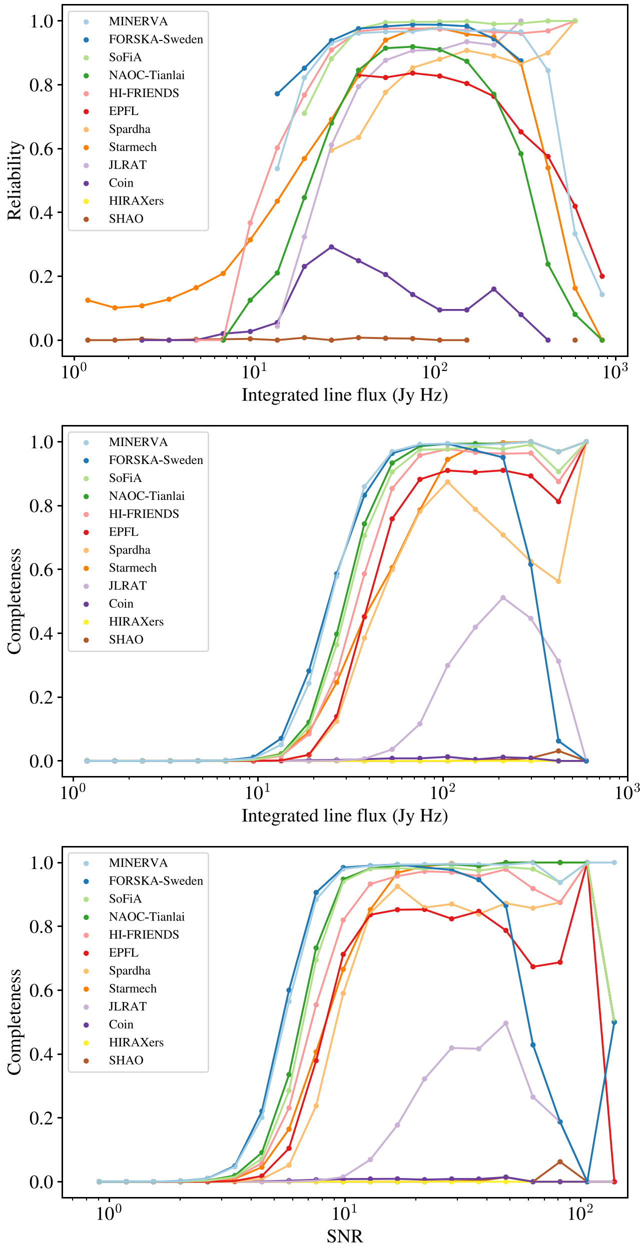

Figure 12. Top: Reliability, defined as the number of matches divided by the number of detections, is plotted for each team as a function of submitted integrated line flux. Middle: Completeness, defined as the number of matches divided by the number of truth catalogue sources, is plotted for each team as a function of true integrated line flux. Bottom: Completeness is plotted for each team as a function of true SNR (see Section 6.2 for a description of the chosen signal-to-noise definition).

calibrate pipeline output against the available truth catalogue. For example, team SOFIA used polynomial fits to affected parameters as a function of flux, in order to derive corrections for flux, H I size, and line width. While corrections can remove the bias, intrinsic scatter, which is likely to be considerable at low SNR, will remain (see e.g. Hogg & Turner 1998). The overestimation of H I size is compounded by the finite resolution of the simulated observation: the fractional error on H I size understandably rises steeply as the true size decreases below the 7 arcsec beam size. Despite this limitation, some teams are significantly more accurate in constraining the source size limit.

7.1.2 *H I mass recovery*

The H I mass distributions presented in Fig. 15 are constructed without making corrections for survey sensitivity, which is a non-trivial task that falls outside the scope of the Challenge. Our analysis is therefore intended to demonstrate the depth of H I mass that can be probed by respective methods, and the discrepancy that may arise between number counts of observed and intrinsic masses of detected sources.

A 50 per cent completeness threshold was chosen to characterize H I mass recovery depths following Rosenberg & Schneider (2002), who, using an H I-selected galaxy sample from the Arecibo Dual-Beam Survey (Rosenberg & Schneider 2000), found a negligible difference between the mass function derived using only sources above the 50 per cent ‘sensitivity limit’ and the function derived using all sources. Fig. 16 demonstrates that the two top scoring teams’ methods are able to probe the H I knee mass with a 50 per cent completeness out to a redshift of approximately 0.45, or 1740 Mpc of comoving distance. For comparison, the ALFALFA survey – ith a footprint of ∼6900 deg^2 – has probed the knee mass out to distances of approximately 200 Mpc.

With the caveat that line width completeness corrections have not been performed on the mass distributions constructed using teams’ submitted values, we use Fig. 15 also to demonstrate the relative error between distributions constructed using the true and submitted values of teams’ detections. The top three scoring teams attain a relatively high degree of accuracy for detected sources, each seeing an overestimation in the mass distribution of less than 0.1 dex at the point where completeness falls below 50 per cent.

7.2 ML vs non-ML

Supervised ML methods, particularly CNN, proved a popular technique during the Challenge, and featured in the pipelines of the two top scoring teams. Of particular note is the significant success by winning team MINERVA in both the finding and characterization parts of the Challenge. The winning technique, which used ML both to find and characterize sources, achieved a 10 per cent improvement over the next team in characterization accuracy across a SNR range ∼4–30. Methods involving traditional signal processing techniques also achieved high scores, including the SOFIA package, which was used not only by the third-placed team of its developers, but also in the source characterization of the second placed team and by several others.

7.2.1 *Generalization*

The results demonstrate the promise of ML in the analysis of very large and complex data sets. As seen in similar community challenges (e.g. Metcalf et al. 2019), ML methods are often able to outperform traditional methods. This success is not without its caveats. In order for supervised ML models to transfer successfully to real data, they must be able to generalize beyond the parameter distribution that has been sampled by the training data (Burges 1998). Overfitting by models with large numbers of parameters can be avoided using a sufficiently large set of training data. A more difficult problem is that of covariate shift: when the distributions of training and real data sets are intrinsically different. This is a common issue for astronomy (see e.g. Freeman, Izbicki & Lee 2017; Luo et al. 2020; Autenrieth et al. 2021), where techniques are often being developed in preparation for data that is yet to be recorded. Models are instead trained using simulated data, which cannot capture unknown characteristics of the

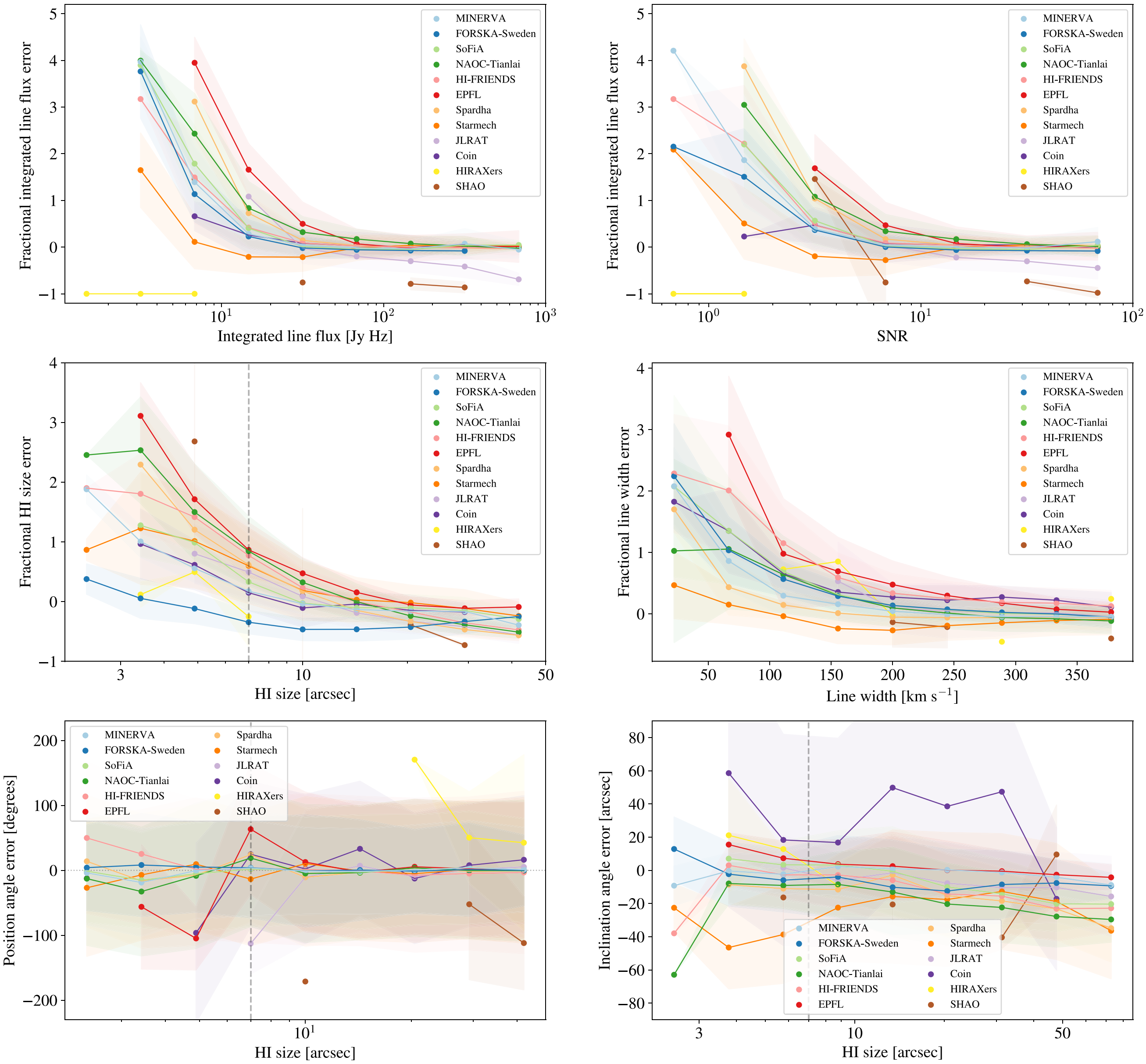

Figure 13. Error terms (see Table 2), calculated without using absolute values, are plotted as a function of true property value, SNR, or spatial source size. Joined circles represent the median error per logarithmic bin, the filled regions represent the standard deviation of the error, and all plots use teams' matched submissions. A dashed line represents the beam size of the simulated observations.

future observations. Limitations to the realism of the SDC2 data products (Section 3.4) are likely in turn to introduce limitations in the ability of SDC2 ML models to transfer to real data. An increased number of real H I observations used to generate the H I emission cube will reduce the risk of model overfitting. Further characterization of RFI and other instrumental effects during the commissioning phase of the SKAO telescopes will enable the simulation of ever more realistic data sets for training purposes, and transfer learning (Pan & Yang 2009; Tang, Scaife & Leahy 2019) could close the gap further still. In future SDCs, the inclusion of a data product produced using a different distribution could provide a test for model robustness to covariate shift.

Non-ML methods, generally making use of far fewer parameters than ML models and less reliant on the availability of training data, may transfer more successfully from simulated to real data. This advantage appears to be evidenced by the comparative successes of team-SOFIA and HI FRIENDS – both of which used the SOFIA software package – at the brighter end of the integrated flux and SNR ranges across reliability, completeness, and characterization accuracy (see Figs 11, 12, and 14). By contrast, the ML-based pipelines used by teams MINERVA and FORSKA-Sweden have produced a number of false positives and false negatives, respectively, in the detection of the very brightest sources. The ML-based pipelines also appear to show a fall in characterization accuracy at the very highest SNR values. It is possible that the paucity of very bright samples in the training data sets has prevented ML methods from modelling very well the features of the brightest sources. On the other hand, it is likely that the small number of bright samples in the Challenge data set has led to the prioritization during pipeline optimization of greater accuracy for fainter populations, since the

Table 4. The H I mass (in units of 10^9 $M_\odot$) above which at least 50 per cent of truth catalogue sources are recovered is reported per redshift interval for the SDC2 finalist teams.

Team name	Redshift interval				
	0.25 −0.30	0.30 −0.35	0.35 −0.40	0.40 −0.45	0.45 −0.50
MINERVA	2.60	3.82	5.27	7.12	10.04
FORSKA-Sweden	2.52	3.80	5.15	6.91	9.57
Team SOFIA	3.32	4.77	6.68	8.52	11.59
NAOC-Tianlai	3.12	4.67	6.33	8.40	11.69
HI-FRIENDS	3.67	5.37	7.51	9.94	13.55
EPFL	4.14	6.10	8.45	11.21	17.60
Spardha	4.78	6.98	9.47	12.55	20.91
Starmech	3.97	6.52	9.41	12.44	20.22
JLRAT	–	46.03	–	46.77	72.57
Coin	–	69.44	–	70.11	72.52
HIRAXers	–	–	–	–	–
SHAO	–	–	–	–	–

large number of fainter sources produce a much greater impact on the score.

7.3 Method complementarity

The strategy employed by winning team MINERVA underscores one of the most important outcomes of the Challenge: that of method complementarity. By combining the outputs of two independent pipelines the teams were able to recover sources from a larger amount of the flux–line width parameter space than by using a single pipeline alone (Fig. 5), and could further exploit the independence of the pipelines to reduce bias and variance in source measurements. The success of this strategy demonstrates that, given a selection of sufficiently independent and well-performing methods, stacking – where the predictions made by a group of independent ML methods are used as inputs into a subsequent learning model – could improve generalization from training data to new data (see also Wolpert 1992; Zitlau et al. 2016; Alves 2017).

The promise of a multimethod approach is further demonstrated by the performance of different methods in different aspects of the Challenge. Teams Starmech and Coin, for example, though occupying the lower half of the leaderboard, performed particularly well in the recovery of line flux and H I size, respectively (Fig. 13). Teams NAOC-Tianlai, HI-FRIENDS, EPFL, though missing out on the top three positions of the leaderboard, all demonstrated a high accuracy in the recovery, variously, of flux, source size, and inclination angle. HI-FRIENDS also achieved highest overall reliability, while Team ForSKA, a very close second on the leaderboard, achieved the highest level overall completeness (Table. 3). If the measurement of source properties is considered a separate problem from source finding, and the measurement of different source properties considered a many-problem task in itself, then a so-called bucket-of-models approach (Kim, Brunner & Carrasco Kind 2015) could harness the capabilities of different methods to further improve performance beyond any individual method.

7.4 Scoring metrics

In the case of SDC2, the scoring algorithm has been designed to evaluate source finding and characterization performance together. We note that the choice of any scoring metric will necessarily have an impact on the analysis that teams will perform. Strategies designed to maximize such a score might not be the best ones for other scientific goals: a search for fewer, highly resolved sources will take a very different approach from one aiming to produce a complete catalogue. The Challenge leaderboard score, if looked at in isolation, can obscure strong performance by teams on source characterization. This is a consequence of the strong penalty for false positives. Given the strong degree of method complementarity, a challenge scoring system that can reflect specialized solutions to a problem may further exploit complementarity as a quality of a collection of independent methods.

7.5 Open Science

The SDC2 reproducibility awards were designed to recognize Open Science best practice in the preparation and dissemination of analysis software pipelines. By providing public access to codes written to address SDC2, six teams were able to enhance the reproducibility and reusability of their methods. Noteable examples of best practice included the use of clear and comprehensive documentation, quick-start examples, command line interface excerpts, open-source licensing, and descriptive variable names. Practices employed by the Gold-standard HI-FRIENDS pipeline included the use of the workflow management system SNAKEMAKE (see Section 4.4) to design the overall workflow and suggest well-structured code directories, to manage the installation of software dependencies, and to generate a workflow graph image, all of which support the reusability and portability of the code. The advantages of well-documented and easily accessible codes are underscored by the popularity during the Challenge of the publicly available and regularly maintained SOFIA package, which was used by six of the participating teams.

Reproducible and reusable analysis pipelines help to address some of the challenges of conducting research under a deluge of data while leveraging the many new technologies available to deal with the data. However, preparing software for public access can require a significant time investment. As we look ahead to the exascale era of data (Scaife 2020), adequate funding to allow for software package maintenance and development will be essential.

7.6 Data handling

Teams were able to handle the large Challenge data set with minimal difficulty thanks to the generous provision of computational resources by the SDC2 partner facilities (Section 2.2). By dividing the data set into smaller portions and running parallelized codes, teams could comfortably process the full Challenge data set in under 24 h of wall clock time. Efficiency savings will become ever more important as volumes of observational data grow and analysis pipelines proliferate; the use of fewer resources to analyse data will not only allow future SKA Regional Centres to support a greater number of researchers, but will also reduce energy consumption during processing.

7.7 Lessons learned

We summarize here the opportunities for improvement in Challenge delivery that would further support the achievement of the overall goals of the SDC series:

(i) Additional guidance for the use of radio astronomy convention and conversions (see Section 6.1.1).

(ii) Consideration of the use of multiple scoring metrics to reflect different aspects of a challenge (see Section 7.4).

Table 5. Reproducibility awards were made to six teams who submitted pipelines demonstrating best practice in the provision of reproducible results and reusable methods.

Team name	Reproducibility award	Pipeline
EPFL	Bronze	https://github.com/epfl-radio-astro/LiSA
FORSKA-Sweden	Silver	https://github.com/FraunhoferChalmersCentre/ska-sdc-2
HI-FRIENDS	Gold	https://github.com/HI-FRIENDS-SDC2/hi-friends
NAOC-Tianlai	Bronze	https://github.com/kfyu/SDC2-tianlai
SHAO	Bronze	https://github.com/astrosumit/SDC2-SHAO
Team SOFIA	Silver	https://github.com/⟨0:sc ⟩Sofia⟨/0:sc⟩-Admin/SKA-SDC2-⟨0: sc ⟩Sofia⟨/0:sc⟩

Entries were evaluated by an expert panel using a pre-defined set of criteria (Section 5.2).

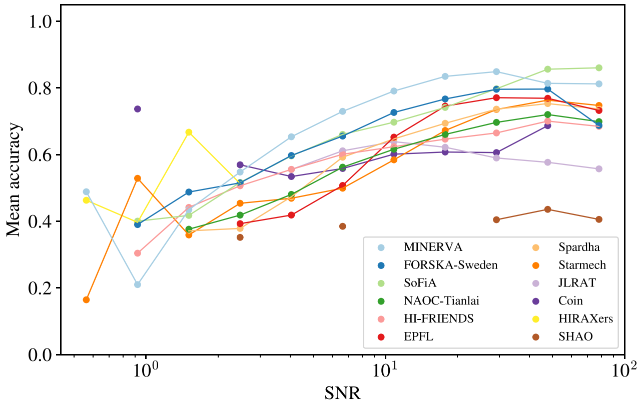

Figure 14. Source characterization as a function of SNR. Source accuracy is determined according to Section 5.1.2, averaged over all properties except position in RA, Dec and central frequency, for all matches per team in the given SNR interval.

(iii) A smaller set of criteria for a reproducibility component of a challenge could prove more accessible for teams to achieve (see Section 5).

8 CONCLUSIONS

The second SKAO Science Data Challenge has brought together scientists and software experts from around the world to tackle the problem of finding and characterizing H I sources in very large SKAO data sets. The high level of engagement coupled with multidisciplinary collaboration has enabled the goals of the Challenge to be met, with over 100 finalists gaining familiarity with future SKAO spectral line data in order to drive forward new data processing methods and improve on existing techniques.

Interpretation of the results from SDC2 is limited by three main factors:

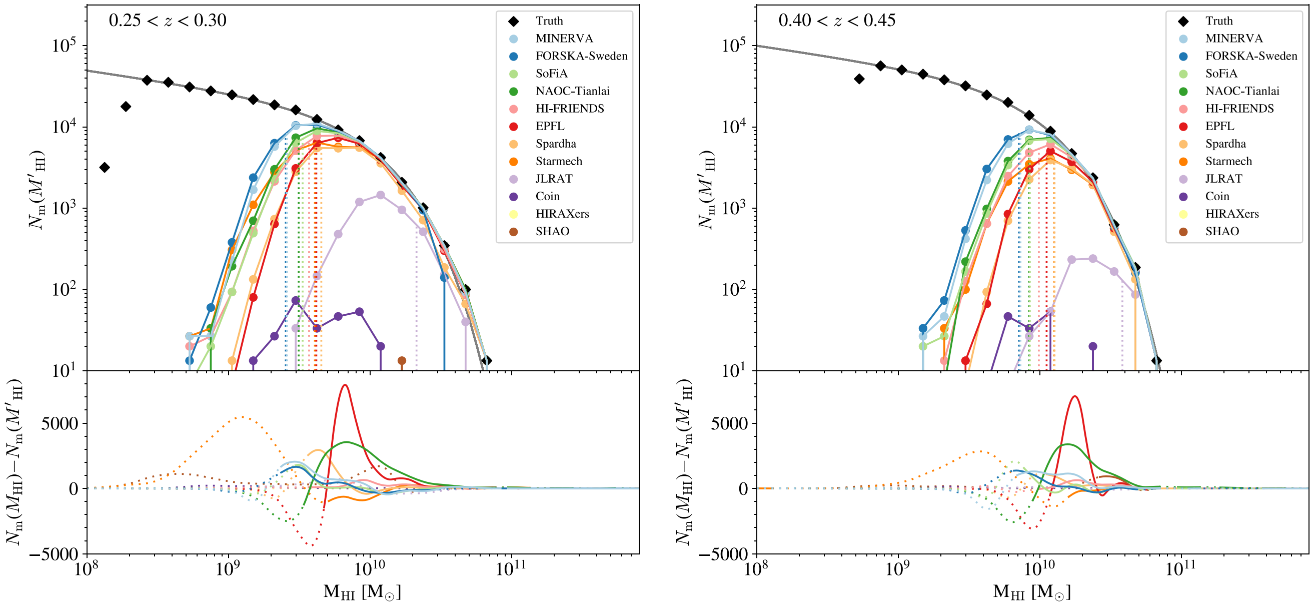

Figure 15. Top panels: H I mass distributions $N_m(M'_{HI})$ are constructed using the true values of integrated line flux and central frequency of each teams' matches (joined circles). The redshift-dependent H I mass function (equation 15), from which truth catalogue sources were drawn, is multiplied by the comoving volume of the given redshift interval and plotted (grey curve). Black diamonds represent the H I mass distribution reconstructed using the full truth catalogue. Dotted lines indicate for each team the H I mass above which completeness exceeds 50 per cent. Bottom panels: The H I mass distribution residual represents the difference between the distribution constructed from the values of teams' submissions and distribution constructed from truth values of teams' matches. Both distributions are interpolated prior to finding the residual. Completeness values are in this case calculated using teams' submitted values, and dotted and solid curves are used to delineate H I masses where completeness falls below and above 50 per cent, respectively.

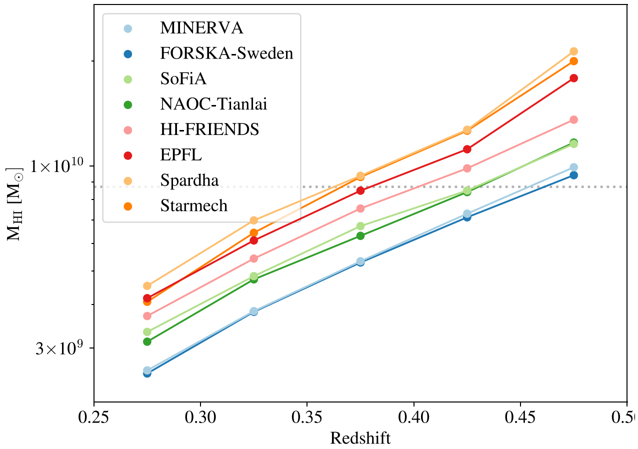

Figure 16. The H I mass above which at least 50 per cent of truth catalogue sources are recovered is plotted against redshift for the eight top scoring teams. The dotted line represents the input H I 'knee' mass, M_* (equation 1), which marks in the H I mass function the exponential decline from a shallow power law.

(i) The Challenge data set is a simulation and cannot fully represent real future SKA observations. Data set realism is limited most significantly by oversimplication of the noise (see Section 3.4).

(ii) The Challenge did not aim to provide a standardized cross comparison of methods; only a single data set was used and no attempt was made to control for team effort or domain expertise.

(iii) Team methods were developed as a means to maximize a score calculated according to the Challenge definition. Depending on the scientific goal, alternative metrics may be measured, for which other strategies may be explored.

With these caveats in mind, the main outcomes from the Challenge are summarized below:

(i) 12 international teams, using a variety of methods (Section 4) were able to complete the full Challenge.

(ii) Simulated data products representing a 2000 h spectral line observation by SKA-Mid telescopes were produced for the Challenge (Section 3), and are now publicly available together with accompanying truth catalogues.[14] We encourage the use of these data products by the science community in order to support the preparation and planning for future SKAO observations.

(iii) The generous contribution from supercomputing partner facilities (Section 2.2) has been integral to the success of the Challenge. Thanks to the provision of resources for hosting, processing, and access to Challenge data, it has been possible to provide a realistically large H I data product in an accessible way. The support has also provided the opportunity to test several aspects of the future SRC model of collaboratively networked computing centres, from web technologies involved in the SDC2 scoring service (Section 5), to the access processes in place for resource users.

(iv) The provision of a realistically large H I data product has allowed participants to explore approaches for dealing with very large data sets. By interacting with the full Challenge data set, finalist teams were able to investigate optimization and efficiency savings in readiness for future SKAO observational data products.

(v) Analysis of teams' submissions (Section 6.1) has shown that sources are recovered with over 50 per cent completeness down to a SNR limit of ~5 and an integrated flux limit of ~20 Jy Hz by the top scoring teams. Keeping in mind the caveats above, this translates to the ability to probe the H I mass function down to $\sim 3\times10^{9}$ $M_\odot$ at $0.25 < z < 0.30$ and to $\sim 1\times10^{10}$ $M_\odot$ at $0.45 < z < 0.50$. The 'knee' mass of the H I mass function can be probed out to $z \sim 0.45$ by the same methods for the chosen redshift evolution.

(vi) The analysis of submitted catalogues also provides a qualitative and quantitative understanding of the biases inherent to sensitivity-limited survey results. Biases arising from the presence of local noise fluctuations resulted in overestimation of flux at SNR$\lesssim$7. Source size and line width also showed a positive bias with fainter objects and smaller sizes.

(vii) Six teams took part in the SDC2 reproducibility awards, which ran alongside the main Challenge and were designed to recognize best practice in the preparation of reproducible and reusable pipelines. All six teams received an award, with team HI-FRIENDS receiving a Gold award for an exemplary software pipeline.

(viii) New applications of ML-based techniques – used by the two top scoring teams – have shown particular promise in the recovery and characterization of H I sources. The results suggest a dependency on sufficient training data, evidenced by a drop in performance at the bright flux end, where a paucity of very bright training sources exists. A more uniformly distributed training sample may address this problem. Further work using real observations from SKAO commissioning activities and from precursor instruments will examine how well ML models can transfer from simulated training data to real observational data.

(ix) The existing SOFIA software package also performed very well, achieving third place in the Challenge and also being used by several other teams, including by the second placed team for source characterization. That the package proved so popular further demonstrates the value of clearly documented and easily accessible codes, in addition to its accuracy and efficiency. This challenge highlights the need for such software packages, built and designed by astronomers to tackle specific problems, to receive the funding to be well maintained.

(x) Perhaps the most important finding of the Challenge is that of method complementarity. Also seen in the first SKAO SDC (Bonaldi et al. 2020), the relative performance of individual teams varied across aspects of the Challenge. It is likely that a combination of methods will produce the most accurate results. This finding is underscored by the strategy employed by the winning team, MINERVA. By optimizing the combined predictions from two independent ML methods, the team was able to record an improvement in score 20 per cent above either method alone (see Fig. 5). The result demonstrates the promise of ensemble learning in exploiting very large astronomical data sets.

SUPERCOMPUTING PARTNER FACILITIES

We would like to make a special acknowledgment of the very generous support from the SDC2 computing partner facilities (Section 2.2), without which a realistic and accessible Challenge would not have been possible. We acknowledge support from the Australian SKA Regional Centre (AusSRC) and the Pawsey Supercomputing Centre. This work was granted access to the HPC/AI resources of IDRIS under the allocations AP010412412, AP010412365, and AP010412404 made by GENCI. The authors acknowledge use of IRIS (https://www.iris.ac.uk) resources delivered by the SCD Cloud at STFC's Rutherford Appleton Laboratory (https://www.scd.stfc.ac.uk/Pages/STFC-Cloud-Operations.aspx). This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS).

[14] https://sdc2.astronomers.skatelescope.org/sdc2-challenge/data

This work used resources of China SKA Regional Centre prototype (An et al. 2019; 2022) funded by the National Key R&D Programme of China (2018YFA0404603) and Chinese Academy of Sciences (114231KYSB20170003). The Enabling Green E-science for the Square Kilometre Array Research Infrastructure (ENGAGE-SKA) team acknowledges financial support from grant POCI-01-0145-FEDER022217, funded by Programa Operacional Competitividade e Internacionalização (COMPETE 2020) and the Fundação para a Ciência e a Tecnologia (FCT), Portugal. This work was also funded by FCT and Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020-UIDP/50008/2020 and UID/EEA/50008/2019. The authors acknowledge the Laboratory for Advanced Computing at University of Coimbra for providing HPC, computing, consulting resources that have contributed to the research results reported within this paper or work. This work used the Spanish Prototype of an SRC (SPSRC) at IAA-CSIC, which is funded by SEV-2017- 0709, CEX2021-001131-S, RTI2018-096228-B-C31 AEI/ 10.13039/501100011033, EQC2019- 005707-P AEI/ 10.13039/501100011033 ERDF, EU, TED2021-130231B-I00 AEI/ 10.13039/501100011033 EU NextGenerationEU, PRTR SOMM17_5208_IAA funded by the Regional Government of Andalusia. We acknowledge the computing infrastructures of ⟨0:funding-source 3:href="http://dx.doi.org/10.13039/100008082"⟩INAF⟨/0:funding-source⟩, under the coordination of the ICT office of Scientific Directorate, for the availability of computing resources and support.

ACKNOWLEDGEMENTS

We would like to thank members of the SKAO HI Science Working Group for useful feedback. We are grateful for helpful discussions with the Software Sustainability Institute. The simulations make use of data from WSRT HALOGAS-DR1. The Westerbork Synthesis Radio Telescope is operated by ASTRON (Netherlands Institute for Radio Astronomy) with support from the Netherlands Foundation for Scientific Research NWO. The work also made use of 'THINGS', the HI Nearby Galaxy Survey (Walter et al. 2008), data products from which were kindly provided to us by Erwin de Blok after multiscale beam deconvolution performed by Elias Brinks. We would like to thank INAF for the hosting of SDC2 data products. LA is grateful for the support from UK STFC via the CDT studentship grant ST/P006809/1. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 679627; project name FORNAX). JMvdH and KMH acknowledge support from the European Research Council under the European Union 7th Framework Programme (FP/2007–2013)/ERC Grant Agreement no. 291531 (HIStoryNU). SSI. The works of the NAOC-Tianlai team members have been supported by the National Key R&D Program grants 2018YFE0120800, 2017YFA0402603, 2018YFA0404504, 2018YFA9494691, The National Natural Science Foundation of China (NSFC) grants 11633004, 11975072, 11835009, 11890691, 12033008, the Chinese Academy of Science (CAS) QYZDJ-SSW-SLH017, JCTD-2019-05, and the China Manned Space Projects CMS-CSST-2021-A03, CMS-CSST-2021-B01. Team FORSKA-Sweden acknowledges support from Onsala Space Observatory for the provisioning of its facilities support. The Onsala Space Observatory national research infrastructure is funded through Swedish Research Council (grant No. 2017–00648). Team FORSKA-Sweden also acknowledges support from the Fraunhofer Cluster of Excellence *Cognitive Internet Technologies*. CH, MB acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2121 'Quantum Universe' – 390833306. MP acknowledges the support of the CEFIPRA foundation under project 6504–3. We acknowledge financial support from SEV-2017-0709, CEX2021-001131-S, AEI/ 10.13039/501100011033. LD, JG, KMH, JM, MP, SSE, LVM, AS from RTI2018-096228-B-C31, PID2021-123930OB-C21 AEI/ 10.13039/501100011033 FEDER, UE. LVM, JG, SSE acknowledge The European Science Cluster of Astronomy and Particle Physics ESFRI Research Infrastructures project that has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 824064. LVM, JG, and JM RED2018-102587-T AEI/ 10.13039/501100011033. LVM, JG, SSE, JM acknowledge financial support from the grant IAA4SKA (Ref. R18-RT-3082) from the Economic Transformation, Industry, Knowledge and Universities Council of the Regional Government of Andalusia and the ERDF from the EU, TED2021-130231B-I00 AEI/ 10.13039/501100011033 EU NextGenerationEU/PRTR. LVM, JG, KMH acknowledges financial support from the coordination of the participation in SKA-SPAIN, funded by the Ministry of Science and Innovation (MCIN). LD from PTA2018-015980-I AEI/ 10.13039/501100011033. MP from the grant DOC01497 funded by the Economic Transformation, Industry, Knowledge and Universities Council of the Regional Government of Andalusia and by the Operational Program ESF Andalucía 2014–2020. MTS acknowledges support from a Scientific Exchanges visitor fellowship (IZSEZO_202357) from the Swiss National Science Foundation. AVS thanks Martin Kunz and Bruce Bassett for the valuable discussions. Team Spardha would like to acknowledge SKA India Consortium, IUCAA and Raman Research Institute for providing the support with the computing facilities. Team Spardha would also acknowledge National Supercomputing Mission (NSM) for providing computing resources of 'PARAM Shakti' at IIT Kharagpur, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (Meity) and Department of Science and Technology (DST), Government of India.

DATA AVAILABILITY

The SDC2 simulated data sets are publicly available from the SDC2 website: https://sdc2.astronomers.skatelescope.org/

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available at: https://www.tensorflow.org/

Alom M. Z., Hasan M., Yakopcic C., Taha T. M., Asari V. K., 2018, preprint (arXiv:1802.06955)

Alves A., 2017, J. Instrum., 12, T05005

An T., Wu X. P.,Hong X.,2019, Nature Astron., 3, 1030

An T., Wu X., Lao B., Guo S., Xu Z., Lv W., Zhang Y., Zhang Z., 2022, Sci. China Phys. Mech. Astron., 65, 129501

Anaconda, 2020, Available at: https://docs.anaconda.com/

Autenrieth M., van Dyk D. A., Trotta R., Stenning D. C., 2021, preprint (arXiv:2106.11211)

Barkai J. A., Verheijen M. A. W., Martínez E. T., Wilkinson M. H. F., 2023, A&A, 670, A55

Baugh C. et al., 2019, MNRAS, 483, 4922

Bentley J. L., 1975, Commun. ACM, 18, 509

Bera A., Kanekar N., Chengalur J. N., Bagla J. S., 2019, ApJ, 882, L7

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Blyth S. et al., 2015, in Proc. Sci., Advancing Astrophysics with the Square Kilometre Array (AASKA14). SISSA, Trieste, PoS#128

Blyth S. et al., 2016, in, Taylor R., Camilo F., Leeuw L., Moodley K., eds, MeerKAT Science: On the Pathway to the SKA, Sissa Medialab, Stellenbosch, p. 4

Bonaldi A. et al., 2020, MNRAS, 500, 3821

Bonaldi A., Bonato M., Galluzzi V., Harrison I., Massardi M., Kay S., De Zotti G., Brown M. L., 2019, MNRAS, 482, 2

Bonaldi A., Hartley P. Ronconi T., De Zotti G., Bonato M., 2023, preprint, (arXiv:2305.10175)

Braun R., 2012, ApJ, 749, 87

Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, preprint (arXiv:1912.12699)

Braun R., Bourke T. L., Green J. A., Keane E., Wagg J., 2015, Advancing Astrophysics with the Square Kilometre Array, Sissa Medialab, p. 174

Broeils A. H., Rhee M. H., 1997, A&A, 324, 877

Burges C. J., 1998, Data Min. Knowl. Discovery, 2, 121

Chen J. et al., 2021, preprint (arXiv:2102.04306)

Chowdhury A., Kanekar N., Das B., Dwarakanath K. S., Sethi S., 2021, ApJ, 913, L24

Chrysostomou A., Taljaard C., Bolton R., Ball L., Breen S., van Zyl A., 2020, in Adler D. S., Robert L., Benn C. R., eds, SPIE Conf. Ser. Vol. 11449, Observatory Operations: Strategies, Processes, and Systems VIII. SPIE, Bellingham, p. 114490X

Crouch S. et al., 2013, Comput. Sci. Eng., 15, 74

de Blok W. J. G., Fraternali F., Heald G. H., Adams E. A. K., Bosma A., Koribalski B. S., the HI Science Working Group, 2015, preprint (arXiv:1501.01211)

Diakogiannis F. I., Waldner F., Caccetta P., Wu C., 2020, ISPRS J. Photogramm. Remote Sens., 162, 94

Dodson R. et al., 2022, AJ, 163, 59

Duffy A. R., Meyer M. J., Staveley-Smith L., Bernyk M., Croton D. J., Koribalski B. S., Gerstmann D., Westerlund S., 2012, MNRAS, 426, 3385

Fernández X. et al., 2016, ApJ, 824, L1

Flöer L., Winkel B., 2012, PASA, 29, 244

Fraternali F., van Moorsel G., Sancisi R., Oosterloo T., 2002, AJ, 123, 3124

Freeman P. E., Izbicki R., Lee A. B., 2017, MNRAS, 468, 4556

Garrido J. et al., 2021, J. Astron. Telescopes Instrum. Syst., 8, 1

Häkansson H. et al., 2023, A&A, 671, A39

He K., Zhang X., Ren S., Sun J., 2016, Deep Residual Learning for Image Recognition. Available at: http://image-net.org/challenges/LSVRC/2015/

Heald G. et al., 2011, A&A, 526, A118

Hogg D. W., Turner E. L., 1998, PASP, 110, 727

Holmberg E., 1946, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 117, 3

Huang H. et al., 2020, in Closas P., Bugallo M., eds, ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2020 IEEE Piscataway, New Jersey, p. 1055

Hubble E. P., 1926, ApJ, 64, 321

Jones M. G., Haynes M. P., Giovanelli R., Moorman C., 2018, MNRAS, 477, 2

Jurek R., 2012, Publ. Astron. Soc. Aust., 29, 251

Katz D. S. et al., 2021, preprint (arXiv:2101.10883)

Khvedchenya E., 2019, PyTorch Toolbelt. Available at: https://github.com/BloodAxe/pytorch-toolbelt

Kim E. J., Brunner R. J., Carrasco Kind M., 2015, MNRAS, 453, 507

Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, Conference Track Proceedings, San Diego

Koribalski B. S. et al., 2020, Ap&SS, 365, 118

Leahy J. P., Bridle A. H., Strom R. G., 2013, in Leahy J. P. et al., eds, An Atlas of DRAGNs. Available at: https://www.jb.man.ac.uk/atlas/#lists

Li W., Wang G., Fidon L., Ourselin S., Cardoso M. J., Vercauteren T., 2017, in International Conference on Information Processing in Medical Imaging, Vol. 10265, Springer, Berlin, p. 348

Luo S., Leung A. P., Hui C. Y., Li K. L., 2020, MNRAS, 492, 5377

McGaugh S. S., Schombert J. M., Bothun G. D., de Blok W. J. G., 2000, ApJ, 533, L99

Metcalf R. B. et al., 2019, A&A, 625, A119

Meyer M., Robotham A., Obreschkow D., Westmeier T., Duffy A. R., Staveley-Smith L., 2017, PASA, 34, 52

Milletari F., Navab N., Ahmadi S. A., 2016, in Su H., ed., Proceedings—2016 4th International Conference on 3D Vision, 3DV 2016, IEEE, U.S., p. 565

Mohan N., Rafferty D., 2015, Astrophysics Source Code Library, record ascl:1502.007

Mölder F. et al., 2021, F1000Research, 533, 7604

Moldon J. et al., 2021a, HI-FRIENDS participation in the SKA Data Challenge 2 (1.0.3). Zenodo. Available at: https://doi.org/10.5281/zenodo.5172930

Moldon J. et al., 2021b, HI-FRIENDS participation in the SKA Data Challenge 2 (1.0.1). Zenodo. Available at: https://doi.org/10.5281/zenodo.5167693

Moldon J. et al., 2021c, HI-FRIENDS HI data cube source finding and characterization (1.0.0). Available at: https://workflowhub.eu/workflows/141?version=1

Morganti R., Sadler E. M., Curran S., 2015, Proc. Sci., Advancing Astrophysics with the Square Kilometre Array (AASKA14). SISSA, Trieste, PoS#134

Oktay O. et al., 2018, preprint (arXiv:1804.03999)

Oosterloo T., Fraternali F., Sancisi R., 2007, AJ, 134, 1019

Pan S. J., Yang Q., 2009, IEEE Trans. Knowl. Data Eng., 22, 1345

Planck Collaboration, 2016, A&A, 594, A13

Popping A., Meyer M., Staveley-Smith L., Obreschkow D., Jozsa G., Pisano D. J., 2015, in Proc. Sci., Advancing Astrophysics with the Square Kilometre Array (AASKA14). SISSA, Trieste, PoS#132

Power C. et al., 2015, in Proc. Sci., Advancing Astrophysics with the Square Kilometre Array (AASKA14). SISSA, Trieste, PoS#133,

Power C., Baugh C. M., Lacey C. G., 2010, MNRAS, 406, 43

Jupyter P. et al., 2018, in Akici F., Lippa D., Niederhut D., Pacer M., eds, Proceedings of the 17th Python in Science Conference. SciPy 2018, Austin, Texas, p. 113

Qin X., Zhang Z., Huang C., Dehghan M., Zaiane O. R., Jagersand M., 2020, Pattern Recognit., 106, 107404

Redmon J., Divvala S., Girshick R., Farhadi A., 2015, preprint (arXiv:1506.02640)

Redmon J., Farhadi A., 2016, preprint (arXiv:1612.08242)

Redmon J., Farhadi A., 2018, preprint (arXiv:1804.02767)

Ronneberger O., Fischer P., Brox T., 2015a, in Navab N., Hornegger J., Wells W., Frangi A., eds, International Conference on Medical image Computing and Computer-assisted Intervention – MICCAI 2015. Springer, Cham, p. 234

Ronneberger O., Fischer P., Brox T., 2015b, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351, 234

Rosenberg J. L., Schneider S. E., 2000, ApJS, 130, 177

Rosenberg J. L., Schneider S. E., 2002, ApJ, 567, 247

Sancisi R., Fraternali F., Oosterloo T., van der Hulst T., 2008, A&AR, 15, 189

Sault R. J., Teuben P. J., Wright M. C. H., 1995, in Shaw R. A., Payne H. E., Hayes J. J. E.eds, ASP Conf. Ser. Vol. 77, Astronomical Data Analysis Software and Systems IV. Astron. Soc. Pac., San Francisco. p. 433

Scaife A. M. M., 2020, Phil. Trans. R. Soc., 378, 20190060

Scherzer O., 2010, Handbook of Mathematical Methods in Imaging, Springer Science and Business Media, Berlin

Serra P. et al., 2015a, MNRAS, 448, 1922

Sha Y., 2021, Keras-unet-collection. Available at: https://github.com/yingkaisha/keras-unet-collection,

Starck J.-L., Fadili J., Murtagh F., 2007, IEEE Trans. Image Process., 16, 297

Staveley-Smith L., Davies R. D., Kinman T. D., 1992, MNRAS, 258, 334

Tang H., Scaife A. M. M., Leahy J. P., 2019, MNRAS, 488, 3358

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R.eds, ASP Conf. Ser. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco. p. 29

Teeninga P., Moschini U., Trager S. C., Wilkinson M. H., 2015, in Angulo J., Velasco-Forero S., Meyer F., eds., Mathematical Morphology and Its Applications to Signal and Image Processing. Springer, Cham, p.157

Tolley E., Korber D., Galan A., Peel A., Sargent M., Kneib J.-P., Courbin F., Starck J.-L., 2022, Astron. Comput., 41, 100631

Vafaei Sadr A., Vos E. E., Bassett B. A., Hosenie Z., Oozeer N., Lochner M., 2019, MNRAS, 484, 2793

van der Hulst J. M., de Blok W. J. G., 2013, in Oswalt T. D., Keel W. C., eds, Planets, Stars and Stellar Systems. Springer, Dordrecht, p. 183

van der Walt S. et al., 2014, PeerJ, 2, e453

Vonesch C., Blu T., Unser M., 2007, IEEE Trans. Signal Process., 55, 4415

Walter F., Brinks E., de Blok W. J. G., Bigiel F., Kennicutt R. C., Jr., Thornley M. D., Leroy A., 2008, AJ, 136, 2563

Wang J., Koribalski B. S., Serra P., van der Hulst T., Roychowdhury S., Kamphuis P., Chengalur J. N., 2016, MNRAS, 460, 2143

Westerlund S., Harris C., 2014, PASA, 31, e023

Westmeier T. et al., 2021, MNRAS, 506, 3962

Whiting M., 2012, Astrophysics Source Code Library, record (ascl:1201.011)

Wilkinson M. D. et al., 2016, Sci. Data, 3, 160018

Wilkinson P. N., Kellermann K., Ekers R., Cordes J., Lazio T. J. W., 2004, New Astron. Rev., 48, 1551

Wolpert D. H., 1992, Neural Netw., 5, 241

Yakubovskiy P., 2020, Segmentation Models Pytorch. Available at: https://github.com/qubvel/segmentation_models.pytorch

Yang J., Huang X., He Y., Xu J., Yang C., Xu G., Ni B., 2021, IEEE J. Biomed. Health Inf., 25, 3009

Zitlau R., Hoyle B., Paech K., Weller J., Rau M. M., Seitz S., 2016, MNRAS, 460, 3152

[1] *SKA Observatory, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK*
[2] *Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, The University of Manchester, Manchester M13 9PL, UK*
[3] *Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, CAS, 80 Nandan Road, 200030 Shanghai, China*
[4] *DIO, Observatoire de Paris, CNRS, PSL, F-75104 Paris, France*
[5] *Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 SHJL, UK*
[6] *Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore, 453552 Indore, India*
[7] *National Astronomical Observatory, Chinese Academy of Sciences, 20A Datun Road, 100101 Beijing, PR China*
[8] *School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, UK*
[9] *Centre for Strings, Gravitation and Cosmology, Department of Physics, Indian Institute of Technology Madras, 600036 Chennai, India*
[10] *Department of Physics & Institute of Astronomy, University of Cambridge, Cambridge, CB3 0HA, UK*
[11] *LERMA, Observatoire de Paris, PSL research Université, CNRS, Sorbonne Université, F-75104 Paris, France*
[12] *Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía s/n, E-18008 Granada, Spain*
[13] *Department of Information Technology and Electrical Engineering, University of Naples Federico II, 21 Via Claudio, I-80125 Napoli, Italy*
[14] *Centre for Astrophysics Research, University of Hertfordshire, Hatfield, Hertfordshire AL10 9AB, UK*
[15] *Centro Brasileiro de Pesquisas Físicas (CBPF), 22290-180 URCA, Rio de Janeiro (RJ), Brazil*
[16] *Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland*
[17] *Faculty of Computational Mathematics and Cybernetics of Lomonosov, Moscow State University, Moscow, 119234, Russia*
[18] *Fraunhofer-Chalmers Centre & Fraunhofer Center for Machine Learning, SE-412 88 Gothenburg, Sweden*
[19] *University of Hamburg, Hamburg Observatory, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[20] *Instituto de Física de Cantabria, CSIC-UC, Av. de Los Castros s/n, E-39005 Santander, Spain*
[21] *ASTRON, the Netherlands Institute for Radio Astronomy, Postbus 2, 7990 AA Dwingeloo, The Netherlands*
[22] *Kapteyn Astronomical Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands*
[23] *Department of Electrical and Electronics Engineering, PES University, 560085 Bangalore, India*
[24] *Department of Physics, School of Mathematics and Physics, The University of Queensland, Brisbane QLD 4072, Australia*
[25] *ICRAR M468, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia*
[26] *INAF – Osservatorio Astronomico di Cagliari, Via della Scienza 5, I-09047 Selargius, CA, Italy*
[27] *Department of Astrophysics, School of Physics and Astronomy, Tel Aviv University, 69978 Tel Aviv, Israel*
[28] *Department of Physics, College of Sciences, Northeastern University, 110819 Shenyang, China*
[29] *Raman Research Institute, C. V. Raman Avenue, Sadashivanagar, 560080 Bengaluru, India*
[30] *Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*
[31] *International Space Science Institute (ISSI), Hallerstrasse 6, CH-3012 Bern, Switzerland*
[32] *Department of Physics, Indian Institute of Technology Kharagpur, 721302 Kharagpur, India*
[33] *CSIRO Space and Astronomy, PO Box 1130, Bentley WA 6102, Australia*
[34] *Australian SKA Regional Centre (AusSRC), Kensington Western, 6151, Australia*
[35] *Special Astrophysical Observatory of RAS, 369167 Nizhny Arkhyz, Russia*
[36] *Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden*
[37] *Département de Physique Théorique and Center for Astroparticle Physics, University of Geneva, Genève 4, 1211, Switzerland*
[38] *CAS Key Laboratory of FAST, National Astronomical Observatories, Chinese Academy of Sciences, 100101 Beijing, China*
[39] *School of Physics and Astronomy, Sun Yat-Sen University, 2 Daxue Road, Tangjia, U1YP8R Zhuhai, China*
[40] *Peng Cheng Laboratory, No.2, Xingke 1st Street, 518000 Shenzhen, PR China*
[41] *Department of Physics & Astronomy, Macalester College, 1600 Grand Avenue, Saint Paul, MN 55105, USA*
[42] *Collège de France, 11 Place Marcelin Berthelot, F-75005 Paris, France*
[43] *Université de Strasbourg, CNRS UMR 7550, Observatoire astronomique de Strasbourg, F-67000 Strasbourg, France*
[44] *Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA*
[45] *South African Radio Astronomy Observatory (SARAO), 2 Fir Street, Black River Park, Observatory 7925, South Africa*
[46] *Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute, D-44780 Bochum, Germany*
[47] *Department of Astronomy, Tsinghua University, 100084 Beijing, PR China*
[48] *Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada*
[49] *Space Research Institute of Russian Academy of Sciences, Profsoyuznaya 84/32, 117997 Moscow, Russia*
[50] *Laboratoire Univers et Particules de Montpellier (LUPM)-CNRS, UNIVERSITÉ DE MONTPELLIER LUPM CC 072 - Place Eugène Bataillon 34095 Montpellier Cedex 5, France*
[51] *GEPI, Observatoire de Paris, CNRS, Université Paris Diderot, 5 Place Jules Janssen, F-92190 Meudon, France*
[52] *Department of Physics & Electronics, Rhodes University, PO Box 94, 6140 Grahamstown, South Africa*

This paper has been typeset from a TEX/LATEX file prepared by the author.