

https://dspace.org/

Looking at the Distant Universe with the MeerKAT Array The HI Mass Function in the Local Universe

Amir Kazemi-Moridani; Andrew J Baker; Marc Verheijen; Eric Gawiser; Sphesihle Makhathini ...

https://hdl.handle.net/10539/46258
DOAJ

Downloaded from a DSpace repository. https://dspace.org/

The Astrophysical Journal, 981:208 (18pp), 2025 March 10 https://doi.org/10.3847/1538-4357/ad9f3f
© 2025. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

Looking at the Distant Universe with the MeerKAT Array: The H I Mass Function in the Local Universe

Amir Kazemi-Moridani[1], Andrew J. Baker[1,2], Marc Verheijen[3], Eric Gawiser[1], Sarah-Louise Blyth[4], Danail Obreschkow[5], Laurent Chemin[6], Jordan D. Collier[7,8,9], Kyle W. Cook[10], Jacinta Delhaize[4], Ed Elson[2], Bradley S. Frank[11], Marcin Glowacki[7,12,13], Kelley M. Hess[14,15], Benne W. Holwerda[10], Zackary L. Hutchens[16], Matt J. Jarvis[2,17], Melanie Kaasinen[18], Sphesihle Makhathini[19], Abhisek Mohapatra[4], Hengxing Pan[17], Anja C. Schröder[20], Leyya Stockenstroom[4], Mattia Vaccari[21,22,23], Tobias Westmeier[5], John F. Wu[24,25], and Martin Zwaan[18]

[1] Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854, USA
[2] Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, Bellville 7535, South Africa
[3] Kapteyn Astronomical Institute, University of Groningen, Landleven 12, 9747 AD, Groningen, The Netherlands
[4] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7700, South Africa
[5] International Centre for Radio Astronomy Research (ICRAR), The University of Western Australia, 35 Stirling Highway, Perth, WA 6009, Australia
[6] Universidad Andrés Bello, Facultad de Ciencias Exactas, Departamento de Ciencias Físicas—Instituto de Astrofísica, Fernandez Concha 700, Las Condes, Santiago, Chile
[7] Inter-University Institute for Data Intensive Astronomy (IDIA), Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[8] School of Science, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia
[9] CSIRO, Space and Astronomy, P.O. Box 1130, Bentley, WA 6102, Australia
[10] Department of Physics and Astronomy, 102 Natural Science Building, University of Louisville, Louisville, KY 40292, USA
[11] The UK Astronomy Technology Centre, Royal Observatory Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK
[12] Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK
[13] International Centre for Radio Astronomy Research (ICRAR), Curtin University, Bentley, WA 6102, Australia
[14] Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-43992 Onsala, Sweden
[15] ASTRON, The Netherlands Institute for Radio Astronomy, Postbus 2, 7990 AA, Dwingeloo, The Netherlands
[16] Department of Physics & Astronomy, University of North Carolina Asheville, 1 University Heights, Asheville, NC 28804, USA
[17] Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
[18] European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching, Germany
[19] Wits Centre for Astrophysics, School of Physics, University of the Witwatersrand, 1 Jan Smuts Avenue 2000, South Africa
[20] Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, D-85748 Garching bei München, Germany
[21] Inter-University Institute for Data Intensive Astronomy, Department of Astronomy, University of Cape Town, Private Bag X3, 7701 Rondebosch, Cape Town, South Africa
[22] Inter-University Institute for Data Intensive Astronomy, Department of Physics and Astronomy, University of the Western Cape, 7535 Bellville, Cape Town, South Africa
[23] INAF—Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy
[24] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218-2410, USA
[25] Center for Astrophysical Sciences, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, USA
Received 2024 August 7; revised 2024 December 11; accepted 2024 December 12; published 2025 March 10

Abstract

We present measurements of the neutral atomic hydrogen (H I) mass function (HIMF) and cosmic H I density ($\Omega_{\rm H\,I}$) at $0 \leqslant z \leqslant 0.088$ from the Looking at the Distant Universe with MeerKAT Array (LADUMA) survey. Using LADUMA Data Release 1 (DR1), we analyze the HIMF via a new "recovery matrix" method that we benchmark against a more traditional modified maximum likelihood (MML) method. Our analysis, which implements a forward modeling approach, corrects for survey incompleteness and uses extensive synthetic source injections to ensure robust estimates of the HIMF parameters and their associated uncertainties. This new method tracks the recovery of sources in mass bins different from those in which they were injected and incorporates a Poisson likelihood in the forward modeling process, allowing it to correctly handle uncertainties in bins with few or no detections. The application of our analysis to a high-purity subsample of the LADUMA DR1 spectral line catalog in turn mitigates any possible biases that could result from the inconsistent treatment of synthetic and real sources. For the surveyed redshift range, the recovered Schechter function normalization, low-mass slope, and "knee" mass are $\phi_* = 3.56^{+0.97}_{-1.92} \times 10^{-3}\,{\rm Mpc}^{-3}\,{\rm dex}^{-1}$, $\alpha = -1.18^{+0.08}_{-0.19}$, and $\log(M_*/M_\odot) = 10.01^{+0.31}_{-0.12}$, respectively, which together imply a comoving cosmic H I density of $\Omega_{\rm H\,I} = 3.09^{+0.65}_{-0.47} \times 10^{-4}$. Our results show consistency between recovery matrix and MML methods and with previous low-redshift studies, giving confidence that the cosmic volume probed by LADUMA, even at low redshifts, is not an outlier in terms of its H I content.

Unified Astronomy Thesaurus concepts: Galaxies (573); Galaxy masses (607); H I line emission (690)

1. Introduction

Hydrogen, the most abundant element in the Universe, exists in a variety of gaseous phases spanning broad ranges of temperature and density. Neutral atomic hydrogen (H I) acts as a crucial intermediate stage in the evolution of gas in galaxies,

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

bridging the gap between ionized hydrogen flowing in from the intergalactic medium and molecular hydrogen, which serves as the primary fuel for star formation (M. P. Haynes et al. 1984; N. M. McClure-Griffiths et al. 2023). H I masses in galaxies can change due to the processes of accretion, consumption/conversion, and expulsion. Consequently, tracking the evolution of H I over cosmic time is vital for understanding galaxy evolution. The number density of galaxies as a function of H I mass (i.e., the neutral hydrogen mass function or HIMF) and its integral, the contribution of galaxies to the cosmic H I density ($\Omega_{\rm H\,I}$), are two key metrics for describing the distribution and abundance of neutral gas across cosmic time. Measurements of the HIMF and $\Omega_{\rm H\,I}$ enable comparisons with semianalytic and numerical models of galaxy evolution (G. Popping et al. 2014; H.-S. Kim et al. 2015; R. Davé et al. 2020).

Beyond cosmic averages, understanding H I properties in different settings can illuminate the dependence of galaxy evolution processes on environmental factors. H I is particularly sensitive to galaxy interactions, as it is affected by hydrodynamic pressure and has a more broadly extended distribution within galaxies compared to stars and other gas phases (e.g., J. C. Mihos 2001). The H I contents and distributions of galaxies reflect histories of interaction and the impacts of multiple evolutionary mechanisms, with H I surveys providing vital data for validating simulations and understanding the environmental variations of galaxy evolution (e.g., A. Chung et al. 2009; B. W. Holwerda et al. 2011; M. G. Jones et al. 2018; T. N. Reynolds et al. 2022).

The H I 21 cm line, resulting from a hyperfine transition in the ground state of a hydrogen atom, serves as an essential tool for investigating neutral hydrogen on Galactic and extragalactic scales. Over the past few decades, observations of the 21 cm line have revolutionized our understanding of H I distributions within the local Universe and beyond. Numerous H I emission line surveys, conducted with single-dish radio telescopes (M. A. Zwaan et al. 2003, 2005; A. M. Martin et al. 2010; M. G. Jones et al. 2018), have played a pivotal role in assessing the global H I properties of nearby galaxies. The two largest untargeted H I surveys, the H I Parkes All-Sky Survey (HIPASS; M. J. Meyer et al. 2004) and the Arecibo Legacy Fast ALFA (ALFALFA; R. Giovanelli et al. 2005) survey, have measured the HIMF and $\Omega_{\rm H\,I}$ in the nearby Universe. These surveys and the subsequent interferometric analysis of A. A. Ponomareva et al. (2023) have been instrumental in establishing our knowledge of the HIMF and its properties, including the contributions of various galaxy populations to the HIMF and its dependence on environment (C. M. Moorman et al. 2014; K. Said et al. 2019; M. G. Jones et al. 2020).

Previous H I surveys have predominantly focused on sources in the local ($z < 0.1$) Universe due to the faintness (i.e., low Einstein A coefficient) of the 21 cm line and the limited capabilities of the telescopes that observe it. Despite significant investments of observing time, so far only a small number of galaxies have been detected in H I beyond the local Universe (X. Fernández et al. 2013, 2016; B. Catinella & L. Cortese 2015; K. M. Hess et al. 2019; A. R. Gogate et al. 2020), with the most distant individual detection at $z \approx 0.42$ (H. Xi et al. 2024). Only two surveys have managed to assess the HIMF beyond the $z < 0.1$ Universe; the Arecibo Ultra-Deep Survey (AUDS; L. Hoppmann et al. 2015; H. Xi et al. 2021) covers $0 < z < 0.16$, while the Blind Ultra-Deep H I Environmental Survey (BUDHIES; Y. L. Jaffé et al. 2013; A. R. Gogate et al. 2020) has, for the first time, used direct H I detections to construct the HIMF and calculate $\Omega_{\rm H\,I}$ at $z \sim 0.2$ (A. Gogate 2022), in two volumes centered on galaxy clusters.

Although compiling a large sample of direct H I detections at higher redshifts for statistical investigations remains an elusive goal, several ambitious surveys are aiming to address this deficiency, using new facilities such as the Australian Square Kilometre Array Pathfinder (ASKAP; A. W. Hotan et al. 2021), the Five-hundred-meter Aperture Spherical radio Telescope (FAST; R. Nan 2008), and the MeerKAT array (J. L. Jonas 2009), and the phased array feed upgrade of the Westerbork Synthesis Radio Telescope known as the APERture Tile In Focus (APERTIF; E. A. K. Adams et al. 2022). This next generation of untargeted H I surveys is set to enhance our understanding of how the HIMF and $\Omega_{\rm H\,I}$ evolve over cosmic time. Surveys such as the MeerKAT International GigaHertz Tiered Extragalactic Exploration (MIGHTEE; M. Jarvis et al. 2016; N. Maddox et al. 2021) and the Deep Investigation of Neutral Gas Origins (DINGO; J. Rhee et al. 2023) are designed to probe H I in galaxies across a broad range of redshifts and environments. Building on the precedent of the deep but narrow COSMOS H I Legacy Extragalactic Survey (CHILES; X. Fernández et al. 2016; K. M. Hess et al. 2019) with the Very Large Array, the Looking At the Distant Universe with the MeerKAT Array (LADUMA; S. Blyth et al. 2018) survey now aims to explore H I in emission up to an unprecedented $z \sim 1.4$, utilizing both direct and stacked detections, and is designed to detect thousands of galaxies at its targeted depth.

Previous analyses have inferred notable discrepancies in the shape of the HIMF in the local Universe (e.g., between ALFALFA's spring and fall volumes; M. G. Jones et al. 2018), but it is unclear whether these variations reflect fundamental differences in galaxy evolution in different environments or are expected consequences of cosmic variance. As we extend our observational reach with the current generation of H I surveys, the catalogs curated at higher redshifts are expected to contain few detections compared to surveys of the local Universe, further increasing the difficulties of drawing robust conclusions and motivating the development of improved analytical methods. To fully leverage the potential of these upcoming samples and to address some of the observed discrepancies at lower redshifts, it is imperative to refine and develop new methods for deriving the HIMF.

The present paper addresses this critical need by proposing a novel approach to improve the analysis and interpretation of HIMF estimates across different redshifts. We demonstrate its application by presenting the first determination of the HIMF using observations from LADUMA. Using LADUMA's Data Release 1 (DR1) data set, we derive the HIMF over roughly the last billion years ($0 < z < 0.088$) and calculate an associated $\Omega_{\rm H\,I}$ for this redshift range. The paper is organized as follows. In Section 2, we describe the LADUMA survey, our processing of the DR1 data, and our definition of a high-purity H I sample. In Section 3, we describe how we measure the HIMF, which is parameterized in terms of a P. Schechter (1976) function; the results of our analysis are presented in Section 4. We discuss the novel aspects of our method and implications for its application at higher redshift in Section 5. A summary and conclusions are presented in Section 6. Throughout this paper, we assume a flat cold dark matter (ΛCDM) cosmology with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$, $\Omega_m = 0.3$, and $\Omega_\Lambda = 0.7$. All reported

values from the literature have been rescaled as necessary for consistency with this cosmology.

2. Observations

2.1. Data Processing

LADUMA is a deep untargeted 21 cm survey of a single pointing using the MeerKAT array, with an area that expands from ~0.8 deg^2 at $z_{\rm H\,I} = 0$ to ~5 deg^2 at $z_{\rm H\,I} = 1.4$ (S. Blyth et al. 2018). The LADUMA pointing encompasses the Chandra Deep Field South and is centered at 03:32:30 −28:07:57 (J2000). The LADUMA data were processed on the ilifu facility operated by the Inter-University Institute for Data Intensive Astronomy (IDIA), using the processMeerKAT pipeline[26] (J. D. Collier et al. 2021) combined with custom scripts. The data used in this paper were collected in 19 nighttime tracks with MeerKAT's *L*-band (0.88–1.67 GHz) receivers, whose average duration was 9 hr. Data were obtained at the native 32k channel resolution (26.123 kHz) of the MeerKAT correlator and processed at 8k resolution in three independent spectral windows (SPWs) of 880–933, 960–1161, and 1304–1420 MHz, in order to avoid regions of strong radio frequency interference. This paper focuses on the data in the highest-frequency (low-*z*) SPW, which corresponds to $0 < z_{\rm H\,I} < 0.088$.

The full width at half maximum (FWHM) of the MeerKAT primary beam[27] expands from 60′.5 at 1420 MHz to 65′ at 1304 MHz. The outstanding sensitivity of MeerKAT allows for the detection of H I emitters well past the FWHM of the primary beam. As a result, in this work we analyze a volume extending to the full width at quarter maximum (FWQM) of the primary beam, which increases from 83′ to 90′ over the frequency range of the SPW. A detailed account of the data reduction will be presented in a forthcoming paper (A. Kazemi-Moridani et al. 2025, in preparation), but we summarize key steps here.

Each track is individually calibrated and imaged, with robust weighting adjusted per track to minimize sidelobes. We combine the tracks after subtracting a continuum sky model from each in the *uv* plane. At its central frequency, the combined data cube has synthesized beam dimensions of 8″.0 × 7″.5 with P.A. = −34° and an rms noise of ~33 μJy beam^{-1} per 104.52 kHz channel. Model subtraction leaves low-level continuum residuals, especially in the vicinity of bright continuum sources, that are not visible in a single-track image. However, as the noise integrates down with the combination of multiple tracks, the remaining continuum residuals emerge above the lower noise. We have developed a spline[28]-fitting algorithm to remove these residuals in our deepest data cube, which we will refer to as pixel-based continuum subtraction from this point on. This latter stage of continuum subtraction affects our measured H I fluxes, as it models the underlying continuum emission at locations where the line sources reside. Overestimation (underestimation) of the continuum level can result in underestimation (overestimation) of the line flux, as can also occur for alternative approaches to continuum subtraction (e.g., M. J. Meyer et al. 2004; M. P. Haynes et al. 2011). For the LADUMA DR1 cubes, this effect is minimal for sources with small velocity widths, as the spline-fitting algorithm only needs to model the continuum level over a few channels. For sources with larger velocity widths, the specific choices made in implementing the pixel-based continuum subtraction algorithm have the potential to systematically affect the total measured line fluxes, albeit at a modest (typically <5%−10%) level.

2.2. Source Catalog

Given the need for an automated source-finder for injection/recovery tests (see Section 3.1.1 below), we make use of the Source Finding Application (SoFiA; T. Westmeier et al. 2021) package, which offers excellent efficiency, flexibility, and reliability. Our source-finding approach is refined by maximizing the fraction of synthetic sources recovered with high fidelity in the entirety of the synthetic source population. To maintain consistency between finding real sources and finding synthetic sources, we apply the same approach to the real data and use a high-purity subsample of all of the line sources detected in the low-*z* SPW. We note that this high-purity subsample is considerably smaller than the full LADUMA DR1 source catalog. That catalog, which will be described in detail elsewhere (A. Kazemi-Moridani et al. 2025, in preparation), includes both a core sample (for which SoFiA parameters were iteratively adjusted to match the results of unguided source-finding with visual and matched-filtering methods) and a supplemental sample (containing the results of visual source-finding guided by prior knowledge of optical redshifts, as well as a more extensive exploration of SoFiA parameter space). Sources in the supplemental sample generally have lower signal-to-noise ratios (S/N) and therefore do not figure in the high-purity subsample used in this paper. The SoFiA parameters we use in this analysis match those used to develop the core DR1 sample and are listed in Table 1 (if different from package defaults). For these parameters, in the low-*z* SPW data cube, SoFiA finds ~190 candidate sources, of which we judge ~140 to be real detections based on visual inspection and cross-matching with optical catalogs. On the basis of previous predictive work by H. Roberts et al. (2021), and given the low redshifts in question, we do not expect any OH megamasers (OHMs) to be present in this list.

Out of our parent sample of ~140 sources, 89 are detected with high enough reliability (>99%; see Section 3.1.1 below) to be included in the high-purity sample. By limiting our analysis to the FWQM of the primary beam, we reduce the number of included detections from 89 to 84, ensuring that only high-S/N massive sources remain. Furthermore, in order to avoid biasing the HIMF by relying on the very small volume associated with the two lowest-mass sources (i.e., sources in the lowest-mass bin; see Section 5.1 for further details), we exclude them from our analysis as well, leaving 82 sources in the final sample used to measure the HIMF. Figure 1 shows the distribution of these 82 sources (see Section 3.1.1) as a function of redshift and on the plane of the sky. Given the rarity of rich clusters (e.g., G. O. Abell 1958) and the deficiency of H I emission in the densest environments (e.g., R. Giovanelli & M. P. Haynes 1985), we expect that our H I-selected sources are primarily located in field or intermediate-density environments. Considering LADUMA's angular resolution, we also expect that spectral line confusion will not significantly affect the recovered source population (e.g., M. G. Jones et al. 2015).

[26] https://idia-pipelines.github.io/docs/processMeerKAT

[27] The primary beam response is calculated using the https://github.com/ska-sa/katbeam package.

[28] We use splines rather than polynomials because splines better capture the smooth, ripple-like behavior in the pixel spectra.

Table 1
SoFiA Parameters Used for This Paper's Injection/Recovery Analysis

Parameter	Default	Used Here
window_x	25	63
window_y	25	63
window_z	15	31
scalenoise.interpolate	false	true
scfind.kernelsXY	0, 3, 6	0, 2, 5
scfind.kernelsZ	0, 3, 7, 15	0, 3, 5, 9
threshold	5	4
linker.minsizeZ	5	3
linker.radiusXY	1	2
linker.radiusZ	1	2
reliability.threshold	0.9	0.7
reliability.minSNR	3	5

Note. Parameters are only listed when adopted values differ from defaults.

3. Analysis

3.1. Recovery Matrix Method

3.1.1. Injecting and Recovering Sources

The HIMF is defined as the number density of galaxies as a function of H I mass per unit comoving volume, which is usually represented as

$$\phi(M_{\rm H\,I}) = \frac{dN_{\rm gal}}{dV\ d\log_{10}(M_{\rm H\,I})} \quad (1)$$

where $dN_{\rm gal}$ is the number of galaxies with H I masses falling in a logarithmic mass bin centered on $M_{\rm H\,I}$ and lying in the comoving volume dV. Determining the intrinsic HIMF from observed number counts is a complex problem, especially when a sample is not volume-limited. To measure the HIMF, it is essential to correct for the survey sensitivity, which requires estimating the completeness of the observed sample. Completeness is typically calculated for each H I mass bin as an estimate of the fraction of all galaxies within the associated mass range that have actually been detected. The main factors affecting completeness in our data are distance, nonuniform sensitivity across our field of view due to primary beam attenuation, random orientations of individual H I emitters on the sky (affecting the observed line widths of those sources), uncertainties in flux measurements, continuum subtraction effects, and source-finding accuracy. The combined effects of these factors bias our observed H I galaxy sample toward nearby, gas-rich, and low-inclination sources near the center of the LADUMA field. To assess the completeness of our survey, we adopt an empirical approach that involves inserting synthetic sources into our processed data cube. We evaluate their recovery rate by requiring that they be detected via the same process as the real detections in our survey. As described in Section 2.2, and given the number of artificial source injections required to accurately sample the multidimensional parameter space of real H I sources, employing this approach requires a highly automated source-finder that can detect galaxies with high reliability.

All injection/recovery methods are based on prior knowledge of the underlying distributions of source parameters, such as mass, inclination, size, and velocity width. To create a catalog of synthetic sources that accurately samples the multidimensional parameter space spanned by H I sources, we follow an approach based on the one detailed in A. R. Gogate et al. (2020). The synthetic H I sources are simulated using the Groningen Image Processing System (GIPSY; J. M. van der Hulst et al. 1992) package's `galmod` task, which uses a tilted ring model (D. H. Rogstad et al. 1974) to create the 3D velocity field of a given H I source. For each galaxy, `galmod` requires the radial H I surface density distribution (in $\rm cm^{-2}$) and the rotational velocity distribution (in $\rm km\ s^{-1}$) as a function of the radius (in arcseconds) of the galaxy, as well as the velocity dispersion (in $\rm km\ s^{-1}$), the inclination angle, and the position angle (in degrees). We use the radial surface density distribution profiles as described in J. L. Sérsic (1968) and T. P. K. Martinsson et al. (2016) and the rotational velocity profiles described by M. Persic et al. (1996) and S. Courteau (1997), as applicable to different mass ranges, to generate synthetic H I sources across the wide mass range of our study. We make use of the known local H I size–mass scaling relation (J. Wang et al. 2016) when generating the parameters required by `galmod` to guarantee that the sources follow this relation. The equation for the amplitude of the rotational velocity profiles (from M. Persic et al. 1996) likewise yields results that are consistent with the Tully–Fisher relation (R. B. Tully & J. R. Fisher 1977). Our final catalog comprises over 150,000 synthetic sources, with randomly chosen inclinations and position angles, covering H I masses in the range $7 \leqslant \log(M_{\rm H\,I}/M_\odot) < 10.75$, such that the vast majority of sources have masses in the lower half of the range, as required to ensure reliable recovery statistics for the entire mass range (see below). To ensure minimal alterations to the noise properties of the data cube, which can influence the source-finding process, the number of sources in any single injection/recovery trial is limited to a maximum of 500 for sources near the low-mass limit and a maximum of 100 for sources near the high-mass limit.

When SoFiA is run on a given cube that includes both real and synthetic sources, it delivers a list of all spectral line detections and—for each detection—a 3D "cubelet" that includes generous spatial and spectral buffers around the pixels in which emission is seen.[29] We calculate the total H I mass of each detection as

$$\left(\frac{M_{\rm H\,I}}{M_\odot}\right) = 49.7\left(\frac{D_L}{\rm Mpc}\right)^2\left(\frac{S}{\rm Jy\ Hz}\right), \quad (2)$$

where D_L is the cosmological luminosity distance to the source, and S is the integrated H I flux density (M. Meyer et al. 2017). The integrated H I flux density is calculated using zeroth moment maps that are created by first smoothing individual source cubelets to a circular beam of $20'' \times 20''$. Each smoothed cubelet is clipped at a 3σ threshold (σ is estimated by measuring the rms noise over an emission-free region) to create a mask that encompasses diffuse low-column-density emission. We remove isolated regions corresponding to noise peaks from the mask by discarding all regions whose areas are smaller than that of the smoothed beam. We then apply the resulting mask to the original-resolution cubelet to generate the zeroth moment map and calculate the integrated H I flux density.

False-positive detections can become a major source of error in recovery rate calculations. Given our perfect knowledge of where

[29] The SoFiA parameters used for this stage of our analysis differ from those used to produce the original source catalog only in allowing the detection of sources with very large angular sizes, as is appropriate for synthetic sources with large H I masses and low redshifts.

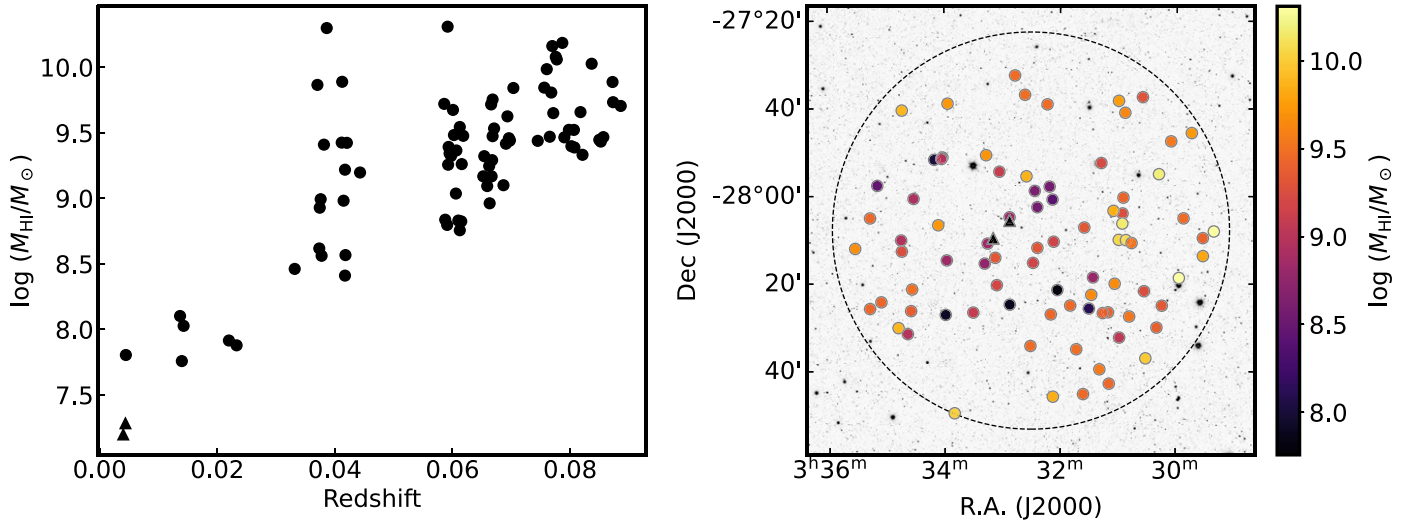


Figure 1. LADUMA H I sources detected by SoFiA with higher than 99% reliability inside the FWQM (two sources with $M_{\rm H\,I} < 10^{7.5}\,M_\odot$, which are not used in our analysis, are plotted as triangles). Left: H I mass as a function of redshift. Right: high-purity subsample of H I sources in the LADUMA field plotted overlaid on the Spitzer/IRAC channel 1 mosaic for the extended CDFS (Euclid Collaboration et al. 2022). The dashed circle represents the FWQM of the MeerKAT primary beam at 1304 MHz (the lower-frequency edge of the spectral window analyzed here).

the synthetic sources are, we can easily identify any false positives that SoFiA "recovers"; however, it is not possible to similarly distinguish false from real sources in our full observed sample. In order to eliminate the complications that would be caused by false-positive detections in our catalog of real sources, we need to work only with a high-purity subset of SoFiA detections in both the observed and synthetic data. Recovery tests on the synthetic sources have shown that using SoFiA's reliability threshold of 99% creates a high-purity subsample (purity $>$ 99%).[30] All of the real sources in the selected subsample are detected in every injection/recovery trial run, which confirms that the injected population in each trial does not alter source-finding accuracy. Based on optical redshift measurements and WISE-based color diagnostics (H. Roberts et al. 2021), our catalog does not include any detections of OHMs. We note that our high-purity sample of 84 H I detections is considerably smaller than the full DR1 catalog of H I detections (A. Kazemi-Moridani et al. 2025, in preparation); this circumstance reflects our deliberate choice to prioritize the accurate determination of sample completeness over a large sample size, which are inescapably in tension with each other.

Each injection/recovery trial includes the following steps:

1. creating synthetic sources in the desired mass range ($M_{\rm H\,I}$ is randomly sampled from a log-uniform distribution, and inclinations are sampled from a $p(i)di = \sin(i)di$ distribution) with `galmod`;[31]
2. convolving the `galmod` output for each source with the point-spread function of our data as appropriate for the assumed source redshift (redshifts are sampled from a $p(z) \propto dV(z)/dz$ distribution, where $V(z)$ represents the volume of the survey as a function of redshift[32]);
3. injecting sources in the pre-pixel-based-continuum-subtraction data cube at randomly selected locations with a uniform cosmological volume density distribution (while avoiding overlap with existing sources in the full LADUMA catalog[33]);
4. running the second-stage (pixel-based) continuum subtraction algorithm on the data cube, including both real and synthetic sources;
5. running SoFiA on the resulting continuum-subtracted line cube to find spectral line sources;
6. cleaning all recovered sources using a custom (J. A. Högbom 1974) CLEAN routine applied to the SoFiA masks; and
7. cross-matching the (R.A., decl., z) tuples of the recovered source catalog with the (R.A., decl., z) tuples of the input source catalog.[34]

The outputs of our 500 trials are then combined to create a reference recovery catalog in which the fate of each injected source and its properties are recorded. We split the total mass range ($7 \leqslant \log(M_{\rm H\,I}/M_\odot) \leqslant 11$) into eight bins in order to investigate the recovery details for each bin. A traditional

[30] Because no analogs of multiwavelength counterparts exist for synthetic sources, in contrast to real sources, we cannot incorporate the use of multiwavelength catalogs into the source-finding process for real sources. Without using multiwavelength data, it becomes much more challenging to determine how many source candidates identified by SoFiA (e.g., above a lower reliability threshold of 95%) are false detections. Therefore, we need to establish a reliability threshold using injection recovery tests alone that delivers high-purity samples (e.g., the high-purity sample from the 99% reliability threshold), which do not require cross-validation against multiwavelength data.

[31] The sources are spectrally smoothed in `galmod` with a Gaussian kernel whose FWHM is twice the channel separation.

[32] Accurately capturing the effects of large-scale structure (LSS) at the point of source injection would (ideally) entail simultaneously solving for both the underlying LSS and the HIMF using our observed mass and spatial distributions. Unfortunately, our sample size is insufficient to support such a joint analysis. Instead, we have assessed source recovery statistics in the presence of LSS analogous to that observed in our real source catalog. These tests were focused on sources with $\log(M_{\rm H\,I}/M_\odot) \geqslant 9$, which represent 75% of our observed sources and sample volumes where a reasonable estimate of LSS can be derived from H I data. Our tests confirm that the presence of LSS in *our* data along the line of sight does not significantly affect source recovery statistics as a function of H I mass, although other surveys might be affected differently.

[33] The spatial position of a given synthetic source is determined in (r, θ) coordinates with respect to the pointing center. Radius r is sampled according to $p(r)dr = rdr$, and θ is sampled from a uniform distribution between 0 and 2π. Each (r_i, θ_i, z) tuple is cross-checked with existing source positions to avoid overlap.

[34] 99% of the injected sources are recovered within a sphere with a 3.5 spaxel radius centered on the injected position.

approach here would calculate the ratio of recovered-to-injected sources in each bin and correct the number of real detections by the recovery factor to determine the underlying distribution of galaxies. However, our simulations show that even with a recovery fraction of close to 100%, a considerable number of galaxies injected in a given mass bin are recovered in a different bin, with a shift into a higher-mass bin much more likely than the alternative. Given that there are significantly more low-mass objects than higher-mass objects in the real Universe, the recovery of sources in higher-mass bins can introduce a bias in the inferred HIMF that needs to be corrected.

3.1.2. Defining a Recovery Matrix

In order to keep track of the migration of recovered sources across mass bins, we create a correction matrix—a 2D array of recovery factors (instead of a correction vector, i.e., a 1D array of recovery factors)—in which the fates of sources injected in each mass bin are recorded as a function of their recovered mass. This "recovery matrix" can be thought of as a function over an $M_{\rm inj(ected)}$ versus $M_{\rm rec(overed)}$ domain, where the function at a given ($M_{\rm inj}$, $M_{\rm rec}$) point represents the fraction of sources with an injected mass in the $M_{\rm inj}$ bin that have been recovered with a mass in the $M_{\rm rec}$ bin. In an ideal case, the recovery fraction matrix would be the identity matrix, signifying that each source has been recovered in the same mass bin in which it was injected. Our tests show that the recovery fraction matrix for our data cube is a band-diagonal matrix with significant nonzero values on the main diagonal and the adjacent diagonals, such that sources are most likely to be recovered in their original injection bin and much more likely to be recovered in the next-higher mass bin than in the next-lower mass bin (Figure 2). The greater likelihood of recovering sources in the adjacent higher-mass bin results from the combined effect of several factors, notably the effects of the source-finding process and image-plane continuum subtraction. Source finding relies heavily on threshold cuts, such that a source near the detection limit might be selected as a detection if it coincides with a positive noise region, but ignored if it coincides with a negative noise region.[35] Estimation of the underlying continuum during pixel-based continuum subtraction in turn could be biased toward underestimating the continuum level for some sources depending on their parameters (such as inclination), leading to under-subtraction of continuum and overestimated line fluxes for those sources. The combination of these factors, along with the effects of other steps in our data processing, leads to an asymmetric bias toward recovery at higher masses.

In order to make sure that we have an accurate estimate of each matrix element, we have injected enough sources to ensure that the fractional uncertainties on the diagonal and below-diagonal matrix elements are <10%. Figure 3 shows the percentage uncertainties on the diagonal matrix elements as a function of the number of injected sources. To calculate the uncertainties on the matrix elements for the recovery of synthetic H I sources, we employ a method inspired by the bootstrapping technique. For each mass bin in the analysis, we start with a large pool of synthetic sources—specifically, at least 15% more sources than the maximum number used in any calculation. We perform multiple random draws of synthetic source samples from this pool, calculate matrix elements for each draw, and then estimate the uncertainties. We begin with a small sample size, $n = 100$, and conduct 4096 random draws of this size from the full synthetic catalog, calculating matrix elements for each draw. The uncertainty for this sample size is then derived from the results by calculating the range that includes 68% of the data around the mean. This procedure is repeated with increasing sample sizes ($n = 200, 400, \cdots$), until we reach a size that is within 15% of the size of the total pool (e.g., 10,000 for a pool of 11,500 sources). Each draw is made without replacement to avoid biases that might arise from reusing the same sources within a single estimation round. This stepwise approach allows us to observe how uncertainties in

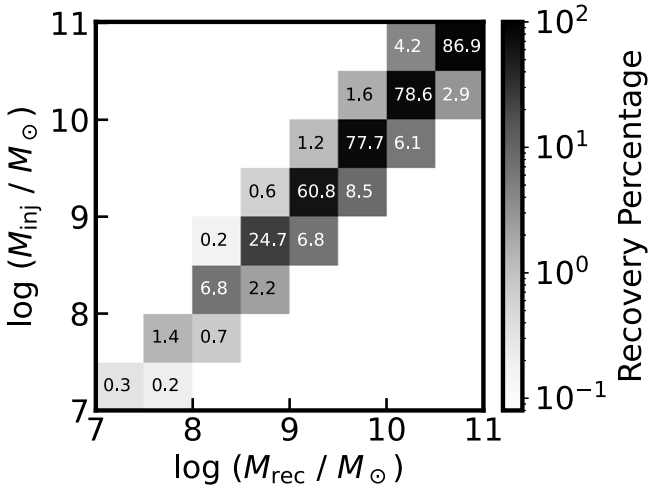


Figure 2. The recovery fraction matrix plotted on the $M_{\rm inj}$ vs. $M_{\rm rec}$ grid. The diagonal elements express recovery fraction as a percentage and increase monotonically as a function of injected mass $M_{\rm inj}$ and recovered mass $M_{\rm rec}$. For the lower-mass bins, the off-diagonal elements are large relative to the on-diagonal elements. If a source is not recovered in the injected mass bin, it is more likely to be recovered with a higher mass than with a lower mass.

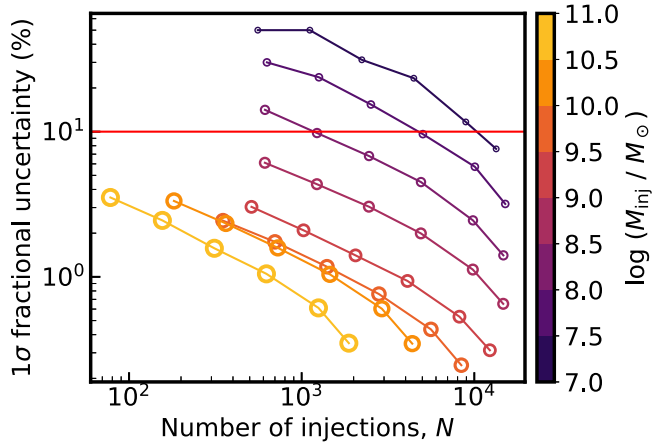


Figure 3. 1σ fractional uncertainty (as a percent) on the diagonal elements of the recovery fraction matrix shown in Figure 2 as a function of the number of injections N. The fractional uncertainty is determined by calculating the recovery matrix elements for 4096 randomly chosen samples of N sources in a given $M_{\rm inj}$ bin from a larger pool of injected sources in that bin. The final 1σ fractional uncertainties on all of the diagonal (shown here) and below-diagonal matrix elements are <10%.

[35] We note that even high-mass sources can lie close to the detection threshold if they lie at large redshifts and/or large offsets from the pointing center. This positive bias is different from the 2%–3% positive bias in SoFiA flux measurements that is noted by T. Westmeier et al. (2021); the latter is not relevant to our analysis here because we do not use SoFiA to measure fluxes (see above).

the recovery matrix elements vary with changes in the sample size (Figure 3), ensuring that the calculated uncertainties are representative of the true variability expected in different sampling scenarios.

3.1.3. Forward Modeling

Numerous studies (e.g., M. A. Zwaan et al. 2005; M. G. Jones et al. 2018) have demonstrated that the HIMF can be effectively characterized by a P. Schechter (1976) function, defined as

$$\phi(M_{\rm H\,I}) = \ln(10)\,\phi_* \left(\frac{M_{\rm H\,I}}{M_*}\right)^{\alpha+1} e^{-\left(\frac{M_{\rm H\,I}}{M_*}\right)}, \quad (3)$$

whose three free parameters are the normalization constant ϕ_*, the "knee" mass M_*, and the low-mass slope α. We adopt a forward-modeling approach to estimate the Schechter function parameters that are consistent with our observed data. For each set of parameters (ϕ_*, M_*, α), we calculate the intrinsic number of H I sources for each bin based on the volume of the survey and the bin widths. The intrinsic numbers for all bins (for a given set of Schechter function parameters) are then multiplied by the rows of the recovery matrix and summed across columns to determine the expected numbers of observed sources in all bins. The likelihood of observing the actual numbers of sources is then calculated based on these expected counts for each set of Schechter parameters during the fitting and Markov Chain Monte Carlo (MCMC) sampling process. By directly estimating the Schechter function parameters using the observed set of detections, we avoid the need to create an unbiased (corrected, binned) histogram of detections for fitting with a Schechter function.

Our forward-modeling approach allows us to convert the numbers of galaxies in the different $M_{\rm inj}$ bins predicted for a given set of Schechter function parameters into numbers of recovered galaxies over the $M_{\rm rec}$ bins. We calculate the total number of recovered galaxies in each $M_{\rm rec}$ bin by summing the number of recovered galaxies from all of the $M_{\rm inj}$ bins. Our tests confirm that the inferred low-mass slope of the Schechter function can be biased if we fail to account for the recovery of sources with higher masses than the bins in which they were injected. Given that nontrivial numbers of sources are recovered with higher-than-injected masses, we have chosen to exclude the lowest $M_{\rm rec}$ bin from the forward-modeling process, as we cannot properly estimate the number of sources with $\log(M_{\rm H\,I}/M_\odot) < 7$ that have been recovered in that bin.

At the high-mass end, our analysis includes the $10.5 \leqslant \log(M_{\rm H\,I}/M_\odot) < 11$ bin, in which we actually detect no real H I sources, in recognition of the fact that a nonnegligible fraction of sources are recovered with higher masses and in order to determine the knee mass more accurately. Including the highest-mass bin in the forward-modeling process constrains the number of sources in the second-highest mass bin to be consistent with the lack of detections in the highest mass bin. We note that the predicted number of sources in a given bin is the mean of a Poisson distribution describing the source count for that bin. The Poisson distribution $P(\mu)$ for mean values $\mu > 7$ can be approximated well with a Gaussian distribution $N(\mu, \sqrt{\mu})$. However, the discrepancy between the two distributions for smaller values of μ introduces a bias in the forward-modeling process by differently weighting the estimated likelihoods for bins with few detections. We choose the Poisson likelihood, as our investigations reveal that choosing the Gaussian or mean-standard-error likelihood instead of the Poisson likelihood significantly underestimates the uncertainties associated with the inferred Schechter function parameters and introduces a bias toward steeper low-mass slopes.[36]

3.1.4. Ensuring Consistency with the Observed Mass Function

It is necessary to maintain consistency between the observed HIMF and the hypothetical HIMF used in completeness calculations. For a given mass bin i, the completeness fraction C_i represents the fraction of all sources in that bin that can be detected within the survey volume. In a traditional injection/recovery method, for bin i, the completeness fraction C_i is defined as the ratio of the number of recovered sources R_i to the number of injected sources I_i in that bin, i.e., $C_i = R_i/I_i$. To account for cross-bin contamination, we can write

$$C_i = \frac{R_i}{I_i} = \frac{\sum_j I_j f_{j\to i}}{I_i} = \sum_j \frac{I_j}{I_i} f_{j\to i}, \quad (4)$$

where $f_{j\to i}$ represents the fraction of sources that are injected in bin j and recovered in bin i. Isolating the term for galaxies both injected and recovered in bin i results in

$$C_i = \frac{R_i}{I_i} = f_{i\to i} + \sum_{j\neq i} \frac{I_j}{I_i} f_{j\to i}, \quad (5)$$

showing that when some fraction of sources recovered in bin i are from bin j (for $j \neq i$), the completeness fraction for bin i depends on the ratio of the number of injected galaxies in bin j to bin i, i.e., I_j/I_i. In a traditional approach, the number of injections in each bin is proportional to the predicted count from the HIMF to preserve relative ratios and enhance statistical reliability. For example, if the HIMF predicts N sources in a given bin, then the number of injected sources is some multiple of N. Given that any HIMF varies significantly—by >2 orders of magnitude—across the mass range of a sample like ours, for every source in the highest-mass bin, about 300 sources need to be injected into the lowest-mass bin. This requirement poses a practical challenge for any traditional approach; for example, to obtain reliable statistics for our highest-mass bin if it were to contain only 100 sources, over 30,000 sources would need to be injected into the lowest-mass bin alone to maintain ratios consistent with the HIMF. Using a recovery matrix, however, eliminates this constraint by incorporating cross-bin contamination directly into the matrix, allowing the number of injections in each bin to be chosen independently of the other bins. We make use of this flexibility when determining the number of injections by requiring a 10% threshold on the uncertainties associated with the diagonal and below-diagonal matrix elements.

3.1.5. Inferring Schechter Function Parameters

In order to calculate the best-fit Schechter function parameters and their associated uncertainties, we first maximize the Poisson likelihood for our observed data to calculate a set of best-fit parameters, which are then used in an MCMC sampling process to estimate the uncertainties on those

[36] For more details on these calculations, please see Appendix A.

parameters. As discussed above, we inject enough sources in each $M_{\rm inj}$ bin (independent from the other $M_{\rm inj}$ bins) to estimate the diagonal and below-diagonal elements of the recovery fraction matrix with $<10\%$ uncertainty. However, the estimated matrix elements and their uncertainties depend in part on the slope of the mass distribution of the injected sources *within* a given bin. A discrepancy between the observed HIMF and the assumed slope of the mass distribution of sources in a given bin, if present, will result in a biased estimation of the Schechter function parameters. To mitigate this effect, we determine the best-fit Schechter function parameters using two iterations. In the first iteration, an initial recovery matrix is calculated using the ALFALFA $\alpha.100$ Schechter function parameters (M. G. Jones et al. 2018). This matrix is used to maximize the Poisson likelihood for our data to find an initial set of best-fit Schechter function parameters. In the second iteration, we update the slope of the mass distribution of sources in each bin to be the slope recovered from the first iteration and recalculate the matrix elements, which are then used to find a revised set of best-fit Schechter function parameters. The Schechter function parameters from the second iteration change by $<1\%$ compared to the results from the first iteration, eliminating the need for any further iterations.[37]

We use the PyMC[38] implementation of the MCMC method to estimate the uncertainties in the best-fit Schechter function parameters and their covariance by sampling an appropriate posterior probability distribution. The prior distributions for the Schechter function parameters are $\log_{10}(\phi_*)$ uniform in $[-4, -2]$, $\log_{10}(M_*)$ uniform in [9, 10.5], and α uniform in $[-1.8, -0.5]$. In the MCMC sampling process, we create 512 different realizations of the recovery matrix using the best-fit Schechter function parameters, in order to account for the uncertainties in the matrix elements, and perform an MCMC sampling with eight chains and 32,768 steps (after burn-in) for each of these 512 matrices. The posteriors from the 512 realizations are then combined to create a full posterior, for estimating the uncertainties on the best-fit Schechter function parameters in a way that includes the uncertainties in the matrix elements.

Taken together, the elements of the recovery matrix method for HIMF determination—starting with source injection and recovery as laid out in Section 3.1.1, and ending with the estimation of Schechter function parameters as described immediately above—could potentially be subjected to a full end-to-end validation test. In such a test, ideally, a single realization of a mock galaxy population with a known HIMF could be simulated into a source-free version of the LADUMA DR1 data cube, and the Schechter function parameters recovered from that mock population could be compared to those of the input HIMF. Unfortunately, as discussed below in Section 5.4 and Appendix A, the appropriate (Poisson) uncertainties in the Schechter function parameters for a sample of the size being analyzed in this paper are very large, such that the parameters recovered for a single realization are likely to be very different from the input parameters a priori. As a result, (in)consistency between input and output HIMF parameters for a single mock data cube cannot be used as a test of the recovery matrix method per se. End-to-end validation of the method could in principle be achieved with an ensemble of mock data cubes; however, creating an ensemble of cubes that is sufficiently large to test the recovery matrix method (beyond the ability of Poisson uncertainties to compromise the test) would be computationally prohibitive and is beyond the scope of this paper.

3.2. Modified Maximum Likelihood Method

In addition to the forward-modeling approach discussed above, we have calculated the HIMF using the modified maximum likelihood (MML) method described in D. Obreschkow et al. (2018). The MML method, which assumes Poisson statistics, recovers the HIMF without any binning while dealing with mass uncertainties and the selection biases present in the data. This method uses as an input a selection function that describes the recovery rate of sources as a function of distance and mass, which is averaged over extra variables such as width (D. Obreschkow et al. 2018). We estimate the selection function for our data cube by creating 2D histograms of the injected and recovered sources on a distance–mass grid. We have chosen 32 distance bins and 16 mass bins spanning the full distance and mass ranges of our data. The selection function is defined using a linear interpolator (in lieu of an analytical expression) on the 2D mass–distance gridded data, where the gridded values are calculated as ratios of the recovered source histogram and the injected source histogram. We use the dftools package developed in R and described in detail in D. Obreschkow et al. (2018) to calculate the corrected number density of H I sources in each bin. We then fit a Schechter function to the corrected number densities and estimate the uncertainties in the parameters of the Schechter function using MCMC sampling.

The MML method offers the flexibility to calculate the Schechter function parameters in several different ways, such as providing the maximum volume $V_{\rm max}$ in which each source can be detected, or alternatively, providing a selection function.[39] While using the maximum volume option results in a reasonable fit to our data, we chose to generate a selection function based on our extensive injection/recovery tests for a fair comparison with the recovery matrix method, as described in Section 3.2.[40] However, this approach results in convergence issues with the MML method, resulting in an unreliable fit. To work around this problem, we use the built-in functionality of the dftools package to extract the corrected HIMF values for our mass bins and subsequently fit a Schechter function to these values. This process closely aligns with previous HIMF studies, where the corrected HIMF is calculated from binned data and a Schechter function is then fit to those values, making this discussion relevant to previous works that follow a similar analytical framework.

4. Results

4.1. HIMF and Schechter Function Parameters

In Figure 4, we compare the Schechter functions preferred by the recovery matrix method and the MML method for the

[37] We have confirmed that starting the iterations with a flat mass distribution across each bin similarly converges after two iterations, eliminating any concern that the use of an ALFALFA Schechter function to calculate the recovery matrix might somehow bias our results.

[38] https://www.pymc.io

[39] For more details, see https://rdrr.io/github/obreschkow/dftools/man/dffit.html.

[40] To allow for a fair comparison between the results of the MML and recovery matrix (RM) methods, we need to use the selection function approach for the former, so that we can make use of the same injection/recovery information used by the latter.

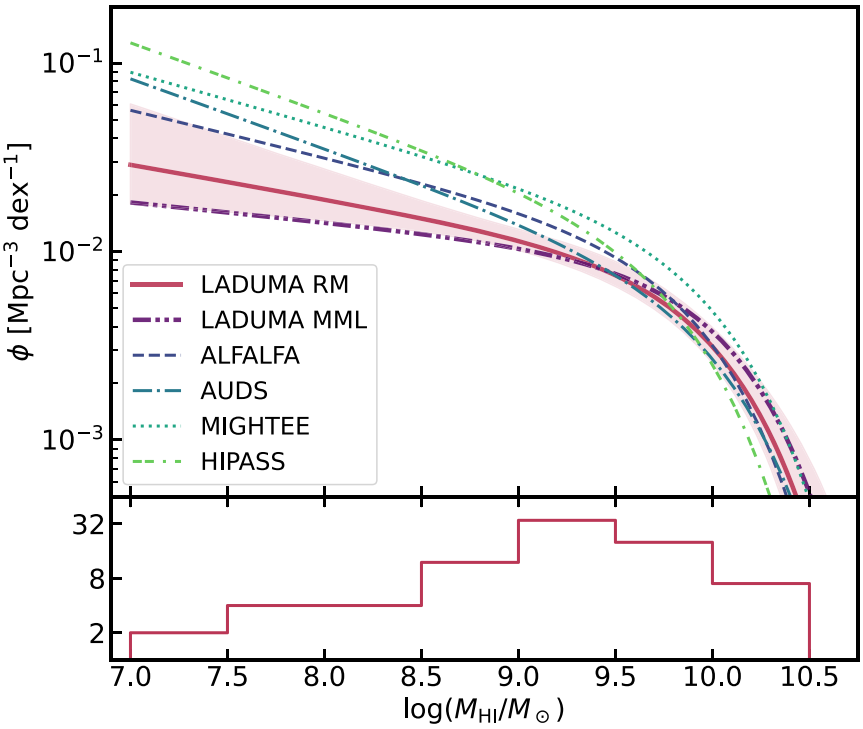


Figure 4. The HIMF for LADUMA as derived via the recovery matrix and the MML methods in comparison with previous measurements in the literature. The best-fit Schechter functions based on the recovery matrix (RM) and the MML methods are shown with the solid and dashed–double-dotted lines. The shaded region represents the 1σ uncertainty in the RM fit, sampled from the MCMC posteriors (Figure 5); the analogous uncertainty in the MML fit is shown in Figure 6. The histogram in the lower panel shows the number of observed sources in the LADUMA high-purity sample as a function of H I mass. The lowest-mass bin shown here (with two detections) was not included in the forward modeling and posterior sampling (see Section 5).

LADUMA DR1 data, along with the actual detected number of galaxies in each bin within our high-purity sample. The best-fit parameters for both methods are determined using a search on a variable resolution grid with a high density of points around the best-fit region. Figure 5 shows the estimated Schechter function parameters and their uncertainties plotted on the marginalized 1D and 2D posterior probability distributions for the recovery matrix method. The associated uncertainties in the parameters are determined directly from the 3D sampled posterior, as none of the three Schechter function parameters are considered nuisance parameters.[41]

Determining the appropriate uncertainties for the MML result is complicated by the fact that the MML method does not provide uncertainties for any bin with zero detections. If we circumvent this limitation by excluding the zero-detection highest-mass bin, we obtain only a weak constraint on the knee mass, with a 1σ upper limit reaching the upper limit ($\log(M_*/M_\odot) = 10.5$) of the MCMC prior range. Excluding the highest mass bin from the fitting process eliminates any loss associated with the separation between the Schechter function value for that bin and its observed value of zero, essentially allowing the fit to extend more freely to inappropriately large values at the high-mass end. In order to improve the Schechter function fit to the MML results, we assign an artificial value and an associated uncertainty to the highest-mass bin when we fit the Schechter function and perform MCMC sampling. As the knee mass and its associated uncertainties are sensitive to the artificial value assigned to the highest mass bin ($10.5 \leqslant \log(M_{\rm H\,I}/M_\odot) < 11$), we assign a value of 0.4 detections in that bin (only modestly higher than the actual value of 0, and encompassing both 0 and 1 within its 1σ Poisson uncertainty). This adjustment to the highest-mass bin leads to

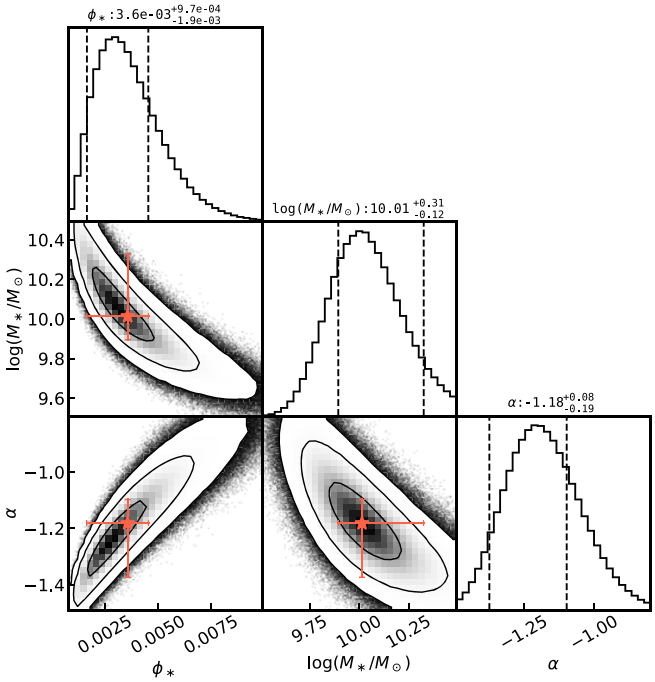


Figure 5. Marginalized 1D and 2D distributions of the sampled posterior for the recovery matrix method. The best-fit values and their associated uncertainties, obtained from the 3D posterior, are shown on the 2D marginalized distributions. The vertical dashed lines on the 1D histograms show the 1σ uncertainty regions for each parameter. The contours show $\{1, 2, 3\}\sigma$ level (i.e., 39th, 86th, and 99th percentile) uncertainties for the 2D marginalized distributions.

[41] The 4D volume enclosed inside the 1σ limit of a 3D Gaussian distribution is 24.91% of the total volume. Estimating the parameters and their uncertainties from the marginalized posterior distributions does not lead to the same results as when parameters are estimated from the 3D distribution (see Figure 5).

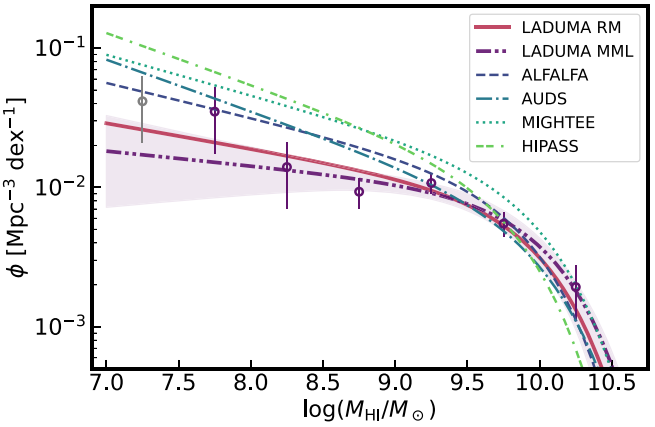


Figure 6. The HIMF for LADUMA from the MML method (see Section 3.2) in comparison with previous measurements in the literature. The best-fit Schechter function is shown as the solid line; the shaded region represents the 1σ uncertainty range on the MML fit. The lowest mass bin data point, shown in gray, is not included in the fitting process.

more reliable estimates of the knee mass and its associated uncertainties. Given that the output of the MML method is the corrected number density and associated uncertainty for each bin, we need to use a Gaussian likelihood for fitting and MCMC sampling. Therefore, the MML results are slightly biased by the effects of bins with low numbers of counts (as explained in Appendix A) compared to the results from the recovery matrix method. The results from the MML method, which are consistent with our recovery matrix method results but less so with the results of previous surveys, are shown in Figure 6.

Comparing our results with those of previous surveys is complicated by the fact that different authors have used different values of H_0 (with which ϕ_*, M_*, and $\Omega_{\rm H\,I}$ scale straightforwardly), different approaches to defining cosmological volume (which are not always stated but will affect ϕ_* and $\Omega_{\rm H\,I}$), and different choices of likelihood that translate to different "1σ" uncertainties. To provide as consistent comparisons with previous results as we can, in Table 2 we have scaled the values of ϕ_*, M_*, and $\Omega_{\rm H\,I}$ reported by the HIPASS, AUDS, and MIGHTEE teams for our choice of H_0 (ALFALFA uses the same value of H_0, so requires no rescaling), and we provide indications in the table notes of how we might expect results to change further on the basis of volume calculations. We also present the best-fit Schechter function parameters for the recovery matrix method and the MML method for LADUMA, along with a "free fit" version of the MML results that illustrates the impact of ignoring the zero-detection highest-mass bin. Figure 7 shows that the estimated Schechter function parameters from the recovery matrix and MML methods are in good agreement with each other. Relative to previous HIMF measurements from the literature, while the individual Schechter function parameters that we recover with the MML method are not consistent with all previous results (e.g., the AUDS measurement for α), those we recover for LADUMA using our (preferred) recovery matrix method agree with those for all previous surveys within the associated 1σ uncertainties when added in quadrature. The situation for the HIMF as a whole is not as clear: due to covariance between the Schechter function parameters, some of the literature HIMFs fall outside the 1σ uncertainty swath shown in Figure 4 across a wide range of masses. A full assessment of the consistency of our derived HIMF with the results of previous surveys, which would require detailed knowledge of the covariances among those surveys' respective Schechter function parameters and perhaps recalculation of their uncertainties using a Poisson likelihood (which then might or might not overlap with the LADUMA uncertainty range), is beyond the scope of this paper.

4.2. Cosmic H I Density ($\Omega_{H\,I}$)

A complete inventory of H I in the local Universe would include neutral hydrogen that lies outside of galaxies, both within and beyond the cosmic web. In this paper, we can estimate the contribution of galaxies to $\Omega_{\rm H\,I}$ based on the comoving H I mass density ($\rho_{\rm H\,I}$) that is calculated by integrating the Schechter function. The density $\rho_{\rm H\,I}$ is computed analytically (M. Meyer et al. 2017) as

$$\rho_{\rm H\,I} = \Gamma(\alpha + 2)\phi_* M_*, \tag{6}$$

where Γ is the Euler gamma function and ϕ_*, M_*, and α are the Schechter function parameters. The contribution of galaxies to $\Omega_{\rm H\,I}$ is then calculated as

$$\Omega_{\rm H\,I} = \frac{8\pi G}{3H_0^2}\,\rho_{\rm H\,I}, \tag{7}$$

where G is the gravitational constant, and H_0 is the Hubble constant. We measure a galactic $\Omega_{\rm H\,I} = 3.09^{+0.58}_{-0.49} \times 10^{-4}$ with the recovery matrix method and a galactic $\Omega_{\rm H\,I} = 3.48^{+0.26}_{-0.93} \times 10^{-4}$ with the MML method. Uncertainties in these measurements encompass 68% of $\Omega_{\rm H\,I}$ values (around the best-fit value) calculated from 4096 sets of Schechter function parameters randomly drawn from the 3D sampled posterior. Our results are in agreement with the results for all of the previous surveys except for MIGHTEE within 1σ uncertainties added in quadrature.

5. Discussion

5.1. Available Volume as a Function of H I Mass

The HIMF we have measured from the LADUMA data is subject to some of the same caveats as mass functions measured from previous surveys, in ways that are instructive about the limitations that apply to any HIMF measurement. One significant caveat is that the detectability of a source depends on its H I velocity width as well as its H I mass; our recovery matrix approach deals with this complication by marginalizing over the expected distribution of disk inclinations (see, e.g., D. Obreschkow et al. 2018). Perhaps the most important caveat is that for a flux-limited H I sample, the sources in the lowest-mass bin are typically confined to a much smaller (nearby) volume than the total survey volume (M. A. Zwaan et al. 1997; see especially their Figure 4). While previous authors have emphasized the importance of the bias of a high-mass galaxy population relative to the underlying dark matter distribution as a contributor to cosmic variance (e.g., B. P. Moster et al. 2011), the limited volume in which low-mass galaxies can be detected for a flux-limited sample can be just as important. Among the three Schechter function parameters, M_* will be least sensitive to "available" volume and most sensitive to population bias, as it is predominantly determined by the high-mass bins. In contrast, and as pointed out by M. G. Jones et al. (2018), α is expected to be most sensitive to available volume (albeit least sensitive to population bias), as it is predominantly determined by the low-mass bins. ϕ_* is expected to fall somewhere in between in terms of susceptibility to cosmic variance, as all bins contribute to the overall normalization within the limits of their associated uncertainties. For our data, the accessible volume associated with

Table 2
Best-fit Parameters of the Schechter Function and Corresponding Galactic $\Omega_{\rm H\,I}$ Values for LADUMA (This Work), ALFALFA, HIPASS, AUDS, and MIGHTEE

Survey	$z_{\rm max}$	Volume (Mpc^3)	$N_{\rm src}$	ϕ_* ($\times 10^{-3}$ Mpc^{-3})	$\log(M_*/M_\odot)$	α	$\Omega_{\rm H\,I}$ ($\times 10^{-4}$)
LADUMA RM	0.088	$\sim 9 \times 10^3$	82	$3.56^{+0.97}_{-1.92}$	$10.01^{+0.31}_{-0.12}$	$-1.18^{+0.08}_{-0.19}$	$3.09^{+0.65}_{-0.47}$
LADUMA MML	〃	〃	〃	$3.78^{+1.81}_{-1.14}$	$10.07^{+0.14}_{-0.17}$	$-1.10^{+0.16}_{-0.12}$	$3.48^{+0.26}_{-0.93}$
LADUMA MML (free fit)	〃	〃	〃	$3.61^{+0.47}_{-2.29}$	$10.10^{+0.41}_{-0.09}$	$-1.12^{+0.05}_{-0.22}$	$3.62^{+0.67}_{-0.97}$
ALFALFA (J18)	0.05	$\sim 6.2 \times 10^6$[a]	22831	$4.50^{+0.82}_{-0.82}$	$9.94^{+0.05}_{-0.05}$	$-1.25^{+0.10}_{-0.10}$	$3.90^{+0.61}_{-0.61}$[b]
HIPASS (Z05)	0.042	$\sim 12 \times 10^6$	4315	$4.88^{+0.81}_{-0.81}$	$9.86^{+0.04}_{-0.04}$	$-1.37^{+0.06}_{-0.06}$	$3.75^{+0.61}_{-0.61}$
AUDS (X21)	0.16	$\sim 39 \times 10^3$	247	$2.65^{+0.48}_{-0.48}$	$10.06^{+0.04}_{-0.04}$	$-1.37^{+0.03}_{-0.03}$	$3.33^{+0.10}_{-0.10}$
MIGHTEE (P23)	0.084	$\sim 21 \times 10^3$[c]	203	$5.12^{+5.58}_{-2.89}$	$10.04^{+0.24}_{-0.24}$	$-1.29^{+0.37}_{-0.26}$	$5.26^{+0.91}_{-0.95}$

Notes. $z_{\rm max}$ shows the upper limit of the redshift range for each survey. For LADUMA, the *disfavored* MML "free fit" shows the best-fit values obtained by excluding the highest-mass bin, which has no detections, from the fitting process. We calculate the comoving volumes using the area and redshift range reported for each survey in the corresponding publication, scaled as needed to our cosmology. All other parameters are rescaled as needed for $H_0 = 70$ km s^{-1} Mpc^{-1}. References in the table are as follows: J18 = M. G. Jones et al. (2018), Z05 = M. A. Zwaan et al. (2005), X21 = H. Xi et al. (2021), and P23 = A. A. Ponomareva et al. (2023).

[a] We list here the comoving volume probed by the ALFALFA α.100 sample for our assumed cosmology, which is slightly smaller than the 6.5×10^6 Mpc^3 actually used in the HIMF calculations of M. G. Jones et al. (2018; see also M. Jones 2024, private communication). Use of our smaller volume would modestly increase ϕ_* and $\Omega_{\rm H\,I}$ for ALFALFA relative to the values reported in this table.

[b] The reported ALFALFA $\Omega_{\rm H\,I}$ is corrected for H I self-absorption; the value without this correction is $3.50^{+0.51}_{-0.51}$.

[c] We list here the comoving volume probed by the MIGHTEE analysis of the COSMOS and XMM-LSS fields for our assumed cosmology, which is (a) larger than the 7×10^3 Mpc^3 "cosmological volume" reported in A. A. Ponomareva et al. (2023) that includes an unnecessary extra factor of h^{-3}, but (b) only slightly different from the volume actually used in that paper's $V_{\rm eff}$ analysis (A. Ponomareva 2024, private communication). Use of our slightly different survey volume in the context of a $V_{\rm eff}$ analysis would modestly change ϕ_* and $\Omega_{\rm H\,I}$ for MIGHTEE relative to the values reported in this table.

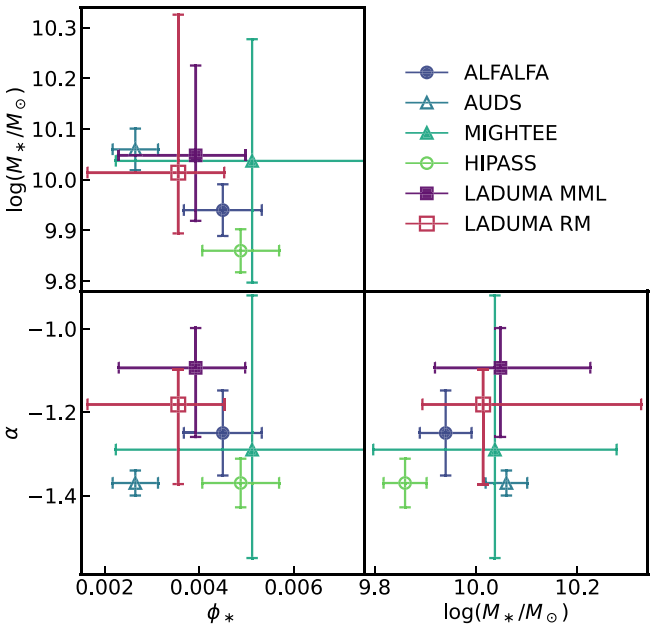


Figure 7. Schechter function parameters with associated uncertainties for different surveys. Square symbols represent LADUMA measurements. LADUMA's ϕ_* and M_* agree with the results for all of the previous surveys within 1σ uncertainties. LADUMA's low-mass slope is shallower than those measured in all previous surveys, although it is in agreement within the 1σ uncertainty with ALFALFA and MIGHTEE results. The values for these parameters are reported in Table 2, along with the literature references for previous surveys. The best-fit parameters from previous surveys have been rescaled to $H_0 = 70$ km s^{-1} Mpc^{-1}.

the lowest-mass bin, in which we recover any sources in our high-purity sample (specifically, two sources), is only about 2% of the total survey volume. We find that including this bin in our calculations for the best-fit Schechter function parameters would result in a significantly steeper value for α. Our decision to exclude this bin from our calculations reduces the impact of the extremely small volume of the lowest mass bin. Appendix B uses the ALFALFA α. 100 galaxy catalog to demonstrate that cosmic variance increases the uncertainty in the number of high-mass ($\log(M_{\rm H\,I}/M_\odot) \geqslant 9.75$) galaxies in the LADUMA volume compared to Poisson uncertainties alone, but the accessible volume subtleties discussed above preclude any extension of this analysis to individual Schechter function parameters.

We note that LADUMA does derive some benefit for HIMF derivation from its greater depth compared to previous surveys: as Figure 8 illustrates, the volume that is available to LADUMA for constraining the HIMF at $\log(M_{\rm H\,I}/M_\odot) \leqslant 8.5$ is only ~1 order of magnitude smaller than those that were available to HIPASS and ALFALFA, in contrast to the ~4 orders of magnitude advantage those surveys have over LADUMA in terms of available volume for the highest ($\log(M_{\rm H\,I}/M_\odot) \geqslant 10$) masses. From this point of view, the fact that LADUMA recovers a faint-end slope α flatter than (but consistent within the errors with) ALFALFA—and indeed matching the ALFALFA measurements of α in certain of that survey's subvolumes (M. G. Jones et al. 2018)—is not an unexpected result. At higher masses, LADUMA benefits in a different way from its greater depth, namely superior resilience against the systematic loss of high-inclination, high-mass sources at large distances within its observed sample, as demonstrated in Figure 9.

5.2. More Complex Models

The behavior of the observed LADUMA HIMF (see Figure 6), particularly the rolloff just below M_* followed by an upturn at lower masses, is reminiscent of features seen in optical/near-IR luminosity functions (LFs). These LFs can be effectively decomposed by galaxy type or color into multiple LFs, each represented by its own Schechter function (A. Sandage et al. 1985; J. Loveday et al. 2012; S. P. Driver et al. 2022). Previous studies (e.g., K. Said et al. 2019; M. G. Jones et al. 2020) have shown a dependency of the Schechter function parameters on environment, suggesting that a more complex

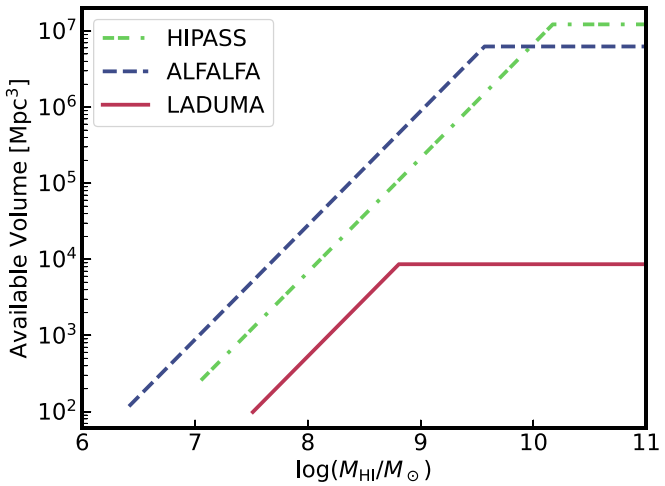


Figure 8. Volumes accessible for the detection of sources with given $M_{\rm H\,I}$ by LADUMA, ALFALFA, and HIPASS, calculated based on their respective detection limits. The maximum volume inside which *all* sources of H I mass $M_{\rm H\,I}$ are detected cannot be precisely defined, as it depends on parameters such as velocity width. However, we can roughly define an "accessible" volume beyond which a given survey in practice detects vanishingly few sources of H I mass $M_{\rm H\,I}$; these are the volumes plotted above. The number density distribution for sources of H I mass $M_{\rm H\,I}$ is sensitive to the distribution of sources (large-scale structure) inside the corresponding "accessible" volume. While all surveys have a much smaller (by several orders of magnitudes) available volume at lower masses compared to higher masses, that discrepancy is less significant for LADUMA compared to HIPASS and ALFALFA. The fact that the flat part of the LADUMA curve (corresponding to its *total* survey volume) extends farther left than do the flat parts of the HIPASS and ALFALFA curves is due to LADUMA's greater depth (and in spite of LADUMA's slightly larger redshift range) relative to the other two surveys.

functional form (e.g., different Schechter functions for high-density versus low-density environments) might be required to describe the overall HIMF accurately. Our forward-modeling approach is designed to accommodate various forms of mass functions beyond the traditional Schechter form, providing a flexible framework for analyzing the HIMF in greater detail for larger data sets. In addition to more complex functional forms, our approach can easily incorporate nonparametric models by directly sampling the intrinsic numbers of galaxies in different bins based on different priors, predicting corrected number densities for all bins. However, our current data set does not permit the fitting of models with more degrees of freedom, as the relatively small number of sources available cannot adequately constrain a more complex model.

5.3. Determining Completeness

This Section and the subsequent Section highlight two challenges in HIMF determination, illustrating the merits and limitations of the recovery matrix approach compared to traditional methods. The completeness of a survey for sources of a certain mass can be assessed through two primary methods. The effective volume ($V_{\rm eff}$) method estimates the survey's completeness for sources of mass M by evaluating the effective volume within which these sources can be observed.[42] Injection/recovery

[42] The effective volume ($V_{\rm eff}$) for sources of mass M is equal to the harmonic mean of the maximum volume ($V_{\rm max}$)—calculated by accounting for the dimming of the signal as the distance to the source increases and taking into account the sensitivity of the survey as a function of distance and position—for each source with mass M detected in the survey (D. Obreschkow et al. 2018). $V_{\rm max}$ is a function of $M_{\rm H\,I}$ and H I velocity width ($\Delta v_{\rm H\,I}$), while $V_{\rm eff}$ is a function of $M_{\rm H\,I}$ only.

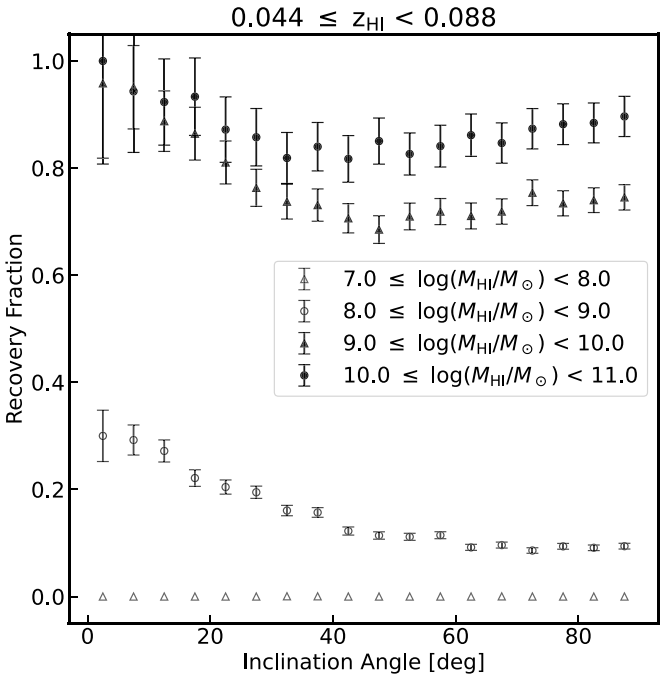


Figure 9. Recovery fraction as a function of source inclination for the higher-redshift half of LADUMA's low-z SPW. While the recovery fraction for low-mass sources decreases significantly at high inclinations, the recovery fraction for high-mass sources remains consistently high at all inclinations.

methods offer an alternative approach by introducing synthetic sources into the data cube and tracking their detection or nondetection using the same processes applied to real sources. These approaches can be encapsulated in expressions for the measured number density of sources of mass M, $\phi(M)$,

$$\phi(M)\, d\log M = \frac{n_o(M)}{C_s(M) V_s} \quad \text{(Injection/recovery)} \tag{8}$$

$$= \frac{n_o(M)}{V_{\rm eff}(M)}, \quad \text{(Effective volume)} \tag{9}$$

where $n_o(M)$ represents the number of sources detected in a bin of width $d\log M$ centered at M, $V_{\rm eff}(M)$ is the effective volume for sources with H I mass M, V_s is the total volume of the survey, and $C_s(M)$ is the completeness of the survey for sources with H I mass M (estimated via injection/recovery tests). By comparing Equations (8) and (9), it becomes evident that the effective volume is equivalent to the completeness multiplied by the total volume of the survey. The completeness C_s can be further decomposed as $C_s(M) = C_p(M) \cdot C_V(M)$, where $C_p(M)$ is the completeness based on the distribution of source properties, such as inclination, and $C_V(M)$ is the fraction of the total volume of the survey available for detection of sources of mass M. As described in Section 5.1 and demonstrated in Figure 8, the accessible volume for sources of mass M can vary drastically, e.g., by orders of magnitude, over the mass range of a survey. The completeness for low-mass sources is largely determined by the fraction of the total volume accessible to those sources and less so by the distribution of their intrinsic properties. The combination of these factors reveals that fluctuations (e.g., in the number of observed sources, which is sensitive to the distribution of the sources) in the small volume available to sources at the low-mass end of the mass range are amplified by the inverse of

the accessible volume fraction in the completeness correction process. Therefore, it is essential to limit the amplification of these fluctuations, which can be done by imposing a minimum accepted volume fraction (10% for this work) for the low versus the high end of the mass range.

Beyond the amplification by the inverse of the volume fraction, as laid out in detail in Appendix C, the accuracy of the completeness correction in the $V_{\rm eff}$ method is sensitive to the distribution of properties in the observed population. To demonstrate this point, we consider two different types[43] of sources of mass M: face-on and edge-on. For unresolved detections, it is clear that edge-on sources can only be detected in a smaller volume than otherwise identical face-on sources, since their larger line-of-sight velocity widths translate to lower peak flux densities. Let us represent the ratio of the volume available to face-on sources to the volume available to edge-on source as $R_{f/e}$. Assuming (unrealistically) that there are intrinsically the same numbers of edge-on and face-on sources in a survey volume, we would expect that for each edge-on source we detect, $R_{f/e}$ face-on sources should be detected. As described in Appendix C, the effective volume calculated for sources of mass M is sensitive to the *observed* ratio of the number of face-on to the number of edge-on sources. However, for an injection/recovery approach, this sensitivity would not affect the inferred completeness of the survey as long as the underlying distribution of sources is well understood,[44] since the completeness is determined through source injection.

5.4. Choice of Likelihood

Another challenge for HIMF determination involves accurately addressing uncertainties in the fitting process, particularly in bins with few or no detections. The data consist of observed source counts across various mass bins, each representing a count randomly drawn from a Poisson distribution. A traditional approach assumes that the counts are Gaussian-distributed with a Poisson uncertainty—computed as the square root of the number of sources in each bin—and uses a completeness correction factor to calculate the corrected number density of H I sources per bin. These values are then used to estimate the best-fit Schechter function parameters and their uncertainties. As outlined in Appendix A, using a Gaussian approximation for the full Poisson likelihood significantly biases the derived best-fit parameters of the Schechter function and tends to underestimate the 1σ uncertainties, especially in bins with few detections. This underestimate affects both bins where completeness is high but galaxies are few in number, i.e., at high masses, and bins where galaxies are plentiful but completeness is low, i.e., at low masses.[45] The root of this discrepancy lies in the asymmetric and positively skewed nature of the Poisson distribution. For a bin with a small number of detections, assuming Gaussian-distributed counts skews the estimated mean toward the observed count, leading to overfitting of the data. For LADUMA, this bias would manifest as an unrealistically steep low-mass slope, since (as is typical for any survey) the lowest-mass bins contain relatively few detections. In contrast to a traditional approach, the recovery matrix method employs forward modeling to predict the mean of the underlying Poisson distribution for each mass bin for any set of Schechter function parameters and implements a Poisson likelihood for model fitting and MCMC sampling. One significant advantage of forward modeling is its ability to integrate bins with zero detections into the fitting process, enhancing the robustness of the fit. As discussed in Section 4.1, for the MML method, we observe significantly larger uncertainties in the knee mass parameter when we exclude the zero-detection highest-mass bin. However, forward modeling naturally accommodates such cases, as it models uncertainties based on the predicted values and is compatible with the assumption of a Poisson distribution for the observed counts.

[43] Type can represent any property of a source that affects the volume within which it can be detected; one example would be inclination.

[44] It is noted that for an injection/recovery method, any bias in the assumed underlying distributions of source parameters would result in a similar effect and introduce additional uncertainty in the completeness fractions.

[45] In practice, surveys with smaller volumes and fewer total detections can mitigate this problem by using coarser mass bins, each containing a larger number of sources, when inferring HIMF parameters.

6. Summary and Conclusions

In this paper, we present a comprehensive analysis of the HIMF and the contribution of galaxies to $\Omega_{\rm H\,I}$ from a portion of the LADUMA survey covering $0 \leqslant z \leqslant 0.088$. Using a high-purity sample of H I detections from LADUMA's DR1, we develop a new "recovery matrix" method and benchmark it against a traditional maximum likelihood approach for measuring the HIMF. Our estimates of the Schechter function parameters are $\phi_* = 3.56^{+1.79}_{-1.51} \times 10^{-3}\,{\rm Mpc}^{-3}\,{\rm dex}^{-1}$, $\alpha = -1.18^{+0.14}_{-0.14}$, and $\log(M_*/M_\odot) = 10.01^{+0.23}_{-0.17}$. These values are in agreement with previous measurements of the HIMF over similar redshift ranges, enhancing confidence in the robustness of our results and the resilience of LADUMA's single-pointing geometry against the effects of cosmic variance.

Our methodology uses extensive synthetic source injections to correct for survey incompleteness and, for the first time, includes the effects of the continuum subtraction process. We account for varying sensitivity and completeness across the survey volume and mass range. In particular, our forward-modeling approach proves beneficial in handling bins with few or no detections (using an appropriate Poisson likelihood during the fitting and MCMC sampling process), thereby minimizing systematic effects on the derived Schechter function parameters. By using a Poisson likelihood instead of a Gaussian likelihood with Poisson uncertainties, we avoid overfitting the low-mass and high-mass bins of the Schechter function, which usually have few detections. Our analysis highlights the importance of cultivating a high-purity sample for reliably estimating survey completeness and avoiding biases in the recovery of mass function parameters. The recovery matrix allows for independent determination of the required number of injections per mass bin, ensuring reliable statistics across all bins—overcoming a challenge faced by traditional injection/recovery methods.

Looking forward, the LADUMA survey's future data releases will allow us to refine these measurements and potentially reveal new aspects of H I evolution across a broader redshift range (e.g., M. Hoosain et al. 2025, in preparation). The continued development and application of new methods like the recovery matrix will be crucial in leveraging the full potential of these data to enhance our understanding of galaxies' H I distributions and their role in galaxy evolution.

Acknowledgments

The MeerKAT telescope is operated by the South African Radio Astronomy Observatory (SARAO; www.sarao.ac.za),

which is a facility of the National Research Foundation (NRF), an agency of the Department of Science and Innovation. The authors thank the members of the SARAO engineering, commissioning, and science data processing teams for building and operating an absolutely superb facility. The MeerKAT data presented in this paper were processed using the ilifu cloud computing facility (www.ilifu.ac.za), a partnership between the University of Cape Town (UCT), the University of the Western Cape (UWC), the University of Stellenbosch, Sol Plaatje University, and the Cape Peninsula University of Technology. The ilifu facility is supported by contributions from the Inter-University Institute for Data Intensive Astronomy (IDIA, which is a partnership between the University of Cape Town, the University of Pretoria and the University of the Western Cape), the Computational Biology division at UCT, and the Data Intensive Research Initiative of South Africa (DIRISA). Data processing used pipelines that were developed at IDIA and are available at https://idia-pipelines.github.io.

The authors thank the anonymous referee for providing a constructive and helpful report and Matt Bershady, Barbara Catinella, Munira Hoosain, Renée Kraan-Korteweg, and Dominik Schwarz for helpful discussions. A.K.M. and A.J.B. acknowledge support from NSF grants AST-1814421 and AST-2308161. A.K.M. thanks the LSST-DA Data Science Fellowship Program, which is funded by LSST-DA, NSF Cybertraining grant No. 1829740, the Brinson Foundation, and the Moore Foundation; his participation in the program has benefited this work. A.J.B. acknowledges support from the Radcliffe Institute for Advanced Study at Harvard University. E.G. acknowledges support from NSF grant AST-2206222. D.O. is a recipient of an Australian Research Council Future Fellowship (FT190100083) funded by the Australian Government. M.G. is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (DP210102103), by UK STFC Grant ST/Y001117/1, and by IDIA. M.J.J. acknowledges support from the Oxford Hintze Centre for Astrophysical Surveys (funded through generous support from the Hintze Family Charitable Foundation), from Science and Technology Council (STFC) consolidated grants [ST/S000488/1] and [ST/W000903/1], and from a United Kingdom Research and Innovation (UKRI) Frontiers Research Grant [EP/X026639/1] selected by the European Research Council. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript (AAM) version arising from this submission.

Appendix A
Likelihood Selection

In order to fit a model to the data or sample the posterior probability distribution of the model parameters using an MCMC method, we need to define a likelihood function. It is common practice to minimize the mean squared error (MSE) between the data and the model, selecting the parameters that minimize the MSE as the best-fit model. Minimizing the MSE is equivalent to maximizing the Gaussian likelihood, under the assumption that the errors on the data are Gaussian-distributed and are determined based on the observed counts (D. W. Hogg et al. 2010). However, this assumption can lead to biased results when dealing with Poisson-distributed data, especially when counts are low. We summarize the inferred Schechter function parameters and their uncertainties for different likelihood functions in Table 3. It is well established that for large mean values, the Poisson distribution can be approximated as a Gaussian distribution with the same mean and variance; hence, for large mean values, maximizing the Gaussian likelihood is an excellent approximation to maximizing the Poisson likelihood. Similarly, the difference between minimizing the MSE and maximizing the Gaussian likelihood becomes small for large mean values. As shown in the likelihood column of Table 3, the difference between the MSE and Gaussian likelihoods can be clarified by comparing how errors are calculated: Poisson error based on observed data[46] ($\sqrt{y}$ for an observed count of y) versus Poisson error based on model mean ($\sqrt{\mu}$ for a predicted mean of μ). For instance, if the observed count deviates from the model mean by one standard deviation, i.e., $y = \mu + \sqrt{\mu}$, the relative error ($\frac{\sqrt{y}}{y}$) estimated from the observed data is smaller by a factor of $\sqrt{1 + \frac{1}{\sqrt{\mu}}}$ compared to the relative error estimated from the mean ($\frac{\sqrt{\mu}}{\mu}$). This computation illustrates that for large mean values, the discrepancy between model-based and data-based error estimates becomes small (e.g., for $\mu = 25$, the discrepancy for an observed count that deviates from the mean by one standard deviation is only ~10%), making the MSE an effective approximation of the Gaussian likelihood. However, for cases with small means, the likelihood of larger-than-mean observations is significantly smaller for the Gaussian distribution compared to the Poisson distribution, and the full Poisson likelihood is required to ensure unbiased results. Using Gaussian or MSE likelihoods instead of a Poisson likelihood disproportionately influences the overall likelihood of the fit due to the impact of overfitting the bins with few detections; we therefore use the Poisson distribution in deriving our results with the recovery matrix method.

Table 3
Inferred Schechter Function Parameters for Different Likelihood Functions

	Negative log Likelihood	$\phi_* \times 10^3$ (Mpc^{-3})	$\log(M_*/M_\odot)$	α
MSE	$\sum_{i=1}^{N_{\rm bins}}(\mu_i - y_i)^2/\mu_i$	$2.87^{+0.49}_{-1.17}$	$10.09^{+0.23}_{-0.07}$	$-1.26^{+0.05}_{-0.12}$
Gaussian	$\sum_{i=1}^{N_{\rm bins}}\ln(\mu_i) + (\mu_i - y_i)^2/\mu_i$	$3.44^{+0.66}_{-1.17}$	$9.99^{+0.18}_{-0.08}$	$-1.21^{+0.06}_{-0.11}$
Poisson	$\sum_{i=1}^{N_{\rm bins}}\mu_i - \ln(\mu_i)y_i$	$3.56^{+0.97}_{-1.92}$	$10.01^{+0.31}_{-0.12}$	$-1.18^{+0.08}_{-0.19}$

Note. μ_i and y_i represent the predicted and observed counts, respectively.

[46] When errors are calculated based on observed data, the error term in the overall likelihood is a constant—$\sum_{i=1}^{N_{\rm bins}}\ln(y_i)$ using the terminology of Table 3—and can be factored out of the likelihood sum.

Appendix B
Uncertainty due to Cosmic Variance

To provide an indication of how cosmic variance affects our results, we use the ALFALFA α. 100 H I source catalog (M. P. Haynes et al. 2018) to estimate how many high-mass H I galaxies would be detected within volumes comparable to that of LADUMA's low-z SPW cube. We begin by choosing a subsample of the ALFALFA catalog with high completeness, by excluding sources with $\log(M_{\rm H\,I}/M_\odot) < 9.75$ and $z > 0.05$. These cuts minimize the impact of survey incompleteness and result in a sample of over 8500 galaxies. To ensure a fair comparison with the low-z SPW for LADUMA, we define 388 independent "pencil-beam" subvolumes within the ALFALFA volume; each of these is matched to the volume of the LADUMA low-z SPW by selecting a larger solid angle on the sky to compensate for the smaller redshift coverage of ALFALFA. Within each defined subvolume, we count the number of galaxies. We then calculate the standard deviation in galaxy counts across these subvolumes, which is found to be ~8.5 (relative to a mean of 16). This observed standard deviation exceeds the expected Poisson uncertainty of $\sqrt{16} = 4$, providing an indication of the impact of cosmic variance on our results. In our LADUMA sample, the observed number of high-mass sources is 15, closely matching the average number derived from the ALFALFA test. While this exercise suggests that Poisson uncertainties represent only ~47% of the total uncertainty when cosmic variance is included, the facts that uncertainties are not symmetrically distributed for the Schechter function parameters and that the ALFALFA sample does not provide a sufficiently robust estimate of cosmic variance for low-mass galaxies (due to the smaller available volumes) mean that extending this analysis to individual Schechter function parameters is nontrivial.

Appendix C
Comparing Maximum Volume and Injection/Recovery Methods

Considering the intrinsic number density, ϕ_t, of sources with H I mass M, i.e., the value of the HIMF at $M_{\rm H\,I} = M$, we can write:

$$\phi_t(M)\, d\log M = \frac{n_t}{V_s}, \tag{C1}$$

where $d\log M$ is the width of a bin centered at M, and n_t is the intrinsic number of sources (not necessarily an integer) within a survey's volume, denoted by V_s.[47] We can express the observed number density of sources with H I mass M in a survey using a $V_{\rm max}$ (VM) framework as

$$\phi_{\rm V\,M}(M)\, d\log M = \frac{n_o}{V_{\rm eff}} \tag{C2}$$

and using an injection/recovery (IR) framework as

$$\phi_{\rm IR}(M)\, d\log M = \frac{n_o}{C_s V_s}, \tag{C3}$$

where n_o represents the integer number of sources detected in an interval of width $d\log M$ centered at M, which can be described as a randomly sampled (integer) value from a Poisson distribution with mean $\mu = C_s n_t$, i.e., $n_o \sim P(\mu = C_s n_t)$; $V_{\rm eff}$ is the effective volume of sources with H I mass M (estimated using the observed sample); and C_s is the overall completeness of the survey for all types of sources with H I mass M (estimated via injection/recovery tests). Comparing Equations (C2) and (C3) shows that the $V_{\rm max}$ approach estimates the completeness of a survey by calculating the effective volume ($V_{\rm eff}$) using the observed source population. It is clear that any error introduced by the randomness of the observed population affects the numerators in Equations (C2) and (C3) identically. For the denominators, this analogy breaks down. The denominator in Equation (C3) is not affected by errors due to the randomness of the observed population, as V_s is fixed and the completeness, C_s, is determined via injection/recovery tests. However, the denominator in Equation (C2)—when determined using the observed population—will be affected by errors due to the randomness in the observed sample. It is helpful to track how random Poisson uncertainties affect the observed number density for these two methods. For this exercise, we assume that the uncertainties introduced by the injection/recovery process are negligible (as they can be reduced by increasing the number of injected sources) compared to the random uncertainties.

First, we need to consider the different "types" of galaxies with $M_{\rm H\,I} = M$. We assume there are N types of galaxies with normalized abundances of x_i, where $\Sigma_{i=1}^{N} x_i = 1$; one example would be galaxies with the same mass but different inclination angles. Assuming a completeness fraction of C_i[48] for each source type, the expected number of sources $\langle n_o \rangle$ in the survey volume would be

$$\langle n_o \rangle = \sum_{i=1}^{N} \phi_t x_i C_i V_s \tag{C4}$$

$$= \sum_{i=1}^{N} \phi_i V_i = \sum_{i=1}^{N} \langle n_i \rangle \tag{C5}$$

$$= \phi_t V_s \sum_{i=1}^{N} x_i C_i = \phi_t V_s C_s = n_t C_s, \tag{C6}$$

where $\phi_i = \phi_t x_i$, $V_i = C_i V_s$, and $C_s = \Sigma_{i=1}^{N} x_i C_i$ are, respectively, the intrinsic number density of sources of type i, the effective volume available for the detection of sources of type i, and the completeness of the survey for sources of H I mass M, and $\langle n_i \rangle$ is the expected number of sources of type i in the survey volume. Equations (C5) and (C6) are two representations of the expected number of sources: the sum of the expected numbers of sources of type i (C5) and the intrinsic number of sources multiplied by the completeness of the survey (C6). The number of observed sources n_o can be expanded as

$$\begin{aligned} n_o = \langle n_o \rangle + \delta n_o &= \sum_{i=1}^{N} \langle n_i \rangle + \sum_{i=1}^{N} \delta n_i \\ &= \sum_{i=1}^{N} (\langle n_i \rangle + \delta n_i) = \sum_{i=1}^{N} n_i, \end{aligned} \tag{C7}$$

where δn_o and δn_i (which may be positive or negative) are, respectively, the random uncertainty in the total number of

[47] We use a t subscript on ϕ_t and n_t to evoke the "true" values of the mass function and the number of sources.

[48] This completeness fraction C_i for sources of type i can be interpreted as the fraction of the total volume available for the detection of sources of type i, i.e., $V_i = C_i V_s$.

observed sources and the random uncertainty in the number of observed sources of type i, and n_i is the number of observed sources of type i.[49]

In an injection/recovery framework, by combining Equations (C3), (C6), and (C7), we have:

$$\phi_{\rm IR}(M)\, d\log M = \frac{n_o}{C_s V_s} = \frac{\langle n_o\rangle + \delta n_o}{C_s V_s} = \frac{\langle n_o\rangle}{C_s V_s} + \frac{\delta n_o}{C_s V_s}$$
$$= \frac{\phi_t C_s V_s}{C_s V_s} + \frac{\delta n_o}{C_s V_s} \quad \text{(C8)}$$
$$= \phi_t + \frac{\delta n_o}{C_s V_s}, \quad \text{(C9)}$$

showing that the uncertainty in the observed number density is proportional to the random uncertainty in the total number of observed sources with $M_{\rm H\,I} = M$. When the number of observed sources is equal to the expected number of sources $n_o = \langle n_o\rangle$, i.e., $\delta n_o = \Sigma_{i=1}^{N}\delta n_i = 0$, the observed number density is equal to the intrinsic number density. Similarly, in a $V_{\rm max}$ framework, we can combine Equations (C2), (C5), and (C7) to get:

$$\phi_{\rm V\,M}(M)\, d\log M = \frac{n_o}{V_{\rm eff}} = \sum_{i=1}^{N}\frac{n_i}{V_i} = \sum_{i=1}^{N}\frac{\langle n_i\rangle + \delta n_i}{V_i} \quad \text{(C10)}$$
$$= \sum_{i=1}^{N}\frac{\langle n_i\rangle}{V_i} + \sum_{i=1}^{N}\frac{\delta n_i}{V_i} = \sum_{i=1}^{N}\frac{\phi_i V_i}{V_i} + \sum_{i=1}^{N}\frac{\delta n_i}{V_i} \quad \text{(C11)}$$
$$= \sum_{i=1}^{N} x_i\phi_t + \sum_{i=1}^{N}\frac{\delta n_i}{V_i} = \phi_t\sum_{i=1}^{N} x_i + \sum_{i=1}^{N}\frac{\delta n_i}{V_i} \quad \text{(C12)}$$
$$= \phi_t + \sum_{i=1}^{N}\frac{\delta n_i}{V_i}, \quad \text{(C13)}$$

which shows that the uncertainty in the measured number density is proportional to the sum of the random uncertainties in the numbers of observed sources with $M_{\rm H\,I} = M$ of type i divided by the effective volumes for sources of type i. Equation (C13) shows that when $\Sigma_{i=1}^{N}\frac{\delta n_i}{V_i} = 0$, the measured number density would be equal to the intrinsic number density. The trivial case is when $\delta n_i = 0$ for all source types, i.e., for all source types, the number of observed sources of type i is equal to expected number of sources of type i, $n_i = \langle n_i\rangle$. The more complicated scenario will be for all of the nonzero $\{\delta n_i\}$ in the sum to cancel each other. An interesting case is to compare Equations (C9) and (C13) in the limiting case of a volume-limited sample for all sources of type i. For a volume-limited sample, the effective volume for each source type is equal to the total volume of the survey, i.e., $V_i = V_s$, and the completeness for sources of type i is equal to 1, as all such sources can be detected throughout the total volume, i.e., $C_i = 1$. Substitution of these values into Equations (C9) and (C13) results in

$$\text{C9} \rightarrow \ \phi_{\rm IR}(M)\, d\log M = \phi_t + \frac{\delta n_o}{C_s V_s} = \phi_t + \frac{\delta n_o}{V_s} \quad \text{(C14)}$$

[49] The sum, $Z = X + Y$, of two independent Poisson random variables, $X \sim \text{Poisson}(\mu_x)$ and $Y \sim \text{Poisson}(\mu_y)$, is also a Poisson random variable, $Z \sim \text{Poisson}(\mu_x + \mu_y)$.

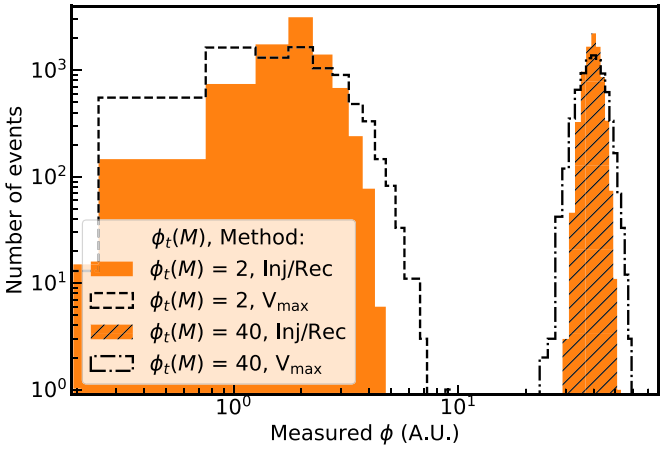


Figure 10. Distributions of observed number density of sources with $M_{\rm H\,I} = M$ for the $V_{\rm max}$ and injection/recovery methods. It is clear that the $V_{\rm max}$ method leads to a wider distribution of observed number densities around the true number density of sources with $M_{\rm H\,I} = M$. This effect is due to the fact that the $V_{\rm max}$ method uses the observed data set to estimate the completeness of the survey and therefore is affected by uncertainties in the observed population to a larger degree than the injection/recovery method. It is important to note that these differences are inherent in the methods themselves, and this figure assumes perfect knowledge of the completeness of the survey and the effective volume for the (two) different types of galaxies of H I mass M.

$$\text{C13} \rightarrow \phi_{\rm V\,M}(M)\, d\log M$$
$$= \phi_t + \sum_{i=1}^{N}\frac{\delta n_i}{V_i} = \phi_t + \sum_{i=1}^{N}\frac{\delta n_i}{V_s}$$
$$= \phi_t + \frac{\Sigma_{i=1}^{N}\delta n_i}{V_s} = \phi_t + \frac{\delta n_o}{V_s}, \quad \text{(C15)}$$

making it clear that when all types of sources of H I mass M are detectable throughout the entire volume of the survey, the two approaches are equivalent. We note that in these derivations we have assumed that the effective volume and the completeness of the survey are perfectly determined, underscoring the fact that the difference between these two approaches arises from the methods themselves and are not due to imperfect calculations. In practice, the effective volume and the completeness fraction calculations are affected by errors that contribute to the overall uncertainties in the measured number density distributions.

Considering a simplified case of two types of galaxies of equal abundance and identical H I mass M, following an approach similar to Appendix C of D. Obreschkow et al. (2018), we demonstrate how random uncertainties affect the observed number density of sources with $M_{\rm H\,I} = M$. Assuming that $\phi_t(M) = 2$, $x_1 = x_2 = 0.5$, $C_1 = 1/8$, $C_2 = 1$, and $V_s = 8$, we can calculate the expected number of galaxies as

$$\langle n_o\rangle = \sum_{i=1}^{N}\langle n_i\rangle \quad \text{(C16)}$$
$$= \sum_{i=1}^{N}\phi_t x_i C_i V_s \quad \text{(C17)}$$
$$= \phi_t x_1 C_1 V_s + \phi_t x_2 C_2 V_s = \langle n_1\rangle + \langle n_2\rangle \quad \text{(C18)}$$
$$= 2\cdot\tfrac{1}{2}\cdot\tfrac{1}{8}\cdot 8 + 2\cdot\tfrac{1}{2}\cdot 1\cdot 8 \quad \text{(C19)}$$

$$= 1 + 8 = 9. \quad \text{(C20)}$$

Therefore, we expect to observe one galaxy of type 1 and eight galaxies of type 2, totaling to nine galaxies of $M_{\rm H\,I} = M$ in the survey volume. We can randomly sample two Poisson distributions with means of $\mu_1 = 1$ and $\mu_2 = 8$ to simulate observations and calculate the observed number density of sources with $M_{\rm H\,I} = M$ for the two approaches using Equations (C10) and (C8). A histogram of the outcome of 2^{13} simulated observations for cases of $\phi_t(M) = 2$ and $\phi_t(M) = 40$ is shown in Figure 10, demonstrating that the $V_{\rm max}$ approach leads to a wider distribution of observed number densities around the true number density of sources with $M_{\rm H\,I} = M$, due to the fact that the $V_{\rm max}$ approach uses the observed data set to estimate the completeness of the survey and therefore is affected by the uncertainties in the observed population to a larger degree compared to the injection/recovery approach.

We close by noting that the preceding analysis treats the parameter (M) whose distribution (the HIMF) we are trying to derive in a fundamentally different way from other parameters (e.g., inclination) that determine which "types" different galaxies have, in the sense of affecting the maximum volumes in which they can be detected. This distinction is not absolute: if we were interested (for example) in deriving the H I velocity width function, we could do so treating M as a parameter that determines "type," as it affects the effective available volume for the detection of sources of a given velocity width, and the limitations of the $V_{\rm max}$ approach relative to the injection/recovery approach would remain the same as we have characterized them above.

ORCID iDs

Amir Kazemi-Moridani https://orcid.org/0000-0003-4496-9553
Andrew J. Baker https://orcid.org/0000-0002-7892-396X
Marc Verheijen https://orcid.org/0000-0001-9022-8081
Eric Gawiser https://orcid.org/0000-0003-1530-8713
Sarah-Louise Blyth https://orcid.org/0000-0002-5777-0036
Danail Obreschkow https://orcid.org/0000-0002-1527-0762
Laurent Chemin https://orcid.org/0000-0002-3834-7937
Jordan D. Collier https://orcid.org/0000-0002-2326-7432
Kyle W. Cook https://orcid.org/0000-0002-4012-779X
Jacinta Delhaize https://orcid.org/0000-0002-6149-0846
Ed Elson https://orcid.org/0000-0001-9359-0713
Bradley S. Frank https://orcid.org/0000-0003-3599-1521
Marcin Glowacki https://orcid.org/0000-0002-5067-8894
Kelley M. Hess https://orcid.org/0000-0001-9662-9089
Benne W. Holwerda https://orcid.org/0000-0002-4884-6756
Zackary L. Hutchens https://orcid.org/0000-0002-8574-5495
Matt J. Jarvis https://orcid.org/0000-0001-7039-9078
Melanie Kaasinen https://orcid.org/0000-0002-1173-2579
Sphesihle Makhathini https://orcid.org/0000-0001-9565-9622
Abhisek Mohapatra https://orcid.org/0000-0001-6047-4521
Hengxing Pan https://orcid.org/0000-0002-9160-391X
Anja C. Schröder https://orcid.org/0000-0002-2017-584X
Mattia Vaccari https://orcid.org/0000-0002-6748-0577
Tobias Westmeier https://orcid.org/0000-0002-5300-2486
John F. Wu https://orcid.org/0000-0002-5077-881X
Martin Zwaan https://orcid.org/0000-0003-0101-1804

References

Abell, G. O. 1958, ApJS, 3, 211
Adams, E. A. K., Adebahr, B., de Blok, W. J. G., et al. 2022, A&A, 667, A38
Blyth, S., Baker, A. J., Holwerda, B., et al. 2018, in MeerKAT2016: On the Pathway to the SKA (Trieste: PoS), 004
Catinella, B., & Cortese, L. 2015, MNRAS, 446, 3526
Chung, A., van Gorkom, J. H., Kenney, J. D. P., Crowl, H., & Vollmer, B. 2009, AJ, 138, 1741
Collier, J. D., Frank, B., Sekhar, S., & Taylor, A. R. 2021, in 2021 XXXIVth General Assembly and Scientific Symp. of the Int. Union of Radio Science (URSI GASS) (Piscataway, NJ: IEEE), 701
Courteau, S. 1997, AJ, 114, 2402
Davé, R., Crain, R. A., Stevens, A. R. H., et al. 2020, MNRAS, 497, 146
Driver, S. P., Bellstedt, S., Robotham, A. S. G., et al. 2022, MNRAS, 513, 439
Euclid Collaboration, Moneti, A., McCracken, H. J., et al. 2022, A&A, 658, A126
Fernández, X., Gim, H. B., van Gorkom, J. H., et al. 2016, ApJL, 824, L1
Fernández, X., van Gorkom, J. H., Hess, K. M., et al. 2013, ApJL, 770, L29
Giovanelli, R., & Haynes, M. P. 1985, ApJ, 292, 404
Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2005, AJ, 130, 2598
Gogate, A. 2022, PhD thesis, Univ. of Groningen 10.33612/diss.206445119
Gogate, A. R., Verheijen, M. A. W., Deshev, B. Z., et al. 2020, MNRAS, 496, 3531
Haynes, M. P., Giovanelli, R., & Chincarini, G. L. 1984, ARA&A, 22, 445
Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49
Haynes, M. P., Giovanelli, R., Martin, A. M., et al. 2011, AJ, 142, 170
Hess, K. M., Luber, N. M., Fernández, X., et al. 2019, MNRAS, 484, 2234
Högbom, J. A. 1974, A&AS, 15, 417
Hogg, D. W., Bovy, J., & Lang, D. 2010, arXiv:1008.4686
Holwerda, B. W., Pirzkal, N., Cox, T. J., et al. 2011, MNRAS, 416, 2426
Hoppmann, L., Staveley-Smith, L., Freudling, W., et al. 2015, MNRAS, 452, 3726
Hotan, A. W., Bunton, J. D., Chippendale, A. P., et al. 2021, PASA, 38, e009
Jaffé, Y. L., Poggianti, B. M., Verheijen, M. A. W., Deshev, B. Z., & van Gorkom, J. H. 2013, MNRAS, 431, 2111
Jarvis, M., Taylor, R., Agudo, I., et al. 2016, in MeerKAT Science: On the Pathway to the SKA, ed. R. Taylor (Trieste: SISSA), 6
Jonas, J. L. 2009, in Proc. IEEEP 97 (Piscataway, NJ: IEEE), 1522
Jones, M. G., Haynes, M. P., Giovanelli, R., & Moorman, C. 2018, MNRAS, 477, 2
Jones, M. G., Hess, K. M., Adams, E. A. K., & Verdes-Montenegro, L. 2020, MNRAS, 494, 2090
Jones, M. G., Papastergis, E., Haynes, M. P., & Giovanelli, R. 2015, MNRAS, 449, 1856
Kim, H.-S., Wyithe, J. S. B., Power, C., et al. 2015, MNRAS, 453, 2315
Loveday, J., Norberg, P., Baldry, I. K., et al. 2012, MNRAS, 420, 1239
Maddox, N., Frank, B. S., Ponomareva, A. A., et al. 2021, A&A, 646, A35
Martin, A. M., Papastergis, E., Giovanelli, R., et al. 2010, ApJ, 723, 1359
Martinsson, T. P. K., Verheijen, M. A. W., Bershady, M. A., et al. 2016, A&A, 585, A99
McClure-Griffiths, N. M., Stanimirović, S., & Rybarczyk, D. R. 2023, ARA&A, 61, 19
Meyer, M., Robotham, A., Obreschkow, D., et al. 2017, PASA, 34, 52
Meyer, M. J., Zwaan, M. A., Webster, R. L., et al. 2004, MNRAS, 350, 1195
Mihos, J. C. 2001, ApJ, 550, 94
Moorman, C. M., Vogeley, M. S., Hoyle, F., et al. 2014, MNRAS, 444, 3559
Moster, B. P., Somerville, R. S., Newman, J. A., & Rix, H.-W. 2011, ApJ, 731, 113
Nan, R. 2008, Proc. SPIE, 7012, 70121E
Obreschkow, D., Murray, S. G., Robotham, A. S. G., & Westmeier, T. 2018, MNRAS, 474, 5500
Persic, M., Salucci, P., & Stel, F. 1996, MNRAS, 281, 27
Ponomareva, A. A., Jarvis, M. J., Pan, H., et al. 2023, MNRAS, 522, 5308
Popping, G., Somerville, R. S., & Trager, S. C. 2014, MNRAS, 442, 2398
Reynolds, T. N., Catinella, B., Cortese, L., et al. 2022, MNRAS, 510, 1716
Rhee, J., Meyer, M., Popping, A., et al. 2023, MNRAS, 518, 4646
Roberts, H., Darling, J., & Baker, A. J. 2021, ApJ, 911, 38
Rogstad, D. H., Lockhart, I. A., & Wright, M. C. H. 1974, ApJ, 193, 309
Said, K., Kraan-Korteweg, R. C., & Staveley-Smith, L. 2019, MNRAS, 486, 1796
Sandage, A., Binggeli, B., & Tammann, G. A. 1985, AJ, 90, 1759
Schechter, P. 1976, ApJ, 203, 297
Sérsic, J. L. 1968, Atlas de Galaxias Australes (Cordoba: Observatorio Astronomico)
Tully, R. B., & Fisher, J. R. 1977, A&A, 54, 661

van der Hulst, J. M., Terlouw, J. P., Begeman, K. G., Zwitser, W., & Roelfsema, P. R. 1992, in ASP Conf. Ser. 25, Astronomical Data Analysis Software and Systems I, ed. D. M. Worrall, C. Biemesderfer, & J. Barnes (San Francisco, CA: ASP), 131
Wang, J., Koribalski, B. S., Serra, P., et al. 2016, MNRAS, 460, 2143
Westmeier, T., Kitaeff, S., Pallot, D., et al. 2021, MNRAS, 506, 3962
Xi, H., Peng, B., Staveley-Smith, L., et al. 2024, ApJL, 966, L36
Xi, H., Staveley-Smith, L., For, B.-Q., et al. 2021, MNRAS, 501, 4550
Zwaan, M. A., Briggs, F. H., Sprayberry, D., & Sorar, E. 1997, ApJ, 490, 173
Zwaan, M. A., Meyer, M. J., Staveley-Smith, L., & Webster, R. L. 2005, MNRAS, 359, L30
Zwaan, M. A., Staveley-Smith, L., Koribalski, B. S., et al. 2003, AJ, 125, 2842