



# Submm/mm galaxy counterpart identification using a characteristic density distribution

Stacey Alberts,<sup>1\*</sup> Grant W. Wilson,<sup>1</sup> Yu Lu,<sup>2</sup> Seth Johnson,<sup>1</sup> Min S. Yun,<sup>1</sup>  
Kimberly S. Scott,<sup>3</sup> Alexandra Pope,<sup>1</sup> Itziar Aretxaga,<sup>4</sup> Hajime Ezawa,<sup>5</sup>  
David H. Hughes,<sup>4</sup> Ryohei Kawabe,<sup>6</sup> Sungeun Kim,<sup>7</sup> Kotaro Kohno<sup>8,9</sup>  
and Tai Oshima<sup>10</sup>

<sup>1</sup>Department of Astronomy, LGRT-B 619E, University of Massachusetts, Amherst, MA 01003, USA

<sup>2</sup>Kavli Institute for Particle Astrophysics and Cosmology, Stanford, CA 94309, USA

<sup>3</sup>North American ALMA Science Center, National Radio Astronomy Observatory, Charlottesville, VA 22903, USA

<sup>4</sup>Instituto Nacional de Astrofísica, Óptica y Electrónica, Calle Luis Enrique Erro 1, Sta. Ma. Tonantzintla, Puebla, Mexico C.P. 72840

<sup>5</sup>ALMA Project Office, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan

<sup>6</sup>Joint ALMA Observatory, Alonso de Cordova 3107 OFC 129, Vitacura, Santiago 763 0355, Chile

<sup>7</sup>Astronomy and Space Science Department, Sejong University, Seoul 143-747, South Korea

<sup>8</sup>Institute of Astronomy, University of Tokyo, 2-21-1 Osawa, Mitaka, Tokyo 181-0015, Japan

<sup>9</sup>Research Center for the Early Universe, University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

<sup>10</sup>Nobeyama Radio Observatory, National Astronomical Observatory of Japan, Minamimaki, Minamisaku, Nagano 384-1305, Japan

Accepted 2013 January 24. Received 2013 January 24; in original form 2012 November 2

## ABSTRACT

We present a new submm/mm galaxy counterpart identification technique which builds on the use of *Spitzer* Infrared Array Camera (IRAC) colours as discriminators between likely counterparts and the general IRAC galaxy population. Using 102 radio- and Submillimeter Array-confirmed counterparts to AzTEC sources across three fields [Great Observatories Origins Deep Survey-North, -South and Cosmic Evolution Survey (COSMOS)], we develop a non-parametric IRAC colour–colour characteristic density distribution, which, when combined with positional uncertainty information via likelihood ratios, allows us to rank all potential IRAC counterparts around submillimetre galaxies (SMGs) and calculate the significance of each ranking via the reliability factor. We report all robust and tentative radio counterparts to SMGs, the first such list available for AzTEC/COSMOS, as well as the highest ranked IRAC counterparts for all AzTEC SMGs in these fields as determined by our technique. We demonstrate that the technique is free of radio bias and thus applicable regardless of radio detections. For observations made with a moderate beam size ( $\sim 18$  arcsec), this technique identifies  $\sim 85$  per cent of SMG counterparts. For much larger beam sizes ( $\gtrsim 30$  arcsec), we report identification rates of 33–49 per cent. Using simulations, we demonstrate that this technique is an improvement over using positional information alone for observations with facilities such as AzTEC on the Large Millimeter Telescope and Submillimeter Common User Bolometer Array 2 on the James Clerk Maxwell Telescope.

**Key words:** methods: data analysis – techniques: photometric – galaxies: high-redshift – infrared: galaxies – radio continuum: galaxies – submillimetre: galaxies.

## 1 INTRODUCTION

The discovery of a large population of bright sources at high redshift through the opening of the submillimetre (submm) and millimetre (mm) wavelength windows continues to have a profound impact on

our understanding of galaxy evolution in the early Universe. These submm/mm-selected galaxies (hereafter SMGs) are characterized by large far-infrared luminosities ( $\gtrsim 10^{12} L_{\odot}$ ), tremendous star formation rates ( $\gtrsim 500 M_{\odot} \text{ yr}^{-1}$ ) and a number density that is high compared to local ultraluminous infrared galaxies (ULIRGs), indicating strong evolution within the population (e.g. Scott et al. 2002). SMGs are thought to represent young, massive systems that exist in an epoch of rapid mass build-up and may be the predecessors

\*E-mail: salberts@astro.umass.edu

of modern-day giant elliptical galaxies (see review by Blain et al. 2002). As a population, SMGs contribute a significant portion of the cosmic energy density at  $z = 2\text{--}3$  through some combination of dusty starbursts and active galactic nuclei (AGN; Hughes et al. 1998; Alexander et al. 2005; Chapman et al. 2005; Kovács et al. 2006; Pope et al. 2006; Coppin et al. 2008; Murphy et al. 2011). Observations at submm/mm wavelengths benefit from a strong negative  $k$ -correction and thus present a unique opportunity for an unbiased view of star formation out to extreme redshifts. Understanding SMGs is vital to any complete view of galaxy evolution.

Currently, thousands of SMGs have been detected in both ground-based surveys at  $450\text{--}2000\ \mu\text{m}$  (e.g. Smail, Ivison & Blain 1997; Barger et al. 1998; Hughes et al. 1998; Cowie, Barger & Kneib 2002; Scott et al. 2002, 2008; Borys et al. 2003; Greve et al. 2004; Laurent et al. 2005; Coppin et al. 2006; Perera et al. 2008; Austermann et al. 2010) and in balloon- and space-based surveys at  $250\text{--}500\ \mu\text{m}$  (Pascale et al. 2008; Devlin et al. 2009; Eales et al. 2010; Oliver et al. 2010). A complete understanding of the physical processes within these systems, however, requires that these observations be matched to multiwavelength data, from radio to X-ray. Such multiwavelength counterpart identification of SMGs is difficult due to the poor angular resolution ( $> 10$  arcsec) of single-dish submm telescopes and the intrinsic faintness of their optical counterparts due to dust obscuration. A great deal of work has been invested into counterpart identification techniques, primarily using radio continuum data from interferometric observations (e.g. Ivison et al. 2002, 2007; Chapman et al. 2005; Pope et al. 2006; Chapin et al. 2009; Yun et al. 2012). Radio observations are particularly useful for counterpart identification due to (1) the low number density of radio sources, which minimizes the probability of a chance association, and (2) the tight correlation between radio and submm/mm emission (e.g. Condon 1992; Yun, Reddy & Condon 2001), both star formation tracers. Direct submm/mm interferometric imaging of SMGs with high angular resolution (e.g. Younger et al. 2007, 2009; Barger et al. 2012; Smolčić et al. 2012) supports this radio–submm association.

Unfortunately, the use of radio detections for counterpart identification has drawbacks. Because of the radio–submm correlation, radio counterparts will preferentially be found for the brightest SMGs, which may not be representative of the full SMG population. Interferometric observations have shown that a small percentage ( $\sim 7\text{--}15$  per cent) of single radio sources near SMGs are not actually associated with the submm emission. In cases where multiple radio detections are found within the submm beam, interferometry reveals that in  $\sim 80$  per cent of these instances, the submm emission is only associated with one of the radio sources (see Section 3.1; Younger et al. 2007, 2009; Barger et al. 2012; Smolčić et al. 2012). Finally, radio observations suffer from a strong positive  $k$ -correction, which results in a bias against high-redshift galaxies, while submm emission remains equally detectable out to  $z \sim 8$  (Blain et al. 2002). Current radio observations that overlap with submm surveys have placed the fraction of SMGs that have a detectable radio counterpart at  $\sim 40\text{--}100$  per cent (Ivison et al. 2002, 2007; Chapin et al. 2009; Weiß et al. 2009; Lindner et al. 2011; Barger et al. 2012; Smolčić et al. 2012; Yun et al. 2012), depending on the depth of the radio observations. For example, given the shallowness of the available radio data in the Cosmic Evolution Survey (COSMOS) field (see Section 2), we can predict that we will detect  $\sim 40$  per cent of SMGs in the radio based on submm counterpart studies of Great Observatories Origins Deep Survey-North (GOODS-N) SMGs, which utilized deep radio data to build a cumulative distribution of radio counterparts as a function of radio depth (Pope 2007). And, in fact, we find that 40 per cent

of SMGs detected with single-dish observations in COSMOS are associated with radio detections (Aretxaga et al. 2011, this work). This is further confirmed by interferometric observations of AzTEC SMGs in COSMOS. Younger et al. (2007, 2009) imaged a sample of 15 SMGs with the Submillimeter Array (SMA) and found that only 6/15 (40 per cent) have a robust radio counterpart. Of the nine with no radio counterpart, seven are also undetected with Multi-band Imaging Photometer for *Spitzer* (MIPS)  $24\ \mu\text{m}$ . On the other hand, 13/15 of these SMGs are detected with *Spitzer* Infrared Array Camera (IRAC; Fazio et al. 2004a). Of the two not detected, one is confused with a foreground source. Additionally, several studies have shown that SMGs are routinely detected in the IRAC bands at  $3.6$  and  $4.5\ \mu\text{m}$  at the  $\geq 1\ \mu\text{Jy}$  level (Iono et al. 2006; Wang et al. 2007, 2011; Hatsukade et al. 2010; Tamura et al. 2010; Ikarashi et al. 2011). The high source density ( $\sim 60\ \text{arcmin}^{-2}$  at the  $1.4\ \mu\text{Jy}$  level at  $3.6\ \mu\text{m}$ ; Fazio et al. 2004b) of IRAC galaxies makes it impossible to use IRAC detections directly for unique identifications (Pope et al. 2006), however, the selection of IRAC counterparts is expected to suffer less bias against high-redshift sources and may be more representative of faint SMGs (see Section 5.2).

More recently, two studies have examined follow-up interferometry of SMGs detected with the Large APEX Bolometer Camera (LABOCA;  $870\ \mu\text{m}$ ; Siringo et al. 2009) and Submillimeter Common User Bolometer Array (SCUBA)/SCUBA-2 ( $850\ \mu\text{m}$ ). Smolčić et al. (2012) analysed a set of 34 LABOCA sources in COSMOS with interferometry from various sources, including 26 new detections from the IRAM Plateau de Bure Interferometer (PdBI). Using a radio catalogue extended to include lower signal-to-noise ratio (S/N) sources, they found that 18/34 ( $\sim 50$  per cent) SMGs were detected at  $20\ \text{cm}$  and the same fraction, though a different subset, were detected with IRAC. This study also determines redshifts to their sources and concludes that the redshift distribution of SMGs may peak higher than previously thought, including a large tail of sources at very high redshift ( $z > 3$ ) that will be difficult to detect in even the deepest radio surveys due to the  $k$ -correction. It should be noted, however, that this study did not require the LABOCA SMGs to be detected at high significance. If we limit the sample to high significance interferometric detections ( $> 4.5\sigma$ ), then 10/16 (62 per cent) SMGs are detected in IRAC, while, if we look only at LABOCA sources which are also detected with AzTEC, 8/10 (80 per cent) have IRAC counterparts. Additionally, Barger et al. (2012) looked at a small, but homogenous sample of 12 SMGs in GOODS-N and found 16 independent, high significance detections with the SMA. Of these 16, all are detected in an ultradeep radio catalogue and 15 are detected in the IRAC bands.

In this study, we develop a new technique for counterpart identification which takes advantage of the high sensitivity of *Spitzer* IRAC and does not rely on radio or MIPS  $24\ \mu\text{m}$  observations. Our starting point is the IRAC colour–colour space diagram first developed by Lacy et al. (2004) for identifying AGN. IRAC colours, and this diagram in particular, have been shown to also be useful for SMG counterpart identification, with  $\lesssim 20$  per cent of SMGs overlapping with AGN in IRAC colour–colour space (Pope et al. 2006; Yun et al. 2008, 2012; Biggs et al. 2011). We take these studies a step farther and, using a large sample of SMGs with secure radio or SMA counterparts, develop an IRAC colour–colour density distribution. This multidimensional colour prior is combined with positional uncertainty information via the likelihood ratio (LR; Sutherland & Saunders 1992), providing a method for ranking potential IRAC counterparts around SMGs. Previous studies utilizing likelihood ratios for counterpart identification can be found in the literature across all wavelength regimes; for example, Chapin et al.

(2011) combined priors based on radio, MIPS and IRAC fluxes and colours with positional information to identify counterparts to Balloon-borne Large Aperture Submillimetre Telescope (BLAST) and LABOCA sources. Smith et al. (2012) and Kim et al. (2012) similarly used individual flux and colour priors to identify counterparts to *Herschel* Spectral and Photometric Imaging Receiver (SPIRE) sources. The methods used in these studies necessarily assume that the various priors used are independent, though some or all may be correlated (Chapin et al. 2011). In this work, we incorporate correlations between different flux bands and colours by using a characteristic density distribution (CDD) rather than individual priors.

Since our method only requires *Spitzer* IRAC observations, it can be applied in numerous fields that lack deep radio observations and is not as strongly biased against high-redshift sources as techniques which rely on radio data. Many of the fields in which deep IRAC data already exists will be prime targets for future submm surveys and we show that the technique developed here is useful not only for existing submm surveys but also for upcoming surveys with the Large Millimeter Telescope (LMT) and SCUBA-2. Though the particular CDD developed in this study is only applicable to source populations with a redshift distribution and spectral energy distribution (SED) similar to sources selected at 850–1200  $\mu\text{m}$  (see Section 3.2.1), we posit that a modified version of this technique can be developed for *Herschel* SPIRE sources, which suffer from similar challenges due to low resolution.

This paper is organized as follows. In Section 2, we describe the data used in this study. In Section 3, we develop the counterpart identification technique, which we use in Section 4 to calculate an LR and its corresponding reliability for all IRAC counterpart candidates across the three fields, including an analysis of the resulting ranking via a blind test using our radio- and SMA-identified counterpart sample. In Section 5, we discuss our technique in the context of future submm surveys and address the issue of possible bias introduced through the use of radio counterparts in developing the IRAC colour–colour CDD. Section 6 contains our conclusions.

## 2 DATA

To develop our new technique for selecting IRAC counterparts to SMGs, we utilize a statistically significant sample of 272 SMGs detected at 1.1 mm with AzTEC (Wilson et al. 2008) from three fields: COSMOS, GOODS-N and GOODS-S. These three fields have extensive multiwavelength data ranging from X-ray to radio, much of which was observed at high resolution. We take advantage of observations from the Very Large Array (VLA) at 1.4 GHz and interferometric observations taken with the SMA to identify secure counterparts for a subset of the AzTEC SMGs; the IRAC counterparts are then used to derive a CDD utilizing the four *Spitzer* IRAC bands (3.6, 4.5, 5.8 and 8.0  $\mu\text{m}$ ), which is described in the next sections. We describe each data set below.

**AzTEC:** The AzTEC observations used here are described in the following papers: Perera et al. (2008) for GOODS-N, Scott et al. (2010) for GOODS-S and Scott et al. (2008) and Aretxaga et al. (2011) for COSMOS. Data reduction was performed using the standard AzTEC pipeline (see Scott et al. 2008; Downes et al. 2012). The reader is referred to Downes et al. (2012) for the most updated source lists for the GOODS fields. Observations in GOODS-N were taken on the James Clerk Maxwell Telescope (JCMT) with uniform coverage to a depth of  $\sim 1.3 \text{ mJy beam}^{-1}$  and a full width at half-maximum (FWHM) of 18 arcsec. GOODS-S and COSMOS were imaged using the Atacama submillimetre Telescope Experiment

(ASTE; Ezawa et al. 2004) to a depth of  $\sim 0.6$  and  $\sim 1.3 \text{ mJy beam}^{-1}$ , respectively. These observations were done at a lower angular resolution and we measure a FWHM of 30 arcsec for GOODS-S and 33 arcsec for COSMOS by fitting the post-filtered point spread functions (PSFs) of each set of observations with a Gaussian. Our sample contains 36, 47 and 189 SMGs in GOODS-N, GOODS-S and COSMOS, for a total of 272 sources.

**VLA:** Radio observations at 20 cm (1.4 GHz) were obtained for all fields using the VLA with an angular resolution of  $\leq 3.5$  arcsec. Our deepest radio catalogue is in GOODS-N, with a sensitivity of  $3.9 \mu\text{Jy beam}^{-1}$  ( $1\sigma$ ) and 1230  $5\sigma$  radio detections (Morrison et al. 2010). Our GOODS-S catalogue was made using the ‘Search and Destroy’ (*SAD*) procedure in *AIPS* to identify all sources greater than  $3\sigma$ , which were then cross-matched with the IRAC SIMPLE (Damen et al. 2011) and GOODS catalogues to exclude noise peaks (see Yun et al. 2012). The GOODS-S catalogue has an rms sensitivity of  $8.5 \mu\text{Jy beam}^{-1}$  and  $\sim 1500$  sources. For COSMOS, we use the  $>4.5\sigma$  catalogue of Schinnerer et al. (2010), which has  $\sim 3000$  sources to a  $1\sigma$  sensitivity of  $10 \mu\text{Jy beam}^{-1}$ . We demonstrate in Section 5.3 that the sensitivity of the radio catalogue has little impact on our results.

**SMA:** 15 AzTEC/JCMT sources in the COSMOS field were the targets of two follow-up observations by the SMA (Younger et al. 2007, 2009). All were detected at 890  $\mu\text{m}$  with positional accuracies of  $\sim 0.2$  arcsec. 6/15 sources are detected in the radio, while nine are not. We include the three radio-undetected sources that were detected in all four IRAC bands in our analysis and in our derivation of the IRAC colour–colour CDD below.

**Spitzer IRAC:** GOODS-N, GOODS-S and COSMOS were imaged with all four IRAC bands with an angular resolution of 2 arcsec and a positional accuracy of  $\sim 0.2$  arcsec. In GOODS-S, the SIMPLE survey (Damen et al. 2011) covers  $\sim 1600 \text{ arcmin}^{-2}$  to a  $5\sigma$  depth of 1.0, 1.2, 6.0 and 6.7  $\mu\text{Jy}$  for bands 3.6, 4.5, 5.8 and 8.0  $\mu\text{m}$ . The SIMPLE IRAC catalogue comprises all 3.6  $\mu\text{m}$  detections with  $S/N > 5$ . Deeper observations were done with a  $5\sigma$  sensitivity of 0.1, 0.2, 1.5 and 1.6  $\mu\text{Jy}$  in GOODS-N (Dickinson et al. 2003; Treister et al. 2006) and shallower observations were taken with depths of 0.9, 1.7, 11.3 and 14.6  $\mu\text{Jy}$  in S-COSMOS (Sanders et al. 2007), with a catalogue containing all sources detected at 3.6  $\mu\text{m}$  at the  $\gtrsim 1 \mu\text{Jy}$  level. Based on studies in the GOODS fields, these sensitivities allow for the detection of galaxies with stellar masses greater than  $10^{10} M_{\odot}$  and a typical age of a few hundred million years up to  $z \sim 6$  (Yan et al. 2006).

## 3 DERIVING THE CHARACTERISTIC DENSITY DISTRIBUTION

Our submm galaxy counterpart identification technique relies on two pieces of information: the positional uncertainty of a given AzTEC detection and the general distribution of IRAC colours for the SMG population. The latter is obtained from our ‘training set’, the largest sample of radio- and SMA-identified counterparts to SMGs selected at 1.1 mm used to date in a counterpart identification study. We describe the selection of secure radio counterparts and the development of the IRAC colour–colour CDD of SMGs in this section.

### 3.1 Identifying radio counterparts to SMGs: the training set

Because of the typically low  $S/N$  ( $\lesssim 10$ ) and low angular resolution of submm/mm observations with single-dish telescopes, the positional offsets between the detected emission and the true position

of the source are large due to noise fluctuations in the submm maps. We can characterize these positional uncertainties through simulations wherein simulated sources of known flux densities are inserted into the map one at a time at random positions. From this we build a cumulative positional probability distribution  $P(>D; S/N)$ , which is the probability that a source will be detected at a distance greater than  $D$  from its true position as a function of  $S/N$ , where the noise is the instrument noise of the observation (see Perera et al. 2008; Scott et al. 2010; Aretxaga et al. 2011). The distributions determined from these simulations are consistent with the analytical solution from Ivison et al. (2007), where

$$P(>D; S/N) = \int D' \exp\left(\frac{-D'^2}{2\sigma^2}\right) dD', \quad (1)$$

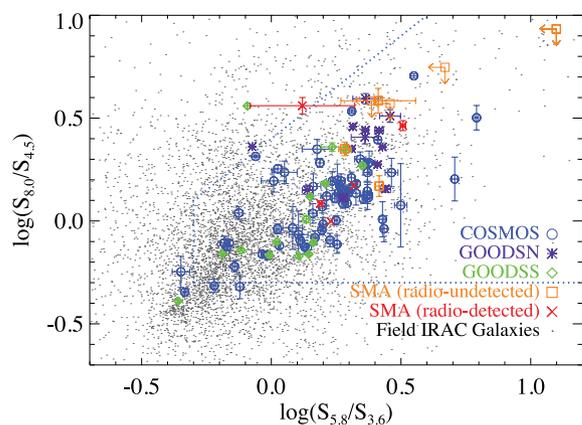
with

$$\sigma \simeq \frac{0.6 \text{FWHM}}{S/N}. \quad (2)$$

Using this formula, we can identify all potential radio counterparts around each SMG using a variable search radius based on the  $S/N$  of the AzTEC detection. To be conservative, we select all radio sources within a search radius where  $P(>D; S/N)$  goes to 1 per cent. For the lowest  $S/N$  sources considered here ( $S/N = 3.5$ ), this corresponds to a search radius of 9.5, 16 and 17.6 arcsec for SMGs in GOODS-N, GOODS-S and COSMOS, respectively. Any AzTEC source that has a single radio source within the appropriate search radius is included in our training set. From interferometric follow-ups to AzTEC and LABOCA SMGs, we expect that the false association rate for single radio sources detected within the submm beam to be  $\sim 7$ –15 per cent (Younger et al. 2007, 2009; Smolčić et al. 2012). Using a variable search radius based on the positional uncertainties, as in this work, should decrease this rate. An analysis of the 12 radio-detected SMGs in our training set that have interferometric data available (AzTEC/GN3, AzTEC/GN5, AzTEC/GN14, AzTEC/C3, AzTEC/C12, AzTEC/C13, AzTEC/C14, AzTEC/C18, AzTEC/C22, AzTEC/C38, AzTEC/C98 and AzTEC/C145) shows that we have selected the correct radio counterpart in all cases except AzTEC/GN14 (see Table A1; Younger et al. 2007, 2009; Barger et al. 2012; Smolčić et al. 2012). AzTEC/GN14 has a faint radio counterpart that is below the detection limits of our catalogue, in addition to an unrelated radio source within its search radius.

Traditionally, the robustness of a radio counterpart is determined by the  $P$ -statistic ( $P$ -stat; Downes et al. 1986), with a  $P$ -stat  $< 0.05$  considered a robust counterpart and  $0.05 < P$ -stat  $< 0.2$  considered a tentative counterpart. We identify a total of 107 isolated radio counterparts to our SMGs, 18 in GOODS-N, 15 in GOODS-S and 74 in COSMOS. Of these, 95 (79 per cent) have  $P$ -stat  $< 0.05$ , while 22 (21 per cent) have  $0.05 < P$ -stat  $< 0.2$ . We elect to include these tentative counterparts in our training set and have found that they do not significantly influence the resulting IRAC colour distribution. Our radio counterparts are in good agreement with those found in Chapin et al. (2009) for GOODS-N and Yun et al. (2012) for GOODS-S. This work presents the formal analysis of the radio counterparts to AzTEC/COSMOS sources described in Aretxaga et al. (2011) (Tables A1 and A2).

IRAC counterparts are obtained by searching within 2 arcsec of the radio positions, which is possible due to the excellent positional accuracy of the VLA observations ( $< 1$  arcsec; Kellerman et al. 2008). For determining the IRAC colour–colour CDD, we require that IRAC sources be detected in all four bands. This leaves 96/107 (90 per cent) radio-detected SMGs with complete IRAC data. Of the remainder, 10 are completely undetected and one is detected in only



**Figure 1.** The radio- and SMA-identified IRAC counterparts to SMGs that are used as a training set in this study. In addition to the 102 IRAC counterparts detected in all four IRAC channels, four counterparts that are missing one or more IRAC detections are plotted as upper limits. These sources are not part of the training set. The dotted lines denote the IRAC colour cuts suggested for SMG counterpart identification in Yun et al. (2008). The grey dots represent a random subsample of  $\sim 10\,000$  IRAC field galaxies.

three bands. To this sample, we add an additional six SMGs with secure IRAC counterparts (detected in all four bands) determined from the SMA imaging that are radio undetected, have multiple radio counterparts, or are not in the current AzTEC/COSMOS catalogue. This gives a total of 102 IRAC counterparts which we use to build an IRAC colour space CDD. The list of radio and IRAC counterparts to radio-detected AzTEC sources<sup>1</sup> can be seen in Table A1, including the search radii used and associated  $P$ -stat. Those AzTEC sources not included in the training set due to lack of an IRAC counterpart are indicated.

We find that 8 (22 per cent), 4 (9 per cent) and 15 (8 per cent) SMGs in GOODS-N, GOODS-S and COSMOS have multiple radio counterparts<sup>1</sup> within their search radii (Table A2). SMGs with multiple radio counterparts are not included in the training set due to ambiguity in the source of the mm emission.

### 3.2 IRAC colour–colour space

Following the lead of Lacy et al. (2004) and Yun et al. (2008, 2012), we create an IRAC colour–colour diagram with the colours  $\log(\frac{S_{5.8}}{S_{3.6}})$  versus  $\log(\frac{S_{8.0}}{S_{4.5}})$ . The 102 IRAC counterparts in our training set can be seen in IRAC colour–colour space in Fig. 1, along with a random sample of 10 000 IRAC galaxies from our three IRAC catalogues. The dotted lines represent the colour cuts presented in Yun et al. (2008), which encase the bulk of IRAC colours for securely identified SMGs.

The following describes the building of the IRAC colour–colour CDD for SMGs, starting with a brief discussion of the physical mechanisms responsible for the IRAC colour–colour properties of SMGs.

#### 3.2.1 The physical mechanism behind the IRAC colour–colour distribution of SMGs

Because of the large redshift range at which SMGs have been detected thus far ( $z \sim 1$ –6 with a median redshift of  $\sim 2$ –2.5; e.g.

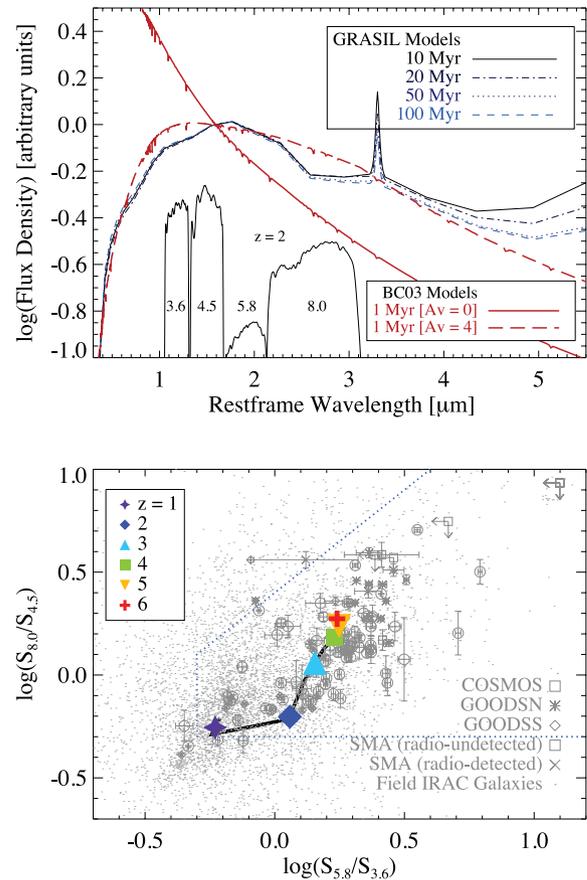
<sup>1</sup>A list of AzTEC sources and radio/IRAC counterparts, including photometry, is available at [www.astro.umass.edu/aztec/CounterpartID/counterpartid.html](http://www.astro.umass.edu/aztec/CounterpartID/counterpartid.html)

Smail et al. 2004; Chapman et al. 2005; Aretxaga et al. 2007, 2011; Chapin et al. 2009; Smolčić et al. 2012; Yun et al. 2012), the IRAC bands probe a handful of spectral features over the rest-frame wavelength range of 0.5–4.0  $\mu\text{m}$ . In a typical star-forming galaxy, this range includes the stellar photospheric feature at rest frame 1.6  $\mu\text{m}$  (the ‘stellar bump’), which is the continuum emission from the old, low-mass stars that form the bulk of the stellar mass of the galaxy, and, at wavelengths  $\gtrsim 3 \mu\text{m}$ , emission from polycyclic aromatic hydrocarbons (PAHs). In the case of extremely young galaxies, emission from OB stars can shift the bulk of the stellar emission to shorter wavelengths, depending on dust content. Systems that are dominated by AGN show power-law emission from hot dust in this wavelength regime.

To determine which of these features predominantly give rise to the observed IRAC colours of SMGs, we generated SMG galaxy models using the GRASIL SED modelling code (Silva et al. 1998) based on the following properties: SMGs are dominated by stellar emission as opposed to dust heated by AGN in the majority of cases (Alexander et al. 2005; Pope et al. 2008; Menéndez-Delmestre et al. 2009; Johnson et al. 2013), the typical time-scale of the SMG phase is  $\sim 40$ –100 Myr (Greve et al. 2005; Narayanan et al. 2010) and submm galaxies are dusty systems (Blain et al. 2002) which contain a pre-existing stellar population (Smail et al. 2004; Borys et al. 2005; Dye et al. 2008; Michałowski et al. 2012). We create several GRASIL models with galaxy ages ranging from 1 to 2 Gyr, each undergoing an exponentially declining starburst with time-scales ranging from 10 to 100 Myr and e-folding times ranging from 20 to 100 Myr. Each model contains a moderate amount of dust with a dust to gas ratio of 5 per cent and an optical depth at 1  $\mu\text{m}$  of 0.01. For extremely young star-forming systems, we compiled a set of models from Bruzual & Charlot (2003, hereafter BC03) with ages 1–100 Myr, both with no extinction and with  $A_v = 4$  using the Calzetti extinction law (Calzetti 2001). Fig. 2 (top) shows GRASIL models with a galaxy age of 1 Gyr with starburst ages of 10, 20, 50 and 100 Myr as well as two 1 Myr BC03 model SEDs, with and without dust extinction. The BC03 1 Myr model without extinction is only for reference as all SMGs are dusty systems. When extinction is added to extremely young systems, the emission from young stars is attenuated, creating a pseudo-stellar bump that is shifted to shorter wavelengths. In both model sets, the stellar bump is prominent by  $\sim 10$  Myr, making it the dominant feature in the IRAC bands at the redshifts typical of SMGs.

Of course, the stellar bump is not unique to SMGs; previous studies have established that the IRAC bands are a useful redshift selection tool for star-forming galaxies at  $z > 1$  due to the pervasiveness and robustness of the stellar bump and its shape (Pope et al. 2006; Farrah et al. 2008; Lonsdale et al. 2009; Sorba & Sawicki 2010). There is an additional complication in that at high redshift ( $z > 4$ ), the IRAC bands begin to sample only one side of the stellar bump, which can mimic a power-law shape, thus making it difficult to determine if these sources are star formation dominated or AGN dominated. We argue that AGN are rare and it is already known that there is some overlap in SMG and AGN colour in IRAC colour-colour space (Yun et al. 2008, 2012). For further discussion on the IRAC colours of AGN, see Donley et al. (2012).

The redshifted colour-colour tracks of our GRASIL models are shown in Fig. 2 (bottom) overlaid on the IRAC colours of our SMG training sample. For comparison, the IRAC colours of a random sample of field galaxies are shown as well. It can be seen from the plots in Fig. 2 that the separation in IRAC colour-colour space between SMGs and the bulk of field galaxies is predominantly a redshift effect. As such, we are essentially using the IRAC colours

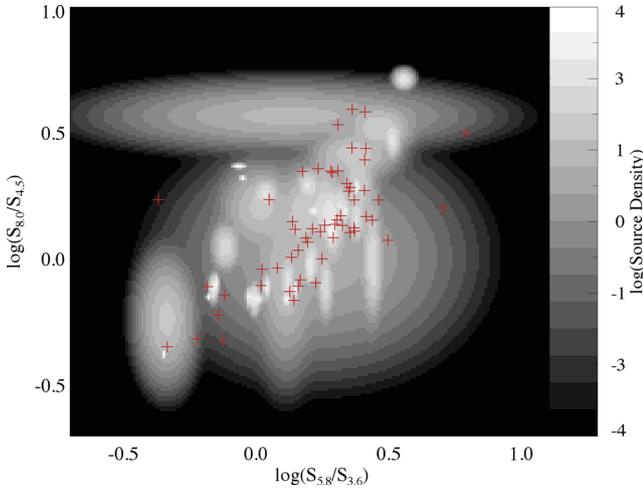


**Figure 2.** Top: starburst models of various ages with the IRAC filters at  $z = 2$  overlaid for reference. The GRASIL models represent a typical SMG with an underlying 1 Gyr old population and a starburst with ages of 10–100 Myr; these models show the ubiquitous stellar bump and a PAH feature at 3.4  $\mu\text{m}$ . The BC03 models represent an extremely young 1-Myr starburst. An extreme starburst with no dust ( $A_v = 0$ ) is not physical for SMGs, which are always dusty systems, so it is only shown for reference. When dust is added ( $A_v = 4$ ), the emission from young stars is attenuated, forming a pseudo-stellar bump even in these extreme young starbursts. In all models we examined, the stellar bump is evident by 10 Myr. Bottom: the redshifted IRAC colour-colour tracks of the GRASIL 50 Myr SMG model between  $z = 1$  and 6 overlaid on the IRAC colours of the training set and a random subsample of IRAC field galaxies.

to separate the high-redshift SMG counterparts from the unrelated low-redshift population.

### 3.2.2 Kernel density estimation

Our CDD is developed using kernel density estimation, which provides a non-parametric estimator of the underlying probability density function of a data set (Parzen 1962; Pisani 1993; Merritt & Tremblay 1994; Vio et al. 1994). Functionally, this process convolves the IRAC colours of our training set with an optimized kernel to create a smoothed distribution which approximates the underlying parent distribution. Cross-validation is used to optimize the kernel, with the optimal kernel width being a function of both the sample size of the data and the broadness (i.e. the standard deviation) of the intrinsic distribution. To build the CDD, we use the 102 radio- and SMA-identified IRAC counterparts of our training set described in Section 3.2. The steps for cross-validation are as follows.



**Figure 3.** An example of one cross-validation realization. The white contours represent the ‘working’ set, a randomly chosen half of the training set. The data points in the working set have been replaced with normalized 2D Gaussians in order to incorporate their photometric uncertainties. The red pluses represent the ‘cross-validation’ set.

(i) Divide the training set in half randomly, forming a ‘working’ set and a ‘cross-validation’ set in IRAC colour–colour space.

(ii) Incorporate the photometric uncertainties into the working set by replacing each data point within IRAC colour space with a normalized 2D Gaussian with a standard deviation defined by the  $1\sigma$  uncertainties on the colours. This can be seen in Fig. 3, overlaid with the data points that form the cross-validation set.

(iii) Next, select a kernel, which of its parameters to vary, and a suitable parameter space. We choose a 2D Gaussian kernel and allow its standard deviations,  $\sigma_{\log\left(\frac{S_{5.8}}{S_{3.6}}\right)}$  and  $\sigma_{\log\left(\frac{S_{8.0}}{S_{4.5}}\right)}$ , to vary independently. Preliminary tests showed that the optimized kernel has negligible rotation and so we do not allow the kernel to rotate in the following steps for computational simplicity. Convolve the working set with the kernel while varying its parameters over the parameter space chosen.

(iv) For each set of parameters, compare the convolved working set to the cross-validation set. The optimized kernel parameters are those which maximize the likelihood that both the convolved working set and the cross-validation set are drawn from the same parent distribution. This likelihood is calculated as

$$L = \prod_i C\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right)_i, \log\left(\frac{S_{8.0}}{S_{4.5}}\right)_i, \sigma_{\log\left(\frac{S_{5.8}}{S_{3.6}}\right)}, \sigma_{\log\left(\frac{S_{8.0}}{S_{4.5}}\right)}\right), \quad (3)$$

where  $C\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right)_i, \log\left(\frac{S_{8.0}}{S_{4.5}}\right)_i, \sigma_{\log\left(\frac{S_{5.8}}{S_{3.6}}\right)}, \sigma_{\log\left(\frac{S_{8.0}}{S_{4.5}}\right)}\right)$  is the value of the convolved working set at the  $i$ th data point in the cross-validation set.

(v) Finally, repeat steps (i)–(iv) to build a representative distribution of optimized kernel parameters. We perform the cross-validation procedure 500 times and find that the distribution of optimized Gaussian widths does not significantly vary in the range of 0.07–0.21 for both colours, which indicates that the final kernel estimation results are not overly sensitive to the exact size of the Gaussian kernel.

We adopt the mean values  $\sigma_{\log\left(\frac{S_{5.8}}{S_{3.6}}\right)} = 0.14$  and  $\sigma_{\log\left(\frac{S_{8.0}}{S_{4.5}}\right)} = 0.15$  for the final convolution of the 102 IRAC counterparts in the training set with an optimized kernel. Our final IRAC colour–

colour  $CDD_{\text{SMG}}^2$  for submillimetre galaxies can be seen in grey-scale in Fig. 4 overlaid with the original IRAC data points used to create it.

To use this CDD to identify counterparts to the rest of our SMGs, we must also determine the IRAC colour distribution for the full IRAC population. We use the same kernel density estimation discussed above on all IRAC galaxies that are detected in all four IRAC bands across the three fields,  $\sim 295\,000$  sources, and find a smaller range in optimal kernel widths, 0.02–0.06. We adopt the mean value of 0.04 for both standard deviations of the Gaussian kernel. The final distribution  $CDD_{\text{IRAC}}$ , shown in red contours in Fig. 4, peaks in a bluer regime of IRAC colour–colour space than the distribution for SMGs. This is expected since the full IRAC sample is dominated by  $z \lesssim 1$  galaxies. At these lower redshifts, the IRAC bands sample the more complex PAH emission features, and the variety in observed galaxy SEDs at low redshifts accounts for the large scatter in IRAC colour–colour space of the general IRAC population. Regardless, the distribution of the field IRAC galaxies is distinct from that of the SMG population, which allows us to make use of this method for counterpart identification.

## 4 IDENTIFYING IRAC COUNTERPARTS TO SMGS

### 4.1 The likelihood ratio and reliability factor

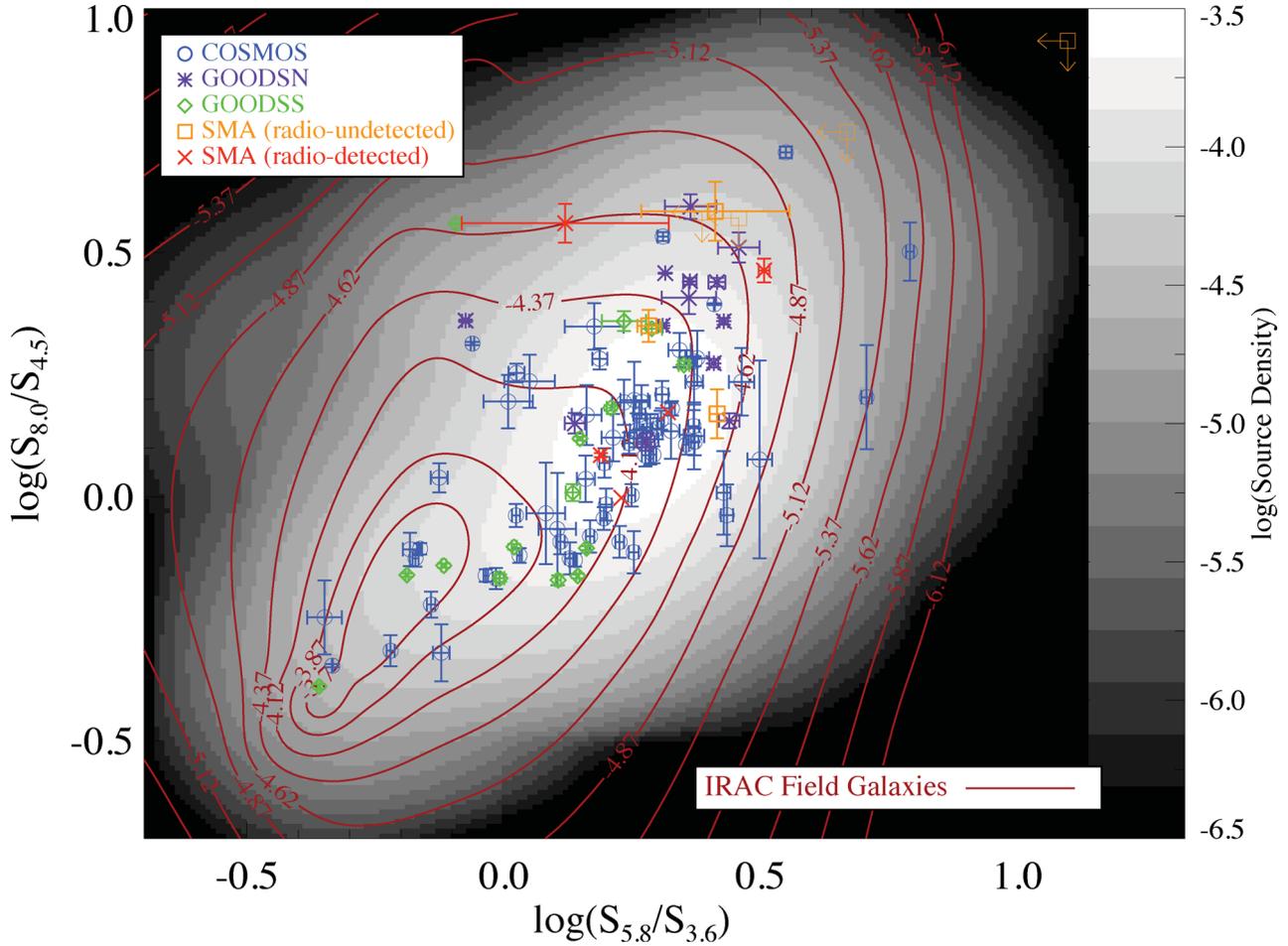
The relative probability that an IRAC galaxy is the counterpart of an SMG as opposed to being an unrelated foreground/background source is given by the likelihood ratio (Sutherland & Saunders 1992):

$$\lambda = \frac{CDD_{\text{SMG}}\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right), \log\left(\frac{S_{8.0}}{S_{4.5}}\right)\right)}{CDD_{\text{IRAC}}\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right), \log\left(\frac{S_{8.0}}{S_{4.5}}\right)\right)} P(D; S/N), \quad (4)$$

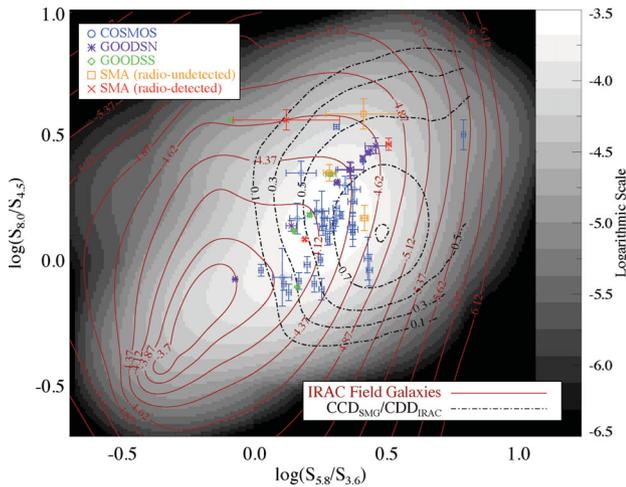
where  $CDD_{\text{SMG}}\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right), \log\left(\frac{S_{8.0}}{S_{4.5}}\right)\right)$  is the value of the CDD of SMGs at the location of a potential IRAC counterpart in IRAC colour–colour space,  $CDD_{\text{IRAC}}\left(\log\left(\frac{S_{5.8}}{S_{3.6}}\right), \log\left(\frac{S_{8.0}}{S_{4.5}}\right)\right)$  is the value of the CDD of the general IRAC population for that potential counterpart and  $P(D; S/N)$  is given by equation (1) integrated over  $D - \epsilon$  to  $D + \epsilon$ , where  $D$  is the distance of a given IRAC galaxy from the centre of the detected submm emission and  $\epsilon = 0.2$  arcsec is the positional accuracy of the IRAC observations.  $P(D; S/N)$  represents the probability that the true SMG position is found at distance  $D$ , based on the S/N and FWHM of the submm detection.

The likelihood ratio provides a means of ranking IRAC galaxies in the proximity of SMGs based on their colours and positions. Since there is no IRAC colour combination that is completely unique to SMGs, this ranking is relative within a set of potential counterparts. To determine the statistical significance of any given likelihood ratio, we calculate its  $1\sigma$  uncertainty by including the measured errors in the IRAC colours. The ratio  $CDD_{\text{SMG}}/CDD_{\text{IRAC}}$  can be seen in Fig. 5 in black, dash–dot contours, overlaid on  $CDD_{\text{SMG}}$  in grey-scale and  $CDD_{\text{IRAC}}$  in red, solid contours. Because of the rapid falling off of both the SMG and IRAC CDDs in IRAC colour–colour space, a floor value of  $\log f = -5.7$  was imposed on both distributions when plotting the ratio to suppress noise features in areas with sparse data. This floor was not used in the calculation of

<sup>2</sup> Available at [www.astro.umass.edu/aztec/CounterpartID/counterpartid.html](http://www.astro.umass.edu/aztec/CounterpartID/counterpartid.html)



**Figure 4.** The final IRAC colour–colour CDD for SMGs ( $CDD_{\text{SMG}}$ ; white image contours) overlaid with the radio- and SMA-identified IRAC counterparts used to create it, as in Fig. 1. Additionally, the CDD of all IRAC galaxies is shown ( $CDD_{\text{IRAC}}$ ; red contours).



**Figure 5.** The ratio  $CDD_{\text{SMG}}/CDD_{\text{IRAC}}$  in black, dash-dot contours, overlaid on  $CDD_{\text{SMG}}$  in the white image contours and  $CDD_{\text{IRAC}}$  in red, solid contours. The data points shown here are the IRAC counterparts in the training set which were identified correctly by their likelihood ratio.

the likelihood ratio and we have verified that these noise features are not affecting our counterpart identification results. The peak of the  $CDD_{\text{SMG}}/CDD_{\text{IRAC}}$  contours shows the area in which the colour prior portion of the likelihood ratio is at a maximum.

Additionally, we adopt the reliability factor, as described in Sutherland & Saunders (1992), which provides a single statistic for a given potential counterpart which incorporates its individual LR, the LRs of all other potential counterparts around that SMG and the probability that the true counterpart is detected. The reliability factor for potential IRAC counterpart  $j$  takes the form

$$R_j = \frac{\lambda_j}{\sum_i \lambda_i + (1 - X)}, \quad (5)$$

where  $\sum_i \lambda_i$  is the sum over the likelihood ratios of all potential IRAC counterparts around a given SMG and  $X$  is the probability that the true IRAC counterpart is included in the four-band IRAC catalogue we are using. Though interferometric studies indicate a high probability that the majority of SMGs are detected at  $\geq 1 \mu\text{Jy}$  at 3.6 and 4.5  $\mu\text{m}$  (e.g. Younger et al. 2007, 2009; Hatsukade et al. 2010; Ikarashi et al. 2011), our technique requires detections at 5.8 and 8.0  $\mu\text{m}$  as well. These bands are less sensitive than the two shorter IRAC channels, though our training set suggests a high percentage (99 per cent) of counterparts are detected in all bands. The overall fraction of IRAC sources not detected in all bands therefore provides an estimate of  $X$ .

To control for possible biases in the properties of IRAC galaxies that are nearby SMGs as opposed to the general field population, we elect to estimate  $X$  only from IRAC galaxies within the vicinity of SMGs, including both our training sample and the radio-undetected SMGs. To do this, we identify all IRAC sources within the search

radii described in Section 3.1 and listed in Tables A1–A3. We find that the fraction of IRAC sources detected in all four bands is 0.83, 0.78 and 0.73 for GOODS-N, GOODS-S and COSMOS, respectively, and we use these as approximations for  $X$  in equation (5). We note that this fraction only changes by  $<10$  per cent if we include all IRAC catalogue sources, suggesting no strong biases to this fraction from redshift selection effects.

In principle,  $\sum_i \lambda_i$  for a given SMG in equation (5) runs over the entire IRAC four-band catalogue. In practice, however,  $\lambda_i$  decreases rapidly with increasing distance between the mm peak emission and the IRAC position (equation 4). For this reason, we limit the sum in the denominator in equation (5) to IRAC galaxies within the search radii for the SMG, defined where  $P(>D; S/N) = 0.01$ , to simplify computations.

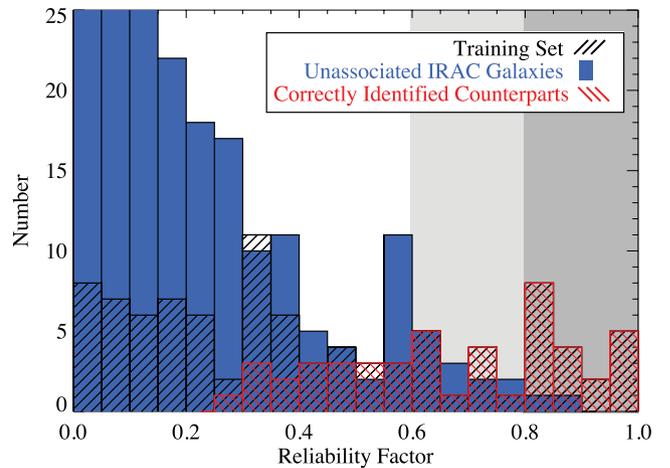
## 4.2 Counterpart identification via the reliability factor

### 4.2.1 Verification of the method: the training set

We evaluate our technique with a blind test against our radio- and SMA-identified training set. All IRAC galaxies within the appropriate search radius of each SMG are ranked using the likelihood ratio (equation 4) and assigned a reliability (equation 5). The likelihood ratio, its  $1\sigma$  errors and reliability of the radio- or SMA-identified counterpart as well as additional top ranked potential IRAC counterparts for the training set can be seen in Table A1.

In GOODS-N, 11/13 (85 per cent) IRAC counterparts are identified correctly, i.e. ranked first among the other possibilities using the likelihood ratio. Three of these sources (AzTEC/GN3, AzTEC/G5 and AzTEC/GN14) are confirmed by interferometry (Barger et al. 2012). In GOODS-S, 5/15 (33 per cent), and in COSMOS, 33/68 (49 per cent), are identified correctly, including one radio-undetected SMG (AzTEC/C4), one multiple radio system (AzTEC/C42, see below) and two radio-detected SMGs (AzTEC/C18 and AzTEC/C22) which were imaged with the SMA. The correctly identified counterparts, as well as SMA-confirmed counterparts (Younger et al. 2007, 2009), can be seen in Fig. 5, overlaid on contours showing  $CDD_{\text{SMG}}$ ,  $CDD_{\text{IRAC}}$  and the ratio  $CDD_{\text{SMG}}/CDD_{\text{IRAC}}$ . The difference in the success rates of GOODS-N versus COSMOS and GOODS-S stems from the difference in beam sizes: 18 arcsec in GOODS-N compared to 33 and 30 arcsec in COSMOS and GOODS-S. Possible explanations for the more subtle differences between COSMOS and GOODS-S are discussed in Section 5.2. The distribution of counterpart rankings (1st, 2nd, 3rd and 3rd+) in each field can be seen in Table 1.

The distribution of reliabilities for the radio- and SMA-identified training set can be seen in Fig. 6. A large  $R$  value indicates a high probability that a given IRAC galaxy is the true counterpart as well as a high probability that only a single galaxy is responsible for the



**Figure 6.** The distribution of reliabilities for the IRAC candidates in the radio- and SMA-identified training set. The entire training set (black) is correctly identified (ranked first via the likelihood ratio; red) above  $R = 0.5$ , though there is contamination from possibly unassociated IRAC galaxies at almost all reliabilities (solid blue). The darker shaded region represents what we consider to be the robust range of reliabilities, while the lighter shaded region defines a tentative range.

submm/mm emission. A low  $R$  value indicates a low probability of being a counterpart and/or that the corresponding SMG consists of multiple galaxies blended within the beam (see Section 5.1). Fig. 6 shows that for  $R \geq 0.5$ , all radio- or SMA-identified IRAC counterparts are identified correctly, whereas for  $R \leq 0.5$ , many radio-identified IRAC counterparts are not. We suggest a conservative cut-off in the reliability factor would be at  $R \geq 0.8$ , where there are only two misidentifications out of 19. This cut-off is the same as that found in similar studies utilizing the reliability factor (e.g. Smith et al. 2012). Additionally, an analysis of the field-by-field reliability factors suggests that extending the cut-off to  $R \geq 0.6$  is appropriate in fields analogous to GOODS-N, where 9/9 (100 per cent) IRAC galaxies assigned  $R \geq 0.6$  are the correct counterpart. The range  $0.6 \leq R \leq 0.8$  is more appropriately named a ‘tentative’ counterpart list in COSMOS with only 18/28 (64 per cent) IRAC galaxies assigned  $R \geq 0.6$  being the correct counterpart. In GOODS-S, this ratio drops to 3/7 (43 per cent) (see Table 1).

### 4.2.2 Multiple radio systems and radio undetected SMGs

The likelihood ratios and reliabilities for SMGs with multiple radio counterparts can be seen in Table A2. Some of these systems may be blended, multiple systems and, as such, the IRAC counterparts for many of these systems have low reliabilities. There are a small number of cases (5/27), however, where a single IRAC candidate has an  $R \geq 0.8$  and is thus likely associated with only one or none

**Table 1.** Statistical results of counterpart identification. Columns 2–5 show the distribution of ranks for the training set per field. Being ranked first is considered a correct identification. Additionally shown are the number of robust and tentative reliabilities in the training set and the radio-undetected set.

Field	Ranked 1st	Ranked 2nd	Ranked 3rd	Ranked 3rd+	$R \geq 0.8$ ( $R \geq 0.6$ )	
					Training set <sup>a</sup>	Radio undetected
GOODS-N	11 (85 per cent)	2 (15 per cent)	0	0	6/6 (9/9)	2 (8)
GOODS-S	5 (33 per cent)	5 (33 per cent)	2 (13 per cent)	3 (20 per cent)	2/3 (3/7)	6 (14)
COSMOS	33 (49 per cent)	15 (22 per cent)	8 (11 per cent)	12 (17 per cent)	11/12 (18/28)	21 (48)

<sup>a</sup>The number of correctly identified counterparts in the training set with a reliability factor above 0.8 (0.6) compared to the total number of IRAC galaxies assigned a reliability above that value.

of the radio sources. One of these cases, AzTEC/C42, was observed with the SMA, which confirmed that the submm emission is only coming from one of the two radio sources within the AzTEC beam. The IRAC galaxy associated with the true radio counterpart was correctly identified with our technique.

The ranking and reliabilities for the IRAC galaxies corresponding to the radio-undetected SMG set<sup>3</sup> can be seen in Table A3. There are 29 (70) IRAC galaxies associated with radio-undetected SMGs with  $R \geq 0.8$  ( $R \geq 0.6$ ) (see Table 1).

## 5 DISCUSSION

### 5.1 Source blending

Multiple studies have established that some subset of SMGs is in fact blends of multiple sources of submm emission which appear as one source in the large beams of single-dish submm observations. Our best estimates of the blending fraction come from interferometric studies, which have determined that 10–30 per cent of known SMGs are multiple systems (Younger et al. 2007, 2009; Barger et al. 2012; Smolčić et al. 2012). A recent study of LABOCA sources imaged with Atacama Large Millimeter/submillimeter Array (ALMA) determined that the overall blending fraction in their sample of 88 is 26 per cent, though this fraction is higher for the very brightest SMGs ( $S_{870\mu\text{m}} \geq 9\text{ mJy}$ ; Karim et al. 2012).

The formulation of the reliability factor is such that it assumes there is only one unique counterpart. In the event of blending, the reliability of a potential counterpart will reflect that there are multiple sources with large LR values by assigning these sources correspondingly low reliabilities. These cases are indistinguishable from those in which an unrelated IRAC galaxy is assigned a high LR due to a chance superposition and a redshift that is consistent with the peak in the IRAC colour–colour distribution of SMGs. For this reason, the reliability factor is biased against the identification of blended submm systems.

Currently, there is no formulation of the reliability factor that accounts for the possibility of multiple counterparts. An alternative to the reliability factor was presented in Chapin et al. (2011), which used simulations to determine a cut-off threshold on the likelihood ratio above which the rate of false positives was  $\sim 10$  per cent. We performed a preliminary analysis to determine the utility of this technique with our data set as follows: in COSMOS and GOODS-N, we choose 10 000 random locations with the areas around known SMGs masked out and calculate the LR of all IRAC galaxies within a search radius as described in Section 3. As the LR incorporates positional information based on the S/N of the submm observation, we repeated this process for a range of S/N bins. A likelihood ratio threshold is determined such that 90 per cent of the IRAC galaxies around random positions are rejected, which corresponds to a 10 per cent false positive rate. The LR of each potential IRAC counterpart is then divided by the appropriate threshold according to the S/N of the corresponding SMG. As in Chapin et al. (2011), a normalized likelihood ratio (nLR) of greater than one indicates a candidate counterpart.

An analysis of the results of using a nLR finds a higher fraction of SMGs with multiple sources assigned  $\text{nLR} > 1$ , indicating blends, than we would expect from interferometric studies. In COSMOS,

we find that 31/68 (46 per cent) of SMGs have more than one counterpart with  $\text{nLR} > 1$  and in GOODS-N this number is 8/13 (62 per cent). In addition, we examined multiple AzTEC sources that correspond to SCUBA sources in Barger et al. (2012), which were determined to not be blended at the resolution of the SMA. All five sources have multiple counterparts with  $\text{nLR} > 1$ .

These results indicate that the normalized likelihood ratio, when used in conjunction with our data set and methods, suffers from contamination from unrelated IRAC galaxies. An additional difficulty with these simulations arises in small fields such as GOODS-N, where there are only  $\sim 100$  unique locations in the map given our typical search radius. While the normalized likelihood ratio is not biased against blended systems, it presents cases where blending is indistinguishable from confusion due to unrelated IRAC galaxies in excess of the 10 per cent false positive rate expected. As such, we find the reliability factor to be more conservative in identifying unambiguous, single counterparts and do not present the normalized likelihood ratios in this work.

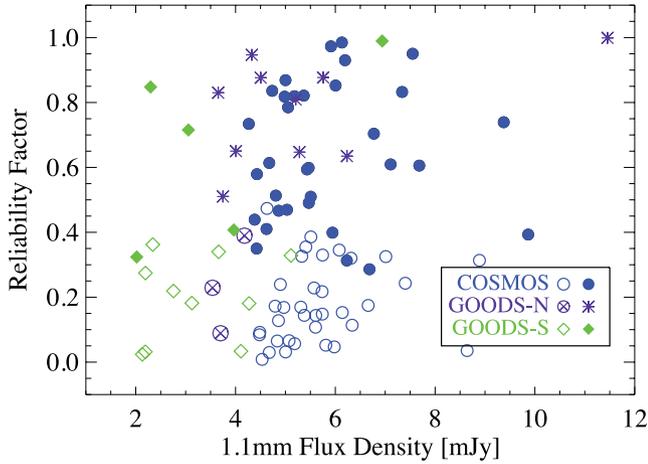
### 5.2 Dependence on the 1.1 mm and IRAC observations

In the previous section, we found that the accuracy of our technique, as blind tested by our training set, improves with decreasing beam size as we would expect. Here we explore additional effects on this technique; particularly we look at differences between the COSMOS and GOODS-S fields, since, despite being observed with similar beam sizes, our technique performs more accurately in COSMOS over GOODS-S by  $\sim 10$  per cent.

The submm number counts reveal that SMGs in GOODS-S are systematically fainter than in COSMOS (Scott et al. 2012). Since our technique depends on the detectability of SMGs in all IRAC bands, we consider whether fainter submm sources implicitly translate to fainter IRAC counterparts. Our training set indicates that this is not the case as we find no correlation between the 1.1 mm flux density and any of the IRAC bands. This is what we would expect, given that the 1.1 mm flux density is powered by dust emission and the IRAC bands primarily probe the mostly dust-free stellar bump at redshifts characteristic of SMGs. Additionally, we considered whether brighter or fainter SMGs have more easily identifiable counterparts due to some external factor which affects their IRAC colours, such as their redshift distribution. Fig. 7 shows the reliabilities of SMGs in our training set versus their 1.1 mm flux density; the two variables have a Pearson correlation coefficient of  $\sim 0.05$ , consistent with being uncorrelated. Cutting the GOODS-S AzTEC catalogue to the depth of the COSMOS AzTEC catalogue similarly produces no improvement in the counterpart identification results in GOODS-S.

Next, we consider what effect the depth of the IRAC catalogue may have on our analysis. To directly compare GOODS-S and COSMOS, we cut the GOODS-S IRAC catalogue to the depth of the COSMOS catalogue, removing any sources with a  $3.6\mu\text{m}$  detection less than  $1\mu\text{Jy}$ . This results in an improved identification rate, with 7/15 (46 per cent) IRAC counterparts to SMGs in GOODS-S correctly identified. This is similar to the 49 per cent identification rate in COSMOS. This effect is due to the large beam size of the AzTEC observations, necessitating a larger search radius for counterparts, and the depth of the IRAC observations in GOODS-S increasing the chances of detecting a high-redshift galaxy which is unrelated to the SMG, but, due to its redshift, has similar IRAC colours. These chance associations can be mitigated by higher resolution submm/mm observations or by cutting the IRAC catalogue, as SMGs will be brighter in the IRAC bands than other typical

<sup>3</sup>A list of AzTEC sources and potential IRAC counterparts, including photometry, is available at [www.astro.umass.edu/aztec/CounterpartID/counterpartid.html](http://www.astro.umass.edu/aztec/CounterpartID/counterpartid.html)



**Figure 7.** The dependence of the reliability factor on the AzTEC flux density for training set counterparts that are (1) ranked first (identified correctly) with our technique (filled symbols) and (2) those that are not ranked first (open symbols). There is no correlation, which suggests that the AzTEC flux density does not directly affect our counterpart identification technique.

high-redshift galaxies due to having larger stellar masses by about an order of magnitude (see Hainline et al. 2011, and references therein).

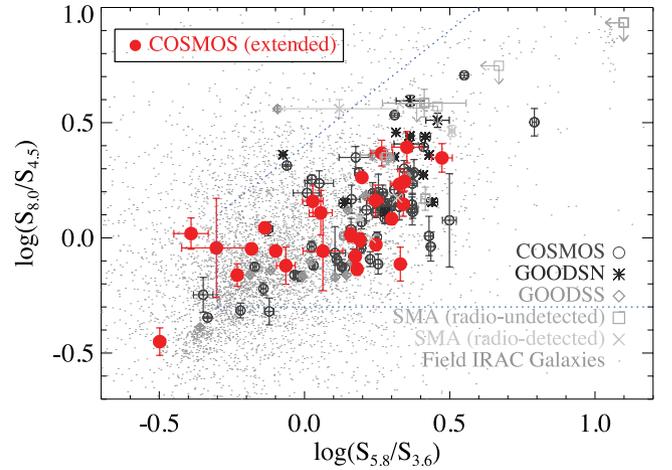
### 5.3 Dependence on the radio flux density

The utility of our technique depends in large part on the applicability of the IRAC CDD to radio-undetected SMGs. We test this in two ways. First, we check that our IRAC colour distribution is robust against the depth of the radio observations. Secondly, we split our training set into bright and faint radio counterparts, to determine if there is a systematic difference in their IRAC colours.

The effects of the radio depth on our technique is tested by extending the COSMOS radio catalogue from the original  $4.5\sigma$  cut-off to  $3\sigma$  (e.g. Chapman et al. 2005; Yun et al. 2012). Since the false positive rate increases dramatically with decreasing S/N, we identify all  $\geq 3\sigma$  peaks above the local noise in the radio map and then match these peaks to the COSMOS IRAC catalogue. Only those radio peaks with a corresponding IRAC counterpart within 2 arcsec are considered.

Originally, we found 74 single radio counterparts to SMGs, plus 15 multiple radio systems. Using the deeper radio catalogue, we find 26 new single radio detections and three new multiple radio detections around previously radio-undetected SMGs. In addition, 7/74 COSMOS SMGs in the original training set were found to have a second radio source within their search radius and one double system was found to have a third radio source (AzTEC/C184). This makes a total of 118/189 (62 per cent) radio-detected SMGs in COSMOS, including 25 (13 per cent) multiple radio systems (Tables A1 and A2). The IRAC colours of the new single radio detections can be seen overlaid on our original IRAC counterpart set in Fig. 8. A 2D Kolmogorov–Smirnov (KS) test (Fasano & Franceschini 1987) returns a value of 0.84, indicating the original training set and extended radio counterparts are consistent with being drawn from the same parent distribution. When analysed using the IRAC CDD built with the original training set, 12/26 (46 per cent) of the new radio sources were ranked first (four with  $R \geq 0.8$ ).

Combining the new single radio counterparts with the previous training set (and removing the new multiple radio systems), we perform the same kernel density estimation as before and apply this



**Figure 8.** IRAC colours of COSMOS radio sources in the extended catalogue ( $S/N = 3\text{--}4.5$ ; red filled circles) overlaid on the original training set used in this study (see Fig. 1).

new IRAC colour–colour CDD in a blind test as in Section 4.1. We find that the addition of 26 new radio counterparts has a negligible effect on the resulting likelihood ratios and reliabilities assigned to each potential IRAC counterpart. This indicates that our IRAC colour–colour CDD is well sampled and robust against the depth of the radio observations.

As an additional check against potential bias on the IRAC colours used in this technique, we split our training set in half to analyse the IRAC colours of radio counterparts with radio flux densities greater than and less than  $65 \mu\text{Jy}$ . A 2D KS test of the two resulting distributions in IRAC colour–colour space gives a value of 0.98, confirming that the IRAC colours of SMGs with radio flux densities above and below  $65 \mu\text{Jy}$  are consistent with being drawn from the same parent distribution. This indicates that there is no dependence on the IRAC CDD on the radio flux density.

Given these two tests, we see no evidence that the IRAC colours of bright and faint radio counterparts are systematically different, which is consistent with the detectability of radio counterparts being a function of the depth of the radio observations and a selection effect due to the strong positive  $k$ -correction in the radio regime. This result gives us confidence that the IRAC colour–colour CDD developed in this study can be applied to our radio-undetected SMG sample.

### 5.4 In the era of warm *Spitzer*, *WISE* and *JWST*

In this study, we have utilized data from the four IRAC bands available during *Spitzer*'s cryogenic mission. This mission surveyed 10s of square degrees on the sky over several wide, deep Legacy Programs, a summary of which can be seen in Table 2. These fields

**Table 2.** Summary of cryogenic mission *Spitzer* Legacy Programs.

Survey	Area	Reference
GOODS	280 arcmin <sup>2</sup>	Dickinson et al. (2003)
SIMPLE	~0.5 deg <sup>2</sup>	Damen et al. (2011)
SpUDS	1 deg <sup>2</sup>	Dunlop et al. (2007)
S-COSMOS	2 deg <sup>2</sup>	Sanders et al. (2007)
SDWFS	10 deg <sup>2</sup>	Ashby et al. (2009)
SWIRE	~60 deg <sup>2</sup>	Lonsdale et al. (2003)

will be natural targets for future submm and mm observations.

In 2009 May, the cryogenic mission of *Spitzer* ended and it began the ‘warm’ mission phase which includes only the 3.6 and 4.5  $\mu\text{m}$  channels. Unfortunately, the close spacing of the two warm IRAC channels means that at  $z \gtrsim 2$ , both bands fall on the same side of the stellar bump and their redshift discriminating power is much diminished. As such, performing a CDD analysis as seen here is not beneficial for SMGs discovered in fields with only warm *Spitzer* coverage, and we must resort to simpler selection methods, such as adopting a straight colour cut or using a colour cut in the formation of an IRAC colour prior or *P*-stat (Yun et al. 2008, 2012; Chapin et al. 2011).

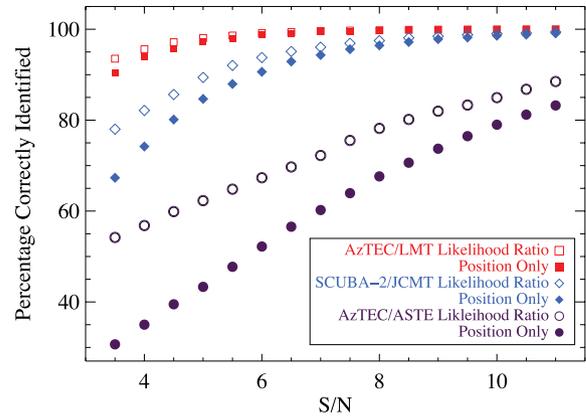
The *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al. 2010) is an infrared survey which will map the entire sky at 3.4, 4.6, 12 and 22  $\mu\text{m}$  to  $5\sigma$  depths of 0.08, 0.11, 1 and 6 mJy. Though this is far too shallow to detect the infrared counterparts of normal SMGs, *WISE* will be able to detect strongly lensed submm background galaxies, with magnification factors up to  $30\times$  (Egami et al. 2010). Though these lensed sources are rare (e.g. Negrello et al. 2010; Swinbank et al. 2010; Conley et al. 2011), surveys with *Herschel* (e.g. Eales et al. 2010; Egami et al. 2010; Oliver et al. 2010), the South Pole Telescope (SPT; Vieira et al. 2010), Atacama Cosmology Telescope (ACT; Marriage et al. 2011) and future large-scale SMG surveys with the LMT and SCUBA-2 will dramatically increase the number of known lensed submm galaxies. The three short *WISE* bands will bracket the stellar bump at  $z > 2$  and have counterpart discriminating power similar to that derived in this study.

*Hubble*’s successor, the *James Webb Space Telescope* (*JWST*), will have two instruments, NIRCam and MIRI, which together will provide imaging in the range 0.6 to 28  $\mu\text{m}$ , covering the wavelength regime of *Spitzer* IRAC and MIPS 24  $\mu\text{m}$ . With sensitivities  $\sim 50\times$  better than IRAC, *JWST* will be able to detect mid-infrared counterparts out to very high redshifts in multiple bands which bracket the stellar bump.

### 5.5 Use in ongoing and future submm/mm surveys

We are currently seeing major advances in single-dish submm/mm observations, with improved resolution for observations at 450–1100  $\mu\text{m}$ . Large surveys are being carried out by SCUBA-2 on the JCMT (450, 850  $\mu\text{m}$ ; Dempsey et al. 2012) and, in the near future, by AzTEC on the LMT (1100  $\mu\text{m}$ ). To put our technique in the context of these surveys, we simulate the counterpart identification results that can be expected with the improved positional accuracy of these observations. We test two beam sizes: 14.5 arcsec, which is the approximate FWHM of SCUBA-2 at 850  $\mu\text{m}$ , and 8.5 arcsec, the FWHM of AzTEC on the LMT in its early 32-m configuration. The latter beam size is also close to the FWHM of SCUBA-2 at 450  $\mu\text{m}$  ( $\sim 7$  arcsec). Additionally, we simulate 33 arcsec, the beam size of AzTEC on ASTE and the COSMOS observations used in this study, which is comparable to the beam size for SPIRE 500  $\mu\text{m}$  observations.

The simulations are done using the 68 SMGs in the COSMOS field that have a known radio and IRAC counterpart. To simulate the effects of random noise in the mm observations, a random position is drawn for each SMG from a distribution given by equation (1) which is centred on the radio counterpart and has a FWHM determined by the observations being simulated and a fixed S/N (equation 2). This random position determines the positional uncertainty, which is then combined with the IRAC colours of all candidate counterparts within an appropriate search radius as per the same blind test performed in Section 4.1. This procedure is re-



**Figure 9.** The simulated percentage of correctly identified IRAC counterparts in COSMOS for beam sizes of 33 arcsec (circles), 14.5 arcsec (diamonds) and 8.5 arcsec (squares), representing AzTEC on ASTE (this study), SCUBA-2 on JCMT and AzTEC on LMT, respectively. The results using this technique (empty symbols) are an improvement over using just positions (filled symbols) in the lowest S/N bins for all beam sizes.

peated 1000 times for each SMG for different S/N bins. The results of the simulation can be seen in Fig. 9, contrasted with the results of using positional information only to predict the counterpart. Our technique is an improvement over using only position in the lowest S/N bins in all cases, and as expected the amount of improvement is largest for increasing beam size. These simulations show that, even with higher resolution in future wide-area surveys, this technique can provide better predictions of SMG counterparts over positional coincidence, as well as a measure of robustness of the prediction through the reliability factor.

## 6 SUMMARY AND CONCLUSIONS

In this study, we have developed a new technique for multiwavelength counterpart identification to submm detections by expanding on previous colour cut techniques to create a CDD of IRAC colours for the SMG population. Our IRAC colour–colour distribution is based on a statistically significant set of 102 SMG counterparts determined through radio and SMA detections and built using kernel density estimation with cross-validation. When combined with positional uncertainty information via the likelihood ratio, our technique provides a method for ranking IRAC galaxies surrounding SMGs according to their relative probability of being the correct counterpart. In addition, we use the reliability factor to weigh the importance of a given likelihood ratio against the likelihood ratios of other candidate IRAC counterparts.

Once the IRAC colour distribution is determined through known counterparts as in this study, it can then be applied to SMGs that lack other ancillary data, such as deep radio or MIPS 24  $\mu\text{m}$  observations. The strength of the CDD for SMGs stems from the ubiquitous nature of IRAC detections of SMGs and the discriminating power of the IRAC colours, which can be used as a redshift indicator for galaxy populations which (1) display a stellar bump in their infrared SEDs and (2) have a redshift distribution which peaks above  $z \gtrsim 1$ . Our overall technique, however, is general and can be applied to any population which occupies a unique colour–magnitude or colour–colour space.

In a blind test using our training set, we show that our technique works well for mm observations with a moderate beam size (e.g. 18 arcsec in GOODS-N or 14.5 arcsec for SCUBA-2 surveys) with an identification rate of 85 per cent, comparable to the current identification rate in GOODS-N, which utilizes extremely deep radio observations (Chapin et al. 2009). Our simulations demonstrate that this technique affords improvement over using only positional information, even for smaller beam sizes (e.g. 8.5 arcsec for AzTEC on the LMT). For larger beam sizes ( $\gtrsim 30$  arcsec), increased positional uncertainties and chance associations with high-redshift non-SMG galaxies reduce the identification rate to 33–49 per cent. These rates are comparable to current identification rates in COSMOS and GOODS-S (Aretxaga et al. 2011; Yun et al. 2012), using moderately deep radio observations. Our technique, however, identifies a different subpopulation of counterparts than previous radio studies by using IRAC data alone to determine counterparts. These counterparts do not necessarily overlap with those identified through radio surveys and may have different properties, such as redshift distribution. Testing reveals no dependence of this method on the radio flux density, indicating it is applicable to radio-undetected SMGs. Future expansions of this technique would include incorporating upper limits in order to allow for IRAC sources that are not detected in all four bands, and the expansion of the CDD beyond two dimensions to include additional information from a wider range of flux densities and colours.

Though the challenges put forth by submm counterpart identification will be eased in the near future by instruments and facilities such as SCUBA-2 on the JCMT, AzTEC on the LMT and the relatively rapid interferometry of ALMA, there will still be a great need for practical and efficient counterpart identification techniques. Single-dish facilities will increase the number of submm sources available for study by orders of magnitude within the next few years, making interferometric follow-up of each source impractical, while the need for multiwavelength analysis remains vital for a complete picture of star-forming galaxies and galaxy evolution. While this study only addresses long-wavelength submm/mm observations, observations at shorter wavelengths in the far-infrared/submm such as with BLAST or *Herschel* suffer from similar counterpart identification challenges, indicating the continuing importance of counterpart identification studies.

## ACKNOWLEDGEMENTS

This work has been funded, in part, by NSF grant AST-0907952 and CONACyT projects CB-2009-133260 and CB-2011-167291. This study is based, in part, on observations made with the *Spitzer Space Telescope*, which is operated by the Jet Propulsion Laboratory at the California Institute of Technology under contract with NASA, and the National Radio Astronomy Observatory, which is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. KSS is supported by the National Radio Astronomy Observatory. SK is supported in part by Mid-career Researcher Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology 2011-0028001. The James Clerk Maxwell Telescope is operated by the Joint Astronomy Centre on behalf of the Science and Technology Facilities Council of the United Kingdom, the Netherlands Organisation for Scientific Research and the National Research Council of Canada. The ASTE project is driven by Nobeyama Radio Observatory (NRO), a branch of National Astronomical Observatory of Japan (NAOJ), in collaboration with University of Chile, and Japanese institutes including University

of Tokyo, Nagoya University, Osaka Prefecture University, Ibaraki University, Hokkaido University and Joetsu University of Education.

## REFERENCES

- Alexander D. M., Bauer F. E., Chapman S. C., Smail I., Blain A. W., Brandt W. N., Ivison R. J., 2005, *ApJ*, 632, 736  
 Aretxaga I. et al., 2007, *MNRAS*, 379, 1571  
 Aretxaga I. et al., 2011, *MNRAS*, 415, 3831  
 Ashby M. L. N. et al., 2009, *ApJ*, 701, 428  
 Austermann J. E. et al., 2010, *MNRAS*, 401, 160  
 Barger A. J., Cowie L. L., Sanders D. B., Fulton E., Taniguchi Y., Sato Y., Kawara K., Okuda H., 1998, *Nat*, 394, 248  
 Barger K. A., Haffner L. M., Wakker B. P., Hill A. S., Madsen G. J., Duncan A. K., 2012, *ApJ*, 761, 145  
 Biggs A. D. et al., 2011, *MNRAS*, 413, 2314  
 Blain A. W., Smail I., Ivison R. J., Kneib J.-P., Frayer D. T., 2002, *Phys. Rep.*, 369, 111  
 Borys C., Chapman S., Halpern M., Scott D., 2003, *MNRAS*, 344, 385  
 Borys C., Smail I., Chapman S. C., Blain A. W., Alexander D. M., Ivison R. J., 2005, *ApJ*, 635, 853  
 Bruzual G., Charlot S., 2003, *MNRAS*, 344, 1000 (BC03)  
 Calzetti D., 2001, *PASP*, 113, 1449  
 Chapin E. L. et al., 2009, *MNRAS*, 398, 1793  
 Chapin E. L. et al., 2011, *MNRAS*, 411, 505  
 Chapman S. C., Blain A. W., Smail I., Ivison R. J., 2005, *ApJ*, 622, 772  
 Condon J. J., 1992, *ARA&A*, 30, 575  
 Conley A. et al., 2011, *ApJ*, 732, L35  
 Coppin K. et al., 2006, *MNRAS*, 372, 1621  
 Coppin K. et al., 2008, *MNRAS*, 384, 1597  
 Cowie L. L., Barger A. J., Kneib J.-P., 2002, *AJ*, 123, 2197  
 Damen M. et al., 2011, *ApJ*, 727, 1  
 Dempsey J. T. et al., 2012, *Proc. SPIE*, 8452, 2  
 Devlin M. J. et al., 2009, *Nat*, 458, 737  
 Dickinson M. et al., 2003, in Bender R., Renzini A., eds, *Proc. ESO/USM Workshop, The Mass in Galaxies at Low and High Redshift*. Springer-Verlag, Berlin, p. 324  
 Donley J. L. et al., 2012, *ApJ*, 748, 142  
 Downes A. J. B., Peacock J. A., Savage A., Carrie D. R., 1986, *MNRAS*, 218, 31  
 Downes T. P., Welch D., Scott K. S., Austermann J., Wilson G. W., Yun M. S., 2012, *MNRAS*, 423, 529  
 Dunlop J. et al., 2007, *Spitzer Proposal ID #40021*  
 Dye S. et al., 2008, *MNRAS*, 386, 1107  
 Eales S., Lilly S., Gear W., Dunne L., Bond J. R., Hammer F., Le Fèvre O., Crampton D., 1999, *ApJ*, 515, 518  
 Eales S. et al., 2010, *PASP*, 122, 499  
 Egami E. et al., 2010, *A&A*, 518, L12  
 Ezawa H., Kawabe R., Kohno K., Yamamoto S., 2004, *Proc. SPIE*, 5489, 763  
 Farrah D. et al., 2008, *ApJ*, 677, 957  
 Fasano G., Franceschini A., 1987, *MNRAS*, 225, 155  
 Fazio G. G. et al., 2004a, *ApJS*, 154, 10  
 Fazio G. G. et al., 2004b, *ApJS*, 154, 39  
 González-Nuevo J. et al., 2012, *ApJ*, 749, 65  
 Greve T. R., Ivison R. J., Bertoldi F., Stevens J. A., Dunlop J. S., Lutz D., Carilli C. L., 2004, *MNRAS*, 354, 779  
 Greve T. R. et al., 2005, *MNRAS*, 359, 1165  
 Hainline L. J., Blain A. W., Smail I., Alexander D. M., Armus L., Chapman S. C., Ivison R. J., 2011, *ApJ*, 740, 96  
 Hatsukade B. et al., 2010, *ApJ*, 711, 974  
 Hughes D. H. et al., 1998, *Nat*, 394, 241  
 Ikarashi S. et al., 2011, *MNRAS*, 415, 3081  
 Iono D. et al., 2006, *ApJ*, 640, L1  
 Ivison R. J. et al., 2002, *MNRAS*, 337, 1  
 Ivison R. J. et al., 2007, *MNRAS*, 380, 199

- Johnson S. P. et al., 2013, MNRAS, preprint (arXiv:1302.0842)
- Karim A. et al., 2012, MNRAS, preprint (arXiv:1210.0249)
- Kellermann K. I., Fomalont E. B., Mainieri V., Padovani P., Rosati P., Shaver P., Tozzi P., Miller N., 2008, ApJS, 179, 71
- Kim S. et al., 2012, ApJ, 756, 28
- Kovács A., Chapman S. C., Dowell C. D., Blain A. W., Ivison R. J., Smail I., Phillips T. G., 2006, ApJ, 650, 592
- Lacy M. et al., 2004, ApJS, 154, 166
- Laurent G. T. et al., 2005, ApJ, 623, 742
- Lindner R. R. et al., 2011, ApJ, 737, 83
- Lonsdale C. J. et al., 2003, PASP, 115, 897
- Lonsdale C. J. et al., 2009, ApJ, 692, 422
- Marriage T. A. et al., 2011, ApJ, 731, 100
- Menéndez-Delmestre K. et al., 2009, ApJ, 699, 667
- Merritt D., Tremblay B., 1994, AJ, 108, 514
- Michałowski M. J., Dunlop J. S., Cirasuolo M., Hjorth J., Hayward C. C., Watson D., 2012, A&A, 541, A85
- Morrison G. E., Owen F. N., Dickinson M., Ivison R. J., Ibar E., 2010, ApJS, 188, 178
- Murphy E. J., Chary R.-R., Dickinson M., Pope A., Frayer D. T., Lin L., 2011, ApJ, 732, 126
- Narayanan D., Hayward C. C., Cox T. J., Hernquist L., Jonsson P., Younger J. D., Groves B., 2010, MNRAS, 401, 1613
- Negrello M. et al., 2010, Sci, 330, 800
- Oliver S. J. et al., 2010, A&A, 518, L21
- Parzen E., 1962, Stochastic Processes. Holden Day Series in Probability and Statistics. Holden-Day, San Francisco
- Pascale E. et al., 2008, ApJ, 681, 400
- Perera T. A. et al., 2008, MNRAS, 391, 1227
- Pisani A., 1993, MNRAS, 265, 706
- Polletta M. et al., 2007, ApJ, 663, 81
- Pope A. A., 2007, PhD thesis, Univ. British Columbia
- Pope A. et al., 2006, MNRAS, 370, 1185
- Pope A. et al., 2008, ApJ, 675, 1171
- Sanders D. B. et al., 2007, ApJS, 172, 86
- Schinnerer E. et al., 2010, ApJS, 188, 384
- Scott S. E. et al., 2002, MNRAS, 331, 817
- Scott K. S. et al., 2008, MNRAS, 385, 2225
- Scott K. S. et al., 2010, MNRAS, 405, 2260
- Scott K. S. et al., 2012, MNRAS, 423, 575
- Silva L., Granato G. L., Bressan A., Danese L., 1998, ApJ, 509, 103
- Siringo G. et al., 2009, A&A, 497, 945
- Smail I., Ivison R. J., Blain A. W., 1997, ApJ, 490, L5
- Smail I., Chapman S. C., Blain A. W., Ivison R. J., 2004, ApJ, 616, 71
- Smith A. J. et al., 2012, MNRAS, 419, 377
- Smolčić V. et al., 2012, A&A, 548, A4
- Sorba R., Sawicki M., 2010, ApJ, 721, 1056
- Sutherland W., Saunders W., 1992, MNRAS, 259, 413
- Swinbank A. M. et al., 2010, MNRAS, 405, 234
- Tamura Y. et al., 2010, ApJ, 724, 1270
- Treister E. et al., 2006, ApJ, 640, 603
- Vieira J. D. et al., 2010, ApJ, 719, 763
- Vio R., Fasano G., Lazzarin M., Lessi O., 1994, A&A, 289, 640
- Wang W.-H., Cowie L. L., van Sadars J., Barger A. J., Williams J. P., 2007, ApJ, 670, L89
- Wang W.-H., Cowie L. L., Barger A. J., Williams J. P., 2011, ApJ, 726, L18
- Webb T. M. et al., 2003, ApJ, 587, 41
- Weiß A. et al., 2009, ApJ, 707, 1201
- Wilson G. W. et al., 2008, MNRAS, 386, 807
- Wright E. L. et al., 2010, AJ, 140, 1868
- Yan H., Dickinson M., Giavalisco M., Stern D., Eisenhardt P. R. M., Ferguson H. C., 2006, ApJ, 651, 24
- Younger J. D. et al., 2007, ApJ, 671, 1531
- Younger J. D. et al., 2009, ApJ, 704, 803
- Yun M. S., Reddy N. A., Condon J. J., 2001, ApJ, 554, 803
- Yun M. S. et al., 2008, MNRAS, 389, 333
- Yun M. S. et al., 2012, MNRAS, 420, 957

## APPENDIX A: DATA TABLES

**Table A1.** Training and radio counterpart set: this table contains the radio counterparts to AzTEC SMGs in GOODS-N, GOODS-S and COSMOS. The full table (385 rows) is available online. Sources are included in the training set if the radio counterpart was detected in the original radio catalogues (non-extended, see Section 5.3) and if there is a corresponding IRAC detection to the radio counterpart that is detected in all four bands. For COSMOS, radio detections which were found in the extended catalogue (see Section 5.3) are indicated. Three additional radio-undetected SMA-confirmed IRAC counterparts are listed and included in the training set. For each source, the IRAC counterpart (if available) is listed on the same line as the radio counterpart, along with its rank, likelihood ratio and reliability factor. The next two highest ranked IRAC galaxies are additionally listed (if available). Some AzTEC sources in COSMOS are cross-listed with Table A2 due to having an additional nearby radio source in the extended radio catalogue, making them multiple systems (see Section 5.3).

AzTEC ID <sup>a</sup>	$R_S$ (arcsec)	Radio coords <sup>b</sup> (J2000)	$P$ -stat <sub>1.4GHz</sub> <sup>b</sup>	IRAC coords <sup>b</sup> (J2000)	Rank <sup>b</sup>	Likelihood <sup>b</sup> ratio	Reliability <sup>b</sup> factor	Other IDs <sup>c</sup>
GOODS-N								
AzTEC/GN1	2.9	<b>J123711.89+622211.8</b>	<b>0.002</b>	<b>J123711.88+622212.5</b>	<b>1/1</b>	<b>182.47 ± 4.7</b>	<b>1.00</b>	
				–	–	–	–	
AzTEC/GN3	5.2	<b>J123633.42+621408.7</b>	<b>0.001</b>	<b>J123633.40+621408.8</b>	<b>1/4</b>	<b>67.17 ± 7.4</b>	0.63	GOODS850-5
				J123633.25+621411.4	2/4	20.49 ± 5.8	0.19	
				J123633.81+621407.5	3/4	16.40 ± 10.3	0.15	
AzTEC/GN4	5.9	<b>J123550.22+621041.3</b>	<b>0.014</b>	<b>J123550.35+621042.2</b>	<b>1/2</b>	<b>152.65 ± 3.8</b>	<b>0.88</b>	
				J123549.89+621043.1	2/2	21.22 ± 3.8	0.12	
				–	–	–	–	
GOODS-S								
AzTEC/GS1	5.0	<b>J033211.37–275212.1</b>	<b>0.026</b>	<b>J033211.35–275213.0</b>	<b>1/1</b>	<b>20.73 ± 0.9</b>	<b>0.99</b>	
				–	–	–	–	
AzTEC/GS5	7.5	<b>J033151.11–274437.6</b>	<b>0.048</b>	<b>J033151.09–274437.0</b>	2/3	3.04 ± 0.1	0.33	
				J033151.06–274432.4	1/3	5.58 ± 4.0	0.60	
				J033151.47–274427.9	3/3	0.41 ± 0.1	0.04	
AzTEC/GS7	8.5	J033213.85–275600.3	0.079	J033213.85–275559.9	4/5	4.10 ± 0.2	0.03	
				J033213.56–275608.3	1/5	100.79 ± 37.5	0.83	
				J033213.91–275607.4	2/5	8.22 ± 3.6	0.07	
COSMOS								
AzTEC/C1 <sup>d</sup>	6.6	<b>J100141.75+022713.4</b>	<b>0.002</b>	<b>J100141.86+022713.7</b>	<b>1/2</b>	<b>8.82 ± 0.7</b>	0.75	
				J100141.51+022714.7	2/2	2.71 ± 0.2	0.23	
AzTEC/C2	6.2	–	–	–	–	–	–	AzTEC8
AzTEC/C3 <sup>f</sup>	6.6	<b>J100008.02+022612.1</b>	<b>0.003</b>	–	–	–	–	AzTEC2

<sup>a</sup>AzTEC IDs correspond to the IDs from Downes et al. (2012) for the GOODS fields and Aretxaga et al. (2011) for COSMOS.

<sup>b</sup>Radio counterparts with  $P$ -stat  $\leq 0.05$  are in bold. IRAC counterparts that are ranked correctly (1st) and reliability factors greater than 0.8 are also in bold.

<sup>c</sup>‘AzTECxx’ correspond to the IDs from Younger et al. (2007, 2009) for SMA observations, ‘COSLA-xx’ correspond to IDs from Smolčić et al. (2012) for PdBI observations of AzTEC sources in the COSMOS field and ‘GOODS850-x’ correspond to IDs from Barger et al. (2012) for SMA observations of SCUBA/AzTEC sources in GOODS-N.

<sup>d</sup>Radio counterpart is identified in the extended  $3\sigma$  radio catalogue (see Section 5.3).

<sup>e</sup>AzTEC source is cross-listed with Table A2.

<sup>f</sup>IRAC counterpart is not detected in all four bands.

<sup>g</sup>IRAC counterparts identified via the SMA (Younger et al. 2007, 2009). These include two radio-undetected SMGs (AzTEC/C4 and AzTEC/C10) and one multiple system (AzTEC/C42) for which the SMA identified the correct counterpart and which is cross-listed with Table A2. Two SMA sources, AzTEC7 (Younger et al. 2007) and AzTEC10 (Younger et al. 2009) do not match any SMGs in our AzTEC catalogues and so their IRAC fluxes were taken from their respective papers.

<sup>h</sup>The true radio counterpart for AzTEC/GN14 is undetected in our radio catalogues. The radio source listed here is unassociated with the submm emission (Barger et al. 2012).

**Table A2.** Multiple radio set: list of SMGs in GOODS-N, GOODS-S and COSMOS that have multiple radio detections. The full table (108 rows) is available online. The IRAC counterpart to each radio detection is listed on the same line with its rank, likelihood ratio and reliability factor. The information for an additional potential IRAC counterpart is often also listed. For COSMOS, radio detections which were found in the extended catalogue (see Section 5.3) are indicated. Some AzTEC/COSMOS SMGs were single systems in the unextended radio catalogue, but are multiple systems in the extended catalogue. These sources were used in the original training set and are cross-listed with Table A1.

AzTEC ID <sup>a</sup>	$R_S$ (arcsec)	Radio coords <sup>b</sup> (J2000)	$P$ -stat <sub>1.4 GHz</sub> <sup>b</sup>	IRAC coords <sup>b</sup> (J2000)	Rank <sup>b</sup>	Likelihood ratio <sup>b</sup>	Reliability factor <sup>b</sup>	Other IDs <sup>c</sup>
GOODS-N								
AzTEC/GN7	6.4	<b>J123711.32+621330.9</b>	<b>0.028</b>	J123711.34+621331.4	3/4	21.40 ± 0.4	0.17	
		<b>J123711.99+621325.5</b>	<b>0.030</b>	J123711.99+621326.0	1/4	48.23 ± 2.4	0.39	
				J123711.86+621333.8	2/4	45.87 ± 3.4	0.37	
AzTEC/GN11	7.4	<b>J123635.87+620707.8</b>	<b>0.008</b>	J123635.85+620708.0	1/8	10.09 ± 0.1	0.37	
		<b>J123635.87+620703.9</b>	<b>0.012</b>	–	–	–	–	
		<b>J123636.32+620706.8</b>	<b>0.034</b>	J123636.37+620707.3	3/8	5.31 ± 0.1	0.20	
AzTEC/GN16	7.9	<b>J123615.94+621514.3</b>	<b>0.024</b>	J123615.82+621515.8	1/5	96.55 ± 3.1	0.51	GOODS850-7
		<b>J123616.10+621513.8</b>	<b>0.027</b>	J123616.10+621514.0 <sup>d</sup>	2/5	74.93 ± 1.2	0.40	
				J123616.84+621514.7	3/5	7.07 ± 0.6	0.04	
GOODS-S								
AzTEC/GS19	10.3	J033222.71–274127.0	0.080	J033222.71–274126.2	3/7	4.26 ± 0.2	0.13	
		J033222.92–274125.3	0.061	J033222.85–274125.1	4/7	4.03 ± 0.1	0.12	
				J033223.77–274131.6	1/7	11.66 ± 0.7	0.34	
AzTEC/GS32	13.7	J033310.14–275124.4	0.195	J033310.12–275124.8	3/4	1.18 ± 0.2	0.20	
		J033308.63–275134.5	0.092	J033308.62–275134.5	1/4	2.03 ± 0.1	0.34	
				J033308.78–275125.2	2/4	1.50 ± 0.4	0.25	
AzTEC/GS35	12.6	<b>J033227.19–274051.5</b>	<b>0.004</b>	J033227.17–274051.7 <sup>d</sup>	4/9	27.92 ± 1.1	0.14	
		<b>J033226.97–274049.9</b>	<b>0.009</b>	J033227.17–274051.7	4/9	27.92 ± 1.1	0.14	
				J033226.83–274056.1	1/9	50.66 ± 8.3	0.26	
COSMOS								
AzTEC/C23	11.0	<b>J100142.36+021836.0</b>	<b>0.013</b>	<b>J100142.34+021835.2</b>	<b>1/2</b>	<b>25.74 ± 1.9</b>	<b>0.90</b>	
		<b>J100142.76+021841.5</b>	<b>0.036</b>	J100142.73+021841.1	2/2	2.62 ± 0.2	0.09	
AzTEC/C43 <sup>f</sup>	12.5	<b>J100003.12+020201.5</b>	<b>0.032</b>	J100003.09+020201.5	5/8	7.91 ± 2.3	0.05	
		J100003.76+020217.6 <sup>g</sup>	0.069	J100003.81+020216.7	7/8	1.25 ± 0.1	0.01	
				J100003.40+020204.4	1/8	65.08 ± 7.8	0.38	
AzTEC/C53	12.8	<b>J100122.47+021209.7<sup>g</sup></b>	<b>0.006</b>	J100122.46+021210.0	4/8	4.82 ± 0.4	0.05	
		<b>J100122.07+021212.5<sup>g</sup></b>	<b>0.038</b>	J100122.04+021213.9	2/8	1.39 ± 0.3	0.02	
				J100122.61+021214.9	1/8	48.66 ± 14.8	0.53	

<sup>a</sup>AzTEC IDs correspond to the IDs from Downes et al. (2012) for the GOODS fields and Aretxaga et al. (2011) for COSMOS.

<sup>b</sup>Radio detections with  $P$ -stat  $\leq 0.05$  are in bold. IRAC detections with reliability factors greater than 0.8 are also in bold.

<sup>c</sup>‘AzTECxx’ correspond to the IDs from Younger et al. (2007, 2009) for SMA observations, ‘COSLA-xx’ correspond to IDs from Smolčić et al. (2012) for PdBI observations of AzTEC sources in the COSMOS field and ‘GOODS850-x’ correspond to IDs from Barger et al. (2012) for SMA observations of SCUBA/AzTEC sources in GOODS-N.

<sup>d</sup>IRAC detection is within 2 arcsec of both radio detections.

<sup>e</sup>IRAC counterpart is not detected in all four bands.

<sup>f</sup>AzTEC source is cross-listed with Table A1.

<sup>g</sup>Radio counterpart is identified in the extended  $3\sigma$  radio catalogue (see Section 5.3).

**Table A3.** Radio-undetected SMG set: list of SMGs without radio detections. The full table (321 rows) is available online. For each AzTEC source, the likelihood ratios and reliability factors for the top three ranked potential IRAC counterparts are listed. IRAC detections with  $R \geq 0.8$  are in bold.

AzTEC ID <sup>a</sup>	$R_S$ (arcsec)	IRAC coords <sup>b</sup> (J2000)	Likelihood ratio <sup>b</sup>	Reliability factor <sup>b</sup>	Other IDs <sup>c</sup>
GOODS-N					
AzTEC/GN2	4.8	<b>J123632.01+621711.1</b>	<b>15.26 ± 2.1</b>	<b>0.99</b>	
		–	–	–	
AzTEC/GN8	6.6	J123645.85+621447.3	12.23 ± 0.4	0.70	GOODS850-11
		J123645.77+621442.6	5.05 ± 0.6	0.29	
		–	–	–	
AzTEC/GN10	7.3	J123627.54+621218.4	59.47 ± 3.7	0.63	
		J123627.01+621217.7	19.76 ± 14.6	0.21	
		J123627.18+621212.9	6.10 ± 11.9	0.06	
GOODS-S					
AzTEC/GS2	5.3	J033218.28–275224.5	11.01 ± 0.6	0.69	
		J033218.84–275223.3	4.71 ± 0.6	0.30	
		–	–	–	
AzTEC/GS3	6.1	J033247.96–275416.3	27.53 ± 1.9	0.55	
		J033247.69–275423.4	22.07 ± 4.6	0.44	
		–	–	–	
AzTEC/GS4	6.4	<b>J033248.93–274251.3</b>	<b>92.64 ± 6.6</b>	<b>1.00</b>	
		–	–	–	
		–	–	–	
COSMOS					
AzTEC/C5	6.9	<b>J095942.85+022938.2</b>	<b>87.20 ± 14.5</b>	<b>1.00</b>	AzTEC1
		–	–	–	
		–	–	–	
AzTEC/C7	7.7	J100015.89+022943.9	8.87 ± 12.4	0.49	
		J100015.74+022939.6	7.44 ± 0.7	0.41	
		J100015.48+021550.7	0.93 ± 0.2	0.05	
AzTEC/C11	8.4	J100141.19+020357.1	0.03 ± 19.2	0.11	
		–	–	–	
		–	–	–	

<sup>a</sup>AzTEC IDs correspond to the IDs from Downes et al. (2012) for the GOODS fields and Aretxaga et al. (2011) for COSMOS.

<sup>b</sup>IRAC detections with reliability factors greater than 0.8 are in bold.

<sup>c</sup>‘AzTECxx’ correspond to the IDs from Younger et al. (2007, 2009) for SMA observations, ‘COSLAXX’ correspond to IDs from Smolčić et al. (2012) for PdBI observations of AzTEC sources in the COSMOS field and ‘GOODS850-x’ correspond to IDs from Barger et al. (2012) for SMA observations of SCUBA/AzTEC sources in GOODS-N.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Table A1.** Training and radio counterpart set.

**Table A2.** Multiple radio set.

**Table A3.** Radio-undetected SMG set (<http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stt155/-/DC1>).

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

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.