MNRAS **000**, 1–25 (2021) Preprint 16 February 2022 Compiled using MNRAS LATEX style file v3.0

SKA Science Data Challenge 2: analysis and results

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

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The Square Kilometre Array Observatory (SKAO) will explore the radio sky to unrivalled 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 in order to extract key science findings. To this end, SKAO is conducting a series of Science Data Challenges, each designed to familiarise 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 characterise neutral hydrogen (HI) sources in a simulated data product representing a 2000 h SKA MID spectral line observation. Through the generous support of eight international supercomputing facilities, participants were able to undertake the Challenge using dedicated computational resources. This model not only supported the accessible provision of a realistically large dataset, but also provided the opportunity to test several aspects of the future SKA Regional Centre network. Sitting alongside the main challenge, 'reproducibility awards' were made in recognition of those pipelines which demonstrated Open Science best practice. The Challenge saw over 100 finalists develop a range of new and existing techniques, in results which highlight the strengths of multidisciplinary and collaborative effort. The winning strategy – combining predictions from two independent machine learning techniques – 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 datasets.

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

1 INTRODUCTION

The Square Kilometre Array (SKA) project was born from an ambition to create a telescope sensitive enough to trace the formation

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

First light at the SKA 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 SKA 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 SKA Science Working Groups – is underway in order to maximise the exploitation of future SKA datasets.

The SKA model of data delivery will provide science users with data in the form of science-ready image and non-image SKA Observatory (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). Particularly challenging is the crowded nature of sky images; the sensitivity of SKA observations will afford unparalleled observation depth, resulting in a very large number of overlapping sources that will require detection and classification.

In order to support the community to prepare for such rich datasets, the SKAO has established a series of Science Data Challenges (SDCs). Each challenge involves some combination of real or simulated datasets designed as closely as possible to represent future SKA data. The purpose of each challenge is then to exercise analysis methods needed to extract science from the data, with the goal of fostering new ideas and methods via wide participation and engagement. The challenges also aim to familiarise the community with the standard products that the SKA will deliver, providing the opportunity to test the validity of scientific proposals and to optimise survey design. For these reasons, all SDC data products are made available publicly for the long term[1]. Science Data Challenge 1 (SDC1, Bonaldi et al. 2020) saw participating teams find and characterise 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. Science Data Challenge 2[2] (SDC2) – the second in the series – has involved simulated spectral line observations designed to represent the SKA view of neutral hydrogen emission (HI) up to $z = 0.5$, again inviting participants to attempt source finding and characterisation within very large datasets.

[1] https://astronomers.skatelescope.org/ska-data-challenges/
[2] https://sdc2.astronomers.skatelescope.org/

Resulting from the 'spin-flip' of an electron in a neutral hydrogen atom, 21cm spectral line emission and absorption traces the distribution of HI across the history of the Universe. This so-called cold gas exists in and around galaxies, fueling star-formation via ongoing infall from the cosmic web. Observations of individual HI 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 study the impact of AGN on the large-scale gas distribution in galaxies (Morganti et al. 2015). HI dynamics also provide a measurement of the dark matter content of galaxies (Power et al. 2015). Deep HI surveys are therefore crucial for our understanding of galaxy formation and evolution over cosmic time (Blyth et al. 2015; Power et al. 2010; Meyer et al. 2017; Dodson et al. 2021). The faintness of HI emission has until now limited survey depths to up to $z \sim 0.25$ (see Sancisi et al. 2008 and van der Hulst & de Blok 2013 for recent reviews of the results so far). The collecting power and high angular resolution of the SKA, however, will enable survey depths of $z \sim 1$ in emission and $z \sim 3$ in absorption, increasing dramatically the number of sources for study and providing statistically robust samples of HI images and spectra. The MeerKAT telescope – a precursor to the SKA – will soon launch the Looking At the Distant Universe with the MeerKAT Array (LADUMA) survey (Blyth et al. 2016), probing HI emission out to z 1.4 using stacking. The size of resulting datasets necessitates the use of automated source finding methods, and several software tools are currently available for HI source detection and characterisation (Flöer & Winkel 2012; Jurek 2012; Whiting 2012; Westerlund & Harris 2014; Serra et al. 2015a; Westmeier et al. 2021).

In line with the Challenge goals, we endeavoured to ensure equal accessibility to the Challenge and its data for all interested teams. As such, we adopted a distributed model for the delivery of data, whereby teams were each able to access Challenge datasets and computational resources at one of eight partner supercomputing facilities, at which each could deploy their own analysis pipelines (Section 2.1). This model also served as a test bed for a number of future SRC technologies that are currently in development. Challenge teams were invited to use analysis methods that were any combination of purpose-built and bespoke to existing and publicly available. The SKA is committed to Open Science values. Throughout the Challenge, therefore, 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.

In this paper we report on the outcome of SDC2. The structure of the paper is as follows: in Section 2 we describe the Challenge, including the methods used to produce the SDC2 datasets, and the Challenge definition; in Section 3 we present some of the methods used by participating teams to complete the Challenge; in Section 4 we describe the scoring procedure and reproducibility criteria used to evaluate Challenge submissions; in Sections 5 we describe the scoring procedures used; in Sections 6 and 7 we present the Challenge results and analysis, before setting out our conclusions in Section 8.

2 THE CHALLENGE

Participating teams were invited to access a 1 TB dataset hosted on dedicated facilities provided by the SDC2 computational resource partners (Section 2.1). The dataset simulates an HI 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 a redshift interval $z = 0.235$–0.495.
(iv) Noise consistent with a 2000 hour total observation.
(v) Systematics including imperfect continuum subtraction, simulated RFI flagging and excess noise due to RFI.

The HI 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.

Together with the full-size Challenge dataset, two smaller datasets were made available for development purposes. Generated using the same procedure as the full-size dataset but with a different statistical realization, the 'development' and 'large development' datasets were provided along with truth catalogues of HI sources. A further, 'evaluation', dataset 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 dataset. The evaluation dataset 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 dataset and matching a required format. The development and evaluation datasets were made available for download prior to and during the Challenge.

2.1 Supercomputing partner facilities

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

AusSRC and Pawsey – Perth, Australia

China SRC-proto – Shanghai, China

CSCS – Lugano, Switzerland

ENGAGE SKA-UCLCA – Aveiro and Coimbra, Portugal

GENCI-IDRIS – Orsay, France

IAA-CSIC – Granada, Spain

INAF – Rome, Italy

IRIS (STFC) – UK

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

2.2 The challenge definition

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

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

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

(i) Integrated line flux (Jy Hz): the total line flux integrated over the signal $\int F \mathrm{d}_\nu$.
(ii) HI size (arcsec): the HI major axis diameter at 1 $\mathrm{M}_\odot$ pc^{-2}.
(iii) Line width (km s^{-1}): the observed line width at 20 percent 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.

Resulting catalogues were submitted via a dedicated scoring service[3] (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 1st February 2021 and closed on 31st July 2021.

2.3 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 generalises 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 HI datacubes involved three steps: source catalogue generation, sky model creation, and telescope simulation. All codes used to generate the dataset are publicly available under open source licence (need to state licence type).[4].

3.1 Source catalogues

Semi-empirical modelling was used to produce a catalogue of sources with both continuum and HI properties. For full details of the procedure, see Bonaldi et al 2021. Here we provide a summary.

Initial catalogues of HI emission sources were generated in FORTRAN by sampling from an HI redshift-dependent mass function derived from the ALFALFA survey (Jones et al. 2018):

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

where the knee mass, $M_* = 8.71 \times 10^9$ $\mathrm{M}_\odot$, marks the exponential decline from a shallow power law parameterised by $\alpha = -1.25$, $\phi_* = 4.5 \times 10^{-3}$ Mpc^{-3} dex^{-1} is a normalisation constant, and a mild redshift dependence is applied by using $\log(M_*(z)) = \log(M_*) + 0.075z$. Conversion from HI mass to integrated line flux followed the relation from Duffy et al. (2012) and source sizes were modelled using the mass-size relation of Wang

[3] https://pypi.org/project/ska-sdc/2.0.0/
[4] link to gitlab repository of SDC2 sim codes

et al. (2016). A lower integrated flux limit of 1 Jy Hz was made. Catalogues of radio-continuum sources – star-forming galaxies (SFGs) and Active Galactic Nuclei (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 and with a flux density limit of 2×10^{-7} Jy at 1150 MHz.

An HI mass was attributed to each AGN source in the continuum catalogue using the relation between dark mass and HI mass from the P-Millennium simulation (Baugh et al. 2019), and to each SFG using the correlation between HI mass and star-formation rates from ALFALFA (Jones et al. 2018). The HI catalogue and the portion of the radio continuum catalogue sharing the same redshift interval were then further processed to identify those that would constitute a counterpart, i.e. be hosted by the same galaxy. To this end, HI galaxies were matched to available radio continuum galaxies by first reserving as 'continuum only' those continuum sources with predicted HI mass below a threshold corresponding the HI integrated flux limit. Both catalogues were then ranked in descending order of HI mass, and HI and continuum sources paired off starting from the highest HI mass. All unmatched HI sources were reserved as 'HI only'. This procedure was chosen over a nearest-neighbour approach due to a mismatch of source distributions between catalogues.

In order to introduce a realistic clustering signal to the sources, the galaxies were associated with dark matter (DM) haloes from the P-Millennium simulation Baugh et al. (2019). Both the mass and environment of host DM halos 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 HI-containing sources.

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 FITSIO for writing to file.

3.2.1 HI datacube

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

(i) Blanking of pixels in order to produce a positive definite noiseless model.
(ii) Measurement of HI major axis diameter at a surface density of 1 $M_\odot$ pc^{-2}.
(iii) Rotation to a common position angle of 0 degrees.
(iv) Preliminary spatial resampling after application of a smoothing filter, such that the physical pixel size of the resampled data would be no lower than required for the lowest redshift simulated sources.
(v) Preliminary velocity resampling after application of a smoothing filter.

For each source from the simulation catalogue, a source from the prepared atlas of sources was chosen from those nearby in normalised HI mass-inclination angle parameter space. In order to exploit the diversity of the limited atlas sample, matches were selected at random from those atlas sources located within a suitable radius. This radius in parameter space was chosen to be large enough to allow a wide spread of matched atlas sources, but small enough not to produce matches which would deviate too far from the desired catalogue HI mass and inclination angle.

Once matched with a catalogue source, atlas sources underwent transformations in size in the three cube dimensions of RA, Dec and frequency ν. An appropriate smoothing filter was applied prior to all scalings, in order to preserve sufficient sampling. Transformation scalings along HI major axis $D_{\rm HI}$, HI minor axis b, and line width w_{20} were determined using the catalogue source properties of HI mass $M_{\rm HI}$, inclination angle i, and redshift z, and making use of the following relations:

$$D_{\rm HI} = 0.51 \log M_{\rm HI} - 3.32, \tag{2}$$

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

$$V_{\rm rot}^2 = \frac{{\rm G}M_{\rm dyn}}{r}, \tag{3}$$

where $V_{\rm rot}$ is the rest frame rotational velocity at radius r and $M_{\rm dyn}$ is the dynamical mass, in order to determine frequency scalings for HI mass;

$$\cos^2(i) = \frac{(b/D_{\rm HI})^2 - \alpha^2}{(1-\alpha^2)}, \tag{4}$$

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

$$V_{\rm rad} = V_{\rm rot}\ \sin(i), \tag{5}$$

where $V_{\rm rad}$ is the rest frame radial velocity, in order to determine frequency scalings for inclination. 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 XIII 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 HI flux, before being placed in the full HI emission field at its designated position in RA, Dec and central frequency.

3.2.2 Continuum datacube

The treatment of continuum counterparts of HI objects was dependent on the full width at half maximum (FWHM) continuum size. A empty datacube with spatial resolution matching the HI 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 synthesised 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 et al. 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 datasets report the continuum size of objects as the Largest Angular Size (LAS) and the exponential scale length of the disk for AGN and SFG populations, respectively

3.2.3 *Calculation of HI Absorption Signatures*

Potential sources of HI absorption were determined by calculating the neutral hydrogen column density associated with every pixel in the HI model cube via

$$N_{\rm HI} = 49.8\, S_{\rm L}(\nu)\, \Delta\nu\, {\rm M}_\odot (1+z)^4 / (N_{\rm p}\, m_{\rm H}\, \Delta\theta^2\, C_{\rm M}^2), \quad (6)$$

where $S_{\rm L}$ is the HI brightness in the pixel in Jy beam^{-1}, $\Delta\nu$ the channel spacing in Hz, $\rm M_\odot$ a solar mass, z the redshift of the HI 21cm 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 HI mass conversion of Duffy et al. (2012). When observed with 100 pc or better physical resolution, the apparent HI column density can be related to an associated HI opacity (Braun 2012), $\tau\Delta V$, as

$$N_{\rm HI} = N_0 e^{-\tau\Delta V} + N_\infty (1 - e^{-\tau\Delta V}), \quad (7)$$

yielding

$$\tau\Delta V = \log[(N_\infty - N_0)/(N_\infty - N_{\rm HI})], \quad (8)$$

where $N_0 = 1.25x10^{20}$ 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 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 HI absorption opacity. The apparent column densities per pixel have therefore been subjected to an arbitrary power law rescaling as

$$N'_{\rm HI} = 10^{19+[\log 10(N_{\rm HI})-19]\beta}, \quad (9)$$

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

$$N''_{\rm HI} = N_\infty [{\rm e}^{2N'_{\rm HI}/N_\infty} - 1]/[{\rm e}^{2N'_{\rm HI}/N_\infty} + 1], \quad (10)$$

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

With the potential HI opacity in hand, two further tests are considered before its application. First, a check is made that the redshift of any continuum emission source along this line of sight is greater than the redshift of the HI for the frequency pixel under consideration. To enable this test, an image of the intensity-weighted emission redshift of the continuum sky model was generated as per Section 3.2.2. Second, a check is made that the brightness temperature of the continuum emission source along this line-of-sight exceeds a threshold, $T_{\rm min} = 100$ K. The corresponding flux density is:

$$S_{\rm min} = 7.35 \times 10^{-4} T_{\rm min} \Delta\phi^2/\lambda^2 \quad (11)$$

with $\Delta\phi$ the beam size in arcsec and λ the observing wavelength in cm, yielding $S_{\rm min}$ in Jy beam^{-1}. At the angular resolution of this data product this test restricts the occurrence of absorption towards only those continuum sources brighter than about 4 mJy beam^{-1}, thereby only to those lines-of-sight where the continuum brightness exceeds a plausible maximum HI brightness temperature. If both tests are passed, then the signature is calculated as:

$$S_{\rm HIA}(\nu) = S_{\rm C}[1 - e^{(-\tau\Delta V/{\rm d}V)}] \quad (12)$$

where $S_{\rm C}$ is the continuum model flux density at this frequency and dV is the actual channel sampling in units of km^{-1}.

3.3 Telescope simulation

The simulation of telescope sampling effects has been implemented by using Python to script tasks from the miriad package (Sault et al. 1995). Multi-processing parallelisation is exploited by applying the procedure over multiple frequency channels simultaneously.

3.3.1 *Preprocessing of the continuum cube*

In order to simulate imperfect continuum emission subtraction within the final HI datacube, a noise cube representing gain calibration error is produced. We first interpolate the simulated continuum sky model, $S_{\rm C}(\nu)$, to a frequency sampling of 10 MHz, before producing for each channel a two dimensional image of uncorrelated noise to represent a r.m.s. gain calibration error of $\sigma = 1 \times 10^{-3}$ and with spatial sampling 515×515 arcsec. The spatial and frequency samplings are 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. Unique random number seeds are used in order to ensure that noise remains uncorrelated between each channel.

The coarsely sampled noise field is 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, $S_{\rm NC}(\nu)$, is 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 are downsampled to the final simulation frequency interval of 30 kHz.

3.3.2 *Net emission and absorption cube*

With all signatures in hand a net continuum-subtracted HI emission and absorption cube is calculated from the sum

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

where the explicit frequency dependence is included to stress that all quantities are evaluated at the final required frequency sampling.

3.3.3 *Calculation of effective PSF and noise level*

The synthesized telescope beam is based on a nominal 8 hour duration tracking observation of the complete SKA MID configuration. A one minute time sampling interval is used in order to make beam calculations sufficiently realistic while avoiding excessive computational overheads. The thermal noise level is based on nominal system performance (Braun et al. 2019) for an effective on-sky integration time of 2000 hours 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, 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 in the data product. The nominal RMS noise level, σ_N, therefore declines linearly with frequency between 950 and 1150 MHz.

Observations the South Celestial Pole 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, $\gamma_{\rm RFI}$, is 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 is applied to all baselines from a minimum $\rm B_{min} = 0$ up to a maximum $B_{\rm max}$ according, in units of wavelength, to

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

which produces 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, is determined, in hours, from

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

where $\gamma_{\rm min} = 1.1$ and $\gamma_{\rm max} = 2$, are used to define the ranges of RFI ratios over which flagging is absent, intermittent or continuous. Intermittent flagging intervals are 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 are generated for each frequency channel. Unique random number seeds ensure that the resulting noise fields are not correlated across frequency. During imaging, a super-uniform visibility weighting algorithm is 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 is followed by a Gaussian tapering of the visibilities to achieve the final target dirty PSF properties, namely the most Gaussian possible dirty beam 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 is 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 is then rescaled to have an RMS fluctuation level, σ_i, corresponding to the nominal sensitivity level of the channel degraded by its RFI noise ratio:

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

3.3.4 *Simulated sampling and deconvolution*

The HI net absorption and emission datacube (Section??) is then subjected to simulated deconvolution and residual degradation by the relevant synthesized dirty beam. All features, both positive and negative, that deviate from zero by more than three times the local noise level, $3\sigma_i$, are extracted as a "clean" image and replaced by that threshold to form a residual sky image. The residual sky image is subjected to a linear deconvolution (via FFT division) with a 7×7 arcsec Gaussian, truncated at 10% of the peak and then convolved with the dirty beam. The final data product cube is 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) Catalogue limitations, arising from the independent generation of HI and continuum catalogues.

(ii) 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.

(iii) HI emission model limitations, arising from the limited number of real HI observations used to generate simulated HI subcubes. An assumption of negligible HI self-opacity is also made which, although widely adopted in current literature, is unlikely to be the case in reality (see e.g. Braun 2012).

(iv) HI absorption model limitations, due to very coarse sampling used to assess physical properties along the line of sight in order to introduce HI 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.

(v) 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 datasets 7.4 PB in size: far exceeding current capabilities .

4 METHODS

SDC2 ran for a duration of six months from February to July 2021. Participating teams made use of a range of methods to tackle the problem, first making use of the smaller development dataset and truth catalogue in order to investigate techniques. 12 teams made a

successful submission entry using the full Challenge dataset. The methods employed by each of those finalist teams are presented below.

4.1 Coin

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

For detection and characterisation of the HI sources in the Challenge datacube, our team implemented and tested a few modern machine learning (ML) algorithms from scratch. In addition, the team developed its own 'classical' baseline detection algorithm based on wavelet filtering for denoising and segmentation, complemented by standard PYTHON routines for source characterisation and/or ML-regression for derivation of source properties.

We considered the following 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 et al. (2015a) to 3D. It was trained on 3D cubic patches that each contain a source without any preprocessing. The source positions, needed to create the training patches, were taken from the provided development catalogue. High (> 90%) rates of false positives could be mitigated to moderate levels (~ 50%; see Fig. 7) by imposing interconnectivity and size cuts on the potential sources, and by discarding continuum-bright areas. Source positions (RA, Dec, central frequency, w20) 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 specialized CNNs (He et al. 2016, type ResNet) applied to the source candidate 3D cutouts. The position angle PA was directly derived from the masks using the SCIKIT-IMAGE package for labelling and ellipse fitting; inclination could not be fitted for most objects. Our last submission during the Challenge was derived with this pipeline, achieving a ~50:50 ratio between true and false positives for 0.25 deg^2 cutouts of the evaluation cube. We tested that this ratio remained roughly constant across the 1 TB cube. For comparison, our 'classical baseline' algorithm detected < 10% true positives for the Challenge data release (and > 90% true for an earlier higher S/N data release), in both cases with an order of magnitude higher rates of false positives. Basic pipeline steps were: Gaussian filtering in frequency direction, wavelet filtering and thresholding, interscale connectivity (Scherzer 2010) and reconstruction. For all approaches the channels affected by residual RFIs (the first 324), as measured by the per-channel signal mean and variance, were discarded.

We conclude that further cleaning and denoising and/or application of techniques from the 'classical' baseline such as wavelet filtering jointly with our machine learning pipeline is needed to improve on our method. Alternatively, further steps that include classification and/or 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 HI data and the need of additional cleaning algorithms jointly with networks and/or multi-step procedures, such as a classification step, when faced with low S/N data.

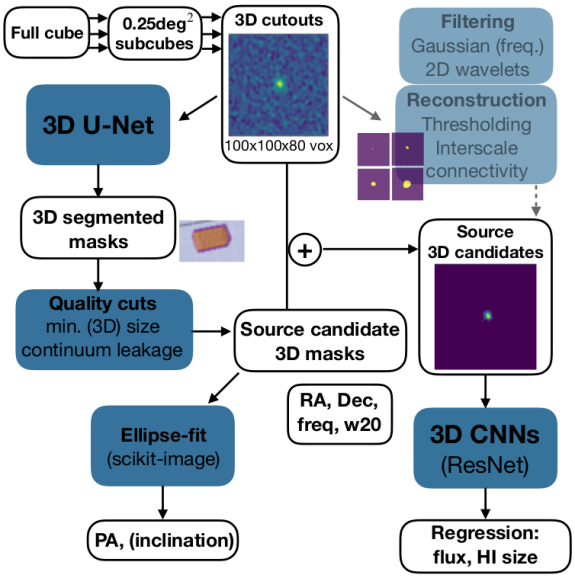

Figure 1. Data processing pipeline used by the coin team.

4.2 FORSKA-Sweden

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

A machine learning-based pipeline was used, trained on the development cube, for which the truth catalogue of the underlying simulation was known. The lower 80% of the development cube, split along the x-axis, was dedicated for training, while the remaining 20% was used for validation and tuning of hyperparameters.

The first step of the production-ready pipeline was a convolutional neural network (CNN) for segmentation of the raw data, which produced a binary mask separating voxels assigned to either a galaxy or not, as visualized in Figure 2. Next, the *merging* and *mask dilation* modules from SoFiA 1.3.2 (Serra et al. 2015b) were employed for post-processing of the mask and extraction of coherent segments into a list of separated sources. The last step of the pipeline was to compute the characterisation 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 segmentation CNN was trained using a binary mask generated from the development truth catalogue, where for each source all voxels within the elliptical cylinder defined by the source's position angle, major axis, minor axis and line width were marked with 1. We used the soft Dice loss as the objective function (Milletari et al. 2016a; Khvedchenya 2019). When training the CNN, batches of 128 cubes of size $32 \times 32 \times 32$ voxels were sampled from the training area. Half of these cubes contained voxels assigned to a source in the target mask, which caused galaxy voxels to be over-represented in a training batch compared to the full development cube. This over-representation made training more efficient.

An U-net architecture (Ronneberger et al. 2015b) was used, with an encoder of a ResNet architecture (He et al. 2016). The initial weights of the model, pretrained from ImageNet, were provided by the PYTORCH-based SEGMENTATION MODELS package (Yakubovskiy 2020). Each 2-dimensional $k \times k$-filter of the pretrained model was converted to a 3-dimensional filter with a procedure similar to Yang et al. (2021). We aligned two dimensions to

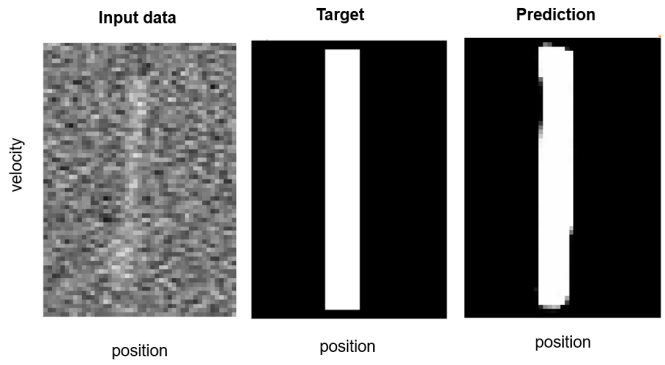

Figure 2. Cross-section images of input data, target and prediction with velocity and one positional dimension for one of the sources in the cube. The position axis is aligned with the major axis of the source.

the spatial plan, 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. Validation was performed regularly during training by using the most recent weights obtained from training and a fixed set of hyperparameters for the post-processing. The score computed from the validation procedure was intended to mimic the scoring of the Challenge. The best model from training was then used as basis for hyperparameter tuning, again using the mimicked scoring.

4.3 EPFL

G. Fourestey, A. Galan, C. Gheller, D. Korber, A. Peel, M. Sargent, E. Tolley

The EPFL team used a variety of techniques developed specifically for the Challenge. The data processing pipeline shown in Figure 3 began with domain decomposition. Overlapping domains are defined by dividing the data cube along RA and Dec. Each of these domains is then analysed by a separate node on the computing system.

First, each domain is denoised using 3D wavelet filtering. To achieve this, different wavelet functions are used in the 2D spatial 2D and 1D frequency dimensions of the data cube. The 2D spatial decomposition uses the Isotropic Undecimated Wavelet Transform Starck et al. (2007), and the 1D frequency axis uses the decimated 9/7 wavelet transform Vonesch et al. (2007).

Next, a joint likelihood model is calculated from the residual noise in the data cube. This model is used to identify HI source candidates through null hypothesis testing in a sliding window along the frequency axis. Voxels with a likelihood score below a certain threshold (i.e. not likely to be noise) are grouped into islands. The size and arrangement of these islands are used to reject data artifacts. Ultimately the location of the voxel with the highest significance is kept as an HI source candidate location.

The rest of the steps of the pipeline use CNNs to classify and characterize the candidate HI sources returned by the likelihood source finder. These networks were trained on the development dataset using extensive data augmentation. First, candidates are distinguished between data artifacts and true HI sources by a classifier CNN. These candidate locations are then used to extract data from the original, non-denoised domain and passed to an Inception CNN which calculates the HI source parameters. The Inception CNN

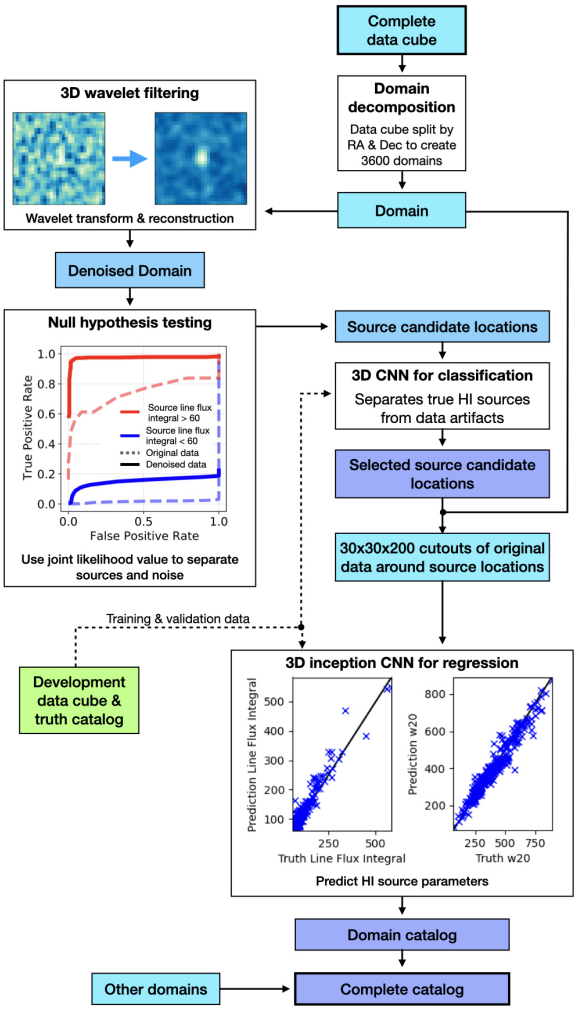

Figure 3. Data processing pipeline used by the EPFL team.

uses multiple modules to examine data features at different scales. Finally, the HI source locations and features for each domain are concatenated to create the full catalog.

4.4 HI FRIENDS

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

The HI-FRIENDS team implemented a workflow (Moldon et al. 2021) 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[5], 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 jobs of the data analysis. SNAKEMAKE manages the installation of the software dependencies of each rule in isolated environments using conda[6] and each rule executes a single program, script, shell command or

[5] https://github.com/HI-FRIENDS-SDC2/hi-friends
[6] https://docs.conda.io/en/latest/

JUPYTER notebook. With this methodology, each step can be developed, tested and executed independently from the others, which benefits modularization and reproducibility of the workflow.

First, the cube is divided in smaller subcubes by finding a regular grid that covers the observed region of the sky. Adjacent subcubes include an overlap of 40 pixels (112 arcsec) to avoid splitting the largest expected galaxy potentially lying at the edge of a subcube. The data cube in fits format is pre-processed using the library SPECTRAL-CUBE[7] from ASTROPY (Astropy Collaboration et al. 2018). Second, a rule executes SOFIA-2 to mask each subcube and to characterize the parameters of the identified sources. 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 optimize 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 source parameters relevant to the SDC2, which are converted to 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 et al. (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 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 HI size (D_{HI}) and HI mass (M_{HI}) of nearby galaxies. Several plots are produced by the different python scripts during the workflow execution, and a final visualization rule generates a JUPYTER notebook with a summary of the most relevant plots.

Our workflow tries 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[8] and WorkflowHub[9], 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 and they can be accessed openly without authentication. To make it interoperable, our workflow can be easily deployed in different platforms and all the dependencies can either be automatically installed (e.g., it can be deployed in a virtual machine instance in myBinder[10]) or executed through singularity, podman or docker containers. Finally, to make it reusable we used an open license, we included workflow documentation[11] 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[12]. 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, M. Kunz, B. Bassett, V. Nistane, N. Oozeer

To address the source characterization problem, we proposed a multi-level deep learning approach. One challenging aspect is that the data is in 3-dimensions. The models should detect 3-dimensional patterns as the region of interest and the corresponding characterization concerning the given set of parameters.

Our proposal extends a similar challenge in two dimensions Vafaei Sadr et al. (2019) where they divided the detection task into image cleaning and source finding. The motivation of this approach comes from the recent progress in the image to image translation techniques and using multi-levels of supervision. One can utilize prior knowledge about source shapes to magnify signals (suppress background). This step applies image-to-image translation techniques and is close to the image cleaning. In the second level, one trains a model to find and characterize the objects.

We focus on training a model that reconstructs the 'clean' image. We investigate two approaches where the first uses 2D cuts through frequency as grayscale images. In this approach, the model learns to retrieve information employing only transverse information. On the second idea, we extend the inputs into 3D to benefit from longitudinal patterns by adding different frequencies as convolutional channels (multichannel image). We also use a 128×128 sliding window to manage memory consumption, the mean squared error loss function, and a decaying learning rate. We used the standard image processor in TENSORFLOW Abadi et al. (2015) for a minimal augmentation, where the ranges are 1 degree for rotation, 1

One can interpret the output as a probability map. The ground truth for that part is a source map that contains masks or probability values according to the selected loss function. Then the initial learning rate is initiated by 1e-3 with a 0.95 decay per 10 epochs using Adam optimizer.

We developed our pipeline to examine different architectures as follows: V-Net Milletari et al. (2016b), Attention U-Net Oktay et al. (2018), R2U-Net Alom et al. (2018), U^2net Qin et al. (2020), UNet3+ Huang et al. (2020), TransUNet Chen et al. (2021), and ResUNet-a **?** where one can find most of the implementations in the KERAS-UNET-COLLECTION Sha (2021) package.

Our results show using the evaluation set, U^2net shows the best performance considering the loss function. U^2net employs residual U-blocks in another U shape architecture. It applies the deep-supervision technique to supervise training in all scales by downgrading the output.

In the second part, we use the 3D output of U^2net to find the objects using a peak finder algorithm. A peak is simply the pixel that is larger than all its 27 neighbors. The "found" catalogue then goes into a modified 8-layers HighRes3DNet Li et al. (2017) as a regressor for characterization and generating the final catalog.

[7] https://spectral-cube.readthedocs.io/en/latest/index.html
[8] https://zenodo.org/record/5172930
[9] https://workflowhub.eu/workflows/141
[10] https://mybinder.org/
[11] https://hi-friends-sdc2.readthedocs.io/en/latest/
[12] https://bestpractices.coreinfrastructure.org/en/projects/5138

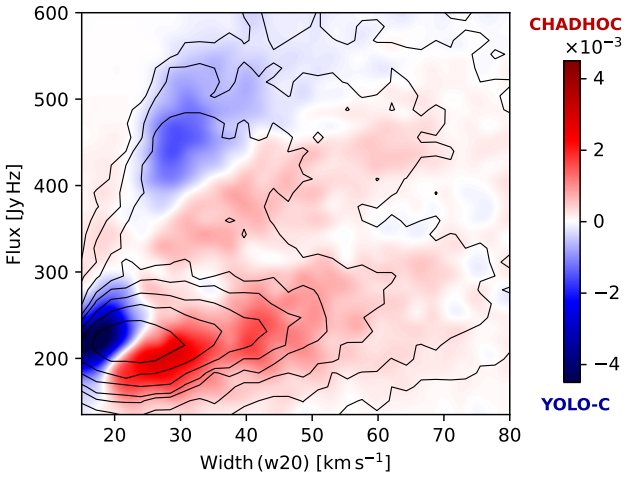

Figure 4. Difference in the number of sources found between CHADHOC and YOLO-C catalogues in a Flux against w20 parameter space. The color encodes the difference in the local number of sources as a proportion of the total merged catalogue size (32652 predicted sources). The contours are the local number of sources averaged between the two catalogues with values: 6,14,30,50,64,92,128,192. The 2D histogram is computed on a 30x30 grid and plotted with interpolation.

4.6 JLRAT

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

The pipeline of the JLRAT team first divided the whole dataset into small cubes (the size is $320 \times 320 \times 160$ for RA, Dec and Freq), then applied a convolutional neural network for the signal of interest detection, which is a fully convolutional layers with a softmax layer. The network mainly finds region proposals that include candidate signals. The network works in the spectrum domain. Its input is a spectrum, and the output is a mask that indicates where a candidate signal is. After that, we compute the correlation among the candidate spectrum with other spectra in the small cube. The correlation is estimated by using the inner product; this conducts the correlation in the space domain. The network gave a three-dimensional cube (position-position-frequency) for a predicted galaxy, in which the approximate position, size, and accurate line-width (w20) of the galaxy were given. A two-dimensional Gaussian function was used to fit the moment zero map to produce an ellipse with the central position (RA and Dec), major axis, position angle, and inclination of the galaxy with an intensity cutoff at 1 $M_\odot$ pc^{-2}. The flux integral was given by integrating the spectra within the ellipse in both space and frequency.

4.7 MINERVA: the YOLO-CHADHOC pipeline

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

For the purpose of the SDC2 we implemented a highly customised version of a YOLO (You Only Look Once, Redmon et al. 2015; Redmon & Farhadi 2016, 2018) network which is a regression based Convolutional Neural Network dedicated to object detection and classification. To train our network we added low level YOLO capabilities in our own general purpose CNN framework CIANNA[13] (Convolutional Interactive Artificial Neural Networks by/for Astrophysicists) which is CUDA GPU accelerated.

Our custom YOLO network works on sub-volumes of $48 \times 48 \times 192$ (RA, Dec, Freq) pixels and performs detection based on a sub-grid of size $6 \times 6 \times 12$. Each grid element (sub-cube of $8 \times 8 \times 16$ pixels) can be associated to a candidate detection with the following parameters: x, y, z the object position inside the sub grid element; w, h, d the 'box' dimension in which the object is inscribed and based on size priors (10, 10, 48 in pixel size); O an objectness score that combines a probability of being a real detection and a box matching score. In addition, YOLO networks are designed to predict classes for each box, but this was not required for SDC2. However, we managed to modify the YOLO loss to add the capability of predicting an arbitrary number of regressed parameter. This allowed us to predict the required HI flux, size, line width, position angle and inclination at the same time for each box.

The definition of the training sample is of major importance to get good results. Most of the sources in the large development truth catalogue are impossible to detect for the network, and tagging them as positive detection would lead to a poorly trained model. For YOLO we used a combination of criteria: i) using the CHADHOC classical detection (see section below), ii) using a volume brightness threshold, and iii) using a local S/N ratio estimation. Our refined training set contains around ∼1500 'true' objects, with 10% kept apart for validation.

Our YOLO network is made of 21 (3D)-convolutional layers which alternate several 'large' filters (usually $3 \times 3 \times 5$) that extract morphological properties and fewer 'smaller' filters (usually $1 \times 1 \times 3$) that force a higher degree feature space and allow to preserve a manageable number of weights to optimise. Some of the layers also include a higher stride value in order to progressively reduce the dimension down to the $6 \times 6 \times 12$ grid and the few last layers include dropout for regularisation and error estimation. The network was trained by selecting either: a sub-volume that contains a true source (at least); or a random empty field to learn to exclude all types of noise aggregation and artifacts. All inputs where augmented using position and frequency offset as well as flips. Despite the fact that YOLO networks are typically much faster than competing networks (Fast R-CNN,...) our customised architecture still requires up to 36 hours of training on a single RTX 3090 GPU using FP16/FP32 Tensor Core mixed precision training (∼3 times faster than a V100). The trained network has an inference speed of 76 input cubes (48x48x192) per second using a V100 GPU on Jean-Zay/IDRIS, but due to necessary partial overlap and to RAM limitations, it still requires up to 20 GPU hours to get the complete prediction on the full ∼1 TB data cube.

4.7.2 CHADHOC

The Convolutional Hybrid Ad-Hoc pipeline (CHADHOC) has been developed entirely to answer the SDC2. It is composed of three steps: a traditional detection algorithm, a Convolutional Neural Network

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

(CNN) for identifying true sources among the detections, and a set of CNNs for source parameter estimation.

The detection step

For detection, a traditional algorithm is used. The signal cube is first pre-processed by smoothing along the frequency dimension (600 kHz width). Then the signal is converted to a signal-to-noise ratio on a per channel basis. Pixels below a fixed S/N ratio (2.2 was found to give good results) are filtered out, and the remaining pixels are aggregated into sources using a simple friend-of-friend linking process (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 sum of the S/N values of the pixels. When dealing with the full cube, we divide the cube in a number of manageable chunks (25 in practice) and produce one catalogue for each chunk.

The selection step

This step is performed with a CNN. A learning sample is built by cross-matching the 10^5 brightest detections in the development cube with the truth catalogue, thus assigning a True/False label to each detection. Unsmoothed S/N cutouts of $38 \times 38 \times 100$ pixels (frequency last) around the position of each detection are the inputs for the network. The learning set is augmented by flipping in all three dimensions, and a test set is isolated made of one third of the detections. The comparatively light network is made of 5 3D-convolutional layers (8, 16, 32, 32 and 8 filters) and 3 dense layers (96, 32 and 2 neurons). Batch normalisation, dropouts and pooling layers are inserted between almost every convolutional and dense layers. 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. For each detection, the output is not a simple True/False statement but a number between 0. (False) and 1. (True). The threshold where the source is labelled as True is a parameter that must be tuned to maximise the metric defined by the SDC2, a procedure that is not equivalent to minimising the network loss, which a simple RMS error. This optimisation is performed independently of the training.

Parameter estimation

A distinct CNN has been developed to predict each of the sources parameters, including a correction to the source position computed during the detection step. The architecture is similar to the one of the CNN for sources selection, with small variations: for example, no dropout is used between convolutional layers for predicting the line flux. Cutouts around the $\sim$ 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 hundreds epochs in a few to 20 minutes each on a Tesla V100 GPU. Training longer results in overfitting and a drop in accuracy.

Many small things 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 in Fig. 4, we can check that they occupy slightly different regions. For example, CHADHOC tends to find a (slightly) larger number of typical sources compared to YOLO, but missed more low-brightness sources because of the hard S/N threshold applied during the detection step. Thus merging the catalogue 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 maximise purity and completeness. Finally the errors on the parameter predictions are at least partially uncorrelated between the two pipelines. Thus averaging the predicted values also improves the resulting catalogue.

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 some trials, the NAOC-Tianlai team used primarily the SoFiA-2 software in the processing of the SDC2 datasets. For the optimal values of the parameters of the SoFiA-2 software, we first used grid search to find some rough estimate, then refined the search using an MCMC simulation in the parameter space. While we are developing a dedicated cosmological simulation to make bottom-up checks, within the time frame of the Challenge, we have mainly used the development and development-large data sets provided by the SDC2 for this search, and then applied the optimization result to the processing of the final dataset.

Due to the memory bottleneck and the consideration of avoiding much division along the spectral axis, the datasets have been split into a number of smaller subsets with the size of about $330 \times 330 \times 3340$ pixels for processing, and the adjacent cubes have an overlap of 10 or 20 pixels/channels along each axis to ensure HI galaxies on the border region will not be missed. The full data set has been divided into 18×18×2 cubes when processing.

Our main procedure of parameters selection is as follows:

(i) Set a list of values to be searched for each parameter of interest, such as `replacement`, `threshold` in *scfind* module, `minSizeZ`, `radiusZ` in *linker* module, and `minSNR`, `threshold`, `scaleKernel` in *reliability* module. We then processed in parallel the data cube of which the true HI galaxy catalog is known, with the different combinations of parameters values.

(ii) Select the optimal parameters combination according to the output catalogs from the previous step. The criteria for choosing the optimal parameter combination include the *total detection number*, the *match rate*(true detection/total detection), the final *score* which is calculated according to the rules explained in Section 5.1, or a combination of them. In our trial, we applied the combination of these three statistics above, and in that order, i.e., we first sorted the results by the total detection number, and dropped small value ones, then sorted the truncated results by the match rate, again drop those with small values, and finally we sorted the remaining results by the final score and selected the optimal one.

(iii) To make the optimal parameters combination found more

robust, different test data cubes are processed following the procedure given above, and the combination which performed well on all data cubes is selected.

A reference parameters setting from our trial is `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 *scfind* module for S+C finder in SoFiA-2, `radiusXY/Z = 2`, `minSizeXY = 5`, `minSizeZ = 20` in *linker* module for merging the masked pixels detected by the finder, and `threshold = 0.5`, `scaleKernel = 0.3`, `minSNR = 2.0` in *reliability* module for reliability calculation and filtering. In our processing, each subcube or instance of parameters combinations took about 5 minutes with 1 CPU thread.

Finally, we applied the optimal parameter combination to the processing of all subsets from the Challenge dataset, and merged the results.

4.9 Team SoFiA

K. M. Hess, R. J. Jurek, S. Kitaeff, P. Serra, A. X. Shen, J. M. 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. Development version 2.3.1 of the software, dated 22 July 2021,[14] was used in the final run submitted to the scoring service. After flagging of bright continuum sources > 7 mJy followed by noise normalisation in each spectral channel, SoFiA's 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.

To minimize processing time, 80 instances of SoFiA were run in parallel, each operating on a smaller region ($\approx$ 11.8 GB) of the full cube. The processing time for an individual instance was just under 25 minutes, increasing to slightly more than 2 hours 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.

Based on tests using the development cube, we improved the reliability of the resulting catalogue 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 signal-to-noise ratio of the detection. Detection counts for the original and filtered catalogue from the development cube are shown in Fig. 5 as a function of SNR. Our final detection rate peaks at SNR $\approx$ 3, with a reliability of close to 1 down to SNR $\approx$ 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%.

It should be emphasised 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 maximise 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″ and an inclination of 57.3° 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 HI disc size.

[14] https://github.com/SoFiA-Admin/SoFiA-2/tree/11ff5fb01a8e3061a79d47b1ec3d353c429adf33

4.10 SHAO

S. Jaiswal, B. Lao, J. N. H. S. Aditya, Y. Zhang, A. Wang, X. Yang

We, the SHAO team, developed a fully-automated pipeline in Python to work on the Challenge dataset. Our method involved the following steps: 1) We first sliced the line data cube into individual frequency channel images and perform the source finding with 2.5 sigma detection threshold (for ~99% detection confidence) and minimum 2 pixels for detection area using the SExtractor software (Bertin & Arnouts 1996) on each of these channel images. 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) We estimated the range of channels for each source detected in at least 2 consecutive channel images and added 1 extra channels on both sides. 4) We extracted subcube, for the channel range obtained in previous step, having 12 pixel spatial size around each identified source. 5) We made moment-0 map for each extracted source using its subcube, after masking negative flux densities. 6) We used the source-finding algorithm (SExtractor) on the moment-0 map of each extracted HI source to estimate the source RA and DEC coordinates, major axis, minor axis, position angle and integrated flux. Inclination angle was estimated using the relation given by Hubble (1926); Holmberg (1946). 7) We estimated the flux densities within a box (of 6 pixels around the source position) on every channel image of each subcube to make global HI profile for each source. 8) We finally fit a single Gaussian model to estimate the central frequency of HI emission and line width at 20% of the peak.

The score obtained by this method is not very satisfactory (Table 2). However, it gave a confidence to us on dealing with large HI 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 future. The use of machine learning techniques could be a good choice for such datasets.

4.11 Spardha

A. K. Shaw, N. N. Patra, A. Chakraborty, R. Mondal, S. Choudhuri, A. Mazumder, M. Jagannath

We, the SPARDHA team, have developed a PYTHON based pipeline which starts with dividing the whole 1 TB data into several small cubelets. We analyze all the cubelets in parallel using an MPI based implementation, where we have run parallel instances of SoFiA-2 on each cubelet to find the sources. We have tuned the parameters of SoFiA-2 to maximize the number of detected sources. A total of 118

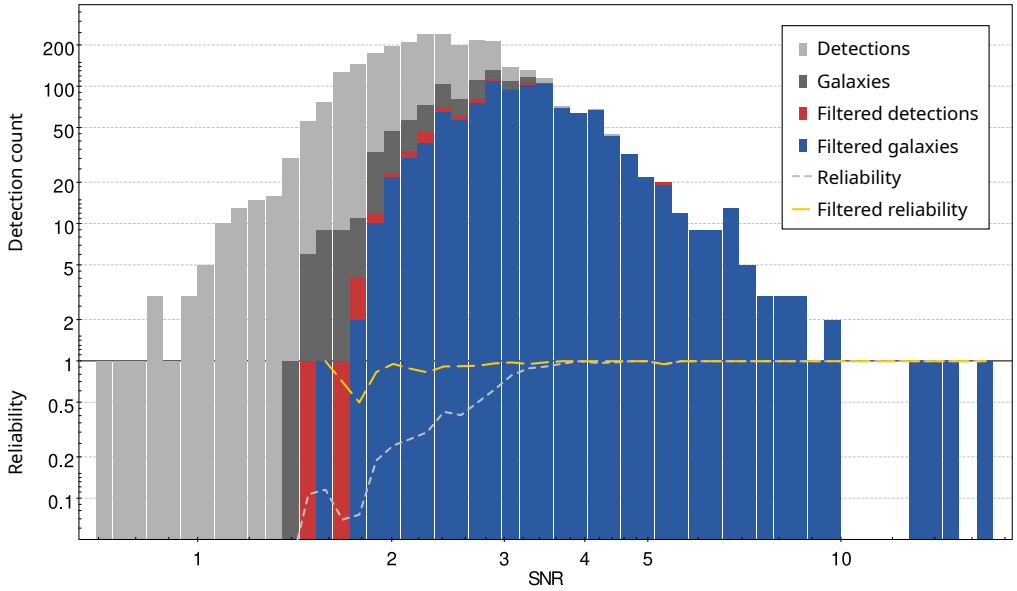

Figure 5. Histogram of total detections (light-grey), real galaxies (dark-grey), detections after filtering (red) and real galaxies after filtering (blue) as a function of integrated signal-to-noise ratio from a SoFiA run on the development cube. The reliability of the original and filtered catalogue is shown as the grey and orange curve, respectively. Parameter space filtering significantly boosts SoFiA's reliability at low SNR.

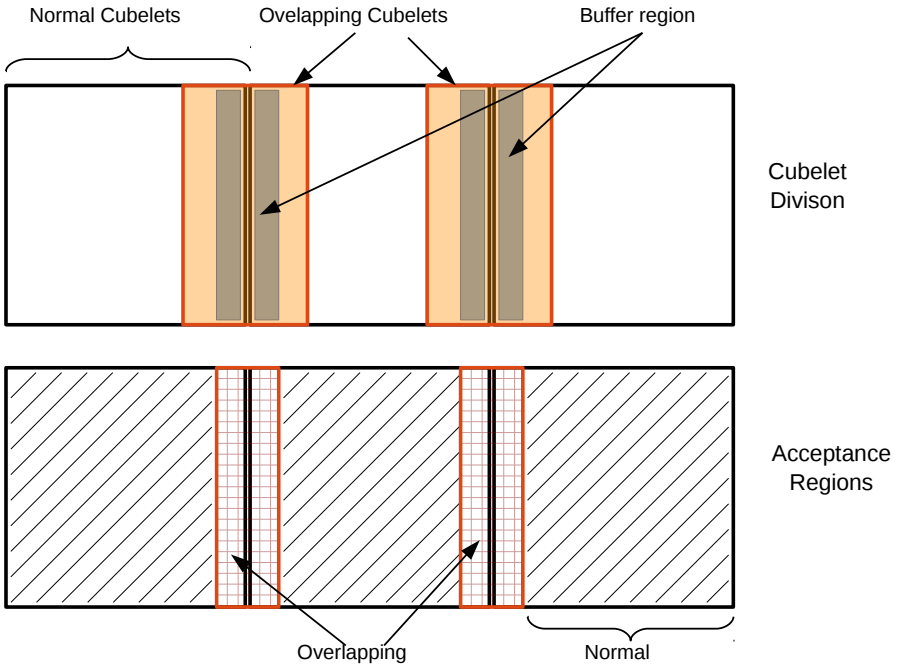

Figure 6. Shows the 2D projection of the schematic division of the data into the *Normal* and *Overlapping* cubelets along one of the axis (top row) and the corresponding Acceptance regions (bottom row).

cubelets were analyzed which can be categorized into two groups namely, (1) Normal cubelet and (2) Overlapping cubelet. We divide the whole data cube into consecutive blocks of equal dimensions which we indicate as Normal cubelets as shown by black outlined boxes in top and bottom rows of Fig. 6.

We separately define and analyze the Overlapping cubelets which encompass regions in the adjacent normal cubelets and are centered at their common boundary as shown by orange boxes in the top row of Fig. 6. This is done for detecting the sources which fall at the common boundaries. The normal cubelets and the overlapping cubelets always have buffer zones (e.g. gray regions for Normal cubelets in top row) around their faces in order to avoid the confusion

between the sources detected near their common boundaries. We conservatively set the width of buffer zones based on the physically motivated values of the spatial (on the sky plane) and frequency extent of typical galaxies scaled at the desired redshifts. We choose the maximum extent of the galaxy on the sky plane to be ∼ 80 kpc (Wang et al. 2016) which roughly corresponds to 10 pixels on the nearest frequency channel. Therefore the buffer region was set to be twice *i.e.* 20 pixels. Hence we make the Overlapping region $4 \times 20 = 80$ pixels wide. Similarly along the frequency direction, galaxies can have extent ≈ 72 channels which corresponds to a line-width ∼ 500 km/s. Therefore the widths of the buffer region and Overlapping region along frequency-axis are 144 and 288 channels respectively. We always accept any source whose center is detected within the cubelet but not in the buffer zone as demonstrated in the bottom row of Fig. 6. The acceptance regions of the cubelets (normal and overlapping) are defined in a particular way so that they span the whole data cube contiguously when arranged accordingly. Although this increases the computation a bit due to analyzing some part of data multiple times (once in normal cubelet and once in associated overlapping cubelet), it ensures that there is no common source present in the list after this step. Analyzing cubelets is the most time consuming part in our pipeline. We analyze 118 cubelets on 472 cores in parallel in around 15 minutes.

Next, we use the physical equations to convert the SoFiA-2 catalogue into the SDC prescribed units and discard the bad detections, *i.e.* sources having *NaN* values in the columns or with negative flux values, etc. In the final stage we put a cap on w_{20} to discard the detections with unusual velocity-width. Motivated by the physical models/observations of the galaxies, we have conservatively accepted the sources having $w_{20} \in [60, 500]$ km/s (McGaugh et al. 2000). We finally arrange the catalogue in the descending order of the flux values. Based on our experience with the 'Development Data Cube', for which the exact source properties are known, we choose around top 35% of total sources to generate the final catalogue for submission.

4.12 Starmech

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

We tackled the Challenge from the point of view of trying to find ways of dealing with the large dataset 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 compute resources are provided as part of shared SKA Regional Centres. Thus our main aim was to use existing source finding tools to process the cube, adapted to the large dataset. We investigated PyBDSF (Mohan & Rafferty 2015), a continuum source finder, and SoFiA and SoFiA-2, two generations of publicly available 3D source finder already optimized for HI (Westmeier et al. 2021).

4.12.1 Approach

While PyBDSF readily generated a catalogue of the continuum sources, and could be run on many slices in frequency space to pick out the brightest emission-line objects even in the full-size data cube, 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, which are optimized for the type of data used in the Challenge. 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.

Roughly speaking, SoFiA-2 requires RAM equal to a little over twice the size of the input data cube, which means that we could not simply run it on the full 850-GB data cube using the resources provided for the Challenge, though it ran without difficulty on the various test data cubes that were available. To use SoFiA-2 we therefore needed to slice the cube either in frequency (z) or spatially (xy) in order to produce sub-cubes that would fit in the available RAM. 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 HI series in the SKA era since the full dataset sizes will exceed any feasible RAM in a single node for the foreseeable future.

4.12.2 Implementation

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 so that each individual sub-cube can fit into the available RAM (or into a specified amount of RAM, for testing purposes). Assuming a tiling of $n \times n$, it then tiles the cube with n^2 overlapping square (or in general 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, using the input.region parameter of SoFiA-2 to select a sub-region; the resulting catalogue is read in using Astropy and the pixel co-ordinates converted back to the native co-ordinates of the full cube. The end result is n^2 overlapping catalogues in memory as Astropy tables.

The final step is to remove the duplicate sources, which may naturally result from the overlapping regions. Considering catalogues from adjacent sub-cubes pairwise, we firstly discard all catalogue entries whose pixel position more than g pixels from the edge of a sub-cube – these should already be present in another catalogue, and are discarded to avoid any edge effects in the data which might cause a mismatch. Then the remaining overlap region, $2g$ pixels in width, height or both, is cross-matched in position and sources whose position and frequency differ by less than user-defined threshold values are considered duplicates and discarded from one of the two catalogues. Finally the resulting n^2 de-duplicated catalogues are merged and written to disk. For our final submission we used SoFiA-2 default parameters with an `scfind.threshold` of 4.5 sigma, $g = 20$ 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, since the source characterization and hence de-duplication will break down at the point where a source is larger than the guard region. 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. As team effort was voluntary and constrained we did not move on to the next obvious step of optimising the pa-

rameters used for SoFiA-2 based on further runs on the test and development datasets.

The conversion of the SoFiA-2 results into the submission format was performed using a trivial Python script where some parameter conversion was done as well (for example, the output for a source flux by SoFiA-2 is in Jy beam^{-1} units but for submission it had to be converted into the line flux integral in Jy Hz).

4.12.3 Other approaches

We would like to have explored the utility of data compression as part of the source finding, for example by using the equivalent of 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.

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, prepared as a pip-installable Python package, could be accessed from any machine by using a simple command line tool. The web-based service submitted each catalogue to a scoring API hosted on a remote server and returned a score. Teams were limited to a maximum submission rate of 30 submissions per 24 hour period.

5.1 Scoring procedure

The scoring API 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

The cross-match procedure considers the position of a source in the 3D cube, identified by RA, Dec and central frequency. All submitted sources with positions within which a truth catalogue source is in range are recorded as matches. For each submitted source, this range in the spatial and frequency dimensions is determined by the beam-convolved submitted HI 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 HI 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 HI 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.

During the cross-matching, 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 multi-parameter error (eq. 5.1.1) will be retained. This procedure ensures that matches in crowded regions will 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 multi-dimensional error with a threshold value, above which any nominally matched submitted sources are discarded and counted as false positives. The multi-parameter error D is calculated using the Euclidean distance between truth and submitted sources in normalised 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 normalised following the definitions in Table 1. The error on HI size is at this stage normalised by the beam-convolved true HI size in order not to lead to the preferential rejection of unresolved sources. The multi-dimensional error threshold is set at 5, i.e. the sum in quadrature of unit normalised 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), HI 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

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

where err_i^j is the error on the attribute and thr_j is a threshold applied to that attribute for all sources. Errors calculated in this step are detailed in Table 1, along with corresponding threshold values. Finally, the weighted scores of submitted sources are averaged over any duplicate matches with unique truth sources.

Property	Error term	Threshold
RA and Dec, x, y	$D_{\rm pos} = \frac{(x-x')^2 + (y-y')^2}{S'}$	0.3
HI size, S	$D_{\rm HI\ size} = \frac{\lvert S - S' \rvert}{S'}$	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

Table 1. Definitions of errors and threshold values for the properties of sources. 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 HI 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 HI 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: $|\theta - \theta'| = \mathrm{atan2}[\sin(\theta - \theta'), \cos(\theta - \theta')]$

5.1.3 *Final score per submission*

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

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

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, and reproducibility awards 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)[15], which was provided to teams for the purposes of self-assessment during the Challenge. The checklist[16] considered the following criteria:

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

(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) Utilise tests

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

[15] https://www.software.ac.uk/
[16] https://sdc2.astronomers.skatelescope.org/sdc2-challenge/reproducibility-awards

6 RESULTS AND ANALYSIS

The final scores of all teams who submitted a catalogue based on the full Challenge dataset are reported in Table 2. Each team's number of detections, $N_{\rm d}$ – composed of matches, $N_{\rm m}$, and false positives, $N_{\rm f}$ – are also listed, along with the number of matches and the overall accuracy of each team's method, defined as the percentage accuracy of source property measurement according to Section 5.1.2, averaged over all properties for all matches per team.

We note that the scoring algorithm (Section 5), designed to penalise 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 reliability is low. 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 characterisation performance of these teams' methods.

Several conventions and conversions are used during the characterisation of HI spectral line data which, without clear and unambiguous specification, can lead to inconsistencies between catalogues and between physical and measured properties. Position angle – spanning 360 degrees for spectral line images – can follow a number of alternative conventions depending on the direction relative to which the angle is measured, the direction of angle rotation, and the choice between receding and approaching sides of the major axis of the source. Velocity – describing both the recessional movement of an object and its rotation – relates directly to rest frequency and observed frequency via the relativistic Doppler effect, and is often approximated using either "radio" or "optical" conventions for objects moving at non-relativistic speeds. When measuring rotational velocities at high redshifts, however, cosmological expansion necessitates the use of a nominal rest frequency obtained by redshifting the HI rest frequency by a factor $1/(1+z)$. Room for error arose during the Challenge 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. Future SKAO Science Data Challenges will benefit from additional instructions and examples where ambiguity or unfamiliarity can be anticipated.

6.1 Source finding

Fig. 7 presents for each team the number of matches and false positives, binned according to integrated line flux along with all sources from the truth catalogue, $N_{\rm t}$. When considering matches, truth catalogue line flux values, F', are used; when considering false positives, the lack of corresponding truth values necessitates the use of submitted line flux values, F. Fig. 8 presents reliability,

Team name	Score	N_d	N_m	Accuracy
MINERVA	23254	32652	30841	81
FORSKA-Sweden	22489	33294	31507	77
Team SoFiA	16822	24923	23486	78
NAOC-Tianlai	14416	29151	26020	67
HI-FRIENDS	13903	21903	20828	72
EPFL	8515	19116	16742	65
Spardha	5615	18000	13513	75
Starmech	2096	27799	17560	70
JLRAT	1080	2100	1918	66
Coin	-2	29	17	60
HIRAXers	-2	2	0	-
SHAO	-471	471	0	-

Table 2. SDC2 finalist teams' scores are reported, rounded to the nearest integer. Also reported are the number of detections N_d and matches N_m (Section 5.1.1), and the source characterisation accuracy (Section 5.1.2).

R, and completeness, C, values as a function of integrated line flux, calculated as follows:

$$R(F) = \frac{N_m(F)}{N_d(F)} = \frac{N_m(F)}{N_m(F) + N_f(F)}; \tag{20}$$

$$C(F') = \frac{N_m(F')}{N_t(F')}, \tag{21}$$

where submitted values are again used in the calculation of reliability due to the absence of corresponding truth values for false positives.

6.2 Source characterisation

In order to investigate the performance of teams' methods in the recovery of source properties, several relationships were investigated. Fig. 9 presents error terms (Table 1) calculated without using absolute values and plotted as a function of true property value for flux, size and line width measurements, and as a function of true size, for position and inclination angle measurements.

Fig. 10 compares HI mass distributions constructed using teams' submissions with the input redshift-dependent HI mass function $\phi(M_{HI}, z)$ (equation 3.1). For each team, an HI mass distribution uncorrected for completeness, $\hat{\phi}(M'_{HI}, z)$, is constructed:

$$\hat{\phi}(M'_{HI}) = \frac{dN_m}{dV\, d\log_{10} M'_{HI}}, \tag{22}$$

where dN_m is the average number of matched sources in the volume dV with true HI masses that fall within a logarithmic bin centred on M'_{HI}. True HI masses are generated during our simulation following the redshift-dependent mass function derived from (Jones et al. 2018) (Section 3). The same masses can be derived from true observable catalogue properties according to the relation from Duffy et al. (2012),

$$M'_{HI} = F' \times 49.8 \times D'^{\,2}_{L}\ M_\odot, \tag{23}$$

using the true luminosity distance, D'_L, obtained via the true central frequency, ν'. A second HI mass distribution, $\hat{\phi}(M_{HI}, z)$, is constructed using eq. 6.2 from submitted property values, F and ν, of teams' detections:

$$\hat{\phi}(M_{HI}, z) = \frac{N_d}{dV\, d\log_{10} M_{HI}}, \tag{24}$$

and is used to plot the residual,

$$\Delta\hat{\phi}(M_{HI}, z) = \hat{\phi}(M_{HI}, z) - \hat{\phi}(M'_{HI}, z), \tag{25}$$

between the submitted and true values of teams' matches and detections, respectively, after applying a second order spline interpolation to both distributions.

For each team, the HI mass distribution derived from true mass values, $\hat{\phi}(M'_{HI}, z)$, is interpolated and compared with the input HI mass function, $\phi(M_{HI}, z)$, in order to identify the HI mass above which at least 50 percent of truth catalogue sources are recovered (Table 3). Fig 11 presents this mass for the top eight scoring teams as a function of redshift and compared with the HI mass function 'knee' mass (equation 3.1). Values of the the residual at the knee mass, $\Delta\hat{\phi}(M_{HI,knee})$, are also plotted as a function of redshift.

6.3 Reproducibility awards

Subsection to be completed after reproducibility award deadline

A total of ?? teams submitted entries for the SDC2 reproducibility awards. Table ?? reports the awards granted to each participating team.

7 DISCUSSION

Challenge teams employed a variety of methods to tackle the simulated SKA MID HI dataset. The results show a wide range of performance both within and between methods. In this section we discuss the findings in terms of individual and collective method capabilities.

7.1 Source finding and characterisation

While reliability and completeness (Fig. 8) generally show an increase with increasing flux, 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.

The analysis of source property recovery (Fig. 9) 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 unresolved sources (I need to quantify this number), 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 3.2.1), and can therefore be approximated by making use of line width, flux and size measurements, does not suffer the same problem. The accuracy of position angle and inclination angle measurements are generally independent of true values; line flux, HI size and line width values are also largely

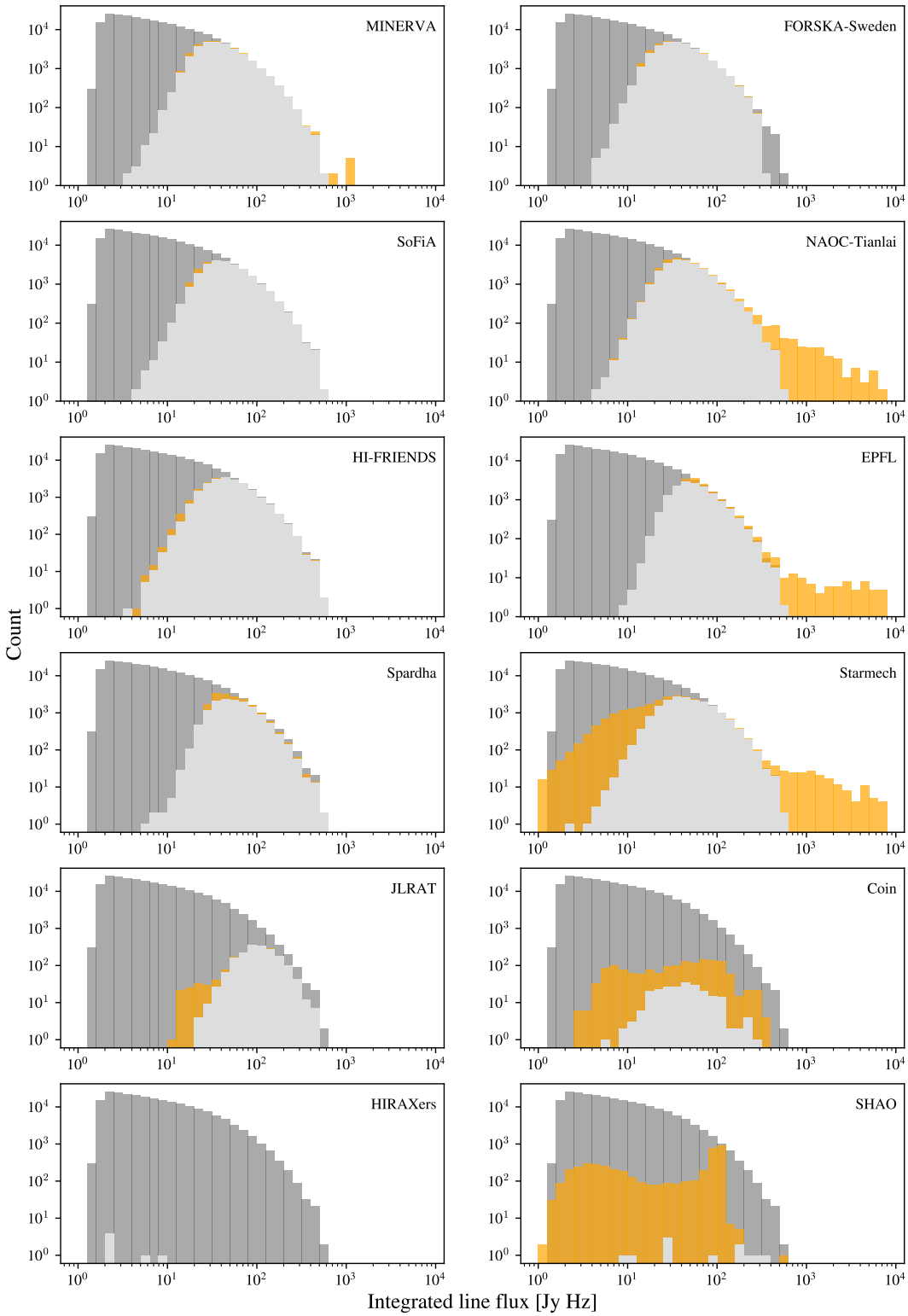

Figure 7. Sources in the full Challenge dataset binned according to integrated line flux value. For each team, all sources in the full truth catalogue (dark grey) are overplotted by the true values of matches (light grey) and by the submitted values of false detections (yellow).

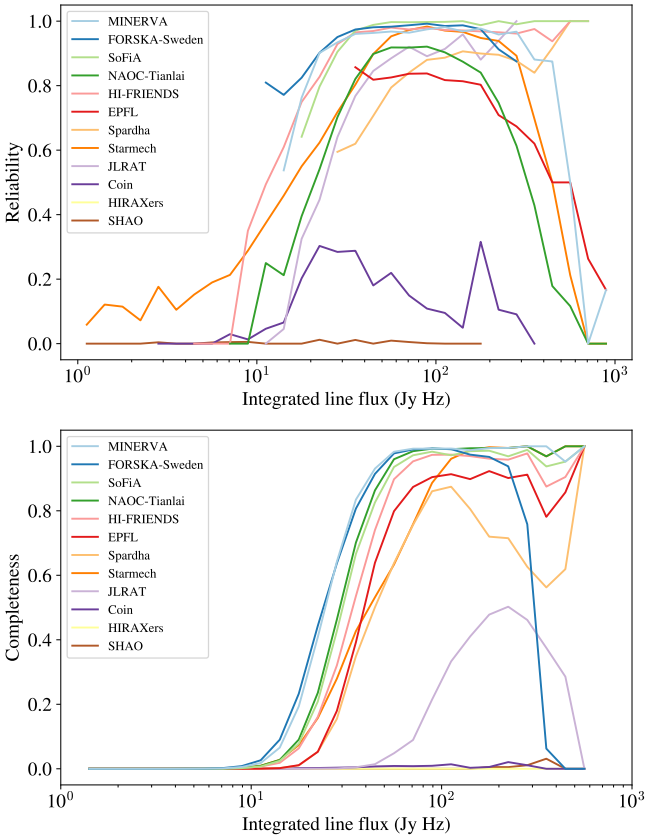

Figure 8. Top: Reliability, defined as the number of matches divided by the number of detections, is plotted for each team as a function of submitted integrated line flux. Bottom: Completeness, defined as the number of matches divided by the number of truth catalogue sources, is plotted for each team as a function of true integrated line flux.

independent above approximate values of 30 Jy Hz, 10 arcsec and 200 km s^{-1}, respectively.

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). Similar noise biases may be apparent in the measurement of HI size and line width, where there is a general tendency to overestimate smaller sizes and underestimate larger sizes. Some teams used the SDC2 development dataset to 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, HI size and line width. The overestimation of HI size is compounded by the finite resolution of the simulated observation: the fractional error on HI 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 HI mass function

The HI mass functions presented in Fig. 10 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 HI 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 percent completeness threshold was chosen to characterise HI mass recovery depths following Rosenberg & Schneider (2002), who, using an HI-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 percent 'sensitivity limit' and the function derived using all sources. Fig. 11 demonstrates that the two top scoring teams' methods are able to probe the HI knee mass with a 50 percent completeness out to a redshift of approximately 0.45, or 1740 Mpc of comoving distance. For comparison, the ALFALFA survey – with a footprint of ~6900 deg^2 – has probed the knee mass out to distances of approximately 200 Mpc.

Two alternative methods are commonly used to convert observations from raw number counts to an intrinsic mass function, accounting for completeness in the process. The $1/V_{\rm max}$ method, originally used to derive the quasar luminosity function (Schmidt 1968), calculates for each galaxy an effective search volume $V_{\rm max}$ based on its HI mass and the flux and distance limits that correspond to that mass. A complete mass function is then constructed by weighting by $V_{\rm max}$ the count for each galaxy. The survey sensitivity itself is a function of source line width, such that detected fainter sources will generally tend towards smaller line widths. An additional correction can be made based on the distribution of observed profile widths (see Martin et al. 2010 for an application of this method to ALFALFA data). The 2-dimensional stepwise maximum likelihood (2DSWML) estimator (Efstathiou et al. 1988; Zwaan et al. 2003) takes an alternative approach, performing non-parametric modelling of the observed mass–line width distribution of sources in order to find the distribution that maximises the joint likelihood of detecting all galaxies in the sample.

With the caveat that line width completeness corrections have not been performed on the mass distributions constructed using teams' submitted values, we use Fig. 10 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 percent. Fig 11 demonstrates that the mass distribution at the knee mass is similarly well recovered by several teams, with a slight trend towards larger errors at higher redshifts. This accuracy would be likely to improve after the application of either of the above HI mass function calculation methods.

7.2 Machine learning vs non-machine learning

Supervised machine learning (ML) methods, particularly convolutional neural networks (CNN), proved a popular technique during the Challenge, and featured in the pipelines of the two top scoring teams. 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 characterisation of the second placed team and by several others.

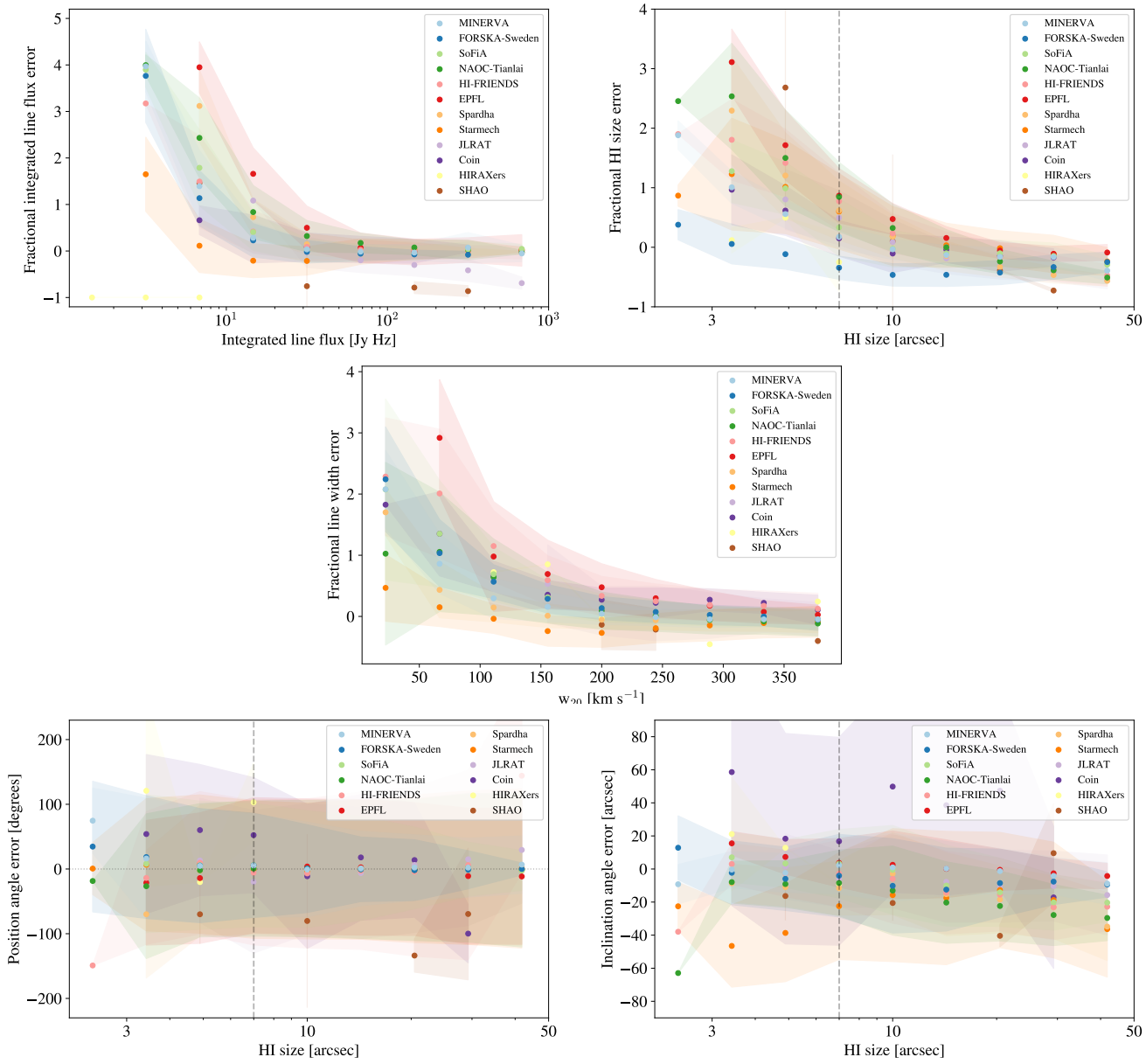

Figure 9. Error terms (see Table 1), calculated without using absolute values, are plotted as a function of true property value (top and middle rows) or spatial source size (bottom row). Circles represent the median error per logarithmic bin, the filled regions represent the standard deviation of the error, and all plots use teams' matched submissions. A dashed line represents the beam size of the simulated observations.

7.2.1 Generalisation

The results demonstrate the promise of ML in the analysis of very large and complex datasets. 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 generalise beyond the parameter distribution that has been sampled by the training data (Burges 1998). Training data distribution may differ from the distribution of real data due to sample selection bias, particularly when the training set is small. The use of regularisation – incorporated as standard into CNN architectures – can avoid this problem by preventing the model from overfitting to the specific training sample. A more difficult problem is that of covariate shift: when the distributions of training and real datasets are intrinsically different. This is a common issue for astronomy (see e.g. Freeman et al. 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 future observations. In the case of SDC2 training data, the characterisation of instrumental features using the MeerKAT precursor has intended to provide participants with the best level of realism possible. Further characterisation of RFI and other instrumental effects during the commissioning phase of the SKAO telescopes will enable the simulation of ever more realistic datasets for training purposes, and transfer learning (Pan & Yang 2009) could close the gap further still.

Also important is the metric used to evaluate the success of

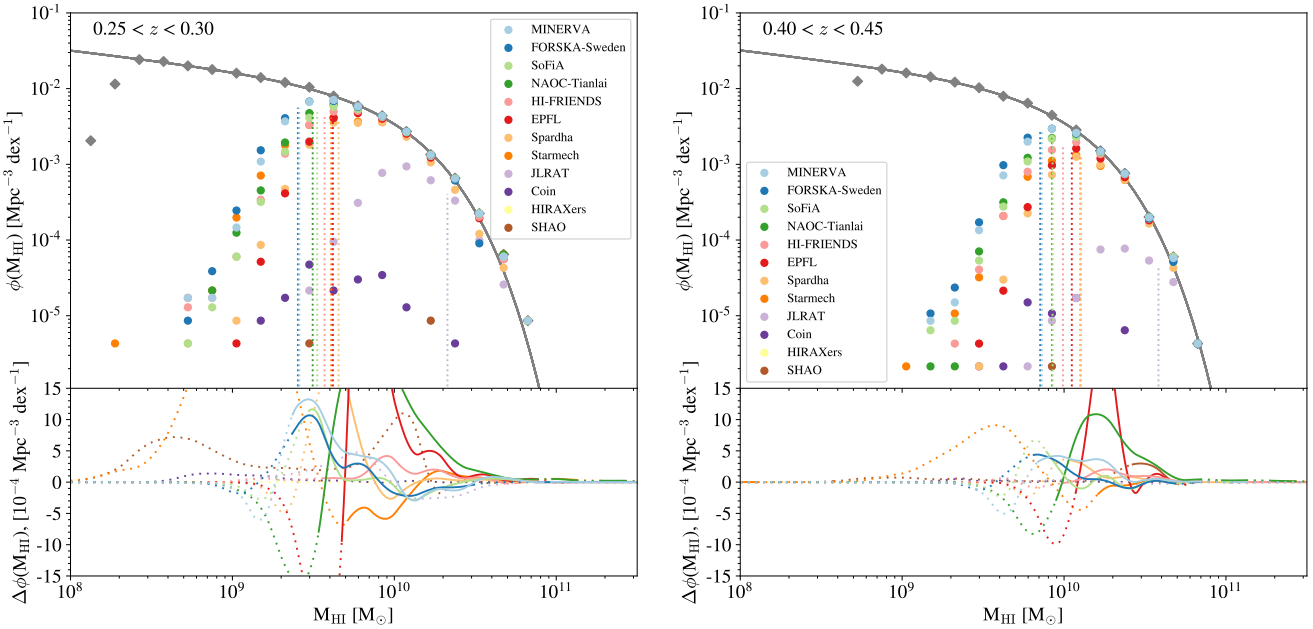

Figure 10. Top panels: HI mass distributions, uncorrected for completeness, are constructed using the true values of integrated line flux and central frequency of each teams' matches (circles). The redshift-dependent HI mass function derived from (Jones et al. 2018), from which truth catalogue sources were drawn (grey curve), is overplotted by the HI mass function reconstructed using the truth catalogue (grey diamonds). Dotted lines indicate for each team the HI mass above which completeness exceeds 50 percent. Bottom panels: the HI 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 again uncorrected for completeness and 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 HI masses where completeness falls below and above 50 percent, respectively.

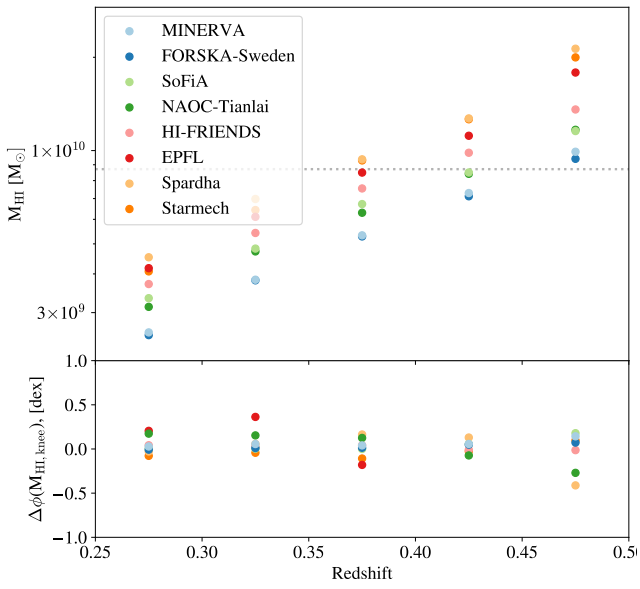

Figure 11. Top: The HI mass above which at least 50 percent of truth catalogue sources are recovered is plotted against redshift for the eight top scoring teams. The dotted line represents the input HI 'knee' mass, M_* (equation 3.1), which marks in the HI mass function the exponential decline from a shallow power law. Bottom: The ratios of the HI mass function measured at the knee mass, expressed in dex, between functions constructed from truth values of teams' matches and functions constructed from the values of teams' submissions.

an ML method: in the case of SDC2, the scoring algorithm has been designed to evaluate source finding and characterisation performance together. This may come at the cost of the performance on either aspect as a standalone task. The measure of success will also depend on the science questions being addressed by the data: a search for fewer, highly resolved sources will take a very different approach from one aiming to produce a complete catalogue.

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 team were able to recover sources from a larger amount of the flux–line width parameter space than by using a single pipeline alone (Fig. 4), 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, so-called ensemble learning (see Opitz & Maclin 1999 for a general review and Kim et al. 2015 for a study of the application of ensemble learning to the problem of star-galaxy classification) has the potential to make predictions that have better performance than any single model in the ensemble. Specifically, stacking – where the predictions made by a group of independent machine learning methods are used as inputs into a subsequent learning model – could improve generalisation from training data to new data (see also Wolpert (1992); Alves (2017); Zitlau et al. (2016)).

The promise of a multi-method approach is further demonstrated by the performance of different methods in different aspects of the Challenge. For example, some teams, though having rela-

Team name	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	-	-	-	-	-

Table 3. The HI mass (in units of 10^9 $M_\odot$) above which at least 50 percent of truth catalogue sources are recovered is reported per redshift interval for the SDC2 finalist teams.

tively less complete and reliable detections (Fig. 7), were able to measure source properties with a relatively high level of accuracy (Fig. 9). Teams Starmech and Coin, for example, though occupying the lower half of the leaderboard, perform particularly well in the recovery of line flux and HI size, respectively. Teams NAOC-Tianlai, HI-FRIENDS, EPFL, though missing out on the top three positions of the leaderboard, all demonstrate a high accuracy in the recovery, variously, of flux, source size and inclination angle. Team ForSKA, a very close second on the leaderboard, achieve the highest levels of reliability and completeness for fainter sources (Figs. 8). 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 et al. 2015) could harness the capabilities of different methods to further improve performance beyond any individual method.

We note that the Challenge leaderboard score, if looked at in isolation, can obscure strong performance by teams on source characterisation. This is a consequence of the strong penalty for false positives. While source characterisation metrics can be presented alongside the overall score, the source characterisation performance of a given method may not be fully understood without delving into lower-scoring submissions of a given team. Reporting on the accuracy of source characterisation while a challenge is live could help teams when evaluating their submissions for this feature. This would be particularly useful in the case of a method that is able to characterise sources very well but less able to find them successfully; given the strong degree of method complementarity, a challenge scoring system that can reflect specialised solutions to a problem may further exploit complementarity as a quality of a collection of independent methods.

7.4 Open Science

Subsection to be completed after reproducibility award deadline

7.5 Data handling

Teams were able to handle the large Challenge dataset with minimal difficulty thanks to the generous provision of computational resources by the SDC2 partner facilities (Section 2.1). By dividing the dataset into smaller portions and applying parallelised codes, teams could comfortably process the full Challenge dataset in under 24 hours of wall clock time. Machine learning methods generally made use of GPU acceleration, which is well suited in particular to CNN architecture due to the very large number of computations that are typically required, particularly during the training stage. GPU acceleration may also boost the efficiency of non-machine learning techniques. 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.

8 CONCLUSIONS

The second SKAO Science Data Challenge has brought together scientists and software experts from around the the world to tackle the problem of finding and characterising HI sources in very large SKAO datasets. 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. Teams approached the challenge with a range of different domain expertise and at different development stages in their pipelines; some teams used the challenge to further their method development, some took the opportunity to improve the quality of their code through reproducibility awards, and others addressed the issue of code efficiency in dealing with a large dataset. Improved performance for all methods can be obtain outside of a time-bound exercise. The main outcomes from the Challenge are summarised below:

(i) 12 international teams, using a variety of methods (Section 4) were successful in completing 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[17]. 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 2.1) 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 HI 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. Findings from the nascent exploration of this model during SDC2 will be summarised in a dedicated report.

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

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

(v) Analysis of teams' submissions (Section 6) has found that sources are recovered with over 50 percent completeness down to an integrated flux limit of ~20 Jy Hz by the top scoring teams. This translates to the ability to probe the HI mass function down to $\sim 3 \times 10^9$ $M_\odot$ at $0.25 < z < 0.30$ and to $\sim 1 \times 10^{10}$ $M_\odot$ at $0.45 < z < 0.50$. The 'knee' mass of the HI mass function can be probed out to $z \sim 0.45$ by the same methods for the chosen redshift evolution. Comparison between property values listed in teams' submitted catalogues and the true values of sources matched with the Challenge truth catalogue finds an error of ≤0.1 dex across HI mass distributions constructed without correcting for completeness.

(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 result in overestimation of flux, source size and line width with fainter objects and smaller sizes. The SKAO spectral line data products can be used to calibrate and remove noise bias effects.

(vii) summary of reproducibility award results to be added after closing date

(viii) New machine learning-based techniques – used by the two top scoring teams – have shown particular promise in the recovery and characterisation of HI sources. Further work using real observations from SKAO commissioning activities and from precursor instruments will examine how well machine learning models translate to real 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 characterisation. That the package proved so popular further demonstrates the value of clearly documented and easily accessible codes, in addition to its accuracy and efficiency.

(x) Perhaps the most important finding of the Challenge is that of method complementarity. Also seen in the first SKAO Science Data Challenge (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 optimising the combined predictions from two independent machine learning methods, the team were able to record an improvement in score 20 percent above either method alone. The result demonstrates the promise of ensemble learning in exploiting very large astronomical datasets. Further work to exploit method complementarity by combining SDC2 finalists' techniques is likely to see an even greater recovery and successful characterisation of HI sources within SKAO spectral line data.

ACKNOWLEDGEMENTS

[To be tidied up a bit:] We would like to thank members of the SKAO HI Science Working Group for useful feedback. 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 multi-scale beam deconvolution performed by Elias Brinks. We would like to thank INAF for the hosting of SDC2 data products. 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 acknowledges support from the European Research Council under the European Union's Seventh Frame- work 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. 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.

SUPERCOMPUTING PARTNER FACILITIES

We would like to make a special acknowledgment of the very generous support from the SDC2 computing partner facilities (2.1), without which a realistic and accessible Challenge would not have been possible.

DATA AVAILABILITY

REFERENCES

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/
Alom M. Z., Hasan M., Yakopcic C., Taha T. M., Asari V. K., 2018
Alves A., 2017, Journal of Instrumentation, 12, T05005
Astropy Collaboration et al., 2018, AJ, 156, 123
Autenrieth M., van Dyk D. A., Trotta R., Stenning D. C., 2021, arXiv e-prints, p. arXiv:2106.11211
Baugh C., et al., 2019, Monthly Notices of the Royal Astronomical Society, 483, 4922
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Blyth S., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 128 (arXiv:1501.01295)
Blyth S. L., et al., 2016, Proceedings of Science
Bonaldi A., Bonato M., Galluzzi V., Harrison I., Massardi M., Kay S., De Zotti G., Brown M. L., 2019, Monthly Notices of the Royal Astronomical Society, 482, 2
Bonaldi A., et al., 2020, Monthly Notices of the Royal Astronomical Society, 500, 3821–3837
Braun R., 2012, ApJ, 749, 87
Braun R., Bourke T. L., Green J. A., Keane E., Wagg J., 2015, in Advancing Astrophysics with the Square Kilometre Array. p. 174
Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, arXiv e-prints, p. arXiv:1912.12699
Broeils A. H., Rhee M. H., 1997, A&A, 324, 877
Burges C. J., 1998, Data mining and knowledge discovery, 2, 121
Chen J., et al., 2021, arXiv preprint arXiv:2102.04306
Chrysostomou A., Taljaard C., Bolton R., Ball L., Breen S., van Zyl A., 2020, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 114490X, doi:10.1117/12.2562120
Dodson R., et al., 2021, arXiv e-prints, p. arXiv:2112.06488

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
Efstathiou G., Ellis R. S., Peterson B. A., 1988, MNRAS, 232, 431
Flöer L., Winkel B., 2012, Publ. Astron. Soc. Australia, 29, 244
Fraternali F., van Moorsel G., Sancisi R., Oosterloo T., 2002, AJ, 123, 3124
Freeman P. E., Izbicki R., Lee A. B., 2017, Monthly Notices of the Royal Astronomical Society, 468, 4556
He K., Zhang X., Ren S., Sun J., 2016, Deep Residual Learning for Image Recognition, 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 ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 1055–1059
Hubble E. P., 1926, ApJ, 64, 321
Jones M. G., Haynes M. P., Giovanelli R., Moorman C., 2018, Monthly Notices of the Royal Astronomical Society, 477, 2–17
Jurek R., 2012, Publications of the Astronomical Society of Australia, 29, 251–261
Katz D. S., et al., 2021, A Fresh Look at FAIR for Research Software (arXiv:2101.10883)
Khvedchenya E., 2019, PyTorch Toolbelt, https://github.com/BloodAxe/pytorch-toolbelt
Kim E. J., Brunner R. J., Carrasco Kind M., 2015, Monthly Notices of the Royal Astronomical Society, 453, 507
Kingma D. P., Ba J., 2014, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings
Leahy J. P., Bridle A. H., Strom R. G., 2013, An Atlas of DRAGNs, An Atlas of DRAGNs
Li W., Wang G., Fidon L., Ourselin S., Cardoso M. J., Vercauteren T., 2017, in International conference on information processing in medical imaging. pp 348–360
Luo S., Leung A. P., Hui C. Y., Li K. L., 2020, Monthly Notices of the Royal Astronomical Society, 492, 5377
Martin A. M., Papastergis E., Giovanelli R., Haynes M. P., Springob C. M., Stierwalt S., 2010, The Astrophysical Journal, 723, 1359
McGaugh S. S., Schombert J. M., Bothun G. D., de Blok W. J. G., 2000, ApJ, 533, L99
Metcalf R. B., et al., 2019, Astronomy & Astrophysics, 625, A119
Meyer M., Robotham A., Obreschkow D., Westmeier T., Duffy A. R., Staveley-Smith L., 2017, Publ. Astron. Soc. Australia, 34, 52
Milletari F., Navab N., Ahmadi S. A., 2016a, Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, pp 565–571
Milletari F., Navab N., Ahmadi S.-A., 2016b, in 2016 fourth international conference on 3D vision (3DV). pp 565–571
Mohan N., Rafferty D., 2015, PyBDSF: Python Blob Detection and Source Finder (ascl:1502.007)
Moldon J., et al., 2021, HI-FRIENDS participation in the SKA Data Challenge 2, doi:10.5281/zenodo.5172930, https://doi.org/10.5281/zenodo.5172930
Morganti R., Sadler E. M., Curran S., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 134 (arXiv:1501.01091)
Mölder F., et al., 2021, F1000Research, 10
Oktay O., et al., 2018, arXiv preprint arXiv:1804.03999
Oosterloo T., Fraternali F., Sancisi R., 2007, AJ, 134, 1019
Opitz D., Maclin R., 1999, Journal of Artificial Intelligence Research, 11, 169–198
Pan S. J., Yang Q., 2009, IEEE Transactions on knowledge and data engineering, 22, 1345
Planck Collaboration et al., 2016, A&A, 594, A13
Popping A., Meyer M., Staveley-Smith L., Obreschkow D., Jozsa G., Pisano D. J., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 132 (arXiv:1501.01077)
Power C., Baugh C. M., Lacey C. G., 2010, MNRAS, 406, 43
Power C., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 133 (arXiv:1501.01564)
Qin X., Zhang Z., Huang C., Dehghan M., Zaiane O. R., Jagersand M., 2020, Pattern Recognition, 106, 107404
Redmon J., Farhadi A., 2016, arXiv e-prints, p. arXiv:1612.08242
Redmon J., Farhadi A., 2018, arXiv e-prints, p. arXiv:1804.02767
Redmon J., Divvala S., Girshick R., Farhadi A., 2015, arXiv e-prints, p. arXiv:1506.02640
Ronneberger O., Fischer P., Brox T., 2015a, in International Conference on Medical image computing and computer-assisted intervention. pp 234–241
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&ARv, 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, Astronomical Society of the Pacific Conference Series Vol. 77, Astronomical Data Analysis Software and Systems IV. p. 433 (arXiv:astro-ph/0612759)
Scaife A. M. M., 2020, Philosophical Transactions of the Royal Society of London Series A, 378, 20190060
Scherzer O., 2010, Handbook of mathematical methods in imaging. Springer Science & Business Media
Schmidt M., 1968, ApJ, 151, 393
Serra P., et al., 2015a, MNRAS, 448, 1922
Serra P., et al., 2015b, Monthly Notices of the Royal Astronomical Society, 448, 1922
Sha Y., 2021, Keras-unet-collection, https://github.com/yingkaisha/keras-unet-collection, doi:10.5281/zenodo.5449801
Starck J.-L., Fadili J., Murtagh F., 2007, IEEE Transactions on Image Processing, 16, 297
Staveley-Smith L., Davies R. D., Kinman T. D., 1992, MNRAS, 258, 334
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29
Vafaei Sadr A., Vos E. E., Bassett B. A., Hosenie Z., Oozeer N., Lochner M., 2019, Monthly Notices of the Royal Astronomical Society, 484, 2793
Vonesch C., Blu T., Unser M., 2007, IEEE Transactions on Signal Processing, 55, 4415
Walter F., Brinks E., de Blok W. J. G., Bigiel F., Kennicutt Robert C. J., 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, Publications of the Astronomical Society of Australia, 31, e023
Westmeier T., et al., 2021, MNRAS, 506, 3962
Whiting M., 2012, Duchamp: A 3D source finder for spectral-line data (ascl:1201.011)
Wilkinson P. N., Kellermann K., Ekers R., Cordes J., Lazio T. J. W., 2004, New Astronomy Reviews, 48, 1551
Wilkinson M. D., et al., 2016, Scientific data, 3, 160018
Wolpert D. H., 1992, Neural Networks, 5, 241
Yakubovskiy P., 2020, Segmentation Models Pytorch, https://github.com/qubvel/segmentation_models.pytorch
Yang J., Huang X., He Y., Xu J., Yang C., Xu G., Ni B., 2021, IEEE Journal of Biomedical and Health Informatics, 25, 3009
Zitlau R., Hoyle B., Paech K., Weller J., Rau M. M., Seitz S., 2016, Monthly Notices of the Royal Astronomical Society, 460, 3152
Zwaan M. A., et al., 2003, AJ, 125, 2842
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, arXiv e-prints, p. arXiv:1501.01211
van der Hulst J. M., de Blok W. J. G., 2013, The Cool ISM in Galaxies. p. 183, doi:10.1007/978-94-007-5609-0_4
van der Walt S., et al., 2014, PeerJ, 2, e453

[1]LERMA, Observatoire de Paris, PSL research Université, CNRS, Sorbonne Université, 75104, Paris, France
[2]DIO, Observatoire de Paris, CNRS, PSL, 75104, Paris, France
[3]Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada
[4]Université de Strasbourg, CNRS UMR 7550, Observatoire astronomique de Strasbourg, 67000 Strasbourg, France
[5]Collège de France, 11 Place Marcelin Berthelot, 75005, Paris, France
[6]GEPI, Observatoire de Paris, CNRS, Université Paris Diderot, 5 Place Jules Janssen, 92190, Meudon, France
[7]Department of Physics & Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa
[8]University of Hamburg, Hamburg Observatory, Gojenbergsweg 11R, R1PR9 Hamburg, Germany
[9]Department of Information Technology and Electrical Engineering, University of Naples Federico II, 21 Via Claudio, I-80125, Napoli, Italy
[10]Faculty of Computational Mathematics and Cybernetics of Lomonosov, Moscow State University, Moscow, Russia
[11]Space Research Institute of Russian Academy of Sciences, Profsoyuznaya 84/32, 117997 Moscow, Russia
[12]Centro Brasileiro de Pesquisas Físicas (CBPF), 22290-180 URCA, Rio de Janeiro (RJ), Brazil
[13]Department of Physics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[14]Raman Research Institute, C. V. Raman Avenue, Sadashivanagar, Bengaluru 560080, India
[15]Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore, Indore 453552, India
[16]Department of Astronomy and Oskar Klein Centre, AlbaNova, Stockholm University, Stockholm SE-10691, Sweden
[17]School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, UK
[18]Department of Electrical and Electronics Engineering, PES University, Bangalore 560085, India
[19]Centre for Astrophysics Research, University of Hertfordshire, Hatfield, Hertfordshire, United Kingdom
[20]Department of Physics & Institute of Astronomy, University of Cambridge, Cambridge, United Kingdom
[21]Special Astrophysical Observatory of RAS, Nizhny Arkhyz, 369167, Russia
[22]Fraunhofer-Chalmers Centre & Fraunhofer Center for Machine Learning, SE-412 88, Gothenburg, Sweden
[23]Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden
[24]ASTRON, the Netherlands Institute for Radio Astronomy, Postbus 2, 7990 AA, Dwingeloo, The Netherlands
[25]Kapteyn Astronomical Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
[26]Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía s/n, 18008 Granada, Spain
[27]Department of Physics, School of Mathematics and Physics, The University of Queensland, Brisbane QLD 4072, Australia
[28]ICRAR M468, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
[29]INAF – Osservatorio Astronomico di Cagliari, Via della Scienza 5, 09047 Selargius, CA, Italy
[30]CSIRO Space and Astronomy, PO Box 1130, Bentley WA 6102, Australia
[31]Australian SKA Regional Centre (AusSRC)
[33]Instituto Astrofísica Andalucía-CSIC, Glorieta de la Astronomía, s/n, E-18008 Granada, Spain
[34]Department of Physics & Astronomy, Macalester College, 1600 Grand Avenue, Saint Paul, MN 55105, USA
[35]South African Radio Astronomy Observatory (SARAO), 2 Fir Street, Black River Park, Observatory 7925, South Africa
[36]Instituto de Física de Cantabria, CSIC-UC, Av. de Los Castros s/n, E-39005 Santander, Spain
[37]Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
[38]Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute, 44780 Bochum, Germany
[39]Laboratoire Univers et Particules de Montpellier (LUPM)-CNRS, UNIVERSITÉ DE MONTPELLIER LUPM CC 072 - Place Eugène Bataillon 34095 Montpellier Cedex 5, France
[40]Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 SHJL, UK
[41]Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[42]CAS Key Laboratory of FAST, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[43]National Astronomical Observatory, Chinese Academy of Sciences, 20A Datun Road, Beijing 100101, P. R. China
[44]Department of Physics, College of Sciences, Northeastern University, Shenyang 110819, China
[45]School of Physics and Astronomy, Sun Yat-Sen University, 2 Daxue Road, Tangjia, Zhuhai, U1YP8R, China
[46]Department of Astronomy, Tsinghua University, Beijing 100084, P. R. China
[47]Département de Physique Théorique and Center for Astroparticle Physics, University of Geneva
[48]African Institute for Mathematical Sciences, Muizenberg, 7945, Cape Town, South Africa
[50]Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, CAS, 80 Nandan Road, Shanghai 200030, China

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