Investigating the Epoch of Galaxy Formation using Artificial Intelligence

by

Bradley Anthony Ward

A Thesis submitted to Cardiff University
for the degree of Doctor of Philosophy

December 2023

Abstract

“ 5. Preliminary Pages 5.1 Each copy of the thesis must contain the following required pages: .1 a title page; **.2 a summary of no more than 300 words;** .3 a list of contents, which includes the page number for each chapter and sub-division listed; .4 acknowledgments, where this is an expectation of your sponsor ”
as taken from Submission-and-Presentation-of-Research-Degree-Theses.pdf — Policy on the Submission and Presentation of Research Degree Theses – v 5.0 – Date of Effect 01.08.2022 – accessed 22/09/2022

Publications

Preface

Keep in mind the wise words of Louise Winter (PGR Admin School of PHYSX (as of 2022))
“if you have material that has been published in journals in your thesis you are advised to seek permission from the relevant journal to publish in your thesis for the final version to be published on ORCA – take a little time to read this page. If you are uncertain about Copyright you should contact Copyright@cardiff.ac.uk as we are not experts on copyright. ”
`https://intranet.cardiff.ac.uk/students/study/postgraduate-research-support/thesis-and-examinations/submitting-your-thesis/copyright-and-your-ethesis`

First Author Publications

authors, et al. year; *Title*, Submitted to MNRAS

Co-Author Publications

, In Prep

Contents

"A dedication quote/sentence"

Acknowledgements

“A quote"

By whom *From what source*

Funding Bodies and Affiliations

Insert thanks to funding bodies, and affiliations (e.g. partner company, CDT, governments, University funders, grants/awards, etc. . . .) There are no strict rules on logos here — use at your discretion.

Personal Thanks

Insert personal thanks, e.g. supervisors, friends, family, colleagues. Keep it proffessional.
“8. Acknowledgements/Dedications 8.1 It is understandable that you will have benefitted from the support of family, friends and peers whilst undertaking your studies and you may wish to express your thanks in an acknowledgements/dedication section of your thesis. If you choose to do this, you should be mindful of the nature and tone of your comments as they may be read widely, including by future employers, funders or other colleagues. 8.2 For data protection purposes, you should not include any private personal details or other confidential information about yourself or others in the acknowledgements section or elsewhere in the thesis.” — University (`https://intranet.cardiff.ac.uk/students/study/postgraduate-research-support/thesis-and-examinat preparing-your-thesis`) Submission-and-Presentation-of-Research-Degree-Theses V5.0 September 2022

Chapter 1

Introduction

1.1 A title

Chapter 2

Herschel-ATLAS Data Release III

2.1 The Herschel-ATLAS

The *Herschel* Astrophysical Terahertz Large Area Survey (H-ATLAS; Eales et al. 2010) was the largest open-time sub-mm survey carried out with *Herschel*. The survey was observed across five photometric bands using two instruments onboard the *Herschel Space Observatory*: the Photodetector Array Camera (PACS, Poglitsch et al. 2010) at 100 and 160 μm, and the Spectral and Photometric Imaging Receiver (SPIRE, Griffin et al. 2010) at 250, 350 and 500 μm. Compared to the first SMGs detected using SCUBA at 850 μm(Smail et al. 1997; Barger et al. 1998; Hughes et al. 1998), the PACS and SPIRE wavebands span the peak of the infrared spectrum for low redshift (z < 1) galaxies. Their intrinsic brightness at the SPIRE wavelengths makes their detection in the thousands more achievable. The main scientific goal of the survey was to estimate the dust masses and dust obscured star formation rates for thousands of nearby galaxies over a large area of sky. While the intention was for a shallow survey, the surprising sensitivity of *Herschel* and the negative k-correction observed at the operating wavelengths of the SPIRE instrument (Blain & Longair 1993) means that many sources were observed at higher redshifts, with a median of z $\sim$ 1. The catalogues of the survey, as detailed below, includes sources with redshifts up to $\sim$ 6 (Amblard et al. 2010; Lapi et al. 2011; Fudamoto et al. 2017; Zavala et al. 2018).
The complete survey covers $\sim 660\,\mathrm{deg}^2$, split into three regions located to avoid emission from Galactic dust and to utilize complimentary spectroscopic surveys including the Sloan Digital Sky Survey (SDSS, York et al. 2000), the 2df Galaxy Redshift Survey (2dfGRS, Colless et al. 2001) and the Galaxy and Mass Assembly (GAMA, Driver et al. 2009). The North Galactic Pole (NGP) region covers $\sim 180\,\mathrm{deg}^2$ of the northern sky, centered at R.A 13^h18^m and declination +29°13' (J2000); three equatorial fields, located at approximately R.A 9^h, 12^h and 15^h coinciding with the GAMA survey (henceforth named GAMA9, GAMA12 and GAMA15 fields), each with an area of approximately $54\,\mathrm{deg}^2$, and the South Galactic Pole (SGP) region, centered at R.A 0^h6^m and declination -32°44' (J2000) with an area of $\sim 318\,\mathrm{deg}^2$.

2.1.1 Detecting Submillimeter Sources on Herschel Images

uation:
uation:
uation:

Due to [...] sub-mm images suffer from two types of noise; instrumental noise [...] and confusion noise which is highly correlated between pixels, most of its contribution coming from the blending together of faint sources. Source confusion is of particular importance to sub-mm surveys [...]. The result of combining instrumental noise with confusion noise is that almost all sources in the Herschel images are unresolved and the optimum filter for detecting these unresolved sources is no longer the point spread function (PSF). Consider a *Herschel* map in which there is only one source of noise: an image with instrumental noise but no confusion noise (i.e. there is only one point source and no fainter, confusing sources), the optimal detection of this source is obtained by convolving the image with the PSF of the instrument. On the other hand, a map with no instrumental noise, but many confused point sources would be optimally detected with its best signal to noise ratio (SNR) by taking the Fourier transform of the image, dividing by the Fourier transform of the PSF and taking the inverse Fourier transform to obtain a perfect deconvolution of the original map (Valiante et al. 2016). For images that have a variable ratio of instrumental to confusion noise like the *Herschel* images of H-ATLAS, Chapin et al. 2011 showed that a convolving function or "matched filter" can be calculated to provide the maximum SNR for an unresolved source.

uation:
tion of
ed filters

To detect H-ATLAS sources from the 250 µmmaps using a matched filter (the 250 µmband is the most sensitive of the SPIRE bands and given the lower sensitivity of the PACS instrument, all sources detected on the PACS images would also be detected on the SPIRE 250 µmimage), Maddox & Dunne 2020 developed a source detection algorithm called the Multi-band Algorithm for Source Detection and eXtraction (MADX). The MADX algorithm works in the following way. Firstly, Galactic dust emission is removed from the images using `Nebuliser`. Next, the images are convolved with the matched filter [...]. The variance map is created by convolving the map of variance in instrumental noise with the matched filter and adding the confusion noise. It is from this map that the SNR of a detected source is determined. The same process is repeated with the 350 and 500 µmmaps and interpolated to the same pixel scale as the 250 µmmaps. The detection map used to extract sources is then generated from a weighted sum of the three SPIRE maps, however, due to the smaller PSF at 250 µmwhich leads to more accurate positions and the increased number of sources when using the 250 µmmaps, zero weighting is given to the 350 and 500 µmimages. This has the effect of making the detection map the same as the 250µmmap.

Sources are identified by peak values $> 2.5\sigma$ in the filtered detection map. Their positions are estimated by fitting a Gaussian to the nearest pixels surrounding the location of the peak. The source is extracted in the other *Herschel* wavebands at the 250 µmposition. Due to the high levels of confusion and high source density on the SPIRE maps, the flux density estimates in each band can be biased by blending with other sources. The MADX algorithm negates some of this problem

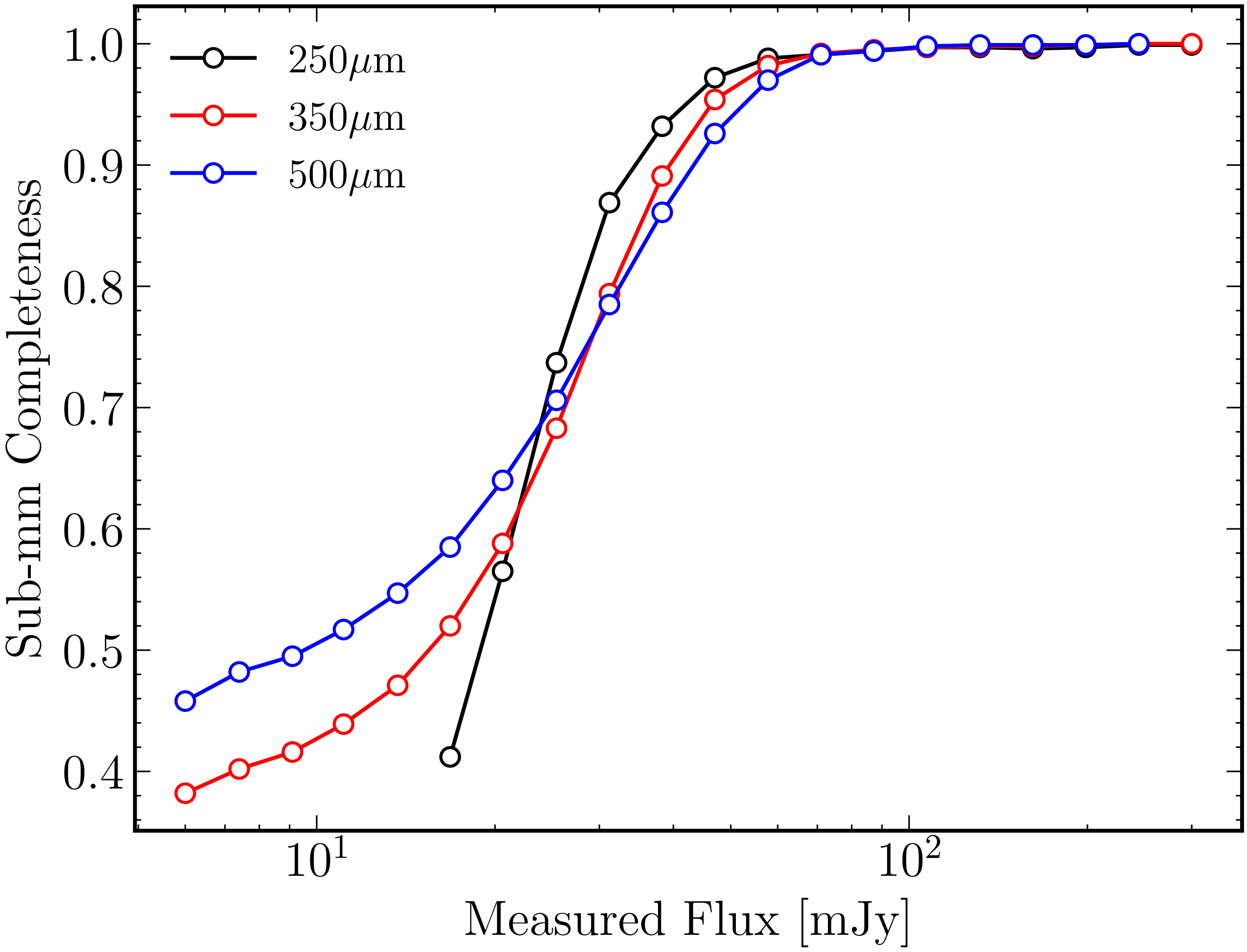

Figure 2.1. Caption

by ordering the sources by their flux density estimates and iteratively fitting and removing a point source from the position of each source, starting with the brightest. The new estimates of the flux densities are then not influenced by contamination from brighter sources.

The catalogue of point sources provided by H-ATLAS come from the extraction of point sources using MADX applied to the SPIRE images of the NGP, SGP and GAMA fields. The final sources list is reduced to those sources with SNR $>$ 4 in any of the SPIRE bands. While the detection method suggests that we may miss sources that are faint at 250 µmbut bright at 350 or 500 µm, due to the weighting of the three images, cataloguing all sources with SNR $>$ 4 in any of the SPIRE bands means that the catalogues are reasonably complete in all bands. The completeness of the sub-mm catalogues as a function of the measured flux density of a source as estimated by Valiante et al. 2016 is illustrated in Figure 2.1.

2.1.2 Data Releases of the H-ATLAS

The first public data release (DR1) of H-ATLAS covered the three equatorial GAMA fields, which span approximately 25% of the total survey area. These fields benefit from multiwavelength coverage from GAMA, SDSS, 2dF, the Galaxy Evolution Explorer (GALEX, Martin et al. 2005), the UKIRT Infrared Deep Sky Survey – Large Area Survey (UKIDSS-LAS, Lawrence et al. 2007), the Wide-field Infrared Survery Explorer (WISE, Wright et al. 2010), the VISTA Kilo-degree Infrared Galaxy survey (VIKING, Edge et al. 2013) and the Kilo-Degree Survey (KiDS, de Jong et al. 2013).

Sources are provided with DR1 if they are detected above the 2.5σ detection limit on the 250 µmmap and have measured flux densities greater than the 4σ flux density limits in one of the three SPIRE bands (29.6 mJy, 37.6 mJy or 40.8mJy at 250, 350 and 500 µm). Across the three fields there are a total of 113,995, 46,209 and 11,011 sources detected at $> 4\sigma$ at 250, 350 and 500 µmas well as detections for 4,650 and 5,685 sources at $> 3\sigma$ at 100 and 160 µm(Valiante et al. 2016). Following the release of the sub-mm sources detected in the GAMA fields, Bourne et al. 2016 used the Likelihood Ratio (LR, Sutherland & Saunders 1992; Ciliegi et al. 2003) method (Section 2.1.4) to identify potential optical counterparts to the 113,995 sources with SNR$_{250} > 4$ from SDSS. Sources with SNR$_{250} < 4$ that were detected by their 350 or 500 µmflux densities were omitted from the matching since these sources have sub-mm colours suggesting a high redshift, and are the most likely sources to be misidentified by SDSS due to the increased probability of chance alignments or gravitational lensing along the line of sight (Negrello et al. 2010; Pearson et al. 2013; Bourne et al. 2014). Bourne et al. 2016 found optical counterparts within 10" of 44,385 (39%) sources with an estimated probability of being the true ID > 80% (the probability of an optical or near-infrared object being the true counterpart to a sub-mm source is defined as the reliability, R, and is derived in Section 2.1.4).

The second public data release (DR2) covered the NGP and SGP, two large fields that together form ∼75% of the total survey area. The NGP was covered in the optical by the SDSS and in the near-infrared by UKIDSS-LAS. Moreover, a small area of 25.93 $\deg^2$ within the NGP was also observed by a deeper K-band survey by the H-ATLAS team using UKIRT (limiting magnitude of K < 19.40 compared to K < 18.69 for UKIDSS-LAS). The SGP is the largest field (approximately half the survey area of H-ATLAS) and was covered by the 2dF spectroscopic survey, KiDS in four optical bands (u, g, r and i) and VIKING in five near-infrared bands (Z, Y, J, H and K_s).

Given that sub-mm sources are only extracted from areas of the *Herschel* maps that have at least two obsersations from the SPIRE instrument, the DR2 catalogues includes sources from the map area reduced by the masking of single *Herschel* scans. The mask reduces the area covered by the NGP point source catalogue to 177.1 $\deg^2$ and the SGP to 303.4 $\deg^2$. As with DR1, sources are included if they are detected on the 250 µmmap above the 2.5σ detection limit by the MADX algorithm and surpass at least one of the 4σ flux density limits at the SPIRE wavelengths. The catalogues contain 118,980 sources for the NGP field (112,069, 48,876 and 10,368 detected at $> 4\sigma$ at 250, 350 and 500 µmand 5,036 and 7,046 at $> 3\sigma$ at 100 and 160 µmrespectively) and 193,527 sources for the SGP field (182,282, 74,096 and 16,084 at 250, 350 and 500 µmand 8,598 and 11,894 at 100 and 160 µm). Furlanetto et al. 2018 applied the Likelihood Ratio method to all counterparts within 10" of the 250 µmsources of the NGP using both the shallower optical and near-infrared catalogue of SDSS and UKIDSS-LAS, and the deeper K-band survey. Of the 112,155 SPIRE sources with SNR$_{250}$ > 4, 77,521 (69.1%) had at least one shallow optical counterpart and 42,429 (37.8%) of these were matched with R > 0.8. In the smaller area observed with WFCAM, Furlanetto et al. 2018 identified 32,041 possible deep near-IR counterparts to 17,247 sources. 10,668 (61.9%) of these sources were matched with an equally high reliability. While this analysis suggests that the inclusion of deeper K-band data drastically increases the fraction of sources matched to their corresponding optical or near-IR counterpart, [...].

continua
we must
consider
differenc
think abo

In the SGP a preliminary counterpart analysis was conducted using the Two Micron All Sky Survey (2MASS, Skrutskie et al. 2006), but no formal LR analysis had yet been applied. A nearest neighbour match within 5" of a 2MASS galaxy gives identifications for 3,444 *Herschel* sources. In the following section we detail the Likelihood Ratio method and apply it to the 250 µmsources detected by *Herschel* in the SGP.

2.1.3 Identifying Optical and Near-IR Counterparts to Herschel Sources

When identifying multiwavelength counterparts across surveys the simplest choice to use the nearest neighbour within a fixed search radius of one of the sources. For surveys conducted at similar wavelengths with a similar resolution and sensitivity this is a suitable approach. However, when matching far-IR/sub-mm surveys to optical/IR data, the poor angular resolution of long wavelength instruments such as SPIRE (the FWHM of 250 µmdetections with SPIRE is $\sim 18"$), which cause large positional uncertainties, force us to increase the search radius around the sub-mm source. This effect, coupled with the intrinsic faintness of optical/near-IR counterparts due to dust obscuration, the relatively flat redshift distribution of sub-mm sources due to the k-correction and the high surface density of objects in optical/IR surveys, means that [...] and it is common for there to be multiple possible counterparts within the search radius from a single sub-mm source. Previously for sub-mm surveys it would be more practical to first match sources with radio or mid-IR sources and then use pre-existing matched catalogues to obtain multiwavelength data (e.g. Ivison et al. 2007; Dye et al. 2009; Biggs et al. 2011, see also Section [...]). However, presently this is not suitable for large surveys such as H-ATLAS as current radio telescopes do not provide the area and depth required to match with more than a small fraction of sub-mm sources. While current and future radio surveys from facilities such as the Square Kilometre Array (SKA), the

continua
identifica
not easy,
reference

add refer
for the fi
chapter

Low Frequency Array (LOFAR) and MeerKAT will increase the radio coverage of the H-ATLAS fields, currently a statitstical identification method is still the preferred way of deciding which objects are associated and which are unrelated foreground/background objects to large samples of sub-mm sources.

2.1.4 The Likelihood Ratio Method

The Likelihood Ratio method assigns a probability (reliability) to all potential matches surrounding low resolution sources to distinguish between likely counterparts and chance alignments and has been used many times to identify counterparts to *Herschel* sources. The LR method was used by Smith et al. 2011 to identify SDSS counterparts in the Science Demonstration Phase (SDP) catalogue (a preliminary data release for H-ATLAS, overlapping with the GAMA9 field), by Kim et al. 2012 to identify Spitzer-IRAC counterparts also in the SDP data, by Fleuren et al. 2012 for VIKING IDs in the Phase 1 catalogue of the GAMA9 field, and as mentioned earlier, by Bourne et al. 2016 and Furlanetto et al. 2018 to find optical and near-IR counterparts in the GAMA fields and NGP field respectively.

The likelihood, L, of a counterpart being the true identification to a *Herschel* source is given by the ratio between the probability that an object observed at a given radius from the source, r, with an optical or near-IR magnitude, m, is the true identifcation and the probability of observing an unassociated object with the same r and m. On the assumption that the distance from the source and the optical/near-IR magnitude are independent on their influence on the probability of being a true counterpart, we find that:

$$L = \frac{P(\mathrm{ID}, r, m)}{P(\mathrm{unassociated}, r, m)} = \frac{P(\mathrm{ID}, r)P(\mathrm{ID}, m)}{P(\mathrm{unassociated}, r, m)} \tag{2.1}$$

Each term in the above equation can be defined in the following way: $f(r) \coloneqq P(\mathrm{ID}, r)$, $q(m) \coloneqq P(\mathrm{ID}, m)$ and $n(m) \coloneqq P(\mathrm{unassociated}, r, m)$, where $f(r)$ represent the radial probability distribution function of positional errors between the source and counterpart, $q(m)$ represents the magnitude probability distribution of true counterparts and $n(m)$ is the magnitude distribution of background objects from the input survey. By using Baye's theorem and the theorem of total probability, we can define the probability that a counterpart is the true ID given it has r and m as:

$$R \coloneqq P(\mathrm{ID}|r, m) = \frac{L}{L+1}. \tag{2.2}$$

es a
ion

Equation 2.2 assumes that there is only a single candidate with a likelihood L. For a source with multiple possible candidates, the reliability R_j of the j^{th} candidate is given by:

$$R_j = \frac{L_j}{\sum_i L_i + (1 - Q)}, \tag{2.3}$$

where i represents the i^{th} counterpart found within the search radius. The Q parameter represents the fraction of all true counterparts that are brighter than the limiting magnitude of the input survey and can therefore be observed. This means that the (1 - Q) term represents the probability that the counterpart is not observed and accounts for the fact that not all counterparts will be detected in the optical/near-IR survey. The value of Q depends on the depth of the survey and the choice of passband used. In the following sections I shall outline the methods used to estimate the functions $f(r)$, $q(m)$ and $n(m)$ and to estimate Q to calculate the likelihood ratios and reliabilities of near-IR counterparts observed on the VIKING images surrounding the 250 µmpositions of *Herschel* sources in the SGP.

2.2 Applying the LR Method to VIKING Galaxies in the SGP

2.2.1 VISTA VIKING Counterparts

2.2.2 True Counterpart Distribution, q(m)

2.2.3 Estimating Q

2.2.4 The Positional Offset Distribution, f(r)

Chapter 2

Todo list

Chapter 3

Conclusion

Chapter 4

(

Derivations) Remember to write down and delete afterwards!

4.1 Equation 2.2

Recall Baye's theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Therefore,

$$P(\mathrm{ID}|r,m) = \frac{P(r,m|\mathrm{ID})p(ID)}{P(r,m)}$$

Now recall the rule of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Therefore,

$$P(\mathrm{ID}|r,m) = \frac{P(ID,r,m)}{P(r,m)}$$

If we assume that r and m are independent of each other then:

$$P(\mathrm{ID}|r,m) = \frac{P(ID,r)P(ID,m)}{P(r,m)}$$

The denominator is the probability that a counterpart has r and m, regardless of whether it is the ID or not. Therefore, now recall the law of total probability:

$$P(A) = P(A \cap B) + P(A \cap B')$$

Therefore:

$$\begin{aligned}
P(\mathrm{ID}|r,m) &= \frac{P(ID,r)P(ID,m)}{P(ID,r,m)+P(\mathrm{unassociated},r,m)} \\
&= \frac{P(ID,r)P(ID,m)}{P(ID,r)P(ID,m)+P(\mathrm{unassociated},r,m)} \\
&= \frac{\frac{P(ID,r)P(ID,m)}{P(\mathrm{unassociated},r,m)}}{\frac{P(ID,r)P(ID,m)}{P(\mathrm{unassociated},r,m)}+1} \\
P(\mathrm{ID}|r,m) &= \frac{L}{L+1}
\end{aligned}$$

Appendix A

An Appendix

A.1 An Appendix

Bibliography

Amblard A., et al., 2010, *A&A*, 518, L9

Barger A. J., Cowie L. L., Sanders D. B., Fulton E., Taniguchi Y., Sato Y., Kawara K., Okuda H., 1998, *Nature*, 394, 248

Biggs A. D., et al., 2011, *MNRAS*, 413, 2314

Blain A. W., Longair M. S., 1993, *MNRAS*, 264, 509

Bourne N., et al., 2014, *MNRAS*, 444, 1884

Bourne N., et al., 2016, *MNRAS*, 462, 1714

Chapin E. L., et al., 2011, *MNRAS*, 411, 505

Ciliegi P., Zamorani G., Hasinger G., Lehmann I., Szokoly G., Wilson G., 2003, *A&A*, 398, 901

Colless M., et al., 2001, *MNRAS*, 328, 1039

Driver S. P., et al., 2009, Astronomy and Geophysics, 50, 5.12

Dye S., et al., 2009, *ApJ*, 703, 285

Eales S., et al., 2010, *PASP*, 122, 499

Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, The Messenger, 154, 32

Fleuren S., et al., 2012, *MNRAS*, 423, 2407

Fudamoto Y., et al., 2017, *MNRAS*, 472, 2028

Furlanetto C., et al., 2018, *MNRAS*, 476, 961

Griffin M. J., et al., 2010, *A&A*, 518, L3

Hughes D. H., et al., 1998, *Nature*, 394, 241

Ivison R. J., et al., 2007, *MNRAS*, 380, 199

Kim S., et al., 2012, *ApJ*, 756, 28

Lapi A., et al., 2011, *ApJ*, 742, 24

Lawrence A., et al., 2007, *MNRAS*, 379, 1599

Maddox S. J., Dunne L., 2020, *MNRAS*, 493, 2363

Martin D. C., et al., 2005, *ApJ*, 619, L1

Negrello M., et al., 2010, Science, 330, 800

Pearson E. A., et al., 2013, *MNRAS*, 435, 2753

Poglitsch A., et al., 2010, *A&A*, 518, L2

Skrutskie M. F., et al., 2006, *AJ*, 131, 1163

Smail I., Ivison R. J., Blain A. W., 1997, *ApJ*, 490, L5

Smith D. J. B., et al., 2011, *MNRAS*, 416, 857

Sutherland W., Saunders W., 1992, *MNRAS*, 259, 413

Valiante E., et al., 2016, *MNRAS*, 462, 3146

Wright E. L., et al., 2010, *AJ*, 140, 1868

York D. G., et al., 2000, *AJ*, 120, 1579

Zavala J. A., et al., 2018, Nature Astronomy, 2, 56

de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Experimental Astronomy, 35, 25