




ORIGINAL RESULTS

Detecting galaxies in a large HI spectral cube

ABINASH KUMAR SHAW^{1,2,*} , MANOJ JAGANNATH³, AISHRILA MAZUMDER⁴,
ARNAB CHAKRABORTY⁵, NARENDRA NATH PATRA^{4,6}, RAJESH MONDAL^{7,8} and
SAMIR CHOUDHURI^{9,10}

¹Department of Physics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

²Astrophysics Research Centre, Open University of Israel, Ra'anana 4353701, Israel.

³Department of Electrical and Electronics Engineering, PES University, Bangalore 560085, India.

⁴Department of Astronomy, Astrophysics & Space Engineering, Indian Institute of Technology Indore, Indore 453552, India.

⁵Department of Physics and McGill Space Institute, McGill University, Montreal H3A 2T8, QC, Canada.

⁶Astronomy and Astrophysics Division, Raman Research Institute, Sadashivanagar, Bengaluru 560080, India.

⁷Department of Astronomy and Oskar Klein Centre, AlbaNova, Stockholm University, Stockholm SE-10691, Sweden.

⁸Department of Astrophysics, School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel.

⁹School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, UK.

¹⁰Department of Physics, Indian Institute of Technology Madras, Chennai 600036, India.

*Corresponding author. abinashkumarshaw@gmail.com.

MS received 28 February 2022; accepted 13 June 2022

Abstract. The upcoming square kilometer array (SKA) is expected to produce humongous amount of data for undertaking HI science. We have developed an MPI-based PYTHON pipeline to deal with the large data efficiently with the present computational resources. Our pipeline divides such large HI 21-cm spectral cubes into several small cubelets, and then processes them in parallel using publicly available HI source finder SOFIA-2. The pipeline also takes care of sources at the boundaries of the cubelets and also filters out false and redundant detections. By comparing with the true source catalog, we find that the detection efficiency depends on the SOFIA-2 parameters, such as the smoothing kernel size, linking length and threshold values. We find the optimal kernel size for all flux bins to be between 3–5 and 7–15 pixels, respectively, in the spatial and frequency directions. Comparing the recovered source parameters with the original values, we find that the output of SOFIA-2 is highly dependent on kernel sizes and a single choice of kernel is not sufficient for all types of HI galaxies. We also propose the use of alternative methods to SOFIA-2, which can be used in our pipeline to find sources more robustly.

Keyword. Methods: statistical, data analysis—techniques: interferometric—cosmology: diffuse radiation.

1. Introduction

The baryonic matter content in the Universe is dominated by hydrogen. In its atomic form, neutral hydrogen (HI) is one of the most reliable tracers of the formation and evolution of different structures in the Universe. The 21-cm line of neutral hydrogen can probe the evolutionary period of the Universe, starting from completely

neutral at very high redshifts to ionize again over cosmic time. The ground state of neutral hydrogen atom consists of a proton and electron in the $1s$ orbital state. The spins of electron and proton are coupled and can be in a parallel or anti-parallel state. Hence, there are two spin states with an energy difference of about $5.87 \mu\text{eV}$, where the higher spin state corresponds to the parallel spins of proton and electron. The hyperfine transition of the atom between these two spin states can be mediated by the absorption or emission of a photon of wavelength ~ 21 cm (1420 MHz). However, the spontaneous emission coefficient for this hyperfine

This article is part of the Special Issue on “Indian Participation in the SKA”.

states is very small, $A_{10} = 2.85 \times 10^{-15} \text{ s}^{-1}$. Due to this, it is challenging to detect the HI line emission from individual galaxies beyond redshift $z \sim 0.4$ with modern telescopes. But, the large fraction of baryons ($\sim 75\%$) is neutral hydrogen, and due to which the 21 cm line signal of neutral hydrogen becomes an excellent tracer of the structure formation. This redshifted HI 21 cm line can be used to map the Universe in 3D.

Galaxy evolution studies in the last decades have majorly focused on understanding different physical mechanisms, which drive the growth of galaxies, starting with cold atomic hydrogen (HI gas) through the cold interstellar medium (ISM; e.g., where the bulk of dense molecular hydrogen, H_2 resides) to star formation (Walter *et al.* 2020). Galaxy evolution includes baryonic processes like gas infalling onto galaxies to form neutral atomic hydrogen (HI gas), which is then converted to the molecular state (H_2), and, finally, the conversion of H_2 to stars (Chowdhury *et al.* 2020). Thus, understanding the evolution of neutral atomic and molecular hydrogen and that of stars is required to understand galaxy evolution. Previous studies have shown that the co-moving star-formation rate density (SFRD) rises towards the end of the epoch of reionization ($z \lesssim 6$) and peaks in the redshift range $z \sim 1\text{--}3$ and then, declines by an order of magnitude from $z \sim 1$ to today (Walter *et al.* 2020). However, the neutral HI mass density (Ω_{HI}) does not show any significant evolution over cosmic time. Chowdhury *et al.* (2020) shows that accretion of gas onto galaxies at $z \leq 1$ may have been insufficient to sustain high star-formation rates in star-forming galaxies and likely to be the cause of the decline in the cosmic SFRD at redshifts < 1 . However, the evolution of cold neutral gas of large samples of galaxies during $z \sim 0\text{--}3$ is still needed to be constrained with more sensitive observation to understand the global flow of gas onto galaxies and probe the history of star formation in the Universe. In addition, the HI 21 cm emission line is being used to measure the rotation curve of the galaxies, which is one of the direct probes of the dark matter (Rubin *et al.* 1970).

There has been significant development in the theoretical understanding of the role and evolution of HI in the Universe using semi-analytical modeling and hydrodynamic simulations (the readers are referred to Section 2.2 of Blyth *et al.* (2015) for a summary of recent advancements). From the observation side as well, there have been surveys like HIPASS (Barnes *et al.* 2001), ALFALFA (Giovanelli *et al.* 2005) that have provided valuable information on global HI properties and galaxies in general. Such surveys have also provided the best

estimates of HI mass function (HIMF). However, sensitivity limitations have confined surveys to the local Universe ($z \sim 0.3$). There have been successful surveys of the 21-cm signal in the local Universe. These have produced HI masses of many galaxies (Zwaan *et al.* 2005; Jaffé *et al.* 2013; Catinella & Cortese 2014; Fernández *et al.* 2016; Jones *et al.* 2018). There have been some studies beyond the low-redshift regime, where either intensity mapping (e.g. Chang *et al.* 2010; Masui *et al.* 2013) or averaging over known galaxy spectra (HI stacking, e.g., Chengalur *et al.* 2001; Lah *et al.* 2007; Kanekar *et al.* 2016; Rhee *et al.* 2017; Bera *et al.* 2019; Chowdhury *et al.* 2020) has been done. The CHILES survey with the VLA is another example of HI survey in the local Universe (Fernández *et al.* 2013). This survey reported the highest redshift detection of HI 21 cm emission signal, so far, from a galaxy at $z \sim 0.376$ in the COSMOS field (Fernández *et al.* 2016).

The square kilometer array (SKA) and its precursor facilities like MeerKAT and ASKAP promise to address the gaps left by current telescopes with their enhanced sensitivities and collecting areas. For instance, the MHONGOOSE survey using the MeerKAT (de Blok *et al.* 2018) is a deep survey using 30 local disk and dwarf galaxies to study HI distribution. This study is expected to address several scientific questions, like the effect of cold accretion, the relation between gas and star formation, dynamics of magnetic fields, etc. On the other hand, the LADUMA survey with the MeerKAT is targeting the observation of the Chandra deep field south (CDFSS) to use HI for studying the evolution of galaxy over $\sim 70\%$ of the age of Universe (Blyth *et al.* 2018). The survey forecasts predict the detection of sources up to $z \leq 1.45$ (Blyth *et al.* 2018). The major LADUMA science questions are: (i) dependence of HIMF on environment density and redshift, (ii) obtain direct measurements of cosmic HI density at high redshifts, (iii) study dependence of galactic HI masses on the stellar mass and host dark matter halo mass as well as other properties as a function of surroundings and redshift, and (iv) time evolution of baryonic Tully–Fisher relations, etc. The survey is ongoing, and the first data release is expected in 2022. MeerKAT is also conducting the HI emission project for the MIGHTEE survey (Maddox *et al.* 2021). It will be a deep blind survey with medium width to study neutral hydrogen emissions up to $z \sim 0.6$. The survey aims to trace how the neutral gas in the galaxies evolved over the past five billion years. The survey volume is designed to cover massive groups and galaxy clusters, while observing low-mass galaxies ($< 10^8 M_{\odot}$) as well. Several surveys— WALLABY, FLASH, DINGO are also being undertaken by the

ASKAP telescope towards HI science. The WALLABY survey (Koribalski *et al.* 2020) is targeting to detect around half a million galaxies in the local Universe ($z \leq 0.26$) with the primary science goals including obtaining a census of gas-rich galaxies in the local group, studying the impact of the environment on galaxy evolution, and refinement of cosmological parameters. The FLASH (Allison *et al.* 2021) survey aims to probe higher redshifts ($0.4 < z < 1.0$) to explore the HI content in the Universe. FLASH will search for absorption lines against the continuum of bright radio sources to produce a database of HI-absorption selected galaxies rich in cool, star-forming gas. It also aims to provide insights into kinematic information for models of gas accretion and jet-driven feedback in radio-loud active galactic nuclei. DINGO survey aims to provide deep legacy data sets for HI emission out to $z \sim 0.4$ (Meyer 2010). The fundamental science objectives for this survey are tracing the evolution of HI mass density, studying the cosmic web and understanding how HI evolves along with other constituents in galaxies over the past 4 Gyr.

The HI surveys with different SKA precursor facilities will create an excellent inventory of the role of HI in various stages of the formation and evolution of galaxies. However, SKA will provide detailed information to characterize the physical processes governing galaxy evolution. SKA will also offer higher angular resolution and sensitivities than currently possible (Blyth *et al.* 2015). SKA will address the issues with the measurement of the HIMF. The HIMF is well measured for the local Universe, and the planned deep surveys with the precursors will enable its measurement up to $z < 0.6$. SKA will push this measured redshift to $z \sim 1$ to provide more stringent constraints on galaxy evolution. It will also enable probing the cosmic neutral gas density to higher redshifts than currently possible. These measurements will also be useful to study the relationship between HI mass and stellar mass, SFR, morphology, etc., down to lower HI masses. Gas inflow from the IGM and its removal from the galaxy plays a crucial role in the formation and evolution of galaxies. Observations for detecting the processes involved in the gas inflow are currently sensitivity-limited. SKA will increase the sensitivity by almost two orders of magnitude (Blyth *et al.* 2015) with high spatial resolution enabling the detection of gas infalling into the galaxy. With the SKA, it will also be possible to study tidal interactions and associated events in a large sample of galaxies—their prevalence, environment and time dependence, etc. The Tully–Fisher relation (TFr) (Tully & Fisher 1977) describes the correlation between a galaxy’s intrinsic

luminosity and the distance-independent width of its global HI profile. It is one of the most valuable tools to obtain distances to galaxies in the local Universe, where peculiar velocities dominate over the Hubble flow. At the same time, its statistical properties are useful for constraining simulations probing galaxy formation and evolution. The SKA precursors will improve observations of the TFr by the improved spatial resolution of the HI kinematics, thus improving TF-based distances in the local Universe. While the SKA itself is unlikely to improve upon this significantly, its sensitivity and angular resolution will make it possible to use TFr for probing galaxy evolution with cosmic time. This will be achieved by using its slope, scatter and zero-point as a function of cosmic environment and time (redshift) and other properties, such as morphological type and SFR.

Thus, with its unprecedented sensitivity, SKA will be the most powerful instrument to study neutral hydrogen in galaxies. However, one limitation is that there are still no clear indications on how far in redshift beyond $z \sim 1$ SKA can probe reliably. Additionally, any SKA survey will come with the inherent problem of data analysis. The high data volume anticipated with SKA observations will make analysis challenging. Powerful supercomputing facilities would be required to process the typical raw data, and these will not be accessible to the community in general. Even after some reduction stages, the data will typically be ~ 1 TB in size for HI galaxy science. Thus, there is a requirement to develop efficient algorithms that will be able to deal with such high volumes of data using reasonable computing power. Towards this end, the SKA usages are presented regularly with science data challenges to familiarize the scientific community with the expectations from this telescope. Recently, the SKA-DC2,¹ where the participants were presented with 1 TB data cube across a 20 square degree sky area to find and characterize the HI content of galaxies.

The challenge saw participation from several teams across the globe and provided some truly novel methods to deal with the HI galaxy science problem. There were several groups from India, and our group, Team Spardha, was able to complete the full challenge. This paper summarizes our efforts and describes the methodology adopted for data analysis. We have used a combination of traditional methods and modern techniques, like machine learning. Within the time limit of the challenge, we were unable to explore all our options completely. Thus, we are currently testing new methods with the ultimate goal of providing an efficient pipeline

¹<https://sdc2.astronomers.skatelescope.org/>.

for performing HI galaxy science. The paper is arranged as follows: In Section 2, we mention the details of the simulated data cube; Section 3 describes the method we have developed so far to detect galaxies from an image cube; in Section 4, we show the results; and finally, we present a summary and conclusions in Section 5.

2. Simulations

To characterize our algorithm and routines, we use simulated data provided by the SKA office (SKAO). This data set was made publicly available as part of the SKA-data challenge 2 (SKA-DC2). In particular, we used their development data set, which has a volume of 40 GB. Note that our analysis routines are generic and implemented using MPI-based parallel algorithm, which can easily be extended to larger data volumes (\sim TB). Here, in this section, we briefly describe how these HI data cubes are generated. For more details, we refer the readers to [Hartley et al. \(2022\)](#).

The simulated HI cubes represent mock data equivalent to SKA observations. In this case, the spectral cubes contain HI signal from galaxies within observing frequency range between 950 and 1150 MHz. This corresponds to a redshift interval of $z = 0.235\text{--}0.495$. A channel width of 30 kHz (velocity width of ~ 6 km s^{-1}) is used, producing ~ 6700 spectral channels across the cube. The development data cube covers a field-of-view ~ 1 square degrees with a beam size of 7 arcsec and a pixel size of 2.8 arcsec. This corresponds to a total co-moving volume $V = 645833.2$ Mpc³ in the sky. The data cube was generated to simulate 2000 h of SKA observation. Accordingly, for the cube, noise, systematic and other imperfections (e.g., radio frequency interference (RFI), residual noise, deconvolution error, etc.) were added.

The HI signal in these cubes originates from the distribution of galaxies in the universe. Hence, to simulate them, a proper prescription is used to mimic their 3D distribution, physical properties, and other observational effects. A detailed description of the same can be found in [Bonaldi et al. \(2021\)](#) and [Hartley et al. \(2022\)](#). Here, we briefly describe the steps adopted to produce the HI cubes.

The HI spectral cubes are simulated in three broad steps. First, considering the distribution of galaxies in the universe, a source catalog is generated. Next, a sky model is produced incorporating model observations from nearby galaxies at the source positions on the catalog. Finally, this sky model is conditioned with telescope response and artifacts to create the final cube.

The catalog of the HI sources is generated by sampling the HI mass function. This provides the average number of galaxies per unit volume within some HI mass range. The HI mass function is adopted from the ALFALFA survey ([Jones et al. 2018](#)) and can be given by

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

where $\phi(M_{\text{HI}}, z)$ represents the number of galaxies per Mpc³ having HI mass of M_{HI} at a redshift z . M_* is the knee mass above which the number of galaxies decreases sharply. α is the power-law index, and ϕ_* is the normalization constant. A redshift-dependent knee mass is used to include mild evolution of the HI mass function as a function of redshift,

$$\log M_*(z) = \log M_* + 0.075 \times z. \quad (2)$$

Several physical properties (e.g., HI diameter, total integrated flux, velocity widths, etc.) of the galaxies are generated using the HI mass from the catalog and different scaling relations observed in the local universe (see e.g., [Duffy et al. 2012](#); [Wang et al. 2016](#)). Note that the source catalog used to generate the 40 GB data cube contains a total of 11091 HI galaxies. These galaxies contain HI masses varying within the range of 1.4×10^8 – $6.6 \times 10^{10} M_{\odot}$, and their velocity width ranges between 30.29 and 891.87 km s^{-1} .

Along with the HI, the SKA observations will also contain signals from continuum sources. These continuum sources could originate in the HI bearing galaxy itself (due to star formation) or through other galactic processes (e.g., AGN activity). However, their presence in the data cube is crucial and should be appropriately modeled. A continuum catalog of extragalactic sources was generated using tiered continuum radio extragalactic continuum simulation (T-RECS, [Bonaldi et al. 2019](#)). A population in this continuum catalog would be associated with the HI sources (originated by SFR). Hence, the continuum catalog is cross-matched against the HI catalog to identify sources compatible with both the HI and continuum properties in the catalogs. Due to the inclusion of full diversity, this cross-identification would leave many sources in both the catalogs, which will not have any match. This results in three categories of sources, ‘continuum only’ (sources do not have an HI counterpart, e.g., AGNs, or HI is too faint to be detected), ‘regular’ (sources having both HI and continuum detected), and ‘HI only’ (sources do not have a continuum, e.g., low SFR galaxies). Further, to mimic the real distribution, these sources are clustered

using the dark matter distribution from P-millennium simulation (Baugh *et al.* 2019) as a backbone.

Model intensity distribution must be built based on the source properties at individual source locations. Thus, actual observations are used from WSRT hydrogen accretion in local galaxies (HALOGAS) survey (Fraternali *et al.* 2002; Oosterloo *et al.* 2007; Heald *et al.* 2011) and the THINGS survey (Walter *et al.* 2008) to prepare an atlas of templates. For every galaxy in the catalog, a galaxy in the atlas is identified whose physical properties match the best. The template was then appropriately smoothed and sampled to maintain the consistency of the data cube at the galaxy location (e.g., pixel size and linear resolution).

After the model observations are placed at the galaxy locations in the cube, telescope response should be incorporated to simulate actual observation. The telescope beam is computed using the SKA-MID configuration. The uv distribution is produced using one minute of sampling over 8 h of observation. Ideally, this uv distribution should be used to produce dirty beams for each channel. However, some of the uv data could be flagged due to artifacts. The noise in the data mainly has two components, thermal noise, and artifacts due to RFI. The thermal noise is calculated considering an equivalent observing time of 2000 h. For RFI contribution, first, a baseline power level (noise floor) is obtained using observation of the MeerKAT telescope. This noise floor is then scaled up by a user-defined value contributing to RFI and distributed amongst all the baselines. To mimic real-life data analysis, a fraction of this RFI is then flagged. This leaves a non-uniform uv distribution at the end, used to estimate the final dirty beam. This dirty beam is used to introduce several imaging-related artifacts, e.g., residual sky image after deconvolution, etc., (see Hartley *et al.* 2022 for more details). The final spectral cube was made by adding the actual HI signals and different artifacts originated due to imperfect observations and imaging.

Using the prescription mentioned above, SKA-DC2 team simulated the 40 GB HI spectral cubes, which are used in characterizing our HI detection techniques. The following section describes the method we use to identify galaxies in an HI spectral cube.

3. Methods

We used the publicly available 3D HI source-finding application (SOFIA) on data cubes to identify galaxies and extract their properties (Serra *et al.* 2015). Although this software has been developed to detect HI galaxies

in widefield ASKAP L-band legacy all-sky blind survey (WALLABY), SOFIA works on any data cube independent of the telescope and observed spectral line (Serra *et al.* 2015). However, the requirement of a large memory footprint and comparatively slow speed of SOFIA makes it challenging to work on large data volumes. To overcome these issues with large data volume, a recently updated version is released for SOFIA-2 (Westmeier *et al.* 2021). It re-implements most of the algorithms of SOFIA in the C-programming language and uses OpenMP for multithreading of the most time-critical algorithms and parallelization to significantly speed up the processing of large data volumes. SOFIA-2 can process a single 800 GB WALLABY data cube in minutes on a modest number of computing nodes (Westmeier *et al.* 2021). We used this latest version SOFIA-2 in our work. Here, we shall briefly describe the algorithms of SOFIA-2, for a detailed description, see Serra *et al.* (2015), Westmeier *et al.* (2021) and references therein.

SOFIA-2 works on the stokes-I data cube in flexible image transport system (FITS) format (Pence *et al.* 2010). It is possible to process different sub-regions of the entire cube simultaneously. SOFIA-2 has an auto-flagging algorithm implemented, which can flag spatial pixels and spectral channels for which the noise deviates from the median noise across all pixels or channels by a user-specified multiple of the RMS as estimated from the more robust median absolute deviation. This helps in identifying corrupted data due to RFI and flag them. Note that SOFIA-2 works only on constant noise across the data cube to apply threshold algorithm to detect the sources. Hence, if the noise level is not constant across the data cube, SOFIA-2 first normalizes the noise and then applies the source-detection algorithm.

There are two source finding algorithms implemented inside SOFIA-2: a simple threshold finder and the smooth and clip (S+C) finder. The threshold algorithm applies an absolute or relative (with respect to the noise) flux-based cut-off to identify sources. This has been rarely used unless the data is preconditioned. The most powerful and widely used algorithm is the S+C finder. This algorithm iteratively smooths the data cube on multiple spatial and spectral scales to obtain significant emission above a user-specified detection threshold on each scale (Serra *et al.* 2012). SOFIA-2 estimates both positive and negative flux values above the threshold. The negative flux values are used to estimate the false detection rate and provide the statistical reliability of the catalog with positive flux values (Westmeier *et al.* 2021).

After detecting the pixels above the threshold, the software uses a simple friends-of-friends algorithm to

link all pixels within a merging radius specified by the user. It assigns a unique identifier to each detection. The linking and merging of pixels happen in 3D, i.e., in spatial and frequency directions. A minimum and maximum filters can also be applied to filter out the noise artifacts, otherwise considered actual sources. SOFIA-2 determine the reliability of each detection and use this information to automatically remove detections from the source catalog that are deemed unreliable (Dickinson *et al.* 2004; Westmeier *et al.* 2021). SOFIA-2 compares the number density of detections with negative and positive total flux in a 3D parameter space made up of the peak flux density, summed flux density and mean flux density across the source and then estimates the reliability of the source based on the user-specified reliability threshold. This helps to reduce the false detections in the source catalog. In addition to the reliability threshold, the user can also set a signal-to-noise threshold. All detections with an integrated signal-to-noise ratio (SNR) below that threshold will be discarded as unreliable, irrespective of their measured reliability. SOFIA-2 also determines the basic source properties, like peak flux, integrated flux density, position angle of the galaxy, line width, etc., and write those in the output catalog (Serra *et al.* 2015; Westmeier *et al.* 2021). The strategy we employed to detect and extract the source properties efficiently from a data cube with SOFIA-2 is discussed below.

We have developed a PYTHON-based pipeline, which starts with dividing the development data cube of size 40 GB 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 (Westmeier *et al.* 2021) 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 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 Figure 1.

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 Figure 1. This is done to detect the sources, which fall at the common boundaries. The normal cubelets and the overlapping cubelets always have buffer zones (e.g., blue regions for normal cubelets) around their faces to avoid confusion between the sources detected near their common boundaries. We conservatively set the width of buffer zones based on the

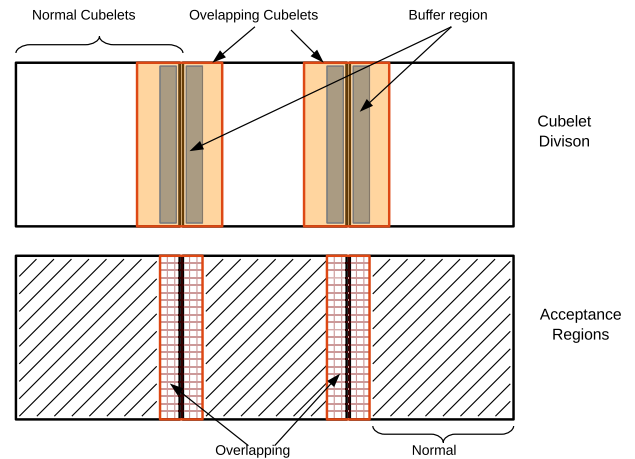


Figure 1. Shows the 2D projection of the schematic division of the data into the normal and overlapping cubelets along one of the axis (top row) and the corresponding acceptance regions (bottom row).

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, 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 correspond to a line-width ~ 500 km s $^{-1}$. Therefore, the widths of the buffer region and overlapping region along the 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 Figure 1. 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 no common source is present in the list after this step. Analyzing cubelets is the most time-consuming part of our pipeline. We analyze 118 cubelets on 32 cores in parallel in around 20 min.

Next, we use the physical equations to convert the SOFIA-2 catalog into the physical 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} (line-width) to discard the detections with unusual velocity-width. Motivated by the physical models/observations of the galaxies, we have conservatively accepted the sources having

$w_{20} \in [60, 500] \text{ km s}^{-1}$. We finally arranged the catalog 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 chose around the top 35% of the sources to evaluate our pipeline's efficiency.

4. Results

Using the method described above, we try to detect HI sources in the spectral cube. For the current study, we use SOFIA-2 as our primary tool for detecting the galaxies in the data cube.

Though SOFIA-2 is a versatile tool to identify HI sources, its detection efficiency depends on several parameter settings. For example, size of the smoothing kernels or the linking length can substantially influence the detection volume in a spectral cube. Moreover, the optimal settings of these parameters could be different for different types of galaxies. For example, a larger smoothing kernel size will have less sensitivity in detecting small galaxies compared to larger galaxies. Further, the flux threshold (the threshold for labeling valid pixels after smoothing) also is a critical parameter, which might have different optimized values for dwarf and spiral galaxies. A higher threshold might recover galaxy properties more accurately in brighter galaxies, whereas the higher threshold might not detect a faint galaxy altogether. Because all types of galaxies exist within a survey volume, different sets of optimized parameters should be used to efficiently recover different kinds of galaxies.

We use the 40 GB data cube and the associated truth catalog to identify the best SOFIA-2 input parameters for detecting sources in the SKA HI spectral cube.

Sizes of the smoothing kernels (in all three dimensions) are one of the most important key parameters influencing the detection of faint signals coming from distant galaxies. A kernel much larger than the galaxy size, would include more noise than signal while smoothing, reducing the overall SNR. On the other hand, a smaller kernel than the galaxy size would include less signal limiting the maximum attainable SNR. Ideally, an optimum kernel size would match the galaxy size (angular), providing maximum SNR. This requires many optimum kernels, almost as many galaxies in a spectral cube (as the angular sizes/spectral extents would be different). Because we do not have priory information on the galaxy position in a spectral cube, one needs to search the whole cube with all sorts of kernels. This is computationally highly expensive.

Instead, we try to identify a set of kernels that reasonably work across different galaxies present in the cube.

To do the same, we run SOFIA-2 on the development data cube with different kernel sizes and estimate the detection rates for each kernel. A detection is when the coordinates of a galaxy recovered by SOFIA-2 matches that of the truth catalog within a beamwidth. The detection rate is determined by the ratio of the number of detections to the total number of galaxies in the truth catalog. This detection rate reflects how efficiently a kernel can detect galaxies in a spectral cube.

As there is no asymmetry in the spatial direction, i.e., the angular galaxy size in RA and DEC direction have no bias due to random position angle, we chose our spatial kernels to be symmetric. The kernel size in the frequency direction is determined independently. We chose spatial kernel sizes in the range between 3 and 9 pixels, which is equivalent to $8.3''$ – $25.2''$ angular size on the sky. We chose kernel sizes between 3 and 15 channels in the frequency direction, equivalent to ~ 18 – 90 km s^{-1} . Our choices are motivated by the range of physical extent (both in spatial and frequency directions) of the galaxies.

In Figure 2, we plot thus obtained detection rates for our chosen kernel ranges. The detection rate and the best-optimized kernel would depend on the brightness of the galaxies. Hence, we divided our whole sample into three flux bins to capture its variation. We evaluate the detection rates for every flux bin (as quoted at the top of every panel in the top row), represented by different horizontal panels in the figure. The color bars indicate detection efficiencies. As can be seen, a spatial kernel size between 3 and 5 and a frequency kernel size between 7 and 15 works reasonably well for all the flux bins.

This exercise was performed using a threshold of 3. As discussed earlier, this threshold decides if a pixel should be identified as a valid one or not after smoothing (see [Serra et al. 2012](#), for more details). Increasing this threshold would increase the accuracy in recovering galaxy properties at the cost of detection rate. To compare how detection rates can change with increasing thresholds, we perform the same exercise with threshold values 4 and 5. The vertical panels of Figure 2 show the results. As we can see, with an increased threshold, the detection efficiencies decrease significantly. This happens because low flux faint galaxies get excluded at a higher threshold. Because of the larger number of faint galaxies (i.e., more number of dwarf galaxies) in an unbiased survey volume, the efficiency dramatically decreases with a slight increase in the threshold.

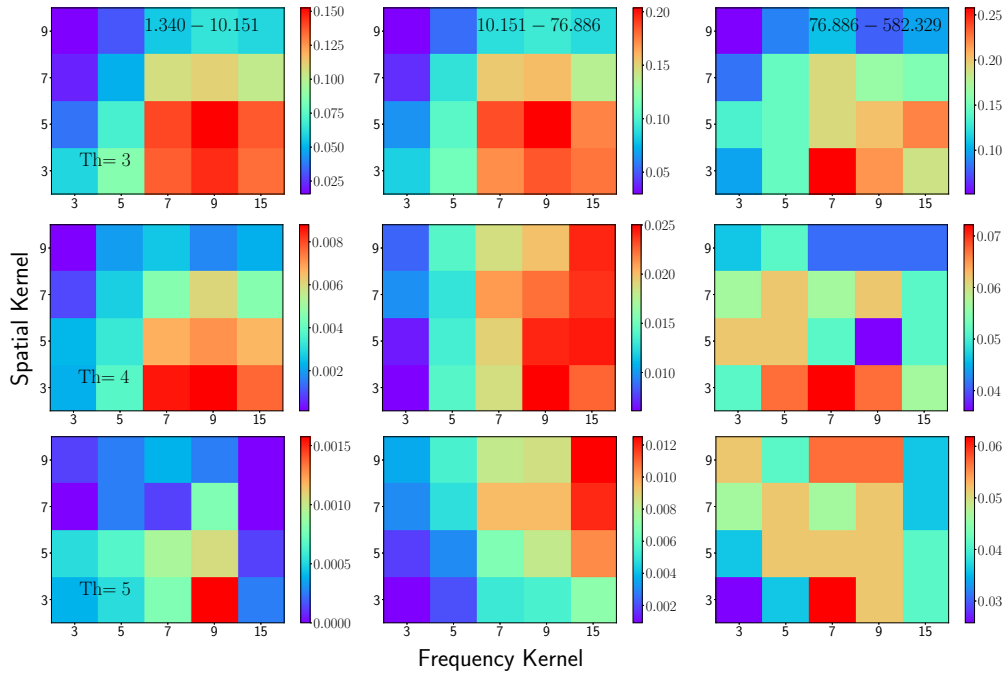


Figure 2. Shows the fraction of detected sources as a function of spatial kernel size and frequency kernel size. The three columns represent the three integrated flux bins with the ranges shown in the top-right corner in the top row. The three rows correspond to the three threshold values as stated in the first column panel.

Further, we notice that unlike threshold 3, no reasonably good kernels can be found for higher thresholds, which works for all flux ranges. Given these results, we chose to use a spatial kernel of size 3 pixels and a frequency kernel of size 9. We also chose to work with a threshold of 3. With these SOFIA-2 settings, we run our routine on the data cube to build the final catalog. Here, we stress that as the truth catalog for the developmental data set is available, we use it to characterize our routine. We will use our routine with these optimized settings for an observed data cube to detect HI sources. We note that we recover only 15–25% of galaxies listed in the truth catalog with these settings. Though it might seem to be a low efficiency, we emphasize that this is relative. Not all the galaxies in the truth catalog need to be detected in the spectral cube. There could be galaxies so faint that they would be buried under noise, and no detection technique would be able to recover them. In that sense, a detection rate of 15–25% could be reasonable (Hartley *et al.* 2022). In this context, we would like to mention that our results are consistent with those obtained by other SKA-DC2 teams.

For cross-matching our detections (i.e., counting a valid detection by our routine), we have only used position, i.e., RA, DEC and frequency. However, the other recovered physical parameters, e.g., total flux, galaxy size, inclination, etc., are also critical for estimating

galaxy properties. We note that identifying a galaxy by an optimized kernel does not guarantee the optimized recovery of these parameters by the same kernel. Further, different kernels might best suit to recover different properties for the same set of galaxies. A small kernel might better identify the center of a galaxy, but a larger kernel might work better to recover the total flux. Ideally, optimized kernels should be identified for every galaxy parameter (starting from size, flux to inclination). An observed cube then should be passed through all these optimized kernels. However, this demands significantly higher computation, as the data cube has to be searched for many times for all the different kernels. Here, we only use a single kernel (i.e., 3 and 9) to detect sources with reasonable confidence. Nonetheless, we characterize the accuracy of our routine in recovering the galaxy properties.

In Figure 3, we plot the probability distribution functions (PDFs) of the deviations of the recovered physical properties of the detected galaxies from the truth values. Horizontal panels refer to different properties. As can be seen from the figure, the fractional deviations in our estimated properties are large, and sometimes it could be a factor of a few. These deviations could originate due to two reasons. First, some of the signals could be lost intrinsically due to noise. This is fundamental and cannot be solved by changing detection techniques, which

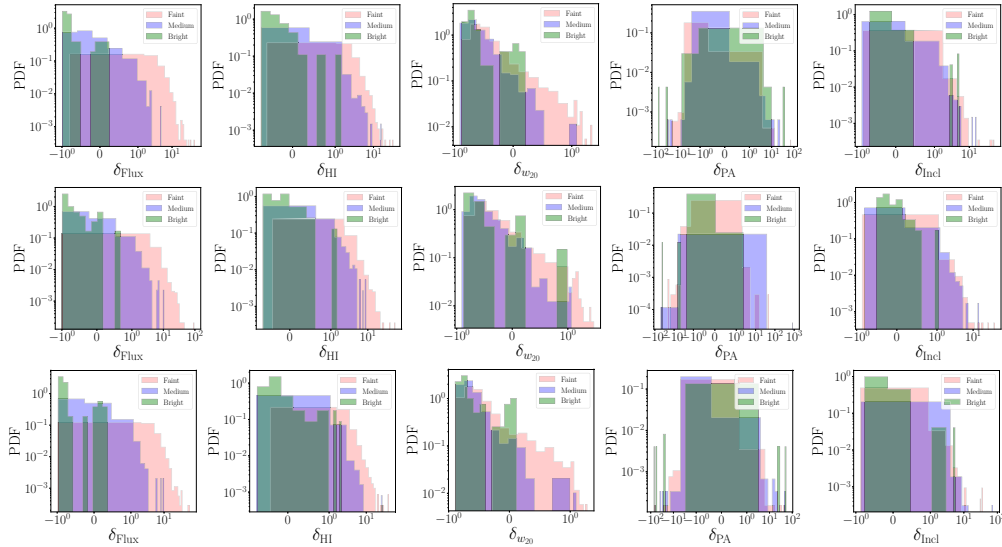


Figure 3. This figure shows the fractional deviations between the recovered and true values (from the truth catalog) of different galaxy properties. Three colors represent three bins having galaxies with different brightness. Horizontal panels show the PDFs for different properties, whereas vertical panels show the PDFs for different kernels. The top, middle and bottom panels show the results for spatial kernels 3, 3, 5 and frequency kernels 9, 15, 9. The deviations shown here are estimated for a 3σ threshold.

is expected to affect the total integrated flux, HI mass and the galaxy size more than other parameters. Second, due to poor choice of kernel, or third, both. As further can be seen from the figure, the deviations are more for fainter galaxies (orange histograms) than the brighter ones. This is expected since a significant fraction of their flux/mass is in the diffused state with low SNR, easily buried under the noise floor. Due to computational limitations, in general, we chose to use a single kernel to operate on the whole cube containing different types of galaxies. However, the accuracy in the recovered galaxy properties would depend on the choice of kernel sizes. We performed the above exercise with two other kernels, i.e., 3, 15 and 5, 9, to test how the PDFs of the deviations change. We show the results in the vertical panels of the figure. As can be seen, with different kernels, the PDFs change considerably. This indicates that the choice of a proper kernel while estimating the galaxy properties is critical.

To investigate this in more detail, we ran our source finding routine over the entire cube with a variety of kernels to search for the best one to recover the galaxy properties more accurately. For these kernels, the deviations would be minimum. In Figure 4, we plot these deviations. The three horizontal panels are for galaxies with different total fluxes (faint, moderate and bright). The top, middle and bottom panels indicate the deviations for flux, w_{20} and inclination. As can be seen, the deviations are very different for different kernels. For

example, a smaller kernel seems better suited for recovering total flux in faint galaxies (top panels). However, the same kernel does not seem to work best to recover w_{20} (middle panels). This further strengthens the idea that optimal kernels are specific to galaxy properties and the best kernels need to be searched for each of them. Thus, a thorough search on the kernel space should be done, while recovering galaxy properties.

The recovered galaxy properties could be sensitive to the inclination for the exercise mentioned above. Hence, considering one individual galaxy in each horizontal panel might bias our result (as the inclinations might be different). Thus, in Figure 4, we calculate the median of the deviations for all the detected galaxies within a flux bin. The considerable number of detected galaxies within a bin is expected to have the full range of inclinations, removing the bias.

Further, this exercise is done with a SOFIA-2 threshold of 3. Though this is a widely used standard cut-off, it can substantially include noise as a signal. This will corrupt the faint galaxies significantly, where the peak flux could be just a little higher than 3σ . Deviations would be large for these galaxies, and no kernel would be best to recover the correct properties. To check the same, we perform the same exercise mentioned above with thresholds of 4 and 5. The results are plotted in Figure 5. The horizontal panels refer to faint, moderate and bright galaxies; the vertical columns indicate the deviation for thresholds of 3, 4 and 5 (top, middle and

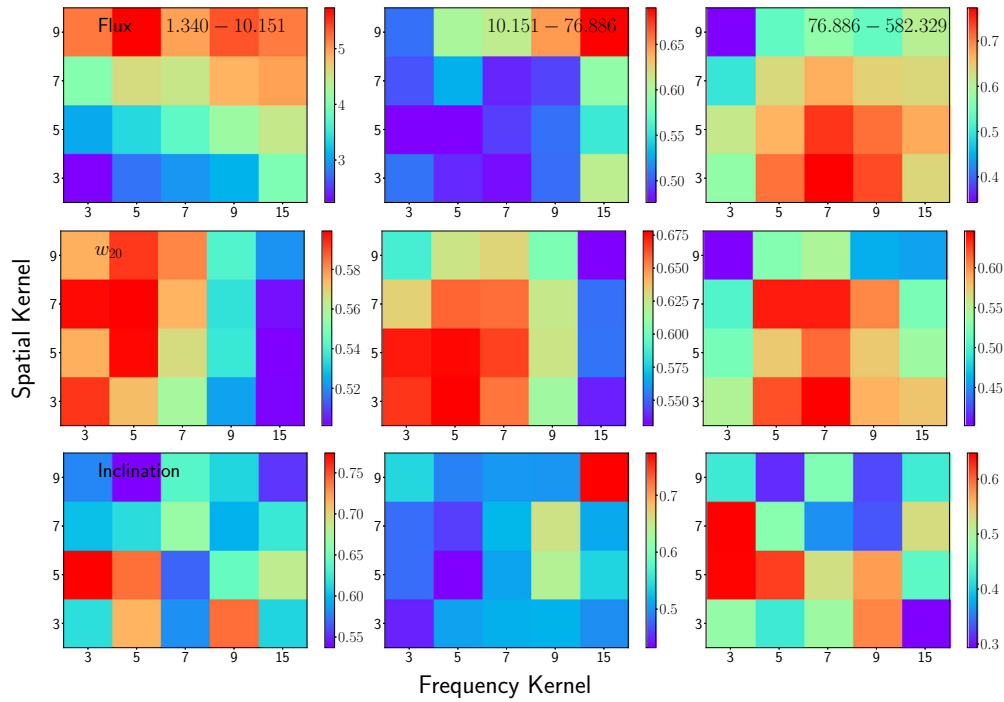


Figure 4. This figure shows the median fractional deviations of the recovered and the true galaxy properties. Horizontal panels show the deviations for galaxies with different brightness. The top, middle and bottom panels are for flux, w_{20} and inclination, respectively. See the text for more details.

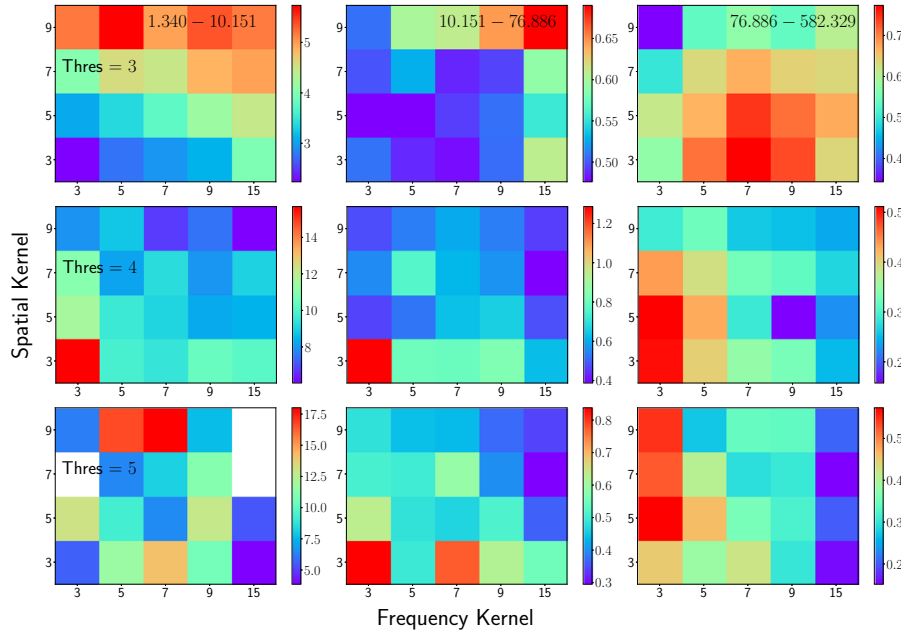


Figure 5. This figure shows the fractional median deviation of the recovered flux as a function of the chosen threshold. The horizontal panels show the results for galaxies with different brightness. The top, middle and bottom panels are for thresholds 3, 4 and 5, respectively. See the text for more details.

bottom). As can be seen from the figure, with increasing threshold, the deviations reduce. For example, for the brightest galaxies (right panels), the fractional deviation decreases from $\sim 40\%$ to $\sim 20\%$ as the threshold

increases from 3 to 5. The same trend can be seen for the moderately bright galaxies (middle panels) as well. However, for faint galaxies, this trend is disrupted. As we discussed, a higher threshold might exclude a large

amount of flux in these galaxies leading to large deviations. This, in turn, can suppress the trend seen for the relatively brighter galaxies.

Next, to investigate the overall detection efficiency of our routine, we calculate the detection rates as a function of redshift. In Figure 6, we plot the number of detected galaxies as a function of redshift. As can be

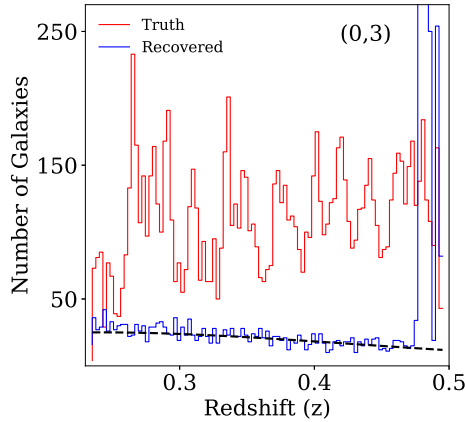


Figure 6. The number of detected galaxies as a function of redshift using our detection method. The red histogram shows the number of galaxies in the truth catalog, whereas the blue histogram represents the number of detected galaxies. The detection rate is estimated for a spatial kernel 0 and a frequency kernel 3. The survey’s selection function is shown by the black dashed line. As can be seen from the figure, our overall number of detection (by the kernel (0,3)) matches well with the survey selection function. See the text for more details.

seen from the figure, the input number of galaxies in the cube is roughly constant (red histogram) as a function of redshift. This is because the (possible) evolution of the HI mass function is not implemented, while simulating the spectral cube. However, the detection rate is expected to decline with redshift due to the change in the detectability of minimum HI mass. This can be quantified using a survey selection function, i.e., the number of expected detections given the survey depth and coverage. For our data set, the minimum detectable HI masses at the nearest ($z = 0.235$) and farthest ($z = 0.495$) redshifts are $\sim 1.3 \times 10^9 M_{\odot}$ and $\sim 1.7 \times 10^9 M_{\odot}$, respectively. The coverage of the simulated survey is 1 square degree. With these survey depth and coverage, we estimate the survey selection function and found that it matches our detection rates reasonably well for the kernel (0, 3). However, we note that the detection rates vary significantly for different kernels. We emphasize that the calculation of our selection function is purely based on the sensitivity of the survey and its coverage. We do not consider the effect of possible imperfections in the data, e.g., RFI, imperfect subtraction of continuum, etc. Further, we note that our routine detects an unusually large number of sources at the highest redshift bins (lowest frequency bins). This is probably due to the inclusion of simulated RFIs in the data at the lowest frequencies (see [Hartley et al. 2022](#), for more details).

To further investigate the detection rates at different mass bins, we use our source catalogs to estimate the HI mass function $\phi(M_{\text{HI}})$, which is shown in Figure 7.

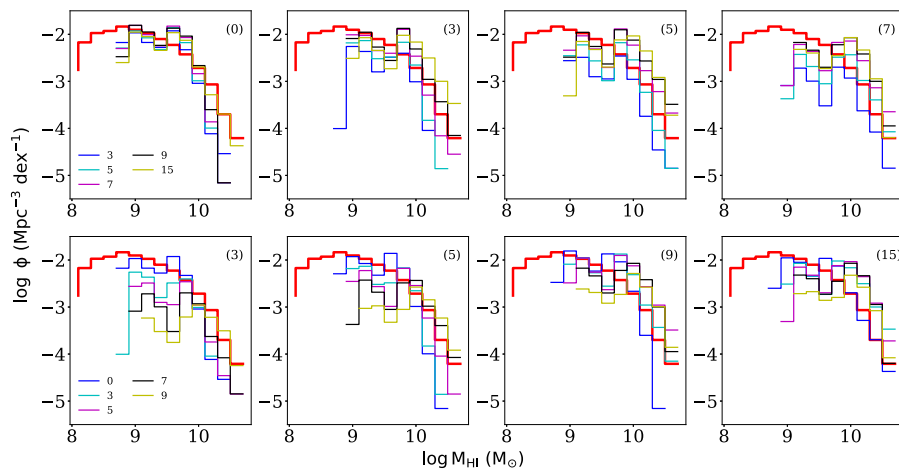


Figure 7. Derived HI mass function for different adopted spatial and frequency kernels. In each panel, the red histogram represents $\phi(M_{\text{HI}})$ derived from the truth catalog. The top panels show the variation of $\phi(M_{\text{HI}})$ with different spatial kernels (as quoted on the top right in each panel). Each individual panel shows the same for different frequency kernels as quoted in the legend. The bottom panels show the variation of $\phi(M_{\text{HI}})$ with different frequency kernels. As can be seen, $\phi(M_{\text{HI}})$ varies significantly with the choice of kernels with the kernel (0, 3) providing a reasonable match to the truth $\phi(M_{\text{HI}})$ (top left panel). See the text for more details.

We found that the recovered mass function significantly depends on the choice of kernels. Smaller kernels tend to detect low-mass galaxies well, whereas larger kernels detect larger galaxies better. We also found that $\phi(M_{\text{HI}})$ obtained using kernel (0,3) matches the truth one better than other kernels (top left panel). However, for this kernel also, the $\phi(M_{\text{HI}})$ shows considerable deviation from the truth value. Hence, a combination of different kernels suited for different types of galaxies would be ideal for galaxy detection and parameter estimation in an HI spectral cube.

As can be seen from these exercises, the choice of kernels is crucial both in detecting and recovering galaxies in an HI spectral cube, especially for fainter galaxies. Ideally, individual galaxy candidates (after the detection) should be passed through a host of kernels to identify property-specific optimized kernels. Thus, a rigorous study is required to create an ideal kernel template for particular sets of galaxies, e.g., with specific flux and inclination range, etc. For our study, we only use a single kernel for all purposes. However, we emphasize that as preliminary work, our routine is mainly focused on detecting potential galaxies in the HI spectral cube. We plan to develop a routine that will inspect these individual detections rigorously and recover the properties with the optimum kernels/SOFIA settings from a predetermined template matrix.

5. Summary and future scope

Hydrogen, being the most abundant baryonic component in the Universe, is the most promising tracer to study galaxy formation and evolution. Its hyperfine transition line of 1420 MHz is an excellent tracer to map HI content and its dynamics within the galaxies at $z \sim 1$. Thus, several surveys such as HIPASS, ALFALFA and CHILES have conducted observations to provide constraints over HI mass function from the galaxies. However, they were limited to the local Universe ($z < 0.4$) due to limited sensitivity. SKA and its precursor facilities like the MeerKAT and ASKAP are expected to have enhanced sensitivity to push the threshold of HI detection from galaxies to higher redshifts. The precursors are currently carrying out several deep surveys, such as MHONGOOSE, MIGHTEE, LADUMA, WALLABY, FLASH, DINGO, etc., to observe galactic HI contents up to redshift $z \lesssim 1$. SKA will add to the findings of these surveys by further observing the HI within the galaxies and the infalling gas with much higher resolution. Thus, it will be able to constrain the HIMF and Tully–Fisher relations of the galaxies with higher fidelity up to $z \approx 1$. However, the data

volume for surveys with SKA will be \sim few TBs even after some real-time processing for data volume reduction. Analyzing such large data volumes is challenging with the currently available tools and computational resources. Hence, there is need in the community for developing more sophisticated and automated data analysis pipelines to handle the expected large data volumes. This paper discusses our efforts to analyze such big data using traditional and modern methodologies.

We have developed our algorithm to analyze large (\sim TB) spectral radio-data cubes. We have tested our methods on a 40 GB (co-moving volume 645833.2 Mpc³) data cube which was made publicly available by the SKA-DC2. Our results in this paper also is based on the analysis of this 40 GB data cube. Note that we have already applied our algorithm to the original 1 TB² data cube of the SDC2 challenge, in which we had taken part as Team Spardha. The cube was simulated between a frequency range of 950–1150 MHz corresponding to the redshift range of 0.235–0.495 with 6668 channels along with the frequency axis. Both cubes were simulated between a frequency range of 950–1150 MHz corresponding to the redshift range of 0.235–0.495 with 6668 channels along with the frequency axis. However, 40 GB and 1 TB data cubes span 1 deg² and 20 deg², respectively, on the sky. The HI signal within the data cubes was simulated to match the HI mass-functions extrapolated from the observations in the nearby Universe. Also, the data cube was infused with system noise corresponding to 2000 h of SKA observations, residual continuum, RFI, systematics, etc., to make the data cube realistic.

We first applied our MPI-based PYTHON pipeline which divides the large data cubes into multiple small cubelets and then find galaxies within each cubelet in parallel. The pipeline uses publicly available SOFIA-2 as its core ingredient to find the HI sources in the cubelets. SOFIA-2 works on a traditional smoothing and clipping (S+C) algorithm to find galaxies from a noisy data. Therefore, the smoothing kernel size and the clipping thresholds are the parameters, which affects the performance of SOFIA-2. Note that our pipeline is designed in a way that takes care of the sources, which lie in the boundaries between the two cubelets. We first explore how the performance of our pipeline depends on the SOFIA-2 kernel sizes and threshold by varying the spatial kernel size and frequency kernel size (in terms of pixels) as well as the clipping threshold step wise. We compare the position of detected sources (RA, DEC and

²1 TB cube was restricted only to the participants through some private computational facilities.

central frequency) with those in the true list of source we already have. We have binned the HI sources into three classes, i.e., faint, moderate and bright, based on their fluxes and present results for them separately. We note that for all flux bins, a spatial kernel size between 3 and 5 pixels and a frequency kernel size between 7 and 15 pixels works reasonably well at a threshold of 3σ . Increasing the threshold level may recover the source properties (e.g., flux, line-width, inclination and HI size) accurately (Figure 5), however, the detection of the sources in each flux bin drops drastically (Figure 2). We also note that there is no such common kernel size combination for which detection efficiency is optimum in all the flux bins. This indicates that multiple sets of kernels are required to detect galaxies with widely varying fluxes. This is also true with other source properties such as flux, line width (w_{20}), inclination, etc. We found that a common kernel size is not sufficient to recover all the source properties as well as the HI mass function $\phi(M_{\text{HI}})$ with a legitimate accuracy (Figures 4 and 7). Therefore, one needs to run the pipeline with multiple kernel sizes to accurately detect sources with widely varying fluxes. This will be a time consuming process when the data volume is approximately a few TB. We plan to improve the pipeline by employing an automated routine that will inspect these individual detections rigorously and recover the properties with the optimum kernels/SOFIA settings from a predetermined template matrix.

The output from SOFIA is not completely reliable due to the possibility of large false detection, as the data we are dealing here have smaller SNR than any typical low-redshift observation. Therefore, we are also developing a hybrid pipeline where we plan to use convolutional neural networks (CNNs) as a classifier to remove the false detections in SOFIA outputs. We plan to make use of the moment-1 maps, obtained from SOFIA, to discard the false detections. We are also in the process of developing modern machine learning-based algorithms on our own to detect the HI sources directly from the noisy data cubes. Our present efforts in this direction indicates that CNNs based on U-net image segmentation algorithm can be effective. Our primary goal here is to recover galaxy extent more accurately besides its location. Additionally, we are also exploring the application of friends-of-friends (FoF)-based algorithm which has potential to find out galaxies using thresholding and linking of data pixels. Note that both the CNN-based as well as FoF-based algorithms are still being developed and tested for their robustness, and it is an ongoing work. We, as a team, are continuously working to develop a data analysis pipeline, which will be efficient enough

to handle large data volumes generated from the future SKA observations.

Acknowledgment

Authors would like to acknowledge SKA India Consortium and IUCAA for the support with the computing facilities. We would also like to thank Nirupam Roy for useful discussions. We acknowledge National Supercomputing Mission (NSM) for providing computing resources of ‘PARAM Shakti’ at IIT Kharagpur, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India. We also acknowledge Raman Research Institute for providing their HPC facility. RM is supported by the Wenner-Gren Post-doctoral Fellowship. AKS acknowledges support by the Israel Science Foundation (grant no. 255/18). AM acknowledges Indian Institute of Technology Indore, for supporting the research through Teaching Assistantship.

References

- Allison J. R., Sadler E. M., Amaral A. D., *et al.* 2021, The first large absorption survey in HI (FLASH): I. Science goals and survey design, 2110.00469
- Barnes D. G., Staveley-Smith L., de Blok W. J. G., *et al.* 2001, Monthly Notices of the Royal Astronomical Society, 322, 486
- Baugh C. M., Gonzalez-Perez V., Lagos C. D. P., *et al.* 2019, Monthly Notices of the Royal Astronomical Society, 483, 4922
- Bera A., Kanekar N., Chengalur J. N., Bagla J. S. 2019, The Astrophysical Journal, 882, L7
- Blyth S., van der Hulst T. M., Verheijen M. A., *et al.* 2015, in Proceedings of Advancing Astrophysics with the Square Kilometre Array—PoS (AASKA14), Vol. 215, p. 128
- Blyth S., Baker A. J., Holwerda B., *et al.* 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA—PoS (MeerKAT2016), Vol. 277, p. 004
- Bonaldi A., Bonato M., Galluzzi V., *et al.* 2019, Monthly Notices of the Royal Astronomical Society, 482, 2
- Bonaldi A., An T., Brügger M., *et al.* 2021, Monthly Notices of the Royal Astronomical Society, 500, 3821
- Catinella B., Cortese L. 2014, Monthly Notices of the Royal Astronomical Society, 446, 3526
- Chang T.-C., Pen U.-L., Bandura K., Peterson J. B. 2010, Nature, 466, 463
- Chengalur J. N., Braun R., Wieringa M. 2001, Astronomy & Astrophysics, 372, 768

- Chowdhury A., Kanekar N., Chengalur J. N., Sethi S., Dwarakanath K. S. 2020, *Nature*, 586, 369
- de Blok W., Adams E., Amram P., *et al.* 2018, in *Proceedings of MeerKAT Science: On the Pathway to the SKA—PoS (MeerKAT2016)*, Vol. 277, p. 007
- Dickinson M., Stern D., Giavalisco M., *et al.* 2004, *The Astrophysical Journal Letters*, 600, L99
- Duffy A. R., Meyer M. J., Staveley-Smith L., *et al.* 2012, *Monthly Notices of the Royal Astronomical Society*, 426, 3385
- Fernández X., van Gorkom J. H., Hess K. M., *et al.* 2013, *The Astrophysical Journal*, 770, L29
- Fernández X., Gim H. B., van Gorkom J. H., *et al.* 2016, *The Astrophysical Journal*, 824, L1
- Fraternali F., van Moorsel G., Sancisi R., Oosterloo T. 2002, *The Astronomical Journal*, 123, 3124
- Giovanelli R., Haynes M. P., Kent B. R., *et al.* 2005, *The Astronomical Journal*, 130, 2598
- Hartley P., Bonaldi A., Braun R., *et al.* 2022, (Manuscript in prep.)
- Heald G., Józsa G., Serra P., *et al.* 2011, *Astronomy & Astrophysics*, 526, A118
- Jaffé Y. L., Poggianti B. M., Verheijen M. A. W., Deshev B. Z., van Gorkom J. H. 2013, *Monthly Notices of the Royal Astronomical Society*, 431, 2111
- Jones M. G., Haynes M. P., Giovanelli R., Moorman C. 2018, *Monthly Notices of the Royal Astronomical Society*, 477, 2
- Kanekar N., Sethi S., Dwarakanath K. S. 2016, *The Astrophysical Journal*, 818, L28
- Koribalski B. S., Staveley-Smith L., Westmeier T., *et al.* 2020, *Astrophysics and Space Science*, 365, 118
- Lah P., Chengalur J. N., Briggs F. H., *et al.* 2007, *Monthly Notices of the Royal Astronomical Society*, 376, 1357
- Maddox N., Frank B. S., Ponomareva A. A., *et al.* 2021, *Astronomy & Astrophysics*, 646, A35
- Masui K. W., Switzer E. R., Banavar N., *et al.* 2013, *The Astrophysical Journal*, 763, L20
- Meyer M. 2010, in *Proceedings of Panoramic Radio Astronomy: Wide-field 1-2 GHz research on galaxy evolution—PoS (PRA2009)*, Vol. 89, p. 015
- Oosterloo T., Fraternali F., Sancisi R. 2007, *The Astronomical Journal*, 134, 1019
- Pence W. D., Chiappetti L., Page C. G., Shaw R. A., Stobie E. 2010, *Astronomy & Astrophysics*, 524, A42
- Rhee J., Lah P., Briggs F. H., *et al.* 2017, *Monthly Notices of the Royal Astronomical Society*, 473, 1879
- Rubin V. C., Ford W., Kent J. 1970, *The Astrophysical Journal*, 159, 379
- Serra P., Oosterloo T., Morganti R., *et al.* 2012, *Monthly Notices of the Royal Astronomical Society*, 422, 1835
- Serra P., Westmeier T., Giese N., *et al.* 2015, *Monthly Notices of the Royal Astronomical Society*, 448, 1922
- Tully R. B., Fisher J. R. 1977, *Astronomy & Astrophysics*, 500, 105
- Walter F., Brinks E., de Blok W. J. G., *et al.* 2008, *The Astronomical Journal*, 136, 2563
- Walter F., Carilli C., Neeleman M., *et al.* 2020, *The Astrophysical Journal*, 902, 111
- Wang J., Koribalski B. S., Serra P., *et al.* 2016, *Monthly Notices of the Royal Astronomical Society*, 460, 2143
- Westmeier T., Kitaeff S., Pallot D., *et al.* 2021, *Monthly Notices of the Royal Astronomical Society*, 506, 3962
- Zwaan M. A., Meyer M. J., Staveley-Smith L., Webster R. L. 2005, *Monthly Notices of the Royal Astronomical Society: Letters*, 359, L30