Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **550,** 1–29 (2026)
Advance Access publication 2026 June 12

https://doi.org/10.1093/mnras/stag1091

MIGHTEE–HI: H I catalogue of 293 sources for the COSMOS field and comparative study of 3-dimensional source finding methods

Michalina Maksymowicz-Maciata,[1★] Natasha Maddox,[1] Catherine Hale,[2,3] Ben Maughan,[1] Matt J. Jarvis,[2] Anastasia A. Ponomareva,[2,4] Ian Heywood,[2,5,6] Hengxing Pan,[7,8] Sushma Kurapati,[9] Tom G. Hardy,[10] Marcin Glowacki,[3,11] Tobias Westmeier,[12] Maarten Baes,[13] Seoyoung Lyla Jung[2] and Andreea A. Vărăşteanu[2]

[1]*School of Physics, H.H. Wills Physics Laboratory, Tyndall Avenue, University of Bristol, Bristol BS8 1TL, UK*
[2]*Sub-Dep. of Astrophysics, Dep. of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[3]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh, EH9 3HJ, UK*
[4]*Centre for Astrophysics Research, School of Physics, Astronomy and Mathematics, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK*
[5]*SKA Observatory, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK*
[6]*Department of Physics and Electronics, Rhodes University, PO Box 94, Makhanda 6140, South Africa*
[7]*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, People's Republic of China*
[8]*Guizhou Radio Astronomical Observatory, Guizhou University, Guiyang 550000, China*
[9]*Netherlands Institute for Radio Astronomy (ASTRON), Oude Hoogeveensedijk 4, NL-7991 PD Dwingeloo, The Netherlands*
[10]*Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[11]*Inter-University Institute for Data Intensive Astronomy, Department of Astronomy, University of Cape Town, Cape Town, Private Bag X3, Rondebosch 7701, South Africa*
[12]*International Centre for Radio Astronomy Research (ICRAR), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia*
[13]*Sterrenkundig Observatorium, Universiteit Gent, Krijgslaan 299, B-9000 Gent, Belgium*

Accepted 2026 June 4. Received 2026 May 29; in original form 2025 November 10

ABSTRACT

We present a catalogue of H I sources extracted from the MIGHTEE survey data cubes covering the COSMOS field. The catalogue contains 293 sources in the redshift range of $0.004 < z < 0.093$. In addition to HI masses and velocity widths, the catalogue includes optical through near-infrared photometry and inferred stellar masses and star-formation rates. The quantity of sources in the HI catalogue acquired through untargeted source finding is greatly influenced by the source finding methods used. This study therefore also provides a well-characterized expected completeness of the detected sample of galaxies based on their properties, informing of any detection biases, inferred through a comparative study of different source finding algorithms. We have tested the performance of widely-used source finders: PYBDSF, PROFOUND, and SOFIA, along with new source finder LESHI, focusing exclusively on H I source detection rather than source characterization in the first instance. The source finders were tested by injecting a sample of simulated galaxies divided into narrow bins of mass, inclination and distance into a MeerKAT data cube. The results inform the source finding strategies for the MeerKAT International GigaHertz Tiered Extragalactic Exploration (MIGHTEE) survey, as well as upcoming SKAO surveys.

Key words: methods: observational – catalogues – software: data analysis – galaxies: abundances – radio lines: galaxies.

1 INTRODUCTION

The 21 cm spectral line emission of H I is one of the primary wavelengths observed in radio astronomy as neutral hydrogen serves as the raw material for the build-up of stellar mass. It can be observed in galaxies, tracing their structure and neutral gas reservoir, and free-floating clouds that have been stripped from galaxies, and is of great scientific interest (see for example J. E. Hibbard et al. 2001, S. Ranchod et al. 2021 and references therein). The H I emission is intrinsically very faint (compared to optical data) and mapping the extragalactic HI Universe therefore requires wide, deep and untargeted surveys, leading to a large number of sparsely populated data cubes.

Several large-area HI surveys have already been undertaken. Among single dish radio telescope surveys is the Arecibo Legacy Fast ALFA Survey (ALFALFA; R. Giovanelli et al. 2005), which uses the observations from the Arecibo Telescope. The whole sky is covered by the southern hemisphere HI Parkes All-Sky Survey (HIPASS; D. G. Barnes et al. 2001), along with the Northern HIPASS extension (NHICAT; O. I. Wong et al. 2006). Other large-

★ E-mail: michalina.maksymowicz-maciata@bristol.ac.uk

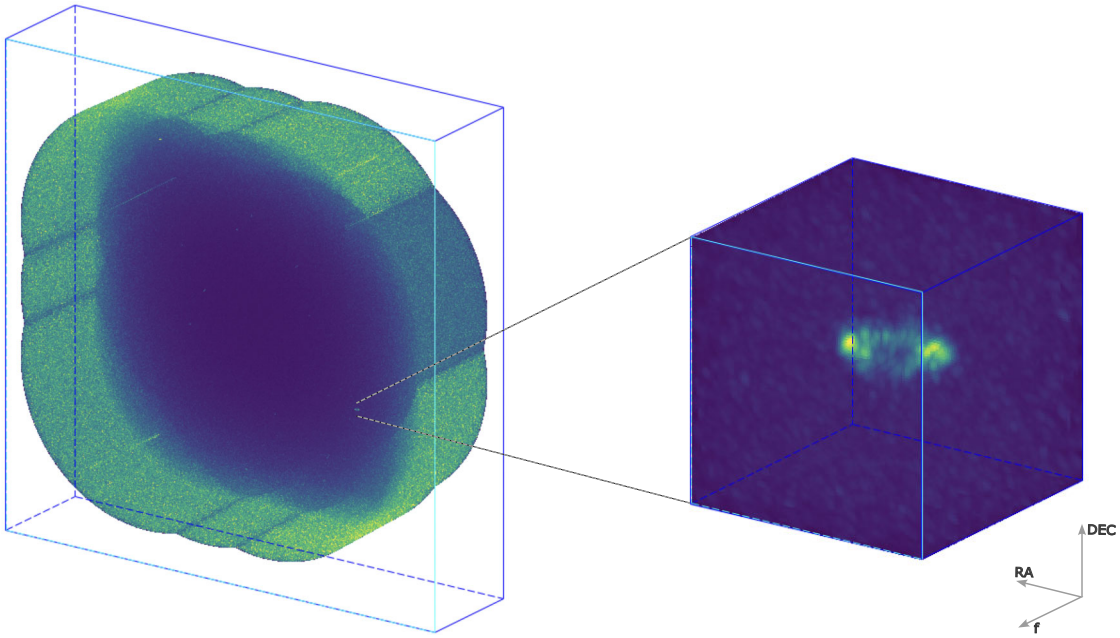

Figure 1. Three-dimensional view of a data cube for the frequency range of 1.3685–1.3961 GHz (spanning 1055 channels) used for source injection in this work (left) and an extracted small volume centred on an example of a real source (right), both adapted from SAOImageDS9 application (W. Joye 2019) visualization. The data cube extract covers the COSMOS field with a mosaic of 15 pointings with a total area spanning $\sim 4\,\mathrm{deg}^2$ (4600 by 4600 pixels), centred on RA= 150.03158 deg, Dec. = 2.208856 deg, taken by the MeerKAT telescope as part of the MIGHTEE survey.

sky surveys include the Commensal Radio Astronomy FAST Survey (CRAFTS; D. Li et al. 2018), and the FAST All Sky H I survey (FASHI; C.-P. Zhang et al. 2024), which utilize the Five hundred meter Aperture Spherical Telescope (FAST; R. Nan et al. 2011). Interferometric and wide surveys include H I surveys with Apertif (E. A. K. Adams et al. 2022), which is a phased-array feed system for the Westerbork Synthesis Radio Telescope, the Widefield ASKAP *L*-band Legacy All-sky Blind surveY (WALLABY; B. S. Koribalski et al. 2020) and the Deep Investigation of Neutral Gas Origins (DINGO; J. Rhee et al. 2023), which use the Australian Square Kilometer Array Pathfinder (ASKAP; A. W. Hotan et al. 2021). Among deeper interferometry surveys is the COSMOS H I Large Extragalactic Survey (CHILES; X. Fernandez et al. 2016) targeting higher-redshift sources with the upgraded Karl G. Jansky Very Large Array (VLA; M. Lacy et al. 2020), and the Blind Ultra Deep HI Environmental Survey (BUDHIES; Y. L. Jaffé et al. 2013), using the the Westerbork Synthesis Radio Telescope (WSRT; J. A. Hogbom & W. N. Brouw 1974). The Karoo Array Telescope (MeerKAT; J. L. Jonas 2009, J. Jonas & MeerKAT Team 2016) is used to study H I by the Looking At the Distant Universe with the MeerKAT Array survey (LADUMA; B. W. Holwerda, S. L. Blyth & A. J. Baker 2012, S. Blyth et al. 2016), and the MeerKAT International GigaHertz Tiered Extragalactic Exploration survey (MIGHTEE; M. Jarvis et al. 2016), which we further describe in Section 4.1. The continuing operation of existing radio telescopes, as well as the upgrade of their resolution and field of view and construction of new telescopes, such as the Square Kilometre Array (SKA; R. Braun et al. 2015), DSA-2000 radio camera (G. Hallinan et al. 2019), or the Canadian Hydrogen Observatory and Radio-transient Detector (CHORD; K. Vanderlinde et al. 2019), will lead to a rapid increase in the number and size of data cubes and their sensitivity. The MeerKAT radio telescope alone can produce 4.7 GB of data per second.

It is therefore a challenge of great importance to find the best approach to optimally and reliably perform source finding. This is particularly challenging for three-dimensional spectral line data (data cubes), which, with the new generation of radio-telescopes, can span a wide range of channels and pixels (e.g. Fig. 1), making the search for the signal much more computationally expensive due to the added frequency dimension. Moreover, the challenge increases for faint sources close to the noise level, as their inclusion may lead to many false positives, forcing trade-offs between sample completeness (fraction of sources found) and reliability/purity (fraction of found sources that are real).

It has been shown that using eyes remains one of the most complete and reliable methods of source finding (R. Taylor 2025); however, applying it to large amount of data is unfeasible. Consequently, a number of automated source finders emerged from the scientific community. Among two-dimensional imaging source finders designed for optical data is PROFOUND, which works well for all kinds of images including radio continuum data (A. S. G. Robotham et al. 2018, C. L. Hale et al. 2019; see Section 2.3.2 for more details). Other two-dimensional source finders are CAESAR (S. Riggi et al. 2019), Python Blob Detector and Source Finder (PYBDSF; N. Mohan & D. Rafferty 2015; see Section 2.3.2 for more details), AEGEAN (P. J. Hancock et al. 2012), and BLOBCAT (C. A. Hales et al. 2012). Even though these source finders were not designed for data cubes, but for two-dimensional images instead, their capabilities can still be applied to three-dimensional data with some post-processing (as done in this work for PYBDSF and PROFOUND). Among the source finders that were designed for three-dimensional data cubes (and therefore particularly relevant for H I source finding) is Source Finding Application (SOFIA; T.

Westmeier et al. 2021, P. Serra et al. 2015; see Section 2.3.2 for more details), DUCHAMP (M. T. Whiting (2012) and LIghtweight Source finding Algorithms (LISA; E. Tolley et al. 2022).

Source finders employ different algorithms, therefore it is essential to match the most suitable tool to the data and scientific requirements. For example, measurements of the H I mass function require reliable source identification with well-known completeness (defined by the ratio of the number of sources found and the number of all sources in the data), while searches for faint sources benefit from high completeness at the cost of reliability (defined by the ratio of the number of found sources that are real and the total number of found sources), which might be aided by ancillary data. Therefore, testing and comparing different tools becomes an important step. Some works have already attempted to tackle this problem: J. A. Barkai et al. (2023) carried out a comparative study of SOFIA, MTOBJECTS, and supervised deep learning algorithm originally designed for medical imaging. Many source finders were also tested during the SKA Science Data Challenge 2 (P. Hartley et al. 2023) on a simulated data product representing a 2000 h SKA-Mid spectral line observation.

In this paper, we test the SOFIA, PROFOUND, and PYBDSF source finders and introduce a new source finder LESHI (Line Emission Source-Hunting Integrator), developed for H I data. We explore how they perform and compare by measuring completeness and purity (reliability) of the resulting samples when run on a cube with simulated injected H I emission line sources and explore their different strengths and weaknesses. In contrast to the past work done on comparing different source finders, we push the testing boundaries by incorporating different ranges of mass, inclination, and frequencies (distances) of the injected sources and investigate how the source finders perform in each parameter bin. Lastly, we derive a function outputting the expected completeness from input mass, inclination and distance bins based on our results.

In the second part of this work, we put the lessons learned from testing the source finders into practice and produce a catalogue of H I sources acquired through untargeted source finding with the LESHI source finder for the MIGHTEE data cubes available for the COSMOS field (I. Heywood et al. 2024). This field has a wealth of deep multiwavelength ancillary data, both spectroscopic and photometric, allowing us to optically confirm the detections and determine the stellar contents and star formation rates. Our catalogue of 293 galaxies includes H I properties, along with photometry, stellar masses and star-formation rates (SFRs) of the optical counterparts.

The paper is structured as follows. In Section 2, we briefly describe the tested source finders and the injecting methods, while in Section 3, we present the results and discuss the strengths and weaknesses of each source finder. Finally, in Section 4, we present the MIGHTEE COSMOS H I catalogue. Section 6 is reserved for summary and conclusions.

Throughout this paper, we assume ΛCDM cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\Omega_\mathrm{M} = 0.3$, and $\Omega_\Lambda = 0.7$. The magnitudes are given in the AB magnitude system and are not extinction corrected.

2 SOURCE FINDERS COMPARISON METHODS

2.1 Source injection

We assess the completeness and reliability of the source finders, by injecting artificial sources into a real cube. This allows us to see how each source finder performs with real data, with noise and artefacts, that are otherwise difficult to emulate. The cube used to test the source finders is a frequency slab of a full data cube produced by the South African MeerKAT radio telescope as part of the MIGHTEE survey (see I. Heywood et al. 2024 and Section 4.1 for more details). It covers the COSMOS field with a mosaic of 15 pointings with a total area spanning $\sim 4\,\mathrm{deg^2}$ and frequency range of 1.3685–1.3961 GHz with channel width of 26.125 kHz. We do not use the full frequency range of the data cube, as we do not need the full volume for the source injection and we avoid frequency ranges impacted by radio frequency interference (RFI). The testing results should hold for all of the frequencies spanned by MIGHTEE data, as the noise does not change significantly across the investigated distances, as can be seen in fig. 6 of I. Heywood et al. (2024). The median channel root mean square (rms) noise is 70 $\mu\mathrm{Jy\,beam^{-1}}$ in the middle of the mosaic rising to 480 $\mu\mathrm{Jy\,beam^{-1}}$ towards the edge of the image. The median synthesized beam is ∼15 arcsec in diameter. The cube has 4600 by 4600 pixels and 1055 channels (see Fig. 1), therefore providing a sparse volume that can be populated with a large number of galaxies without the risk of source confusion.

The cube contains real sources (barely visible 'specs' in the centre region of the cube in Fig. 1) and continuum artefacts ('lines' along the frequency direction in the cube in Fig. 1). To account for the real sources, we have run each source finder on the data (without injecting any new sources) and investigated each source found. The detections that had obvious emission and were within 8 arcsec of sources in the Sloan Digital Sky Survey (SDSS) optical catalogue were treated as real. Those were not masked, but instead a catalogue of all obvious real sources was created and their detections (or non-detections) were subsequently ignored.

The artificial sources were simulated using the ^{3D}BAROLO package (E. M. Di Teodoro & F. Fraternali 2015) that uses 3D tilted-ring models of line emission from galaxies. A sample of six simulated galaxies is shown in Fig. 2. The sources were simulated using the standard H I mass–luminosity conversion (M. Meyer et al. 2017) and HI mass-size relation (J. Wang et al. 2016, S. H. A. Rajohnson et al. 2022), which governs their physical size given an input H I mass. The rotation curves were derived using the baryonic Tully–Fisher relation (R. B. Tully & J. R. Fisher 1977, A. A. Ponomareva et al. 2021) for dwarf galaxies after assuming the baryonic mass being equal to HI mass and for massive galaxies the rotation velocity was randomly chosen between 100 and 350 $\mathrm{km\,s^{-1}}$. The simulated galaxies were convolved with the synthesized beam of the cube used for injection (∼15 arcsec). We injected 600 uniformly-spaced galaxies per run for low mass galaxies ($< 10^8\,\mathrm{M_\odot}$) and to avoid overcrowding the cube, 150 per run for high mass galaxies. We divide them into equal groups of different distances and explore the full mass-inclination plane by repeating the runs for different mass and inclination bins of the injected galaxies, resulting in the total of 24 000 injected sources. For all of the runs, the cube was at least 99.3 per cent empty. We note that the cube used for injection spans only the distance range of 75–165 Mpc. However, we simulate the galaxies as if they were at luminosity distances of 50, 100, 150, and 350 Mpc (accordingly lowering the flux and projected size) and inject them centred at a random frequency channel within the cube, and so the frequency of the injected galaxy no longer corresponds to the distance. The galaxies were injected in the central region of the data cube, where the noise level can be approximated as uniform. We note however, that higher noise has the same effect as higher

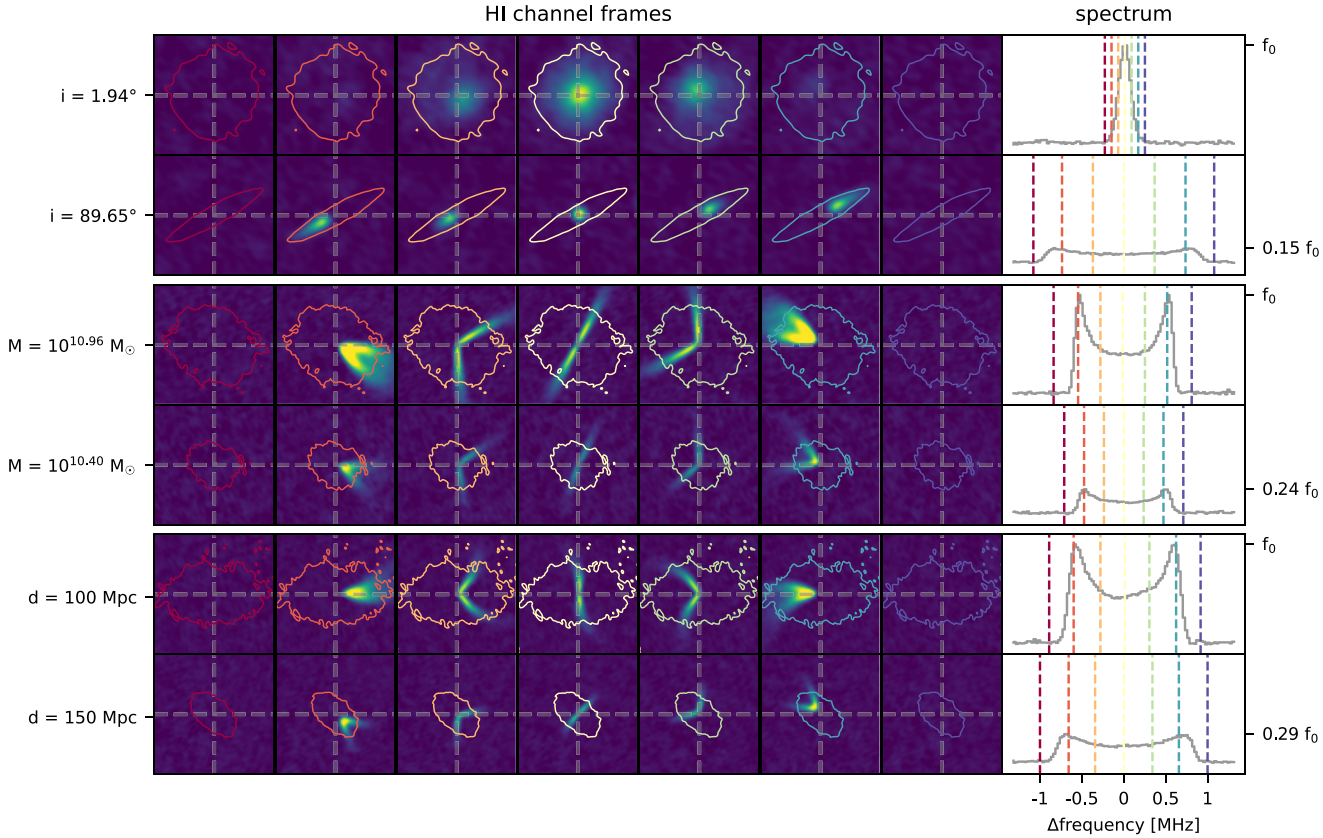

Figure 2. H I emission channel frames and spectra for a sample of simulated galaxies. The top two panels show galaxies of similar mass and distance, but varying inclination, the middle two panels show galaxies of similar inclination and distance, but varying mass, and the bottom two panels show galaxies of similar inclination and mass, but varying distance. Contours enclose the 3σ level of the simulated H I emission and are colour-coded (along with the vertical dashed lines on the spectrum plots) by the frequencies for which the channel images are shown.

distance, consequently lowering the SNR of the emission. We note that the increased noise towards the field edges would need to be accounted for in statistical studies such as measuring the HI mass function (A. A. Ponomareva et al. 2023).

2.2 Injected source parameters

The observed signal from a galaxy is greatly influenced by its properties, such as mass, inclination and distance. Fig. 2 shows how each of these parameters changes the observations in the spatial and frequency dimensions, by plotting the channel frames and spectra for a sample of injected galaxies. As can be seen, the spectrum's height, width and shape change dramatically for different parameters. It is therefore expected that the corresponding inputs for the injected galaxies should dictate how the source finders perform in each of the mass-inclination-distance bins (see Fig. 4). To investigate the response of the source finders to the parameters of the injected galaxies, we inject the galaxies in narrow bins of mass, inclination and frequency (distance). The ranges and bin-widths are elaborated on in the following subsections.

2.2.1 Inclinations

For randomly oriented galaxies, the distribution of the cosines of the inclination angles should be uniform. Our inclination bins therefore span the range of $0 < \cos(i) < 1$ in bins of width 0.2, with $\cos(i) = 0$ for edge-on galaxies and $\cos(i) = 1$ for face-on ones. As can be seen in the first panel of Fig. 2, varying the inclination has an impact on how the measured flux is spread across spatial pixels (if the source is resolved) and spectral channels. For the face-on galaxy (low inclination angle), the width of the spectral line comes mainly from the velocity dispersion of H I clouds, while for the edge-on galaxy (high inclination angle), the width of the emission (and the two peaks) is a result of the galaxy's rotation and the effect of Doppler's shift. While the shape of the spectrum changes significantly, the total flux is conserved, therefore for greater inclinations, while the emission line is wider, the peak flux also must be lower, pushing it towards the noise level, which leads to the galaxy being more challenging to detect. Hence, it is expected that the completeness of source finders should be greater for galaxies of a given mass with lower inclinations (face-on galaxies). However, if the spectral resolution of the data is low enough, emission from face-on galaxies might have the width of only one channel, leading to some source finders to reject it. It is also worth noting that if a high-inclination galaxy is faint enough, the middle section of its double-peaked spectrum can fall below the noise level, which leads to the spectrum looking like two separate peaks, possibly confusing the source finders if their separation is too large to be associated.

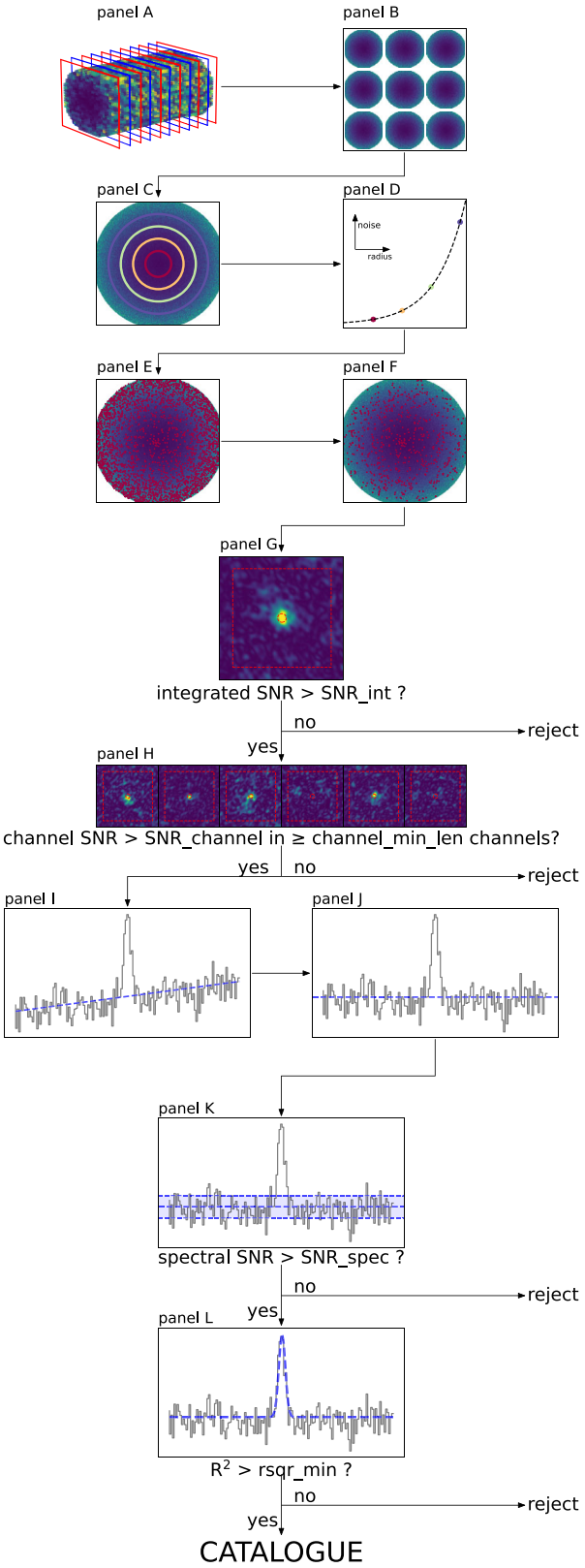

Figure 3. Flowchart visualizing each step of the LESHI script. See the main text for more details.

2.2.2 *Masses*

The typical HI masses in galaxies span from 10^6 $M_\odot$ to 10^{11} $M_\odot$ (J. Wang et al. 2016). We have therefore adopted the same range for the mass bins starting at 10^6 $M_\odot$ and incrementing by 1 in logarithmic space until 10^{11} $M_\odot$, in agreement with previous MIGHTEE–H I masses at $z < 0.1$. We increase the resolution down to 0.5 and 0.25 dex, for masses where the completeness is expected to change most rapidly. We increase the minimal mass tested to 10^8 $M_\odot$ at the distance of 350 Mpc, as lower masses would be below the detection threshold at this distance. As can be seen in the second panel of Fig. 2, H I mass has the expected impact on the form of the observed signal. For higher masses, the peak flux is greater and the emission line is slightly wider. The general shape of it remains similar to that from lower mass galaxies, as it is mostly dictated by the inclination.

2.2.3 *Distances*

The influence of the distance to the galaxy is very straightforward. Since measured flux is inversely proportional to the luminosity distance squared, the galaxy will simply appear dimmer (and smaller), as can be seen in the bottom panel of Fig. 2, lowering the signal-to-noise ratio of the detection in an unsophisticated way. This parameter space therefore does not require involved investigation. We simulate the galaxies as if they were at the luminosity distances of 50, 100, 150, or 350 Mpc (accordingly lowering the flux and projected size) and inject them into the full volume of our cube.

2.3 Source finders

2.3.1 *LESHI*

LESHI[1] (Line Emission Source-Hunting Integrator) is at the core a very straightforward PYTHON script, which is available from github[2]. By design, it does not have the sophistication of other source finders, making it simpler to use and install, at the cost of reduced functionality, as it does not attempt any source characterization and focuses solely on source finding. We note, however, that the exact capabilities of LESHI might change, as it is undergoing active development. As it is relatively fast, it can be used as a first round of source finding to be followed by more advanced tools. LESHI was designed based on the source finding methods done by eye, mimicking its simple tests and checks, as visual source finding has been proven to still be one of the most accurate methods (R. Taylor 2025).

The source finder works as follows. As its name suggests, it relies greatly on signal-enhancing integration. First, the cube is divided into thin slabs in frequency of width defined by the int_image_len input (panel A on Fig. 3), which are then integrated, summing all the signals from each channel, creating a number of images (panel B). This is then repeated on the cube divided into thin slabs with frequency moved by half-slab width, to ensure that no weak signal, that happened to be divided into two separate slabs at first, would be missed. To estimate the background noise for each integrated image, the script samples the noise using elliptical apertures (panel C), with the major and

[1] LESHI (also known as Leshy or Leshen) is a guardian deity of the forest and hunting in Slavic mythology.
[2] https://github.com/misia-mm/LESHI

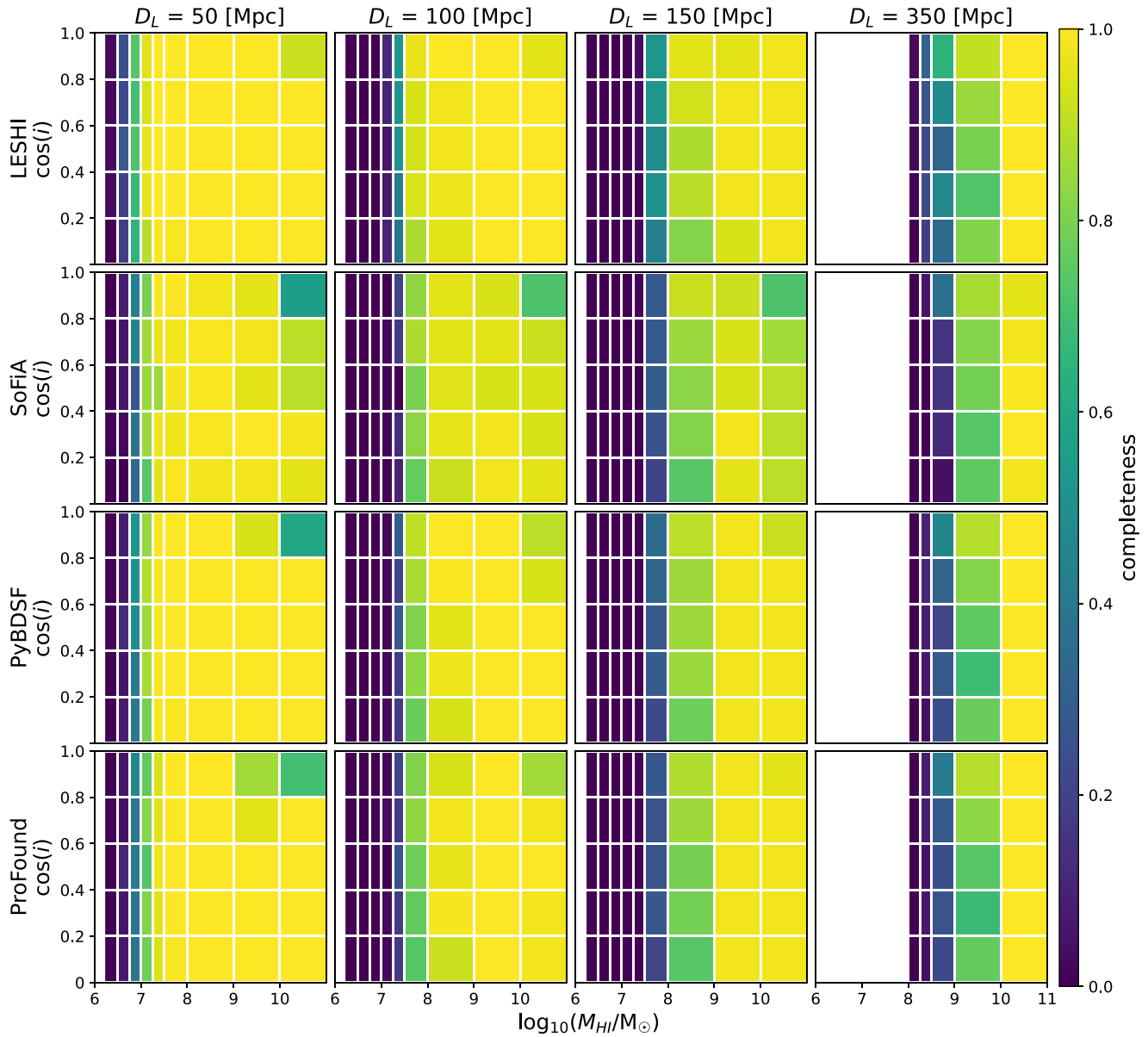

Figure 4. An array of two-dimensional histograms colour-coded by the completeness for each bin of mass (*x*-axis) and cosine of inclination (y-axis) of injected galaxies, achieved by each of the LESHI, SOFIA, PYBDSF, and PROFOUND source finders (accordingly first, second, third, and fourth row of histograms) for each simulated luminosity distance of 50, 100, 150, and 350 Mpc (accordingly first, second, third and fourth column of histograms).

minor axis of the ellipse calculated based on the width and height of the datacube, and fits an exponential function (panel D) to the sampled noise, which gives the estimated noise based on the position relative to the centre of the image. After estimating background noise, all created images are then subjected to the FIND_PEAKS() ASTROPY PYTHON package (Astropy Collaboration 2022) function, which produces a catalogue of all sources detected on the integrated images. The function finds sources above a single specified threshold, which is set based on the noise level at the centre of the image, leading to many false detections towards the edges, where the noise is higher (panel E), due to the primary beam correction. Then, for every source found, the fitted exponential function (from panel D) is used to guess the noise level around the source based on its position. This method is faster than checking the local noise level for every detection candidate (or for a grid); however, it is not fault-proof, since it assumes elliptical symmetry, which is not always the case (especially for mosaicked data). We treat it therefore as a first round of higher-tolerance filtering, to mostly weed out the false detections at the edges of the data (panel F). Then, for every detection that has passed, we estimate the local background more accurately by using a box of size specified by the bg_box_size input parameter around the source and check if the emission's SNR on the integrated image passes the threshold specified by the SNR_integ input parameter (panel G). The catalogue of sources that have passed this test is still greatly contaminated by noise peaks, however by investigating each source in the frequency space, those can be identified and excluded.

The first test in the frequency space checks if at the coordinates of the initially detected source, there is a signal greater

than a specified threshold, that persists across channels and is at least of the size equal to the synthesized beam (panel H). If the source is not detected above an SNR threshold specified by the SNR_channel input in a number or more channels (with the number being specified by the channel_min_len input), it is excluded. Next, the spectrum is created from the integrated spectral profile of central pixels within the beam's FWHM at the coordinates of the detected source (panel I) and the script attempts to subtract the spectrum baseline if the sloped_baseline parameter is set to true (panel J). The second test calculates the SNR of the spectrum (panel K), if the SNR of a given source is below the threshold specified by the SNR_spec input, it is excluded. Lastly, the code fits a Gaussian function to the spectrum (panel L), using the EMCEE (D. Foreman-Mackey et al. 2013) package, since a real spectral line is expected to have Gaussian-like shape within the beam size. If the detected spectral line does not fit the Gaussian within a minimum value of coefficient of determination R^2 specified by the rsqr_min input, it is excluded. Each source is then cross-matched and associated with sources that are within a specified number of channels (given by the max_dist_channel input) and angular separation (given by the max_dist_pix input). Then, the final catalogue of the sources that passed all the checks is outputted.

The LESHI script is parallelized and the number of cores used can be modified using the core_no input parameter. Its runtime is relatively low (see Section 3.3) and can be further reduced by using more cores. Its memory footprint is also relatively low and only depends on the spatial size of the data cube (not the frequency length), since the script works on a number of integrated images at a time (specified by int_image_load_no input). The memory footprint can be therefore further reduced by lowering the int_image_load_no input parameter, at the cost of increased runtime.

It is worth noting that, since the last Gaussian-fitting test is the most rigorous one, the earlier tests' purpose is to mainly weed out the sample as much as possible, as they are much less computationally expensive than the last test. It is therefore important to find a good combination of input thresholds for the earlier tests to allow through weaker sources, but not too many of them to not make the runtime prohibitively high. A good combination of input parameters, that we have used in this work, is given in Appendix A4.

2.3.2 *SOFIA, PROFOUND, and PYBDSF*

SOFIA (T. Westmeier et al. 2021, P. Serra et al. 2015) is a flexible software application for the detection and parametrization of sources in data cubes. SOFIA presents the choice of a variety of techniques for data filtering and 3D source-finding and has been developed to be independent of the type of emission line data used. It utilizes the smooth and clip algorithm by iteratively applying spatial and spectral smoothing to the cube on different scales and clipping emission below a certain threshold at each smoothing level. Currently, this is the most commonly-used pipeline for source finding in HI emission data. SOFIA's input parameters we use are given in Section A1.

PROFOUND (A. S. G. Robotham et al. 2018) is a source finding and image analysis package for two-dimensional data written in the R programming language. It provides tools to generate segmentation maps to discern blended sources and perform photometry. It was mainly developed for optical and infra-red images, but has proven to work well for all types of image data, including radio continuum (C. L. Hale et al. 2019). The input parameters used here for PROFOUND are given in Section A2.

PYBDSF (PYTHON Blob Detector and Source Finder, N. Mohan & D. Rafferty 2015) is a source-finding PYTHON package developed for radio interferometric two-dimensional data, and was adopted for MIGHTEE continuum data (C. L. Hale et al. 2025). It decomposes images into sets of Gaussians, shapelets or wavelets, and has tools to measure the point-spread function (PSF) variation across an image and calculate spectral indices and polarization properties of sources. PYBDSF's input parameters used in this work are given in Section A3.

The choice of parameters has big impact on the performance of the source finders, as they can be tuned for different uses cases. In this work we use either default or sensible parameters aiming to optimize the completeness and reliability.

PYBDSF and PROFOUND are two-dimensional source finders, however their capabilities can be extended into three dimensions with a cross-matching step. Both source finders were first run on every image of each channel of our data cube creating a catalogue of detections. Then, to find the sources, every detection was cross-matched with the others using STILTS (M. B. Taylor 2005, M. B. Taylor 2006) and detections within 8 arcsec radius (which is approximately equal to the beam radius) and one channel away from each other were associated, since we expect real sources to persist in neighbouring channels at the velocity resolution of the cube. Detections that were present in three or more channels were accepted and the rest were rejected as probable noise peaks, since we would not expect any real sources with spectral width of less than three channels, which is $\sim$16.5 km s^{-1} for our data.

3 SIMULATION RESULTS

3.1 Completeness

After completing injection and source finding for each investigated parameter-space bin, we have computed the completeness that each source finder has, by dividing the number of found sources by the number of injected sources, which can be seen in Figs 4 and 5. As anticipated, mass and distance have the greatest influence on the number of sources found, shifting the detection threshold, with the completeness dropping from 1 to 0 within approximately 1 dex of mass. Inclination, although to a much lesser degree, also has an impact, which can be seen in Fig. 4 particularly for the distance of 350 Mpc and the mass bin of 10^9-10^{10} $M_\odot$: all source finders perform better for more face-on galaxies, as expected. However, inclination has less impact at smaller distances, where the detection threshold shifts to lower mass galaxies, for which the rotation velocity is much smaller (as a result of Tully–Fisher relation that we adopt) and width of the spectral line is less affected by inclination. Another interesting inclination feature is that for the highest mass bin (10^{10}-$10^{11}M_\odot$) and face-on galaxies (cos(i) range of $0.8 - 1.0$), all source finders are underperforming, with this effect worsening for smaller distances. This is due to the galaxies at these masses and distances having large projected angular sizes, and for face-on inclination all of the emission appears in very few channels, covering a large region (as opposed to each channel showing only a fraction of the area of the galaxy, as can be seen in the middle panel of Fig. 2). If the emission region is of comparable size to the defined regions used for calculating the noise level, it is overestimated and the galaxy is missed. On the other hand, the background region size

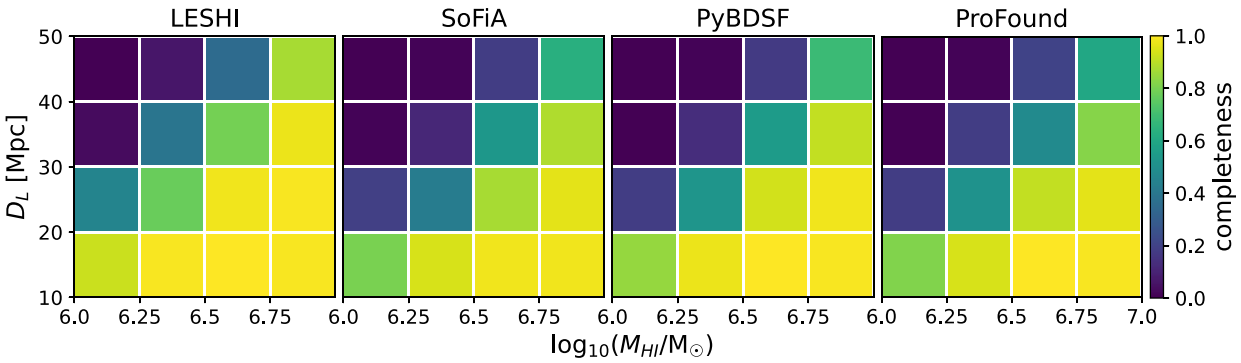

Figure 5. An array of two-dimensional histograms colour-coded by the completeness for each bin of mass (*x*-axis) and luminosity distance (y-axis) of injected galaxies, achieved by each of the LESHI, SOFIA, PYBDSF, and PROFOUND source finders (accordingly first, second, third, and fourth histogram).

should not be too large, as this can lead to over-smoothing of local background variations. This effect should be kept in mind when initializing the source finders. It is important to note that PYBDSF and SOFIA have an option for an adaptive box size, that changes its size based on measured local brightness; however, this greatly increases computational time.

As the completeness drops to zero for masses below 10^7 $M_\odot$ in the nearest distance bin in Fig. 4, we have separately explored the parameter space of HI masses 10^6-10^7 $M_\odot$ and distances 10–50 Mpc for a constant inclination of 45°, since inclination is much less important for low-mass galaxies, which can be seen in Fig. 5. The lowest mass galaxies are found only for the distances < 30 Mpc.

3.2 Completeness function

As expected, completeness depends strongly on the mass and distance of the galaxies. To quantify this relation for the LESHI source finder, we averaged the completeness over all inclinations for given luminosity distance and mass bins, and show this as a function of mass for the four different luminosity distances. We plot the errors accounting for the number of injected sources, assuming Poisson statistics, and errors accounting for the width of the investigated mass bins, calculated by taking the plus minus quarter of the width of the bin (left panel of Fig. 6). For each luminosity distance, we fit an error function to the completeness data in the form of:

$$c = \begin{cases} \mathrm{erf}(s \cdot (\log_{10}(\frac{M_{\mathrm{HI}}}{M_\odot}) - \log_{10}(\frac{M_{\mathrm{min}}}{M_\odot}))), & \text{for } M_{\mathrm{HI}} > M_{\mathrm{min}} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\mathrm{erf}(z)$ is the error function defined by $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t$, c is the completeness, M_{HI} is the H I mass, s and M_{min} are free parameters that depend on the luminosity distance: s is responsible for the slope, while M_{min} is the minimal detectable mass (M_{HI} for which the completeness function reaches zero). These functions are overplotted in the left panel of Fig. 6.

Next, we plot the best fitting M_{min} and s parameters against the luminosity distance, which can be seen in the right panels of Fig. 6. Since the minimal detectable mass should be proportional to the luminosity distance squared, to find the relation between M_{min} and the distance, we fit a function of the form:

$$M_{\mathrm{min}}(D_{\mathrm{L}}) = a_1 \cdot D_{\mathrm{L}}^2 \quad (2)$$

where D_{L} is the luminosity distance in Mpc and we find the best fitting value for the parameter a_1 to be 1320 ± 50 [$M_\odot$Mpc^{-2}] (top right panel of Fig. 6). Next, we plot the best-fitting s parameter for a given luminosity distance against that distance (bottom right panel of Fig. 6), and we observe that their relation is very well fitted by a linear function of the form:

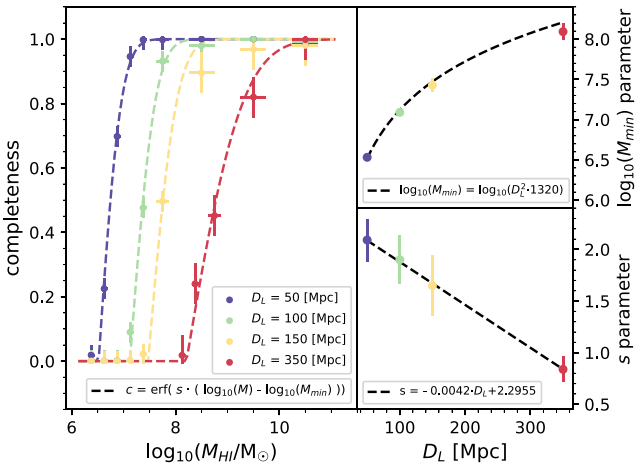

Figure 6. Left panel: inclination averaged completeness achieved by the LESHI source finder *vs* the injected H I mass for different luminosity distances, with the best-fitting error function represented by the dashed lines. Right panels: best-fitting parameters characterizing the fitted error function *versus* the luminosity distance.

$$s(D_{\mathrm{L}}) = a_2 \cdot D_{\mathrm{L}} + b_2 \quad (3)$$

and we find the best fitting parameters to be $a_2 = -0.0042 \pm 0.0007$ [Mpc^{-1}] and $b_2 = 2.30 \pm 0.18$ (unitless).

It is important to note that a decreasing linear function will eventually reach zero. Therefore, for the fitted a_2 and b_2 parameters, the s parameter (characterizing the slope of the completeness function) would reach zero for the distance of ~ 550 Mpc, which is not physical, as we expect to detect H I sources at this distance. Therefore, while the function describing how s parameter changes with distance looks linear for the distance range we have investigated, we would expect its true form to asymptotically reach a certain value at high distances (but never zero). To find the function, a similar study might be conducted in the future for higher distances.

We have attempted to find a form of the completeness function depending on the inclination; however, the sample size was not good enough to find a well constrained relation. Instead, we study

Table 1. Fitted parameters of the completeness function for the completeness averaged over all inclinations, for face-on sources [cos(i) range of 0.9 to 1.0] and for edge-on sources [cos(i) range of 0.0 to 0.1].

	a_1 ($M_\odot Mpc^{-2}$)	a_2 (Mpc^{-1})	b_2
average	1320 ± 50	-0.0042 ± 0.0007	2.30 ± 0.18
face-on	1260 ± 90	-0.0027 ± 0.0014	2.21 ± 0.20
edge-on	1380 ± 110	-0.0039 ± 0.0008	2.10 ± 0.17

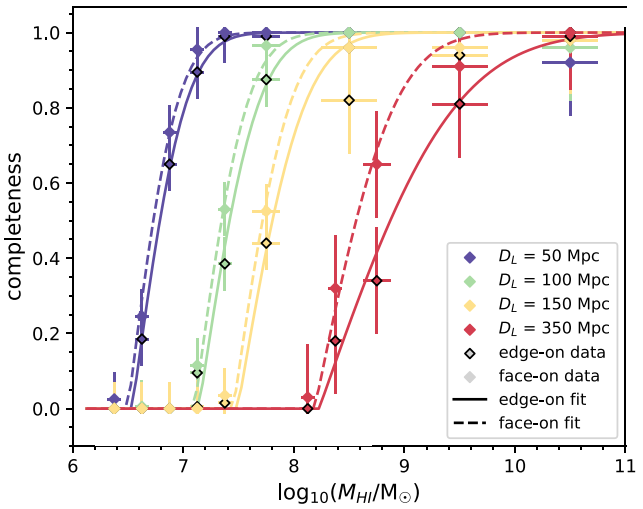

Figure 7. Completeness achieved by the LESHI source finder *vs* the injected H I mass for different luminosity distances for face-on sources (represented by coloured diamonds) and edge-on sources (represented by coloured diamonds with dark edge) and the best-fitting error functions represented by the dashed lines (for face-on sources) and solid lines (for edge-on sources).

the impact of the inclination by considering the two extreme cases: face-on galaxies (cos(i) range of 0.9 to 1.0) and edge-on galaxies (cos(i) range of 0.0 to 0.1). We follow the same procedure described above to fit for the parameters describing the completeness function for each case, which can be found in Table 1. Fig. 7 (analogously to left panel of Fig. 6) shows the completeness data and the fitted function for edge-on and face-on sources. We can clearly see that, while the $M_{\rm min}$ is very comparable for the two cases, the slope is lower for the edge-on sources, especially for the distance of 350 Mpc, effectively lowering the completeness and confirming what we have already found from Fig. 4. Analogous figures and equations for the SOFIA, PROFOUND, and PYBDSF source finders can be found in Appendix B.

All the fits were done using the EMCEE (D. Foreman-Mackey et al. 2013) PYTHON package. We used 32 walkers, 5000 steps and burn in phase of 200 steps to discard and a Gaussian likelihood function.

3.3 Source finder comparison

All source finders could have been optimized differently to satisfy particular use cases. They could be tuned to be more complete at the expense of reliability and vice versa. Therefore, the following results are not representative of the source finders as whole, but

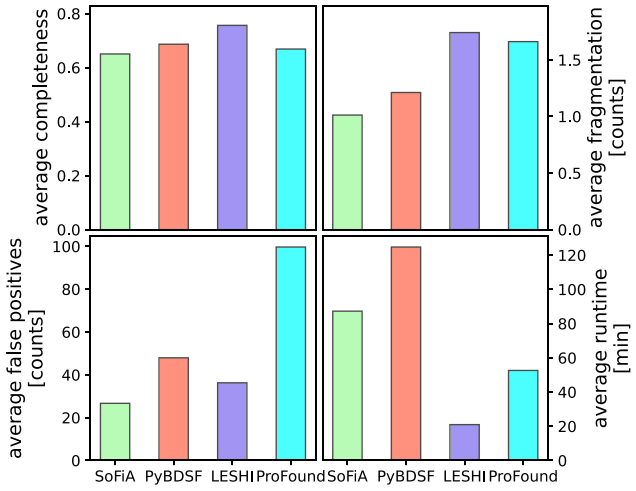

Figure 8. Comparison between the source finders, top left panel shows the average completeness achieved by each source finder, bottom left panel shows the average number of false positives output by each source finder, top right panel shows the average fragmentation of each source finder (defined by the number of sources that were assigned to a single source on average), and lastly, bottom right panel shows the average runtime of each source finder in minutes.

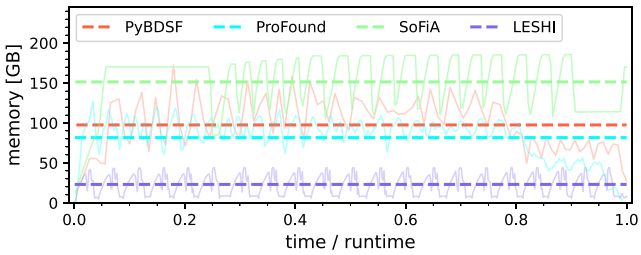

Figure 9. Memory footprints of each source finder run on the 89.3 GB data cube marked by solid lines, with their average marked by the dashed lines.

rather of the particular setup settings used. In this work, we have run the source finders with parameters found through attempting to optimize completeness and reliability for the given data. The input parameters for all source finders and their justification can be found in Appendix A.

Fig. 8 showcases a comparison between the source finders: their average completeness (counting only the runs where at least one source finder had at least 5 per cent completeness), average number of outputted false positives (characterizing reliability), average fragmentation (defined as the average number of found sources output by the source finder that belonged to one injected source) and average runtime of the source finders. The number of false positives did not change between the runs, since it should not depend on the injected sources, but on the quality of data and its noise, which was the same for all the runs (since the same real cube has been used).

As can be seen in Fig. 8, all of the source finders have comparable completeness, with LESHI having the highest value. When it comes to purity, SOFIA performs best, detecting fewest false positives, followed closely by LESHI. SoFiA has the best source fragmentation performance, proving its source characterizing capabilities, while LESHI often detects emission belonging to the same galaxy as separate sources, especially for high-mass ex-

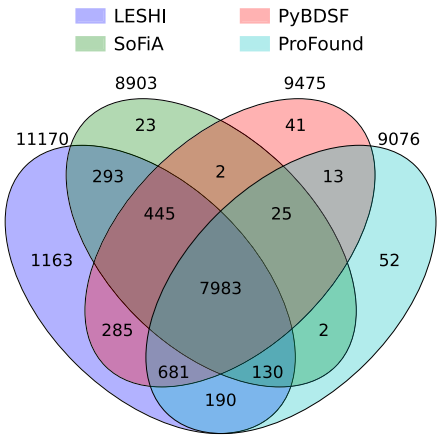

Figure 10. Venn diagram of injected sources that were found by each source finder, representing the overlap between the different source finders (11328 sources were found by at least one source finder out of total of 24 000 injected sources).

tended sources, with large projected sizes. Finally, when it comes to average runtime, LESHI was the fastest. It is important to note that all of the investigated source finders can use parallelization. In this work we have run the source finders on a computer cluster with 40 cores. Fig. 10 shows how many of the injected sources were detected by each source finder and their overlap. Notably, LESHI has the highest number of injected sources detected by no other source finder, while missing only 1.4 per cent of the sources that were detected by at least one source finder.

Fig. 9 shows the memory usage of each source finder run on the 89.3 GB data cube. It is worth noting that since PYBDSF, PROFOUND, and LESHI work on the channel images separately (with LESHI working on the integrated images), their memory footprints are mostly influenced by the number of parallel threads used, rather than the frequency range of the data cube. Their memory footprint can be therefore adjusted based on the available resources.

While SOFIA is very versatile and can achieve excellent results in the hands of an expert, it has to be optimized differently for each data set. LESHI offers an alternative option, at the expense of source characterization, and is straightforward to set up even for non-expert users. A collaborator not involved in LESHI's development ran it on different (with respect to pointing centre, spectral resolution, and depth) MeerKAT data sets with recommended settings. This additional testing achieved good results, with new, previously missed, and convincing sources reported by LESHI (private communication).

4 MIGHTEE-HI DR1 H I CATALOGUE

In the second part of this work, we perform H I source finding with LESHI and produce a catalogue of new H I sources from the MIGHTEE survey. We measure the HI and stellar properties of the galaxies and compile them into a publicly available catalogue.

4.1 The MIGHTEE survey

The MIGHTEE (M. Jarvis et al. 2016) survey is a medium-deep, radio continuum, spectral line and polarization survey, covering total area of 20 deg^2 spread over four well-studied extragalactic fields: COSMOS, XMM–LSS, ECDFS, and ELAIS–S1. It uses the MeerKAT radio telescope (J. L. Jonas 2009, J. Jonas & MeerKAT Team 2016), which is located in South Africa and consists of 64 offset-Gregorian dishes, each comprising of a 13.5 m diameter main reflector, covering the baselines up to 8 km. Three receivers operate in the *UHF* band, *L* band, and *S* band. The science goals of the MIGHTEE survey focus on three primary aspects: radio continuum (I. Heywood et al. 2022, C. L. Hale et al. 2025), polarization (A. R. Taylor et al. 2024), and spectral lines (N. Maddox et al. 2021).

Table 2. Brief description of the MIGHTEE-HI DR1 data used in this paper.

Area covered	∼4 deg^2
Observation time	94.2 h
Right ascension range	∼148.89–151.17 deg
Declination range	∼1.20–3.23 deg
Frequency range	1290–1520 MHz (z = 0–0.1 for H I)
Channel width	26.126 kHz (5.5 km s^{-1} at z = 0)
Pixel size	2 arcsec
Median synthesized beam	15.53 arcsec × 15.53 arcsec (Ranging 14.74–16.33 arcsec)
Sky rms	70 μJy beam^{-1} at the centre of the mosaic up to 480 μJy beam^{-1} towards the edge

This work is part of the MIGHTEE-HI working group which focuses on neutral hydrogen sources. The H I sources from the Early Science data (A. A. Ponomareva et al. 2021) were found using unguided visual source finding, covering areas in the COSMOS and XMMLSS fields. In this work, we use the LESHI source finder to expand the H I catalogue within the COSMOS field to lower SNR sources and characterize completeness.

The data used in this work is from the first Data Release (DR1) of the MIGHTEE-HI project (see I. Heywood et al. 2024 for more details). It covers the COSMOS field and the frequency range 1290–1420 MHz within the *L* band, corresponding to $0.0 < z_{HI} < 0.1$ for HI. The data is summarized in Table 2.

4.2 Ancillary data

The COSMOS field is very well studied and has a plethora of available ancillary data. To confirm our H I detections, we have cross-matched the catalogue with the Sloan Digital Sky Survey sources (SDSS; S. Alam et al. 2015) and The Dark Energy Spectroscopic Instrument catalogue (DESI; DESI Collaboration 2026). Photometry for the *g, r, i, z, y* optical bands was measured using the Hyper Suprime-Cam images (HSC; H. Aihara et al. 2019, S. Miyazaki et al. 2018). Photometry for *Y, J, H, K* infrared bands was measured from UKIRT Infrared Deep Sky Survey (UKIDSS; A. Lawrence et al. 2007) Large Area Survey images, which uses the UKIRT Wide Field Camera (WFCAM; A. Lawrence et al. 2007). We note that the COSMOS field has also available UltraVISTA data (H. J. McCracken et al. 2012) for the infrared bands; however, our HI data extend beyond the UltraVISTA coverage footprint. Therefore, we have used the UKIDSS photometry to make consistent measurements. We note that the COSMOS field is also covered by the DEVILS survey (L. J. M. Davies et al. 2018, 2025); however, while very complete, it only targets the very centre of the field of view of our datacube.

4.3 Source finding on real data

Catalogue sources were first identified through untargeted source finding with LESHI with set-up parameters given in Appendix A4. Every source was then cross-matched with SDSS and DESI catalogues with sources within 8 arcseconds (which is approximately equal to the radius of the beam) and redshift difference less than 0.001 (corresponding to 1.3 MHz or 285 km s^{-1} at the median redshift of 0.05), for systems with spectroscopic redshifts, or within redshift errors for systems with photometric redshifts., producing a sample of 530 detections with an optical counterpart. All sources with optical counterparts, integrated SNR greater than five and outside of frequency range with many artefacts ($\gtrsim$ 1306 MHz) were accepted. Sources with optical counterparts, integrated SNR of 3–5 and/or at frequencies with many artefacts ($\lesssim$ 1306 MHz) were checked by eye. Sources with the spatial distribution of the emission not consistent with the optical counterpart and with the spectral line being surrounded by strong noise peaks and artefacts resulting from RFI were either excluded or flagged as dubious (see for example Fig. D3). All sources were cross-matched with radio continuum data (C. L. Hale et al. (2025)) and their coordinates checked for a 'returning emission' in frequency characteristic of continuum artefacts to rule out their possibility. After removing duplicates (arising from overlapping frequency slabs or overfragmentation of sources), this resulted in 292 H I sources with plausible optical counterparts and consistent redshifts. The spatial and frequency distribution of all sources can be seen on Fig. C1.

We have also performed source finding on all the data cubes with the SOFIA, PROFOUND, and PYBDSF source finders, to check if they find any sources that were missed by LESHI. We filtered the candidates following the procedure described above and cross-matched them with the sources found by LESHI. SOFIA found 99 sources, PYBDSF found 155 sources and PROFOUND found 168 sources with SOFIA finding one additional lower SNR source (see Fig. D1). The sources missed by the three source finders (see for example Fig. D2) were predominantly lower mass, where (as shown by Figs 4 and 5) LESHI's completeness is higher. We note that different settings for all of the above may lead to higher completeness, but introduce lower reliability. Our final catalogue combines the 292 sources detected with LESHI and the one additional source found by SOFIA, totalling in 293 sources.

While the COSMOS field has a wealth of ancillary data facilitating targetted source finding, not every galaxy in the field has optical spectroscopic redshift measured. Therefore, we perform untargeted source finding, which also made it possible to detect any H I emission without an optical counterpart, such as free-floating dark HI clouds. Those detections were noted, however, not included in the catalogue.

4.4 H I mass

To measure the total H I flux, we have employed a different approach to the one previously used for MIGHTEE-HI galaxies in A. A. Ponomareva et al. (2021), which involved manually smoothing each cubelet and generating an emission mask at a specified sigma level. With increasing numbers of sources, we are aiming to minimize manual intervention. Instead, to extract the emission volume, we start with plotting a spectrum for the central pixels within the beam size at the coordinates of the detection, as this is the minimum spatial size of the emission for unresolved galaxies. We identify the H I line and measure its width by fitting a busy function (described in Section 4.5), then we create a moment-0 map by integrating all the channels within the measured width of the spectral line, as determined by the busy function fit. Next, we define the 3σ contour on the initial moment-0 map and generate a spectrum from the integrated flux within that contour. We repeat these steps of generating a spectrum, fitting the busy function, creating a moment-0 map, and generating a new spectrum until the contour changes by less than 5 per cent of the contour area and all the emission is captured by the found contour within the measured channel width of the global profile.

Using the resulting contour and line profile width, we create a 3-dimensional source mask which we apply to the data cubelet, masking out all pixels that do not belong to the source. We then measure the total H I emission flux by summing the emission of the masked cube. We subtract the local continuum level found through fitting the busy function and we correct the measured flux for the size of the synthesized beam and channel width resulting in flux values in the units of Jy Hz. To convert from flux density pixel values in the units of Jy beam^{-1} to total measured flux in Jy Hz, we have used the formula:

$$S_{\rm tot} = \sum_{\rm pix}^{N_{\rm pix}} S_{\nu,{\rm pix}} \cdot \delta\nu \cdot \frac{\delta w^2}{\frac{\pi \cdot b_{\rm min} \cdot b_{\rm maj}}{4\ln(2)}} \tag{4}$$

where $S_{\rm tot}$ is the total measured flux in the units of Jy Hz, $N_{\rm pixel}$ is the number of pixels in the detection volume, $S_{\nu,{\rm pix}}$ is the flux density for each pixel in Jy beam^{-1}, $\delta\nu$ is the width of the channels of the data in Hz, δw is the angular size of one pixel (spatial pixel resolution of the data), $b_{\rm maj}$ is beam major axis and $b_{\rm min}$ is beam minor axis, all in consistent units to cancel out.

To estimate the errors on the H I flux measurement, we create a 7 by 7 grid of spatial coordinates surrounding the source location and central frequency, and offset spatially by the source contour diameter. We apply the source 3-dimensional mask to the data at these offset coordinates and frequency range equal to the frequency range of the source and measure the total flux within the masks, acquiring a set of 49 measured total flux values: one for the source in the centre and 48 for the background around the source. We then calculate the sigma-clipped standard deviation of these 49 values and we take one standard deviation as the error of the measured total flux. We also compute an integrated SNR of the sources by taking the ratio of the measured source total flux and the calculated one standard deviation.

To calculate the total HI mass from the measured flux for each galaxy, we use the formula from M. Meyer et al. (2017), assuming optically thin HI gas:

$$\left(\frac{M_{\rm H\,I}}{\rm M_\odot}\right) = 49.7 \left(\frac{D_{\rm L}}{\rm Mpc}\right)^2 \left(\frac{S_{\rm tot}}{\rm Jy\ Hz}\right) \tag{5}$$

where $M_{\rm HI}$ is the total H I mass in units of solar masses, $D_{\rm L}$ is the luminosity distance in Mpc, and S_{tot} is the total observed flux in Jy Hz ($= 10^{-26}$W m^{-2}). We note that D_L depends on the redshift, which in turn greatly depends on assumed peculiar motions of the emission source, especially for local galaxies (R. B. Tully et al. 2008). Here, we make the simplifying assumption that the spectroscopic redshift represents the distance, ignoring effects from bulk flows. We provide the flux measurements in redshift independent units of Jy Hz, if more accurate calculation of the HI mass are desired.

We test this mass-measuring method on injected galaxies and compare the measured HI masses to the true ones. As can be seen

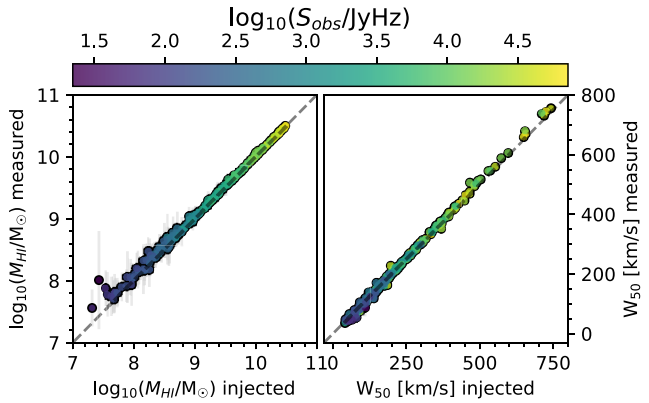

Figure 11. Comparison between H I mass (left panel) and velocity width (right panel) of measured emission (y-axes) and injected emission (x-axes), colour-coded by the observed H I flux.

on the left panel of Fig. 11, our method accurately measures the HI mass within the estimated uncertainties.

4.5 H I profile widths

To characterize the H I emission spectra and measure their velocity widths in an automated manner, we fit a busy function (equation 4 in T. Westmeier et al. 2014) to the profile of the spectrum, as done in A. A. Ponomareva et al. (2021). The function is given by:

$$B(x) = \frac{a}{4} \times (\mathrm{erf}(b_1 \times (w + x - x_e)) + 1) \times (\mathrm{erf}(b_2 \times (w - x + x_e)) + 1) \times (c \times |x - x_p|^2 + 1) + C \quad (6)$$

where erf(z) is the error function defined by $\frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{dt}$ and a, b_1, b_2, x_e, x_p, w, c, C are the free parameters changing the exact form of the function[3]. The function is able to fit well even the most complicated double-horned profiles (see bottom left panel of Fig. 12), but at the cost of many free parameters. To fit for the eight free parameters we use the EMCEE PYTHON package, using 32 walkers, 5000 steps and burn in phase of 500 steps that were discarded and a Gaussian likelihood function. We then measure the velocity width W_{50} at 50 per cent of the average peak flux density of the fitted function, found by identifying the two peaks of the spectral profile and calculating their average value. Since none of the fitted parameters are solely responsible for the width of the spectral profile, we cannot use the output posterior plots to estimate the error on the width. Instead, we measure the widths for each fitted function tried out by the sampler and calculate the error of W_{50} from the 16th and 84th percentiles of the histogram of the measured widths. Some of the output errors are very small, so we combine this error in quadrature with the velocity width of the channels (calculated for the given redshift), equal to ~ 5.5 km s^{-1} for $z = 0$, accounting for the velocity resolution of our data.

We test this method on injected galaxies and compare the measured velocity widths to the true ones. As can be seen on the right panel of Fig. 11, our method estimates the width within the measurement uncertainties.

[3]The busy function can be explored using the Desmos online graphing calculator: https://www.desmos.com/calculator/bbc9ed8be4

4.6 Optical photometry and SED fitting

The galaxies from our sample are at relatively low redshift, meaning they are very well resolved in optical images, complicating the measurement of their photometry. To extract the source fluxes in an automated way, we have used the Detector of astRonomical soUrces in optIcal and raDio images (DRUID) source detection software (R. A. Shaw et al. 2025), which utilizes persistent homology to detect sources and their components. We have utilized DRUID here as it performs very well at associating complex morphology sources while keeping track of substructure, which often characterizes clumpy, star-forming, well-resolved galaxies, like the ones from our sample. We perform the initial photometry on the HSC g-band images, as our galaxies tend to be blue and actively star-forming. To extract the galaxy and reject any background/foreground sources, each detected component was compared against the local flux level and colour of the main galaxy by measuring their brightness and colour from the emission in the g and i band images within the contour of the component. If its brightness differed by more than specified value (a factor of 2 for sources redder than the main galaxy and a factor of 10 for sources bluer than the main galaxy), it was rejected. An example contour can be seen on the upper right panel of Fig. 12 with rejected components marked in red. Every galaxy has been then checked by eye, and the contours modified if needed. The contours were then used to create a mask, which was applied to each of the *g, r, i, z, y, Y, J, H, K* photometric band images and the emission within the mask was summed and background subtraction applied. To convert the raw pixel values from the band images into magnitudes, we have used the following formula:

$$m_{\mathrm{obs}} = -2.5 \cdot \log_{10}(S_{\mathrm{tot}}) + m_{\mathrm{zp}} + m_{AB} \quad (7)$$

where m_{obs} is the observed magnitude in the AB magnitude system, S_{tot} is the total pixel value for the HSC images and the total pixel value divided by the exposure time for the UKIDSS images, m_{zp} is the magnitude zeropoint equal to 27 for the HSC images, and for the UKIDSS images the zeropoint values were taken from the header for each band image, m_{AB} is the magnitude offset to convert magnitudes between the AB and Vega system for the UKIDSS, taken from P. C. Hewett et al. (2006).

To calculate the errors of the measured flux, we have used the following formula:

$$\sigma^2 = \frac{1}{g} \sum_n^{N_A} (S_n - \overline{B}) + N_{\mathrm{A}} \cdot \sigma^2_{\mathrm{B/pix}} \quad (8)$$

where σ is the calculated error, g is the gain of the detector in electrons per pixel data unit, S_n is signal in pixel n in image data units, $\overline{B}$ is estimated median background per pixel, N_{A} is number of pixels in source aperture and $\sigma_{\mathrm{B/pix}}$ is pixel standard deviation of the sky background. Following N. J. Adams et al. (2021) and R. G. Varadaraj et al. (2023) we adopt a minimum flux uncertainty of 5 per cent to mitigate against issues around mismatch between the synthetic galaxy templates and the observations of real galaxies, whilst also accounting for colour-dependent zero-points for the individual filters. To convert the errors to magnitude system, we have used the standard formula:

$$\Delta m_{\mathrm{obs}} = \frac{2.5}{\ln 10} \cdot \frac{\Delta S_{\mathrm{tot}}}{S_{\mathrm{tot}}} \quad (9)$$

where Δm_{obs} is the error in magnitudes and ΔS_{tot} is the measured flux error. For the errors taken to be equal to 5 per cent of the

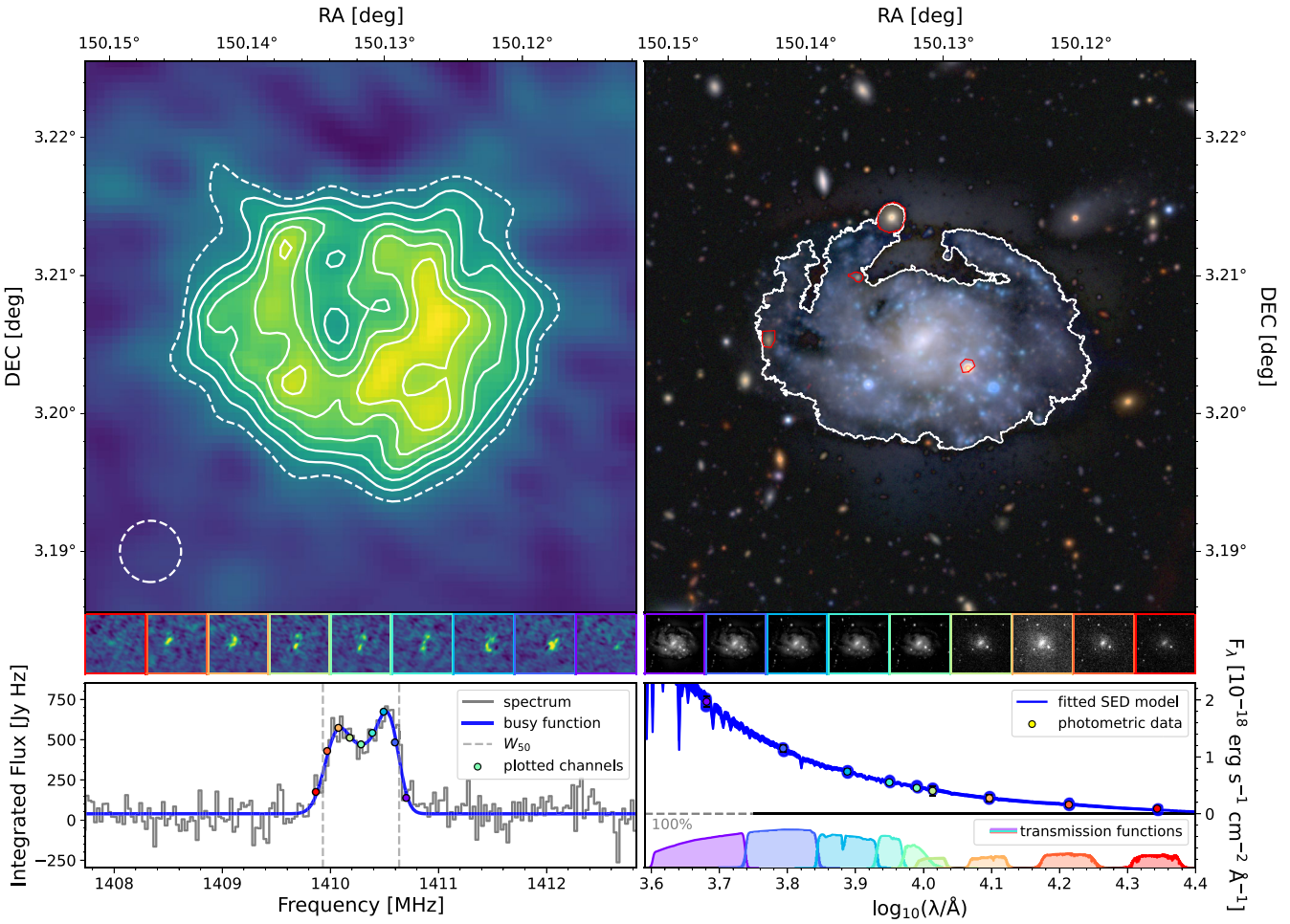

Figure 12. Top left: H I moment-0 map with contours over-plotted on top for an example galaxy from our sample, marking column densities of 3.0, 4.3, 5.7, 7.0, 8.4, 9.7 $\times$ 10^{20} cm^{-2} with the size of the synthesized beam marked by a dashed circle. The row of 9 channel maps at the bottom show the H I emission for the individual channels, which are marked by points on the spectrum below with their colour matching the border of the channel map. Bottom left: Observed H I emission spectrum (marked in grey) with fitted busy function (marked in blue) with the measured velocity width W_{50} marked by vertical dashed lines. The spectrum is for the integrated flux within the first contour on moment-0 map (marked by dashed line) tracing 3σ emission. Top right: Hyper Supreme Cam *gri* composite optical image for the example galaxy from our sample with *g*-band flux contours from DRUID (R. A. Shaw et al. 2025) in white, and contours surrounding rejected components marked in red. The row of 9 images at the bottom show the emission for each of the *g, r, i, z, y, Y, J, H, K* photometric bands. Bottom right: multiwavelength flux measurements for photometric bands (coloured points) with the best-fitting SED model from BAGPIPES (A. C. Carnall et al. 2018) marked in blue, and the transmission functions for each of the *g, r, i, z, y, Y, J, H, K* photometric bands plotted below in colour.

measured flux, their value in magnitude system is equal to the constant value of 0.054.

To obtain estimates of the stellar masses and star formation rates, we employ the Bayesian Analysis of Galaxies for Physical Inference and Parameter Estimation code (BAGPIPES; A. C. Carnall et al. 2018) to perform spectral energy distribution (SED) fitting based on our measured multiwavelength fluxes. The code uses a Chabrier Initial Mass Function (G. Chabrier 2003), coupled with Stellar Population Synthesis models based on G. Bruzual & S. Charlot (2003) with several dust extinction and star formation history models. In this work, we use the delayed star formation history model along with the dust attenuation model using the Calzetti law (D. Calzetti et al. 2000), by allowing the extinction to vary between A_V = 0-2. During the fitting, we keep the redshift fixed to that measured from the H I emission. The bottom right panel of Fig. 12 shows an SED fit for an example galaxy from our sample. The results of the fitting can be found in Table 3, with the uncertainties based on the posterior distributions output by BAGPIPES calculated from the 16th and 84th percentiles.

The inclination was estimated by measuring the axis ratio of isophotes fitted with the PHOTUTILS Python package (L. Bradley et al. 2025) to the emission in the *g* band and derived using the standard relation:

$$\cos(i)^2 = \frac{\left(\frac{b}{a}\right)^2 - q_0^2}{1 - q_0^2} \tag{10}$$

where $\frac{b}{a}$ is the axis ratio of the fitted outermost isophote and q_0 is the intrinsic axis ratio of the disc (with typical values between 0 and 0.4; P. Fouque et al. 1990), which we assumed to be equal to 0.2. Each isophote ellipse was checked by eye and corrected where needed.

Table 3. Table containing first 42 columns from the catalogue for a sample of eight galaxies; five omitted columns contain flags with values of either 0 or 1. See the main text for more details. The full table is available as supplementary material for this paper.

ID catalogue	RA (deg)	Dec. (deg)	Freq. (MHz)	z_{HI}	z_{spec}	D_L (Mpc)	S_{HI} (Jy Hz)	δS_{HI} (Jy Hz)	$\log_{10}(M_{HI})$ ($M_\odot$)	$\delta\log_{10}(M_{HI})$ ($M_\odot$)
MGTH_J100256.4+023440	150.735	2.578	1360.267	0.0442	0.0442	195.6	1471.09	109.80	9.45	0.03
MGTH_J095642.7+022509	149.178	2.419	1306.161	0.0875	0.0874	399.1	202.78	35.36	9.21	0.08
MGTH_J095951.4+014224	149.964	1.707	1385.660	0.0251	0.0249	109.3	1074.15	68.53	8.81	0.03
MGTH_J100006.8+022245	150.028	2.379	1403.974	0.0117	0.0117	50.6	193.29	16.14	7.39	0.04
MGTH_J095741.1+020149	149.422	2.031	1376.726	0.0317	0.0317	138.7	132.05	22.94	8.10	0.08
MGTH_J100156.7+030737	150.486	3.127	1321.157	0.0751	0.0751	339.7	596.37	125.30	9.54	0.09
MGTH_J095914.8+021131	149.812	2.192	1386.287	0.0246	0.0245	107.4	703.72	70.68	8.61	0.04
MGTH_J100357.1+022505	150.988	2.418	1383.623	0.0266	0.0266	116.7	12782.92	672.47	9.93	0.02

SNR_{3D}	W_{50} (km s^{-1})	δW_{50} (km s^{-1})	incl. (deg)	Axis ratio	$\log_{10}(M_{stel})$ ($M_\odot$)	$\delta\log_{10}(M_{stel})$ ($M_\odot$)	SFR ($M_\odot$ yr^{-1})	δSFR ($M_\odot$ yr^{-1})	m_g (mag)	δm_g (mag)	m_r (mag)	δm_r (mag)
13.4	167	7	44	0.73	9.11	0.07	0.24	0.12	18.37	0.05	17.96	0.05
5.9	191	20	41	0.77	9.59	0.10	1.65	0.79	18.18	0.05	17.91	0.05
16.0	101	7	61	0.51	7.54	0.12	0.07	0.03	19.22	0.05	19.14	0.05
12.0	40	6	49	0.67	6.75	0.11	0.00	0.00	20.32	0.05	20.07	0.05
5.4	100	16	57	0.57	8.48	0.08	0.05	0.03	19.25	0.05	18.85	0.05
4.4	280	27	59	0.54	10.10	0.10	3.20	1.78	18.03	0.05	17.41	0.05
10.3	103	8	71	0.37	7.54	0.11	0.02	0.01	20.32	0.05	20.02	0.05
19.1	392	12	75	0.32	10.33	0.10	4.44	2.64	15.07	0.05	14.47	0.05

m_i (mag)	δm_i (mag)	m_z (mag)	δm_z (mag)	m_y (mag)	δm_y (mag)	m_Y (mag)	δm_Y (mag)	m_J (mag)	δm_J (mag)	m_H (mag)	δm_H (mag)	m_K (mag)	δm_K (mag)
17.75	0.05	17.64	0.05	17.50	0.05	17.98	0.09	18.35	0.18	17.40	0.10	17.84	0.19
17.66	0.05	17.61	0.05	17.46	0.05	17.73	0.05	17.54	0.06	17.49	0.07	17.58	0.09
19.14	0.05	19.09	0.05	19.00	0.05	18.81	0.17	18.67	0.17	19.45	0.35	26.93	99
20.03	0.05	19.95	0.05	19.90	0.05	21.73	1.62	19.95	0.41	21.36	1.95	20.38	0.95
18.64	0.05	18.52	0.05	18.40	0.05	18.77	0.16	19.48	0.38	18.59	0.16	19.14	0.39
17.02	0.05	16.82	0.05	16.57	0.05	16.74	0.05	16.44	0.05	16.35	0.05	16.23	0.05
19.90	0.05	19.80	0.05	19.71	0.05	20.04	0.36	19.56	0.31	20.49	0.93	19.41	0.41
14.17	0.05	13.93	0.05	13.70	0.05	13.75	0.05	13.64	0.05	13.40	0.05	13.54	0.05

4.7 Multiwavelength HI source catalogue

The H I, optical and NIR photometry, as well as derived parameters for the full catalogue can be found online as explained in Section 6. Example measurements for a sample of 10 galaxies can be found in Table 3. The table's keywords and description are:

(i) Column 1 – ID_catalogue – a unique identifier computed based on sky coordinates of each galaxy.

(ii) Column 2 – RA_deg – right ascension (J2000) of the optical counterpart of the HI detection, in degrees.

(iii) Column 3 – Dec_deg – declination (J2000) of the optical counterpart of the HI detection, in degrees.

(iv) Column 4 – freq_MHz – frequency centred on the H I detection, in MHz.

(v) Column 5 – z_HI – redshift calculated using the formula:

$$z = \frac{\nu_{\rm rest} - \nu_{\rm obs}}{\nu_{\rm obs}} \tag{11}$$

where $\nu_{\rm obs}$ is the measured frequency of the HI detection and ν_{rest} is equal to 1420.40575 MHz.

(vi) Column 6 – z_spec – spectroscopic redshift of galaxies crossmatched from the DESI catalogue, values of -99 mark the galaxies not found in the DESI catalogue.

(vii) Column 7 – D_L_Mpc – luminosity distance calculated assuming ΛCDM cosmology with $H_0 = 70$ km s^{-1}Mpc^{-1}, $\Omega_{\rm M} = 0.3$ and $\Omega_\Lambda = 0.7$ and ignoring any possible cosmic flows, in Mpc.

(viii) Column 8 (9) – S_HI_Jy_Hz (S_HI_Jy_Hz_err) – total observed flux (and its error) of the detected H I emission, in Jy Hz. For full description of the measurement method see Section 4.4.

(ix) Column 10 (11) – log_M_HI (log_M_HI_err) – base 10 logarithm of the measured total H I mass (and its error) in the units of solar masses, calculated from the total HI flux using equation (5).

(x) Column 12 – SNR_3D – integrated SNR of the H I detection, calculated as described in Section 4.4.

(xi) Column 13 (14) – W_50_km_s (W_50_km_s_err) – H I detection spectral profile velocity width (and its error) measured at 50 per cent of the flux peak, in km s^{-1}. Note that it is not inclination corrected. For full description of the measurement method see Section 4.5.

(xii) Column 15 – incl_deg – inclination of the optical disk, in degrees, with edge-on galaxy having an inclination of 90° and face-on 0°, measured as described in Section 4.6.

(xiii) Column 16 – axis_ratio – ratio between the minor and major axis of the optical disc, measured as described in Section 4.6.

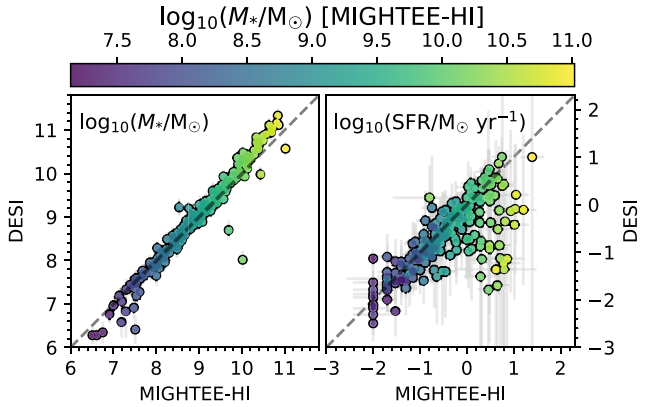

Figure 13. Comparison between stellar mass (left panel) and star-formation rate (right panel), determined through SED fitting in this work (*x*-axes) and DESI (*y*-axes) for crossmatched galaxies present in both catalogues.

(xiv) Column 17 (18) – log_M_stel (log_M_stel_err) – base 10 logarithm of the total stellar mass (and its error) in the units of solar masses, derived through SED fitting, as described in Section 4.6.

(xv) Column 19 (20) – sfr_M_sol_year (sfr_M_sol_year_err) – star formation rate (and its error) in the units of solar masses per year, derived through SED fitting, as described in Section 4.6.

(xvi) Columns 21–38 – m_g – m_K (m_g_err—m_K_err) – observed magnitudes (and their errors) of the flux measured for the *g, r, i, z, y, Y, J, H, K* optical and infrared bands in the AB magnitude system, as described in Section 4.6.

(xvii) Column 39 – low_confidence_flag – flag marking low-confidence detections. Sources with flag 0 have high signal-to-noise and there is little doubt that the detection is genuine. Detections with flag 1, while having an optical counterpart, were judged to be unreliable due to their spatial distribution and spectral profile. An example is given in Fig. D3. This flag does not disqualify the detection, but reflects lower confidence in it and its measurements.

(xviii) Column 40 – blended_flag – flag marking blended sources (flag 1) for which the galaxies have small separation and their HI contents can not be separated and resolved. Sources with flag 0 are isolated sources. An example is given in Fig. D4.

(xix) Column 41 – confused_flag – flag marking potentially confused sources (flag 1) for which the galaxies have small separation (similar to the blended sources), however the morphology of the HI distribution does not strongly suggest a blended source. An example is given in Fig. D5.

(xx) Column 42 – bad_ellipse_flag – flag marking irregularly shaped sources (flag 1) in the optical for which a confident axis ratio (and therefore inclination) could not be determined.

(xxi) Column 43 – contaminated_source_flag – flag marking galaxies for which the optical photometry is contaminated by foreground/background sources.

To retrieve the 'golden sample', the sum of all flags should add up to zero.

In Fig. 13, we compare the stellar population properties inferred through SED fitting in this work and in DESI (H. Zou et al. 2024, DESI Collaboration et al. 2026). As can be seen in the two panels, stellar masses and star-formation rates follow the one-to-one relation very closely with the exception of the most massive systems for which DESI masses are slightly higher and star-formation rates lower. This might be attributed to the different measurement methods and photometric bands used. For example, stellar population properties in H. Zou et al. (2024) were inferred without the use of near-infrared data and utilizing spectroscopy. A similar discrepancy was observed in fig. 6 in H. Zou et al. (2024), where the DESI stellar masses were on average larger across all masses than the ones inferred in the Cosmic Evolution Survey (COSMOS; J. R. Weaver et al. 2022).

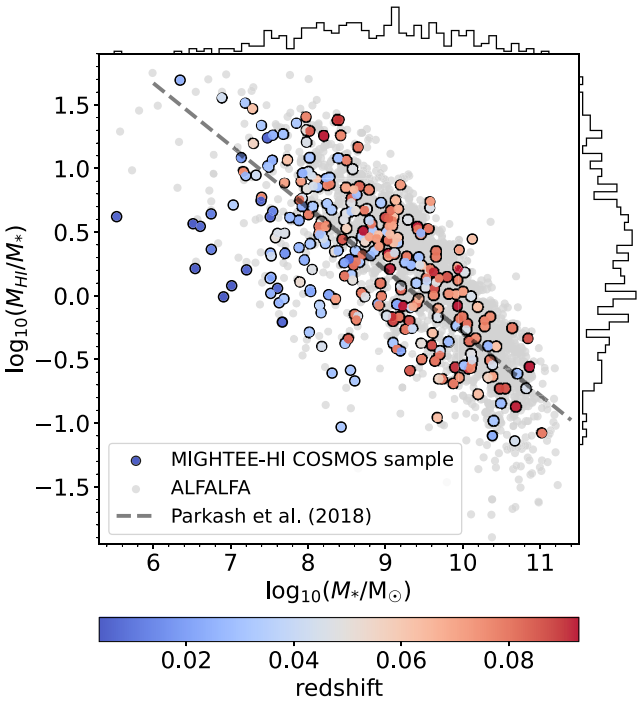

Figure 14. The gas fraction (ratio of H I mass to stellar mass) as a function of the stellar mass, colour-coded by redshift for galaxies from our catalogue and sample distribution histograms plotted on top and on the right of the figure. Galaxies from the ALFALFA survey are plotted in grey. Dashed line marks the relation adapted from equation (8) in V. Parkash et al. (2018).

Fig. 14 shows the gas fraction (ratio of H I mass to stellar mass), as a function of the stellar mass for the galaxies from our catalogue, colour-coded by their redshift. Our sample agrees well with the relation between HI and stellar masses from Eq. 8 in V. Parkash et al. (2018), derived for HI selected sample of galaxies, and the relation from A. B. Romeo (2020), suggesting that more massive galaxies have lower H I gas fractions. In the figure, we can see that higher redshift galaxies tend to be above the relation, since at those redshifts for a given stellar mass, we will detect the most H I-rich galaxies.

We compare our derived HI masses to the ones from public catalogue from A. A. Ponomareva et al. (2023) for galaxies present in both works (see Fig. E1). The measurements agree well with each other and we note that any differences most likely arise due to the MIGHTEE data used in this work having higher spectral resolution and sensitivity than the Early Science data.

5 DISCUSSION

Combining the results from the two main parts of this work, we use equation (1) for the completeness $c(M_{\rm HI}, D_{\rm L})$ to plot a

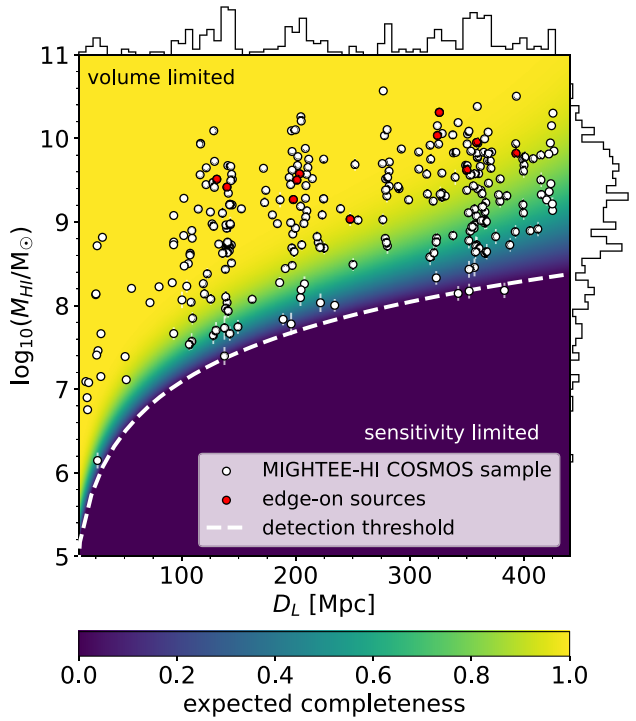

Figure 15. H I mass of the catalogue galaxies plotted against their luminosity distance represented by white points, with most edge-on sources plotted in red and sample distribution histograms plotted on top and on the right of the figure. The theoretical completeness map calculated using equation (1) is plotted as the background in colour with the detection threshold (where the expected completeness drops to zero given by equation 2) marked by the white dashed line.

completeness map by colour-coding every point in the $M_{\rm HI}$-D_L parameter space. We then overplot the real detections from our galaxy sample and see how the results from the two parts of this work compare. As can be seen on Fig. 15, at all distances and masses, our detections lie above the expected detection threshold, where the derived expected completeness drops to zero. Moreover, most edge-on detections [$\cos(i) < 0.1$] marked in red, all lie much above the detection threshold, consistent with our findings that high inclination sources are more difficult to detect.

Computing the completeness through source finding on injected sources is critical in deriving the H I mass function by correcting the sample densities based on their estimated completeness. For this purpose, LESHI is ideal as its relatively low runtime and memory footprint allow for many iterations of source finding on simulated sources to reach high number statistics and thus accurate completeness estimates.

However, one of the caveats of our artificial source injection approach is that the galaxies modelled using ^{3D}BAROLO are idealized; they have symmetric spectra and do not posses any irregularities characteristic of real sources. However, we do not expect it to influence the results by much, as the source finders rely on the emission being above the noise level, rather than it being symmetric or not.

An important thing to note is that source finding done on artificial sources is expected to be more complete and deeper compared to source finding on real data. For injected sources, we know exactly where they are and even very low SNR sources will be accepted. For real systems, we need their emission to have large enough SNR to accept them as a genuine. However, this issue is somewhat diminished in our work by the availability of optical data, since even very low-SNR sources might be accepted if they have an optical counterpart with consistent spectroscopic redshift.

6 SUMMARY AND CONCLUSIONS

In this work, we have introduced a catalogue of 293 H I sources found in the MIGHTEE survey data cubes, through untargeted source finding. The sources lie in the COSMOS field in the redshift range of $0.004 < z < 0.093$. This area has a wealth of ancillary data available, making it possible to crossmatch each of our H I detections to optical counterparts with measured spectral redshifts, resulting in high expected reliability of our sample. In addition to H I masses and velocity widths of the H I emission lines, the catalogue includes optical through near-infrared photometry and stellar population properties, inferred through SED fitting. As the contents of H I catalogue acquired through untargeted source finding greatly depend on the source finding methods used, this study also provides a well-characterized expected completeness of the detected sample of galaxies based on their properties, inferred through a comparative study of different source finding algorithms.

In the first part of this work, we have compared the SOFIA, PROFOUND, PYBDSF, and LESHI source finders, by injecting a sample of simulated systems into the MIGHTEE data cube. We note that the study's aim was not to compare the capabilities of optimized and maximally fine-tuned source finders, but rather their capabilities our collaboration would be able to realistically achieve, by aiming to find the best balance between the reliability and completeness within reasonable amount of time, for the purposes of catalogue creation. We have found that for the set-up settings used, SOFIA had the lowest number of false positives, proving its reliability and did not over-fragment the sources, while LESHI proved to be the most complete (at a cost of slightly worsened reliability) and fastest. PYBDSF and PROFOUND perform very well even though they were designed for two-dimensional data, with their completeness and reliability being comparable to the completeness of LESHI and SOFIA. LESHI has very low memory footprint equal to around 20 per cent of the size of the data cube on average (with jumps of up to 50 per cent), which can be lowered further depending on the input number of integrated images to be analysed at once, at the expense of a longer runtime. Similarly, since PYBDSF and PROFOUND work on separate channel images in parallel, their memory footprint can also be adjusted. They can be therefore run on very large data sets without the need of dividing them into subcubes. With this in mind, a possible automated source finding strategy could include a first round performed with one of the fast, low memory footprint, and very complete source finders, followed by targeted source finding with SOFIA to properly characterize the sources. It is important to note that the way each source finder performs depends greatly on the setup settings used. Each could have been made more complete at the expense of reliability and vice versa.

We have also investigated how the completeness of the source finders depends on the inclination, mass and distance of simulated galaxies. We have observed that the inclination has an impact on the completeness, especially for higher distances, where high mass galaxies are at the detection threshold, making it more challenging to find edge-on sources. This is likely due to the fact that higher-mass galaxies rotate faster, and the spectral line

broadening for high inclinations (as explained in Section 2.2.1), is more severe than for lower masses. We then averaged the completeness over the investigated inclinations and quantified the relation between completeness and mass and distance with a functional form (Fig. 6 and equation 1). We have shown that the completeness changes rapidly with mass, changing from 0 per cent to 100 per cent in the span of 0.5 dex for lower distances and 1 dex for higher distances. The slower rate of change of the completeness function for higher distances can be attributed to the inclination effect on completeness for higher mass galaxies. The mass where the completeness drops to zero is proportional to the logarithm of the distance squared, consistent with the standard relation between the observed flux and luminosity distance. The constant of proportionality for this relation (a_2 in equation 2) should depend on the sensitivity of the data and equation (1) could be potentially applied to other surveys, after changing that constant accordingly. We have also learnt to be cautious of face-on, high-mass and nearby galaxies, which may be missed by the source finders due to improper estimation of the background level.

As the era of the SKAO comes closer, the lessons learned can be used in devising source finding strategies for the anticipated amounts of new data. Moreover, the relationships determined between completeness and galaxy properties for the MIGHTEE data can be now used in statistical studies, such as determining the HI mass function.

ACKNOWLEDGEMENTS

We thank the reviewer for prompt, detailed and insightful comments, adding considerably to the quality of the paper. We thank Prof. Marc Verheijen for sharing his experience and advice, which greatly contributed to the creation of this paper. This work is based on observations made by the MeerKAT telescope, which is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. We acknowledge use of the InterUniversity Institute for Data Intensive Astronomy (IDIA) data intensive research cloud for data processing. IDIA is a South African university partnership involving the University of Cape Town, the University of Pretoria and the University of the Western Cape. The study described in the paper made use of ASTROPY (Astropy Collaboration 2022), Cube Analysis and Rendering Tool for Astronomy (CARTA; A. Comrie et al. 2021, TOPCAT (M. B. Taylor 2005), NUMPY (S. Van Der Walt, S. C. Colbert & G. Varoquaux 2011), SCIPY (P. Virtanen et al. 2020), MATPLOTLIB (J. D. Hunter 2007) and NASA's Astrophysics Data System.

TGH acknowledges support from UNIQ + scholarship. MB gratefully acknowledges the financial support from the Flemish Fund for Scientific Research (FWO-Vlaanderen) and the South African National Research Foundation (NRF) under Bilateral Scientific Cooperation programme (grant G0G0420N). He also acknowledge the support of networking activities by NRF and the Belgian Science Policy Office (BELSPO) under grant BL/02/SA12 (GALSIMAS). SLJ acknowledges the support of a UK Research and Innovation (UKRI) Frontiers Research Grant [EP/X026639/1], which was selected by the European Research Council, and the Science and Technology Facilities Council (STFC) consolidated grants [ST/S000488/1] and [ST/W000903/1]. MG is supported through UK STFC Grant ST/Y001117/1. MG acknowledges support from the Inter-University Institute for Data Intensive Astronomy (IDIA). IDIA is a partnership of the University of Cape Town, the University of Pretoria and the University of the Western Cape. CLH acknowledges support from the Science and Technology Facilities Council (STFC) through grant ST/Y000951/1. CLH and MJJ acknowledge support from the Oxford Hintze Centre for Astrophysical Surveys which is funded through generous support from the Hintze Family Charitable Foundation. MJJ, IH, HP, SK, AAV, and SLJ acknowledge the support of a UKRI Frontiers Research Grant [EP/X026639/1], which was selected by the European Research Council. MJJ, IH, AAP, and SLJ acknowledge support from the STFC consolidated grant [ST/W000903/1].

DATA AVAILABILITY

The catalogue is available online as the supplementary material for this paper.

The MIGHTEE–H I spectral cubes are available from https://doi.org/10.48479/jkc0-g916 (I. Heywood et al. 2024). The optical and near-infrared data images used in this work are all in the public domain (please see the references cited in the main text). Other data underlying the article are available on request to the first author.

REFERENCES

Adams E. A. K. et al., 2022, A&A, 667, A38
Adams N. J., Bowler R. A. A., Jarvis M. J., Häußler B., Lagos C. D. P., 2021, MNRAS, 506, 4933
Aihara H. et al., 2019, PASJ, 71, 114
Alam S. et al., 2015, ApJS, 219, 12
Astropy Collaboration, 2022, ApJ, 935, 167
Barkai J. A., Verheijen M. A. W., Talavera E., Wilkinson M. H. F., 2023, A&A, 670, A55
Barnes D. G. et al., 2001, MNRAS, 322, 486
Blyth S. et al., 2016, Proc. Sci., LADUMA: Looking at the Distant Universe with the MeerKAT Array. SISSA, Trieste, PoS(MeerKAT2016)004
Bradley L. et al., 2025, astropy/photutils: 2.2.0, Zenodo. Available at: https://doi.org/10.5281/zenodo.14889440 (April 18)
Braun R., Bourke T., Green J. A., Keane E., Wagg J., 2015, Proc. Sci., Advancing Astrophysics with the Square Kilometre Array. SISSA, Trieste, PoS(AASKA14)174
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, 682
Carnall A. C., McLure R. J., Dunlop J. S., Davé R., 2018, MNRAS, 480, 4379
Chabrier G., 2003, PASP, 115, 763
Comrie A. et al., 2021, CARTA: The Cube Analysis and Rendering Tool for Astronomy. Zenodo. Available at: https://doi.org/10.5281/zenodo.4905459 (accessed February 3).
DESI Collaboration, 2026, AJ, 171, 285
Davies L. J. M. et al., 2018, MNRAS, 480, 768
Davies L. J. M. et al., 2025, MNRAS, 544, 3005
Di Teodoro E. M., Fraternali F., 2015, MNRAS, 451, 3021
Fernandez X. et al., 2016, ApJL, 824, L1
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Fouque P., Bottinelli L., Gouguenheim L., Paturel G., 1990, ApJ, 349, 1
Giovanelli R. et al., 2005, AJ, 130, 2598
Hale C. L., Robotham A. S. G., Davies L. J. M., Jarvis M. J., Driver S. P., Heywood I., 2019, MNRAS, 487, 3971
Hale C. L. et al., 2025, MNRAS, 536, 2187
Hales C. A., Murphy T., Curran J. R., Middelberg E., Gaensler B. M., Norris R. P., 2012, MNRAS, 425, 979

Hallinan G. et al., 2019, BAAS, 51, 255
Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, Astrophysics Source Code Library, record ascl:1212.009
Hartley P. et al., 2023, MNRAS, 523, 1967
Hewett P. C., Warren S. J., Leggett S. K., Hodgkin S. T., 2006, MNRAS, 367, 454
Heywood I. et al., 2022, MNRAS, 509, 2150
lwHeywood I. et al., 2024, MNRAS, 534, 76
Hibbard J. E., van der Hulst J. M., Barnes J. E., Rich R. M., 2001, AJ, 122, 2969
Hogbom J. A., Brouw W. N., 1974, A&A, 33, 289
Holwerda B. W., Blyth S. L., Baker A. J., 2012, in Tuffs R. J., Popescu C. C., eds, Proc. IAU Symp., Vol. 284, The Spectral Energy Distribution of Galaxies—SED 2011. Preston, UK, p. 496
Hotan A. W. et al., 2021, PASA, 38, e009
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jaffé Y. L., Poggianti B. M., Verheijen M. A. W., Deshev B. Z., van Gorkom J. H., 2013, MNRAS, 431, 2111
Jarvis M. et al., 2016, Proc. Sci., The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey. SISSA, Trieste, PoS(MeerKAT2016)006
Jonas J., MeerKAT Team, 2016, Proc. Sci., The MeerKAT Radio Telescope. SISSA, Trieste, PoS(MeerKAT2016)001
Jonas J. L., 2009, IEEE Proceedings, 97, 1522
Joye W., 2019, SAOImageDS9/SAOImageDS9 v8.0.1 (v8.0.1). Zenodo. Available at: https://doi.org/10.5281/zenodo.2530958
Koribalski B. S. et al., 2020, Ap&SS, 365, 118
Lacy M. et al., 2020, PASP, 132, 035001
Lawrence A. et al., 2007, MNRAS, 379, 1599
Li D. et al., 2018, IEEE Microwave Magazine, 19, 112
Maddox N. et al., 2021, A&A, 646, A35
McCracken H. J. et al., 2012, A&A, 544, A156
Meyer M., Robotham A., Obreschkow D., Westmeier T., Duffy A. R., Staveley-Smith L., 2017, PASA, 34, 52
Miyazaki S. et al., 2018, PASJ, 70, S1
Mohan N., Rafferty D., 2015, Astrophysics Source Code Library, record ascl:1502.007
Nan R. et al., 2011, Int. J. Mod. Phys. D, 20, 989
Parkash V., Brown M. J. I., Jarrett T. H., Bonne N. J., 2018, ApJ, 864, 40
Ponomareva A. A. et al., 2021, MNRAS, 508, 1195
Ponomareva A. A. et al., 2023, MNRAS, 522, 5308
Rajohnson S. H. A. et al., 2022, MNRAS, 512, 2697
Ranchod S. et al., 2021, MNRAS, 506, 2753
Rhee J. et al., 2023, MNRAS, 518, 4646
Riggi S. et al., 2019, PASA, 36, e037
Robotham A. S. G., Davies L. J. M., Driver S. P., Koushan S., Taranu D. S., Casura S., Liske J., 2018, MNRAS, 476, 3137
Romeo A. B., 2020, MNRAS, 491, 4843
Serra P. et al., 2015, MNRAS, 448, 1922
Shaw R. A., Fotopoulou S., Birkinshaw M., Maddox N., Stewart H., 2025, RAS Tech. Instrum., 4, rzaf006
Taylor A. R. et al., 2024, MNRAS, 528, 2511
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Taylor M. B., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, ASP Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV. Astron. Soc. Pac., San Francisco, p. 666
Taylor R., 2025, A&A, 696, A113
Tolley E., Korber D., Galan A., Peel A., Sargent M. T., Kneib J.-P., Courbin F., Starck J.-L., 2022, Astron. Comput., 41, 100631
Tully R. B., Fisher J. R., 1977, A&A, 54, 661
Tully R. B., Shaya E. J., Karachentsev I. D., Courtois H. M., Kocevski D. D., Rizzi L., Peel A., 2008, ApJ, 676, 184
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Vanderlinde K. et al., 2019, The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD). Zenodo. Available at: https://doi.org/10.5281/zenodo.3765414
Varadaraj R. G., Bowler R. A. A., Jarvis M. J., Adams N. J., Häußler B., 2023, MNRAS, 524, 4586
Virtanen P. et al., 2020, scipy/scipy: SciPy 1.5.3. Zenodo
Wang J., Koribalski B. S., Serra P., van der Hulst T., Roychowdhury S., Kamphuis P., Chengalur J. N., 2016, MNRAS, 460, 2143
Weaver J. R. et al., 2022, ApJS, 258, 11
Westmeier T., Jurek R., Obreschkow D., Koribalski B. S., Staveley-Smith L., 2014, MNRAS, 438, 1176
Westmeier T. et al., 2021, MNRAS, 506, 3962
Whiting M. T., 2012, MNRAS, 421, 3242
Wong O. I. et al., 2006, MNRAS, 371, 1855
Zhang C.-P., Cheng C., Zhu M., Xu J.-L., Jiang P., 2024, ApJ, 971, 131
Zou H. et al., 2024, ApJ, 961, 173

SUPPLEMENTARY MATERIAL

Supplementary data are available at *MNRAS* online.

MIGHTEE_HI_COSMOS_catalogue.csv.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: SOURCE FINDERS INPUT PARAMETERS

In this section, we state the input parameters used for each source finder along with the justification for the chosen values of parameters that have an impact on the source finding. We have attempted to find a set of input parameters giving good results for the MIGHTEE data and the best balance between the completeness and reliability.

A1 SOFIA

We have run the SOFIA source finder with the following input parameters:

```
# Global settings
pipeline.verbose = false
pipeline.pedantic = false
pipeline.threads = 8
# Input
input.data = data_cube
input.gain =
input.noise =
input.weights =
input.mask =
input.invert = false
# Flagging
flag.region =
flag.catalog =
flag.radius = 5
flag.auto = false
flag.threshold = 5.0
flag.log = false
# Continuum subtraction
contsub.enable = true
```

```
contsub.order = 0
contsub.threshold = 2.0
contsub.shift = 4
contsub.padding = 3
# Noise scaling
scaleNoise.enable = true
scaleNoise.mode = local
scaleNoise.statistic = mad
scaleNoise.fluxRange = negative
scaleNoise.windowXY = 51
scaleNoise.windowZ = 9999
scaleNoise.gridXY = 0
scaleNoise.gridZ = 0
scaleNoise.interpolate = true
scaleNoise.scfind = false
# Ripple filter
rippleFilter.enable = false
rippleFilter.statistic = median
rippleFilter.windowXY = 21
rippleFilter.windowZ = 21
rippleFilter.gridXY = 0
rippleFilter.gridZ = 0
rippleFilter.interpolate = false
# S + C finder
scfind.enable = true
scfind.kernelsXY = 0, 8, 16
scfind.kernelsZ = 0, 3, 5, 7, 15, 31
scfind.threshold = 3.8
scfind.replacement = 2.0
scfind.fluxRange = negative
scfind.statistic = mad
# Threshold finder
threshold.enable = false
threshold.threshold = 1e-30
threshold.mode = absolute
threshold.statistic = mad
threshold.fluxRange = negative
# Linker
linker.enable = true
linker.radiusXY = 2
linker.radiusZ = 2
linker.minSizeXY = 8
linker.minSizeZ = 6
linker.maxSizeXY = 0
linker.maxSizeZ = 0
linker.minPixels = 0
linker.maxPixels = 0
linker.minFill = 0.0
linker.maxFill = 0.0
linker.positivity = false
linker.keepNegative = false
# Reliability
reliability.enable = true
reliability.parameters = peak, sum, mean
reliability.threshold = 0.8
reliability.scaleKernel = 0.2
reliability.minSNR = 3.0
reliability.minPixels = 0
reliability.autoKernel = true
reliability.iterations = 30
reliability.tolerance = 0.05
reliability.catalog =
reliability.plot = true
reliability.debug = false
# Mask dilation
dilation.enable = false
dilation.iterationsXY = 10
dilation.iterationsZ = 5
dilation.threshold = 0.001
# Parameterisation
parameter.enable = true
parameter.wcs = true
parameter.physical = true
parameter.prefix = SoFiA
parameter.offset = true
# Output
output.directory = output_dir
output.filename = output_file
output.writeCatASCII = true
output.writeCatXML = true
output.writeCatSQL = false
output.writeNoise = true
output.writeFiltered = true
output.writeMask = true
output.writeMask2d = true
output.writeRawMask = false
output.writeMoments = true
output.writeCubelets = true
output.marginCubelets = 10
output.thresholdMom12 = 0.0
output.overwrite = true
```

The input parameters of the SOFIA source finder with an impact on the completeness and reliability are:

(i) scfind.threshold – sets the flux threshold used by the smooth + clip source finder in the units of background rms relative to the noise level. Values in the range of 3–5 are recommended in SOFIA'S manual. We have set this parameter to 3.8, as we wanted to be sensitive towards weaker sources and which leads to a good compromise between the completeness and reliability.

(ii) scfind.kernelsXY – list of spatial Gaussian kernel sizes used to smooth the data in the spatial domain. We have set it to "0, 8, 16" since the synthesized beam's diameter of our data is equal to 8 pixels.

(iii) reliability.enable – if set to true then the reliability calculation and filtering will be performed, which determines the reliability of each detection and discards any that are below the specified reliability threshold. We have set it to true following the recommendation from SOFIA'S manual, as the value of the scfind.threshold is in the lower range, which leads to good achieved reliability. We have run the source finder with reliability.enable set to false (without changing any other parameters), which led to $\sim$ 7000 false positives.

(iv) reliability.threshold – sets the reliability threshold sources below which are discarded. The default value is 0.9; we have set it to 0.8 to be more tolerant to weaker sources.

A2 ProFound

We have run the ProFound source finder with the following input parameters:

```
out_pro = profoundProFound(
channel_image,
```

```
  plot = FALSE,
  skycut = 3.0,
  pixcut = 16,
  tolerance = 1,
  ext = 1,
  box = c(100,100),
  rotstats = TRUE,
  boundstats = TRUE,
  nearstats = TRUE,
  groupstats = TRUE,
  verbose = FALSE)
```

The input parameters of the PROFOUND source finder with an impact on the completeness and reliability are:

(i) skycut – sets the threshold for the object in number of sky rms above the mean. We have tried values in the range of 2–4 and found that 3 leads to a good compromise between the completeness and reliability,

(ii) pixcut – sets the minimum number of pixels of an object. We have set it to 16, as this is approximately the area of the synthesized beam of our data and we would expect the emission size to be no smaller than the beamsize.

(iii) tolerance – sets the minimum height of the object in number of sky rms above the mean between its highest point and the point where it contacts another object and is responsible for combining different components. Recommended range is 1–5, we have set to 1 to avoid associating very close separate sources.

(iv) box – size of the box in pixels used to estimate the background noise. We have set it c(100,100) to not over-smooth any spatial noise variations. Size of 100 pixels is also much larger than most of the sources expected to be found.

A3 PYBDSF

We have run the PYBDSF source finder with the following input parameters:

```
  out_pyb = bdsf.process_image(
  channel_image,
  thresh_isl = 2.5,
  thresh_pix = 3.0,
  adaptive_rms_box = False,
  advanced_opts = True,
  group_by_isl = True,
  group_method = 'intensity',
  group_tol = 1.0,
  ini_gausfit = 'nobeam',
  minpix_isl = 16,
  peak_fit = True,
  rms_value = None,
  split_isl = False,
  atrous_do = False,
  flagging_opts = True,
  flag_minsize_bm = 1,
  flag_minsnr = 0.7,
  flag_smallsrc = True,
  beam = (hdr_im['BMAJ'],hdr_im['BMIN']
,hdr_im['BPA']),
  frequency = hdr_im['CRVAL3'],
  interactive = False,
  mean_map = 'default',
  multichan_opts = False,
  output_opts = True,
  quiet = True,
  polarisation_do = False,
  psf_vary_do = False,
  rms_box = (45,15),
  rms_map = None,
  shapelet_do = False,
  spectralindex_do = False,
  thresh = 'hard'
  )
```

The input parameters of the PYBDSF source finder with an impact on the completeness and reliability are:

(i) thresh_isl – threshold for the island boundary in number of sky rms above the mean. The default value is 3, however we have set it to 2.5, to not reject the weaker sources, which also leads to larger number of false positives, most of which is discarded at the crossmatching with other channels step. We have tried values in the range of 2–4 and found that 2.5 leads to a good compromise between the completeness and reliability.

(ii) thresh_pix – threshold for the island peak in number of sky rms above the mean. The default value is 5, however similarly to thresh_isl we have set it to the lower value of 3. We have tried values in the range of 2.5–5 and found that 3 leads to a good compromise between the completeness and reliability.

(iii) minpix_isl – minimum number of pixels with emission per island. We have set it to 16, as this is approximately the area of the synthesized beam of our data and we would expect the emission size to be no smaller than the beamsize.

A4 LESHI

We have run the LESHI source finder with the following input parameters:

```
  LESHI.source_finder(
  data_file,
  path_to_results = './',
  SNR_integ = 3.5,
  SNR_channel = 2.5,
  channel_min_len = 3,
  SNR_spec = 3,
  rsqr_min = 0.35,
  int_image_len = 10,
  int_image_load_no = 10,
  channel_start = 0,
  channel_end = None,
  beam = None,
  bg_box_size = 100,
  max_dist_pix = 10,
  max_dist_channel = 10,
  test_hist = True,
  sloped_continuum = False,
  core_no = 40)
```

The input parameters of the LESHI source finder with an impact on the completeness and reliability are:

(i) int_image_len – length of the frequency slab to be integrated into a moment-0 map on which the first phase of source finding is performed. We have set it to 10 channels, as this is approximately equal to the width of the weakest expected signals in the MIGHTEE data ($\sim$50 km s).

(ii) SNR_integ – threshold SNR of the source on the integrated image to be accepted. We have set it to 3.5, as this is high enough

to filter out most of the noise peaks, but low enough to allow for weak sources.

(iii) SNR_channel – threshold SNR of the source on the single channel image to be accepted, we have set it to 2.5 to allow for any weaker sources.

(iv) channel_min_len – minimum number of consecutive channels the sources has to persist in (with SNR specified by SNR_channel) to be accepted, we have set it to 3, as we would not expect any real source to have the width narrower than 3 channels ($\sim$15 km s^{-1}).

(v) SNR_spec – threshold SNR of the source in the spectral dimension, we have set it to 3 to filter out most of the noise peaks, while allowing weaker sources.

(vi) rsqr_min – minimum values of the R^2 parameter characterizing how well a Gaussian function fits to the spectrum of the detection. We have set it to 0.35, as it filters out most of the remaining noise peaks and leaves out the genuine 'bumps' in the spectrum, leading to a good compromise between the completeness and reliability.

APPENDIX B: COMPLETENESS FUNCTION

Table B1 contains the parameters describing the fitted completeness function (equation 1) for the SOFIA, PYBDSF, and PROFOUND source finders, derived analogously to Section 3.2. Figs B1, B2, and B3 contain plots showing the fitted completeness function analogously to Figs 6 and 7 for the SOFIA, PYBDSF, and PROFOUND source finders. Qualitatively, the fitted completeness functions for all of the sourcefinders are very similar. Fitted $M_{\rm min}$ parameter follows well the $M_{\rm min} \sim D_{\rm L}^2$ relation and the fitted slope parameter s is lower for the edge-on sources, especially for higher distances, making the completeness for highly inclined sources lower, as discussed in the main text.

Table B1. Fitted parameters of the completeness function for the different source finders for the completeness averaged over all inclinations, for face-on sources [cos(i) range of 0.9 to 1.0] and for edge-on sources [cos(i) range of 0.0 to 0.1].

Source finder	a_1 ($M_\odot$Mpc^{-2})	a_2 (Mpc^{-1})	b_2
SOFIA			
average	2095 ± 57	-0.0054 ± 0.0007	2.54 ± 0.19
face-on	1854 ± 147	-0.0035 ± 0.0011	2.24 ± 0.21
edge-on	2126 ± 167	-0.0048 ± 0.0008	2.35 ± 0.21
PYBDSF			
average	1845 ± 74	-0.0048 ± 0.0007	2.41 ± 0.19
face-on	1672 ± 130	-0.0033 ± 0.0012	2.26 ± 0.21
edge-on	1905 ± 165	-0.0041 ± 0.0009	2.21 ± 0.21
PROFOUND			
average	1872 ± 77	-0.0040 ± 0.0007	2.12 ± 0.19
face-on	1800 ± 146	-0.0028 ± 0.0011	2.05 ± 0.20
edge-on	1913 ± 178	-0.0039 ± 0.0008	2.10 ± 0.17

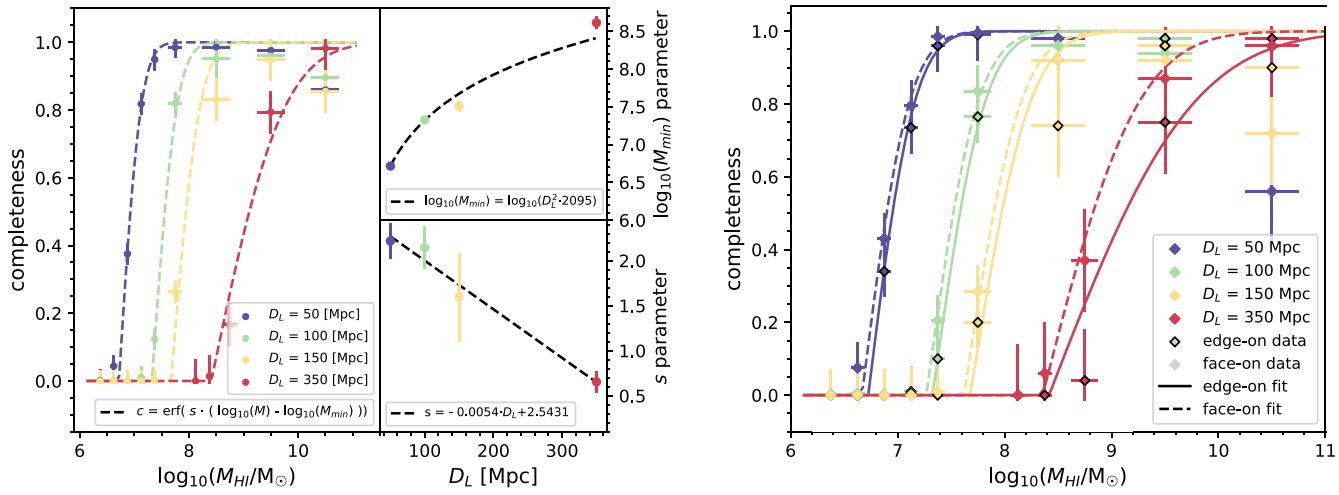

Figure B1. Figures analogous to Figs 6 and 7 for the SOFIA source finder.

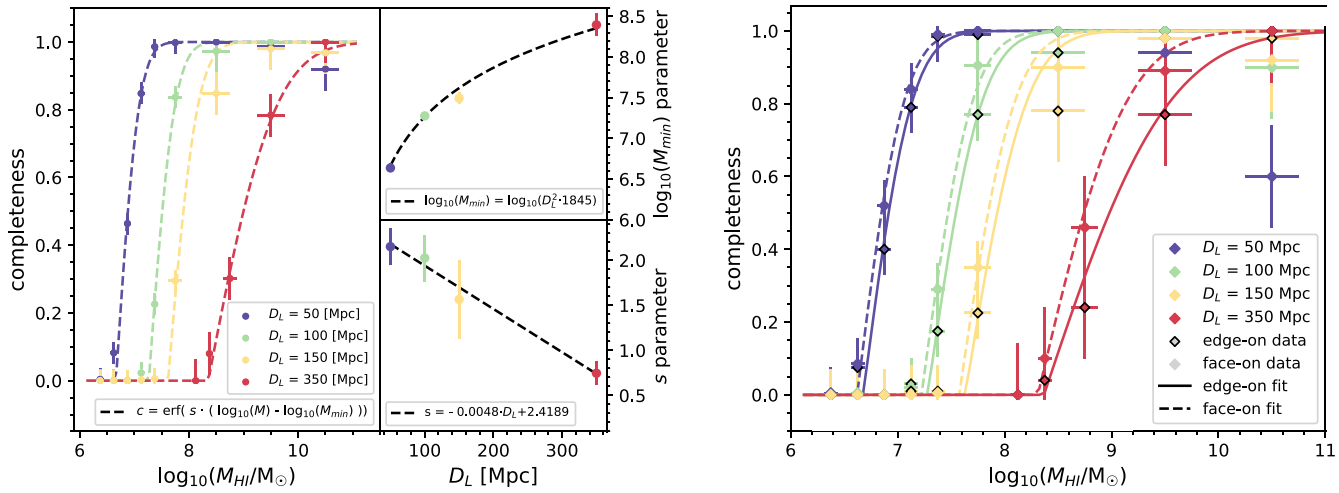

Figure B2. Figures analogous to Figs 6 and 7 for the PYBDSF source finder.

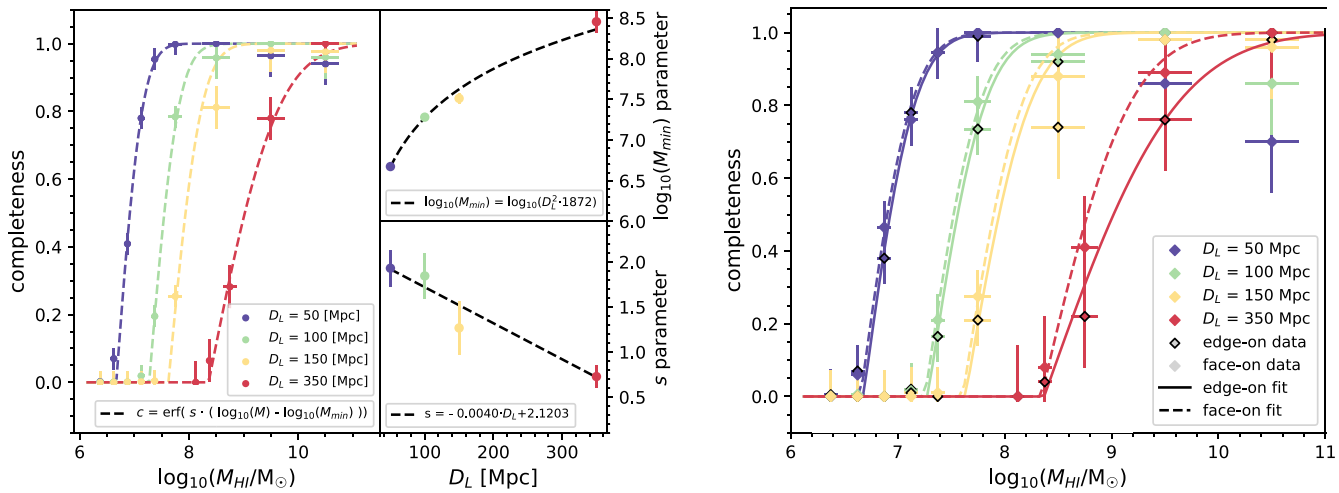

Figure B3. Figures analogous to Figs 6 and 7 for the PROFOUND source finder.

APPENDIX C: CATALOGUE SOURCES MAP

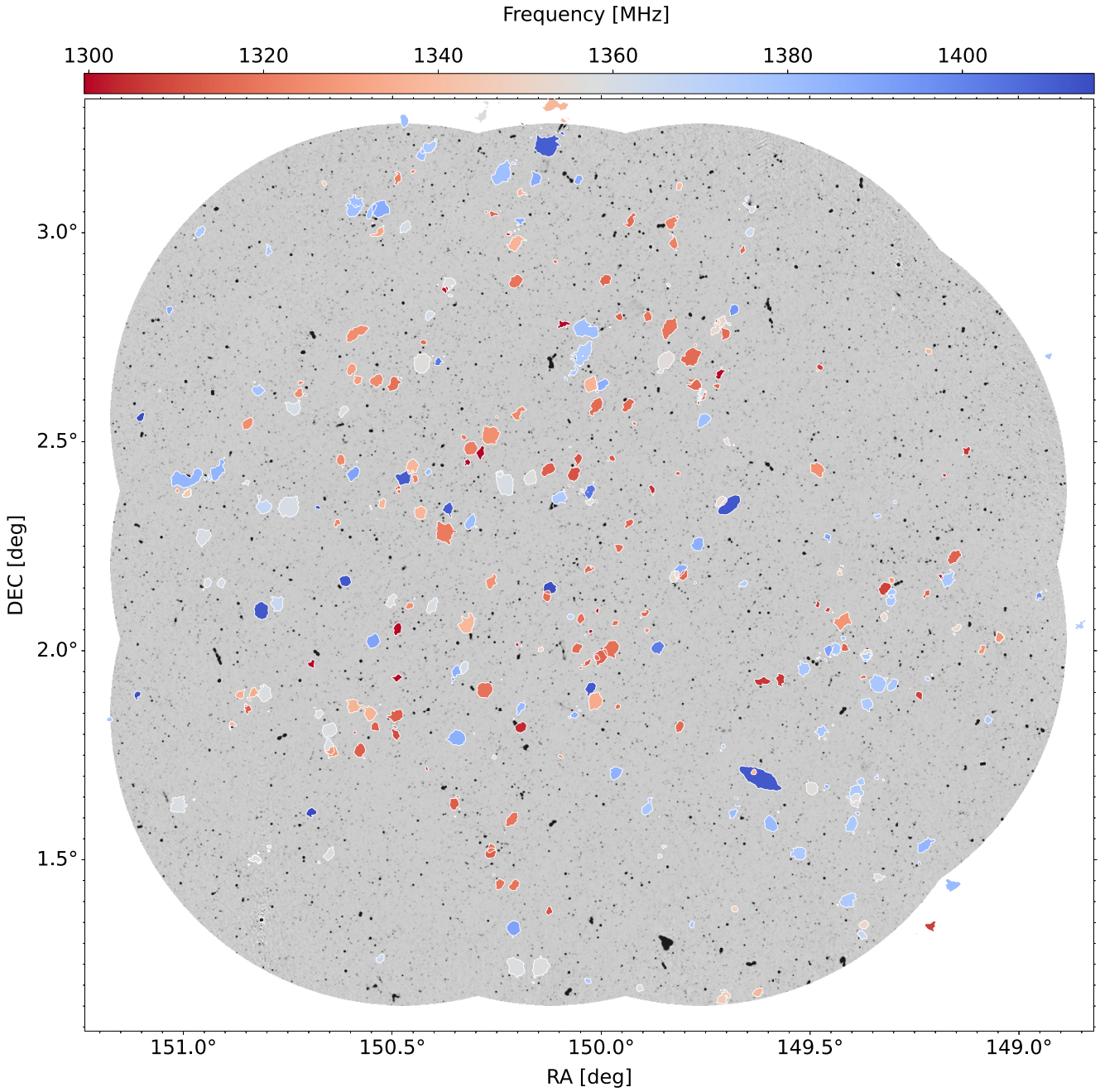

Figure C1. Contour islands of the H I distribution found by LESHI in the COSMOS field for a sample of 293 galaxies, colour coded by detection frequency, overplotted over radio continuum image. The contours are scaled up by a factor of 2 for better visibility and do not represent real sizes.

APPENDIX D: EXAMPLES OF DETECTIONS

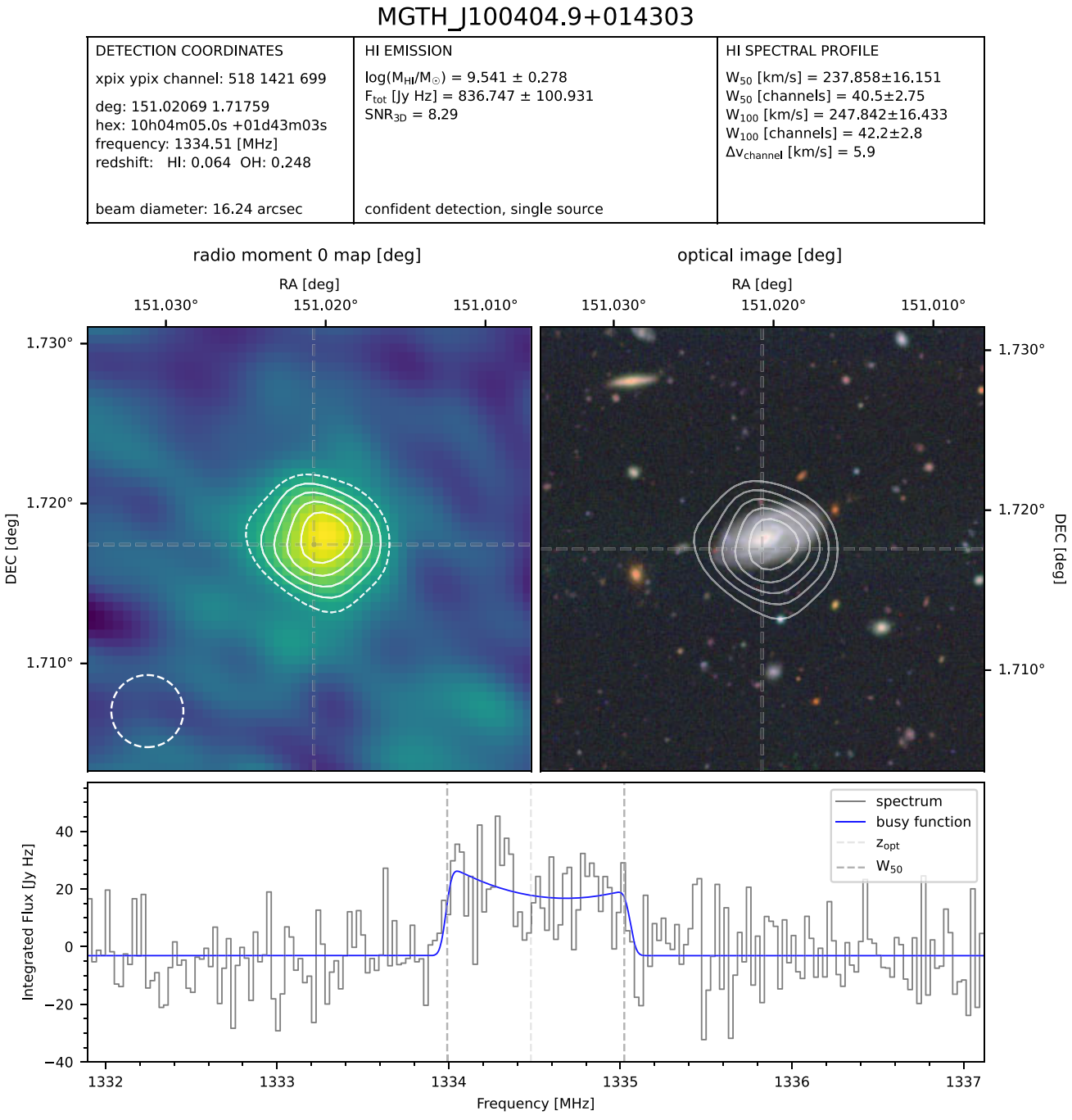

Figure D1. The only detection that was found by SOFIA and missed by LESHI.

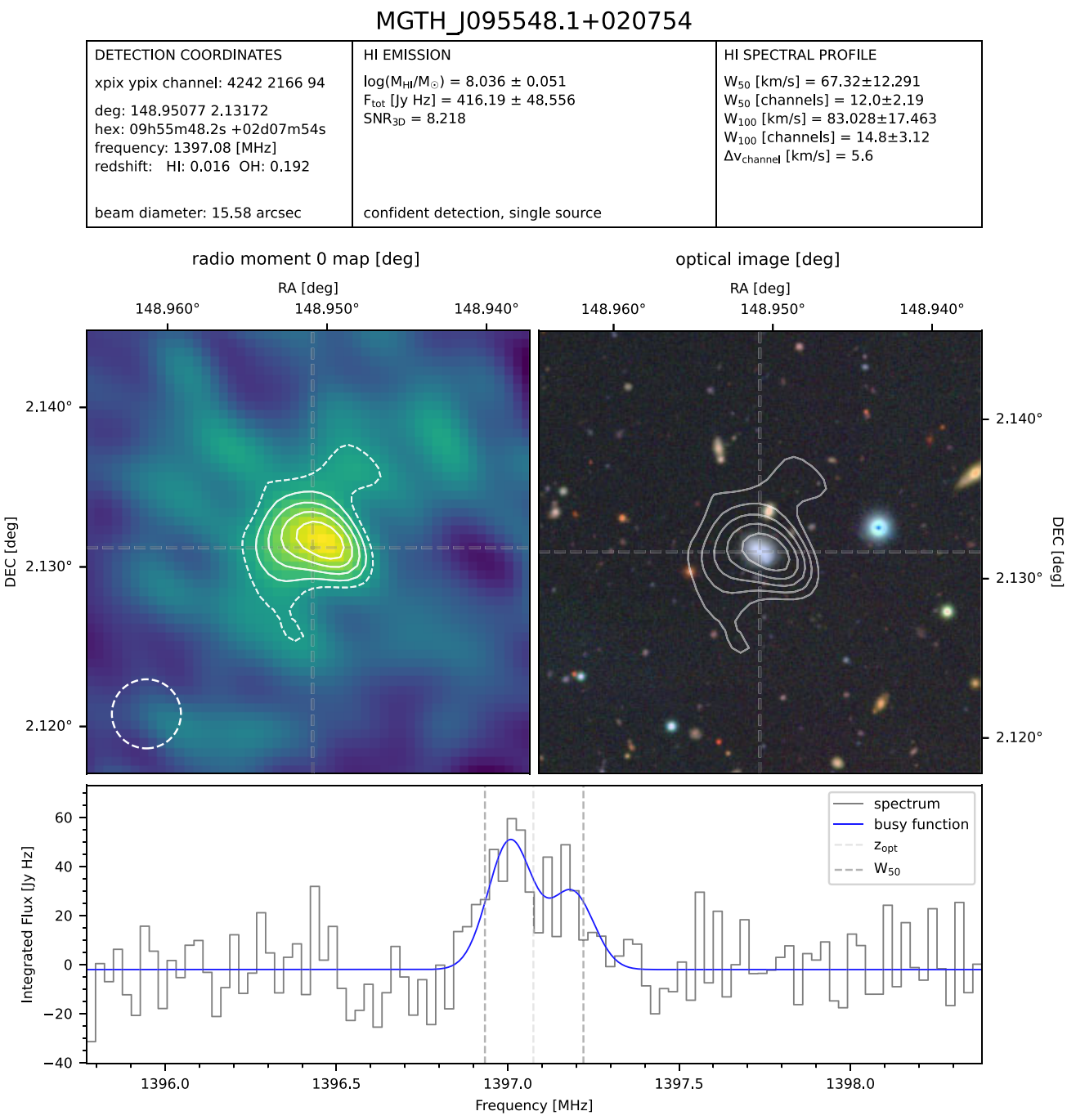

Figure D2. Example of a detection that was found by LESHI, but not by any other source finder.

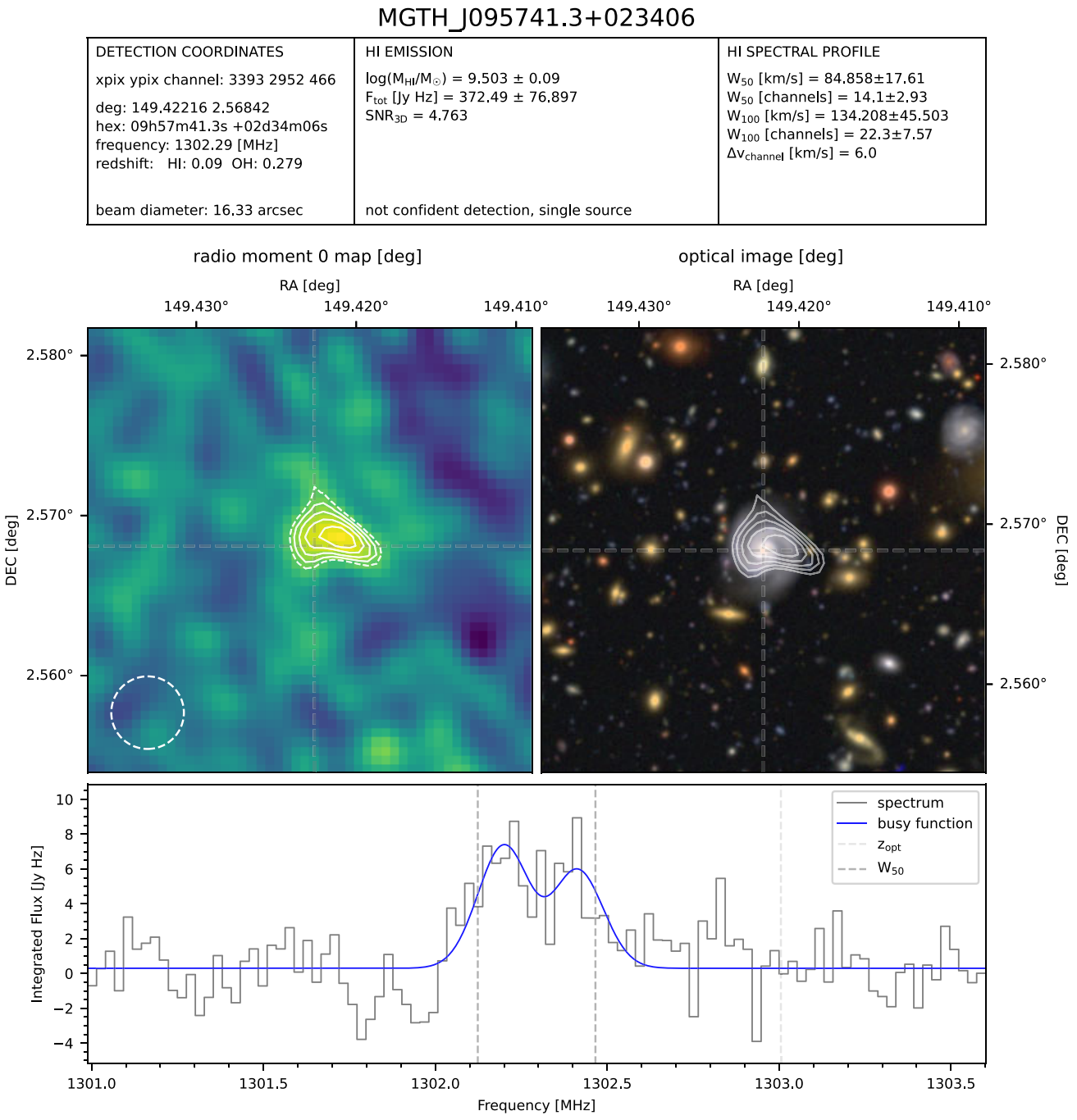

Figure D3. Example of a detection with the 'low confidence' flag. The H I detection has an optical counterpart; however, the detection has low signal-to-noise and its distribution does not completely follow the optical data.

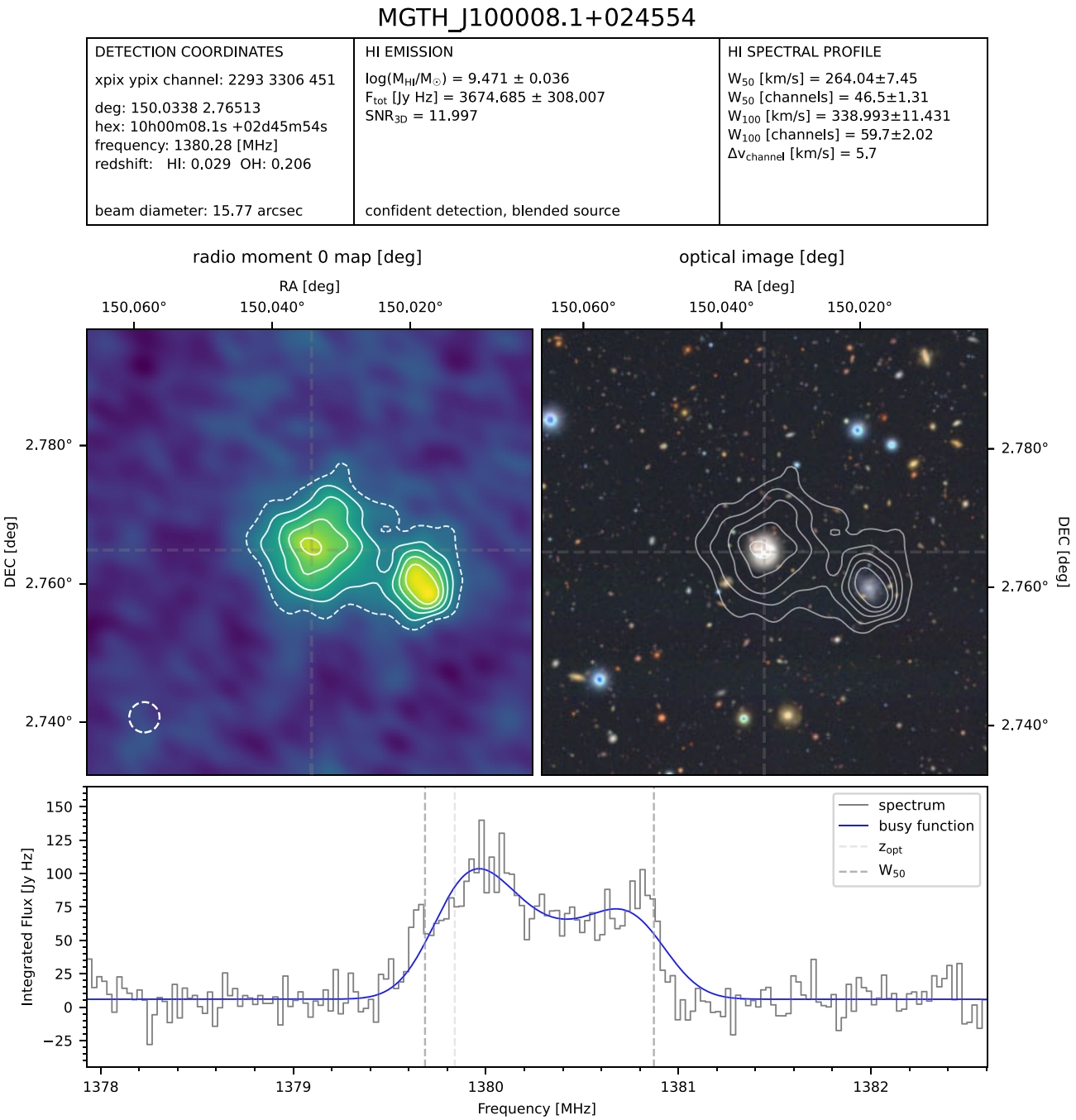

Figure D4. Example of a detection with the 'blended source' flag. The detection contains emission belonging to two galaxies which are interacting and their H I contents cannot be discerned.

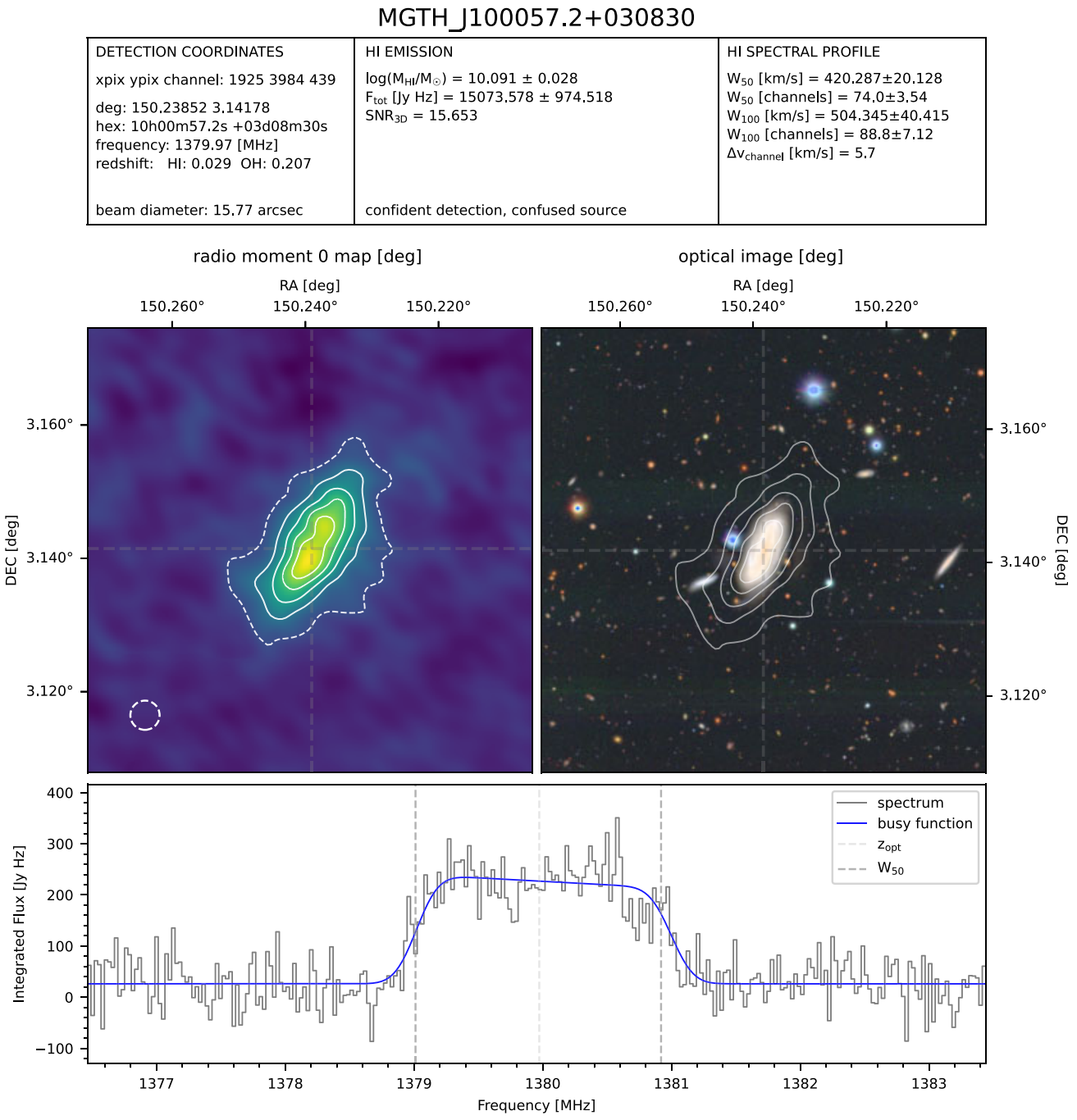

Figure D5. Example of a detection with the 'confused source' flag. Within the distribution of the detected emission there is more than one galaxy with similar redshift in the optical data; however, it is unclear from the H I emission whether it is a blended source or overlapped projection.

APPENDIX E: COMPARISON WITH EARLY SCIENCE DATA

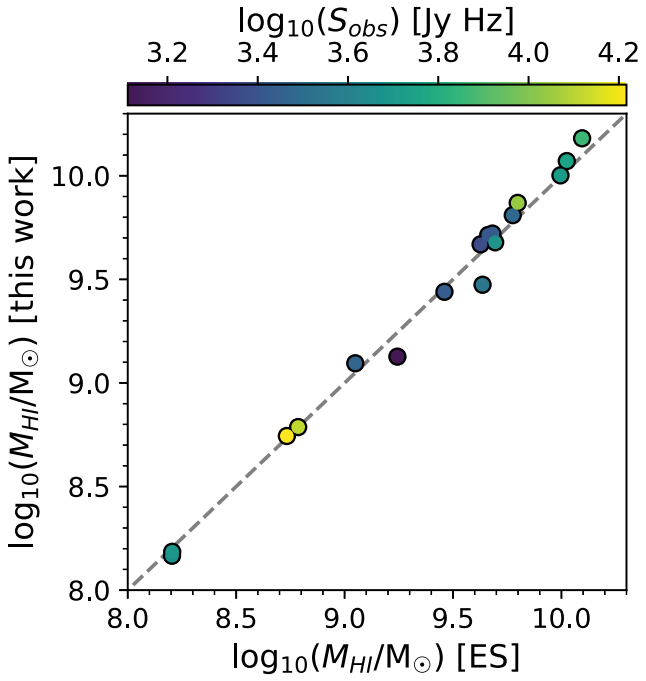

Figure E1. H I mass measured in this work plotted against the H I mass measured in A. A. Ponomareva et al. (2021), colour coded by the observed flux.

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