The infrared-radio correlation of star-forming galaxies is strongly M*-dependent but nearly redshift-invariant since z ~ 4

Delvecchio, I.; Daddi, E.; Sargent, M.T.; Jarvis, M.J.; Elbaz, D.; Jin, S.; ... ; Zamorani, G.

Astronomy & Astrophysics manuscript no. qpaper_delvecchio_v2
October 16, 2021

The infrared-radio correlation of star-forming galaxies is strongly $M_\star$-dependent but nearly redshift-invariant since z∼4

I. Delvecchio[1,2⋆], E. Daddi[2], M. T. Sargent[3], M. J. Jarvis[4,5], D. Elbaz[2], S. Jin[6,7], D. Liu[8], I. H. Whittam[4,5], H. Algera[9], R. Carraro[10], C. D'Eugenio[2], J. Delhaize[11], B. Kalita[2], S. Leslie[9], D. Cs. Molnár[12], M. Novak[8], I. Prandoni[13], V. Smolčić[14], Y. Ao[15,16], M. Aravena[17], F. Bournaud[2], J. D. Collier[18,19], S. M. Randriamampandry[20,21], Z. Randriamanakoto[20], G. Rodighiero[22], J. Schober[23], S. V. White[24], and G. Zamorani[25]

[1] INAF - Osservatorio Astronomico di Brera, via Brera 28, I-20121, Milano, Italy & via Bianchi 46, I-23807, Merate, Italy
[2] CEA, Irfu, DAp, AIM, Universitè Paris-Saclay, Universitè de Paris, CNRS, F-91191 Gif-sur-Yvette, France
[3] Astronomy Centre, Department of Physics & Astronomy, University of Sussex, Brighton, BN1 9QH, England
[4] Astrophysics, Department of Physics, Keble Road, Oxford, OX1 3RH, UK
[5] Department of Physics & Astronomy, University of the Western Cape, Private Bag X17, Bellville, Cape Town, 7535, South Africa
[6] Instituto de Astrofísica de Canarias (IAC), E-38205 La Laguna, Tenerife, Spain
[7] Universidad de La Laguna, Dpto. Astrofísica, E-38206 La Laguna,Tenerife, Spain
[8] MPI for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany
[9] Leiden Observatory, Leiden University, P.O. Box 9513, 2300RA Leiden, the Netherlands
[10] Instituto de Física y Astronomía, Universidad de Valparaíso, Gran Bretaña 1111, Playa Ancha, Valparaíso, Chile
[11] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[12] INAF - Osservatorio Astronomico di Cagliari, Via della Scienza 5, I-09047 Selargius (CA), Italy
[13] INAF - Istituto di Radioastronomia, Via P. Gobetti 101, 40129 Bologna, Italy
[14] Department of Physics, University of Zagreb, Bijenička cesta 32, 10002 Zagreb, Croatia
[15] Purple Mountain Observatory & Key Laboratory for Radio Astronomy, Chinese Academy of Sciences, Nanjing, China
[16] School of Astronomy and Space Science, University of Science and Technology of China, Hefei, Anhui, China
[17] Núcleo de Astronomía, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile
[18] The Inter-University Institute for Data Intensive Astronomy (IDIA), Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
[19] School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
[20] South African Astronomical Observatory, P.O. Box 9, Observatory 7935, Cape Town, South Africa
[21] A&A, Department of Physics, Faculty of Sciences, University of Antananarivo, B.P. 906, Antananarivo 101, Madagascar
[22] University of Padova, Department of Physics and Astronomy, Vicolo Osservatorio 3, I-35122, Padova, Italy
[23] Laboratoire d'Astrophysique, EPFL, CH-1290 Sauverny, Switzerland
[24] Department of Physics and Electronics, Rhodes University, PO Box 94, Makhanda, 6140, South Africa
[25] INAF - Osservatorio Astronomico di Bologna, via P. Gobetti 93/3, 40129 Bologna, Italy

Received

ABSTRACT

Several works in the past decade have used the ratio between total (rest 8-1000μm) infrared and radio (rest 1.4 GHz) luminosity in star-forming galaxies (q_{TIR}), often referred to as the "infrared-radio correlation" (IRRC), to calibrate radio emission as a star formation rate (SFR) indicator. Previous studies constrained the evolution of q_{TIR} with redshift, finding a mild but significant decline, that is yet to be understood. For the first time, we re-calibrate q_{TIR} as a function of *both* stellar mass ($M_\star$) and redshift, starting from an $M_\star$-selected sample of >400,000 star-forming galaxies in the COSMOS field, identified via (NUV-r)/(r-J) colours, at redshifts 0.1<z<4.5. Within each ($M_\star$,z) bin, we stack the deepest available infrared/sub-mm and radio images. We fit the stacked IR spectral energy distributions with typical star-forming galaxy and IR-AGN templates, and carefully remove radio AGN candidates via a recursive approach. We find that the IRRC evolves primarily with $M_\star$, with more massive galaxies displaying systematically lower q_{TIR}. A secondary, weaker dependence on redshift is also observed. The best-fit analytical expression is the following: $q_{TIR}(M_\star,z)=(2.646\pm0.024)\times(1+z)^{(-0.023\pm0.008)}-(0.148\pm0.013)\times(\log\ M_\star/M_\odot-10)$. Adding the UV dust-uncorrected contribution to the IR as a proxy for the total SFR, would further steepen the q_{TIR} dependence on $M_\star$. We interpret the apparent redshift decline reported in previous literature as due to low-$M_\star$ galaxies being progressively under-represented at high-redshift, as a consequence of binning only in redshift and using either infrared or radio-detected samples. The lower IR/radio ratio in more massive galaxies could be possibly linked to higher SFR surface density, which induces larger cosmic-ray scale heights that boost radio synchrotron emission. Our findings highlight that using radio emission as a proxy for SFR requires novel $M_\star$-dependent recipes, that will enable us to convert detections from future ultra deep radio surveys into accurate SFR measurements down to low-SFR, low-$M_\star$ galaxies.

Key words. galaxies: star formation – radio continuum: galaxies – infrared: galaxies – galaxies: active – galaxies: evolution

1. Introduction

For nearly fifty years astronomers have studied the observed correlation between total infrared (TIR; rest-frame 8-1000 μm, i.e. L_{IR}) and radio (e.g. rest-frame 1.4 GHz, i.e. $L_{1.4\ GHz}$) luminosity arising from star formation, usually referred to as the "infrared-radio correlation" (IRRC, e.g. van der Kruit 1971). This tight (1σ ~0.16 dex, e.g. Molnár et al. 2020 submitted) correlation is often parametrized by the IR-to-radio luminosity ratio $q_{\rm TIR}$, defined as (e.g. Helou et al. 1988; Yun et al. 2001):

$$q_{TIR} = \log\left(\frac{L_{IR}\ [W]}{3.75\times10^{12}[Hz]}\right) - \log(L_{1.4\ GHz}\ [W\ Hz^{-1}]) \quad (1)$$

where 3.75×10^{12} Hz represents the central frequency over the far-infrared (FIR, rest-frame 42-122 μm) domain, usually scaled to TIR in the recent literature. In the local Universe, the IRRC (or its parametrization q_{TIR}) appears to hold over at least three orders of magnitude in both L_{IR} and $L_{1.4\ GHz}$ (e.g. Helou et al. 1985; Condon 1992; Yun et al. 2001). Broadly speaking, this is because the infrared emission comes from dust heated by fairly massive ($\gtrsim$5 $M_\odot$) OB stars, while radio emission arises from relativistic electrons accelerated by shock waves produced when massive stars ($\gtrsim$8 $M_\odot$) explode as supernovae. As a consequence, within typical star formation timescales ($\gtrsim$100 Myr, e.g. Kennicutt 1998), these two sources of emission are expected to correlate.

Surprisingly enough, nearly all local star-forming late-type galaxies and even merging galaxies appear to follow the IRRC (e.g. Helou et al. 1985, but see Algera et al. 2020a for high-z sub-mm galaxies). This has been a strong motivator for using radio-continuum emission as a dust-unbiased star formation rate (SFR) tracer also in the faint radio sky (e.g. Condon 1992; Bell 2003; Murphy et al. 2011, 2012). Moreover, measuring the offset from the IRRC has been widely used to indirectly identify radio-excess active galactic nuclei (AGN; e.g. Donley et al. 2005; Del Moro et al. 2013; Bonzini et al. 2015; Delvecchio et al. 2017).

These applications, however, deeply rely on a proper understanding of whether and how the IRRC evolves over cosmic time and across different types of galaxies. Despite its extensive application in extragalactic astronomy, the detailed physical origins of the IRRC and the nature of its cosmic evolution have long been debated (e.g. Harwit & Pacini 1975; Rickard & Harvey 1984; de Jong et al. 1985; Helou et al. 1985; Hummel et al. 1988; Condon 1992; Appleton et al. 2004; Jarvis et al. 2010; Sargent et al. 2010; Ivison et al. 2010a,2010b; Bourne et al. 2011; Smith et al. 2014; Magnelli et al. 2015; Calistro Rivera et al. 2017; Delhaize et al. 2017; Gürkan et al. 2018; Molnár et al. 2018; Algera et al. 2020a).

For example, some studies of local star-forming galaxies (SFGs), ranging from dwarf (e.g. Wu et al. 2008) to ultra-luminous infrared galaxies (ULIRGs; L_{IR} >10^{12} $L_\odot$; e.g. Yun et al. 2001) concluded that the IRRC remains linear across a wide range of L_{IR}. Conversely, other studies have argued that at low luminosities the IRRC may break down, consistent with a non-linear trend of the form $L_{IR} \propto L_{1.4\ GHz}^{0.75-0.90}$ (e.g. Bell 2003; Hodge et al. 2008; Davies et al. 2017; Gürkan et al. 2018).

Several models have attempted to explain this non linearity. On the one hand, *calorimetric models* assume that galaxies are optically thick in the ultraviolet (UV), so that UV emission is fully re-emitted in the IR, likewise cosmic ray electrons (CRe) radiate away their total energy through synchrotron emission before escaping the galaxy (e.g. Voelk 1989). These conditions might hold in the most massive (stellar mass $M_\star$ $\gtrsim10^{10}$ $M_\odot$) SFGs, because of their increasing compactness (i.e. the size-mass relation $R_e \propto M_\star^{0.22}$, van der Wel et al. 2014), that might enhance their ability to retain the gas ejected by stars. However, this is likely to break down towards lower $M_\star$ galaxies, due to smaller sizes and lower obscuration (e.g. Bourne et al. 2012). On the other hand, *non-calorimetric* models or the optically thin scenario (Helou & Bicay 1993; Niklas & Beck 1997; Bell 2003; Lacki et al. 2010; Lacki & Thompson 2010), argue that several physical mechanisms cancel each other out, creating a sort of *conspiracy* that keeps the IRRC unexpectedly tight and linear. Indeed, both TIR and radio luminosities should underestimate the total SFR in low $M_\star$ and low SFR surface density galaxies (Bell 2003), inducing a departure of the IRRC from linearity. This is however not observed. Radio synchrotron models postulate that such small galaxies are not able to prevent CRe from escaping, causing a global deficit of radio emission at fixed SFR. Similarly, the TIR domain becomes less sensitive to SFR in low-$M_\star$ galaxies (e.g. Madau & Dickinson 2014), generating an IR deficit of a similar amount that might counter-balance the radio and keep the IRRC linear. Understanding the discrepancy between model predictions and observations is crucial, since the linearity (or not) of the IRRC has direct implications for using radio emission as a SFR tracer.

From an observational perspective, it is widely recognized that a tight relation links SFR and $M_\star$ in nearly all SFGs, namely the "main sequence" of star formation (MS, scatter~0.2–0.3 dex). This relation holds from z~5 down to the local Universe (e.g. Brinchmann et al. 2004; Noeske et al. 2007; Elbaz et al. 2011; Whitaker et al. 2012; Speagle et al. 2014; Schreiber et al. 2015; Lee et al. 2015), showing a flattening at high $M_\star$ and an evolving normalization with redshift. Because the SFR is directly linked to L_{IR}, especially in massive SFGs (Kennicutt 1998), the existence of the MS gives an additional argument that studying q_{TIR} as a function of $M_\star$ could be of the utmost importance for our understanding of what drives the IRRC in galaxies.

Recent studies have corroborated the idea that the IRRC slightly, but significantly, declines with redshift (Ivison et al. 2010b; Magnelli et al. 2015; Calistro Rivera et al. 2017; Delhaize et al. 2017), in the form of $q_{\rm TIR} \propto(1+z)^{[-0.2:-0.1]}$, although the physical explanation for such evolution is still uncertain. Somewhat different conclusions were reached by other works (e.g. Appleton et al. 2004; Ibar et al. 2008; Jarvis et al. 2010; Sargent et al. 2010; Bourne et al. 2011) which ascribe this apparent evolution to selection effects. For instance, these include comparing flux-limited samples, each with a different selection function.

In this regard, we note that any selection method is sensitive to brighter, i.e. more massive galaxies towards higher redshifts. By binning in redshift, only a restricted range in galaxy $M_\star$ will be detectable at each redshift for any flux limited sample, thus inducing a bias as a function of z. Therefore, it is timely to examine the evolution of the IRRC as a function of $M_\star$ and redshift *simultaneously*. We emphasize that our approach is fully empirical. However, a possible $M_\star$ dependence of the IRRC is expected from some synchrotron emission models (e.g. Lacki & Thompson 2010; Schober et al. 2017), and might reflect some combination of the underlying physics originating the IRRC (see Sect. 5).

The main goal of the present paper is to re-calibrate the IRRC for the first time as a function of *both* $M_\star$ and redshift over a wide range. To this end, we start from an $M_\star$-selected sample of >400,000 galaxies at 0.1<z<4.5 collected from deep UltraVISTA images in the Cosmic Evolution Survey (Scoville et al. 2007) (centered at RA=+150.11916667;

⋆ email: ivan.delvecchio@inaf.it

Dec=+2.20583333 (J2000)). Then we leverage the new de-blended far-IR/sub-mm data (Jin et al. 2018) recently compiled in COSMOS, which allow us to circumvent blending issues due to poor angular resolution and measure L_{IR} for typical MS galaxies out to z~4. In addition, we exploit the deepest radio-continuum data taken from the VLA-COSMOS 3 GHz Large Project (Smolčić et al. 2017b), the VLA-COSMOS 1.4 GHz Deep project mosaic catalogue (Schinnerer et al. 2010) and the 1.3 GHz MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE, Jarvis et al. 2016). This golden radio legacy allows us to test our analysis against varying depths, angular resolutions and radio frequencies. Individual detections will be combined with stacked fluxes of non-detections, at both IR and radio frequencies to assess the average q_{TIR} as a function of $M_\star$ and redshift.

The layout of this paper is as follows. A description of the sample selection and multi-wavelength ancillary data is given in Sect. 2. We describe the stacking analysis in Sect. 3, including measurements of L_{IR} (Sect. 3.1) and $L_{1.4\,GHz}$ (Sect. 3.3). The average q_{TIR} as a function of $M_\star$ and redshift is presented in Sect. 4, where we perform a careful subtraction of radio AGN at different $M_\star$ via a recursive approach. Our main results are discussed and interpreted in Sect. 5 in the framework of previous observational studies and theoretical models. The main conclusions are summarized in Sect. 6. In addition, we test our total 3 GHz fluxes in Appendix A. In Appendix B we discuss how the final IRRC is sensitive to our AGN subtraction method. Finally, in Appendix. C we quantify how different assumptions from the literature would change our main results.

Throughout this paper, magnitudes are given in the AB system (Oke 1974). We assume a Chabrier (2003) initial mass function (IMF) and a ΛCDM cosmology with $\Omega_m = 0.30$, $\Omega_\Lambda = 0.70$, and $H_0 = 70$ km s^{-1} Mpc^{-1} (Spergel et al. 2003).

2. Multi-wavelength data and sample selection

In this Section we describe the creation of a K_s prior catalogue that we used to select our parent sample in the COSMOS field.

The COSMOS field (2 deg^2) boasts an exquisite photometric data set, spanning from the X-rays to the radio domain[1]. The most recent collection of multiwavelength photometry comes from the COSMOS2015 catalogue (Laigle et al. 2016), that contains 1,182,108 sources extracted from a stacked $YJHK_s$ image(blue dots in Fig. 1). In particular, this catalogue joins optical photometry from Subaru Hyper-Suprime Cam (2 deg^2; Capak et al. 2007) and from the Canada-France-Hawaii Telescope Legacy Survey (CFHT-LS, central 1 deg^2; McCracken et al. 2001); near-infrared (NIR) bands Y, J, H, and K_s from UltraVISTA DR2 (down to K_s<24.5 in the central 1.5 deg^2, of which 0.6 deg^2 are covered by ultra-deep stripes with limiting K_s<25.2; McCracken et al. 2012) and from CFHT H and K_s observations obtained with the WIRCam (K_s<23.9 outside the UltraVISTA area; McCracken et al. 2001). Over the full 2 deg^2 area, mid-infrared (MIR) photometry was obtained from the *Spitzer* Large Area Survey with Hyper-Suprime-Cam (SPLASH; Steinhardt et al. 2014; P. Capak et al. in prep.) using 3.6–8μm data from the Infrared Array Camera (IRAC). We refer the reader to Laigle et al. (2016) for more details.

In order to obtain a homogeneous galaxy selection function, we limited our study to the inner UltraVISTA DR2 area, also excluding stars and masked regions in the COSMOS2015 catalogue with less accurate photometry, which reduces the initial sample to 45% of its size (524,061 sources). Following Jin et al. (2018), we partly fill up these blank regions by adding 22,838 unmasked K_s–selected sources from the UltraVISTA catalogue of Muzzin et al. (2013) (3σ limit of K_s<24.35 with 2" aperture). This ensures a more complete coverage within the UltraVISTA area, with fluctuations in prior source density of only 2.5%. This builds our K_s prior sample of 546,899 galaxies. Given the similar selection, we confirm that excluding the slightly shallower ~4% subsample from Muzzin et al. (2013) leaves our results unchanged, and thus we keep them throughout this work.

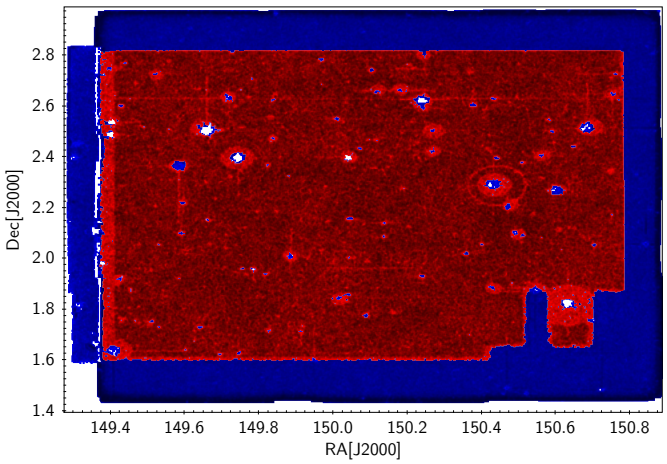

Fig. 1: Distribution of the full COSMOS2015 (Laigle et al. 2016) source list over the COSMOS area (blue dots). The subset of 413,678 $NUVrJ$–based star-forming galaxies analyzed in this work (red dots) includes sources from Laigle et al. (2016) and Muzzin et al. (2013) within the UltraVISTA area, with the exception of masked regions due to saturated or contaminated photometry. See Sect. 2 for details.

Photometric redshifts and $M_\star$ estimates were retrieved from the corresponding catalogues, by fitting the optical-MIR photometry using the stellar population synthesis models of Bruzual & Charlot (2003). Both redshift and $M_\star$ values represent the median of the corresponding likelihood distribution. Laigle et al. (2016) report an average photometric redshift accuracy of $\langle|\Delta z/(1+z)|\rangle = 0.007$ at z<3, and 0.021 at 3<z<6. A similar accuracy of 0.013 is reached in the catalogue of Muzzin et al. (2013) at z<4. We further inspected a subset of 5,400 sources showing a skewed redshift probability distribution function (with ≳5% chance to be offset from the median by >0.5×[1+z_p]). However, we verified that removing such potential redshift interlopers does not have any impact on our results. As in Jin et al. (2018), publicly available spectroscopic redshifts were collected from the new COSMOS master spectroscopic catalog (courtesy of M. Salvato, within the COSMOS team), and were prioritized over photometric measurements if deemed reliable (z_s quality flag >3 ∧ $|z_s - z_p| < 0.1\times(1+z_p)$).

Infrared/sub-mm fluxes were de-blended and re-extracted via the prior-based fitting algorithm presented in Jin et al. (2018), that we briefly describe in Sect. 2.2.

2.1. Selecting star-forming galaxies via (NUV-r)/(r-J) colours

We aim to study the infrared-radio correlation within an $M_\star$-selected sample of star-forming galaxies. To this end, we make use of the rest-frame, dust-corrected ($NUV - r$) and ($r - J$) colours available in the parent catalogues (hereafter $NUVrJ$).

[1] An exhaustive overview of the COSMOS field is available at: http://cosmos.astro.caltech.edu/

As opposed to the widely used UVJ criterion, the (NUV–r) colour is more sensitive to recent star formation (10^6–10^8 yr scales, Salim et al. 2005; Arnouts et al. 2007; Davidzon et al. 2017). Therefore, this criterion enables us to better distinguish between weakly star-forming galaxies (with specific-SFR, sSFR=SFR/$M_\star$ $\sim10^{-10}$yr^{-1}) and fully passive systems (sSFR<10^{-11}yr^{-1}).

We further selected galaxies with redshift 0.1<z<4.5 and 10^8 <$M_\star/M_\odot$ <10^{12}. This leaves us with a final sample of 413,678 star-forming galaxies (red dots in Fig. 1), out of which 22,238 (5.4%) are spectroscopically confirmed. Such a sizable sample enables us to bin galaxies as a function of both $M_\star$ and redshift, while maintaining good statistical power. Fig. 2 shows our sample in the $M_\star$–redshift diagram, highlighting the chosen grid. We note that the $M_\star$ uncertainties taken from the parent catalogues incorporate the covariant errors on stellar population ages and dust reddening. These average $M_\star$ uncertainties are 0.2 dex at 10^8 <$M_\star/M_\odot$ <10^9 and 0.1 dex above, which is far smaller than the corresponding $M_\star$ bin width, thus not impacting our results. The 90% $M_\star$ completeness limit (orange solid line, Laigle et al. 2016) indicates that our sample of SFGs is mostly complete down to 10^{10} $M_\odot$ out to z∼4. Although we acknowledge the increasing incompleteness towards less massive galaxies in the early Universe, we believe that including them brings a valuable addition for constraining the infrared and radio properties of galaxies down to a poorly explored regime of $M_\star$. This will become particularly relevant for the next generation of telescopes, such as JWST and SKA, which will routinely observe such faint sources. In addition, as we will discuss in Sect. 3.2, a very good agreement is observed between our stacked L_{IR} and those extrapolated from the MS relation (Schreiber et al. 2015) also at $M_\star$ <$10^{9.5}M_\odot$, suggesting that even in this incomplete, low-$M_\star$ regime our galaxies are still representative of an $M_\star$-selected sample. We emphasize that the overall conclusions of this work are unchanged if we limit ourselves to z<3 and $M_\star$ >$10^{9.5}M_\odot$, in which our sample is highly complete. Moreover, in light of our main result, i.e. q_{TIR} *decreases* with $M_\star$, we anticipate that including galaxies within an incomplete $M_\star$ regime would at most amplify the final $M_\star$ dependence, thereby reinforcing our findings.

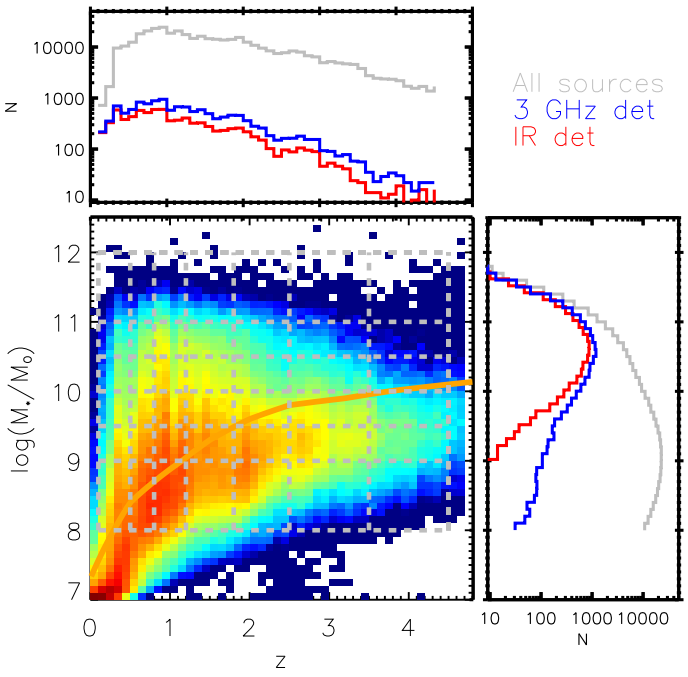

Fig. 2: Distribution of *NUVrJ* star-forming galaxies as a function of $M_\star$ and redshift. The colour scale in the central panel indicates the underlying sample size, increasing from blue to red. The grey dashed grid encloses the 42 $M_\star$–z bins into which we split our sample (413,678 objects). Galaxies within that grid are projected on the upper-left and bottom right histograms with redshift and $M_\star$, respectively (grey lines). The blue and red histograms represent to subsample with S/N>3 at 3 GHz and across all IR bands, respectively (see Sects. 2.2 and 2.3). The orange solid line marks the 90% $M_\star$ completeness limit of Laigle et al. (2016) for comparison.

2.2. Infrared and sub-mm de-blended data

We complemented the existing COSMOS optical-to-IRAC photometry with cutting-edge de-blended photometry from Jin et al. (2018), based on the de-blending algorithm developed in Liu et al. (2018) for the GOODS-North field.

The dataset includes *Spitzer*-MIPS 24 μm data (PI: D. Sanders; Le Floc'h et al. 2009); *Herschel* imaging from the PACS (100-160 μm, Poglitsch et al. 2010) and the SPIRE (250, 350, and 500 μm, Griffin et al. 2010) instruments, as part of the PEP (Lutz et al. 2011) and HerMES (Oliver et al. 2012) programs, respectively. In addition, JCMT/SCUBA2 (850 μm) images are taken from the S2CLS program (Cowie et al. 2017; Geach et al. 2017), the ASTE/AzTEC (1.1 mm) data are nested maps from Aretxaga et al. (2011) over a sub-area of 0.72 deg^2. Finally, Jin et al. (2018) also included MAMBO data (Bertoldi et al. 2007) at 1.2 mm over an area of 0.11 deg^2.

Briefly, Jin et al. (2018) used K_s-selected sources from the UltraVISTA survey (Sect. 2) as priors to perform PSF fitting of MIPS 24 μm, VLA-3 GHz (Smolčić et al. 2017b) and VLA-1.4 GHz (Schinnerer et al. 2010) images down to the 3σ level in each band. This procedure yields a total of 85,171 detections with signal-to-noise ratio S/N>3 within the UltraVISTA area. Nevertheless, adopting a similar approach for extracting FIR/sub-mm fluxes of *all* $M_\star$-selected galaxies, i.e. using the full list of K_s priors, would identify up to 50 sources per beam at the resolution of the FIR/sub-mm wavelengths, causing heavy confusion. Therefore, following the tight correlation between $M_\star$ and SFR (linked to IR luminosity), only an $M_\star$-complete subset of 106,153 priors was considered, which ultimately prioritizes IR brighter sources. Together with MIPS 24 μm/VLA detections, this overall sample of 191,624 K_s+MIPS 24 μm+VLA priors was used to de-blend and extract the *Herschel*, SCUBA2 and AzTEC fluxes. Within our final sample of 413,678 star-forming galaxies, 20,782 (5%) have a combined S/N>3 stacked across all FIR/sub-mm bands (10,289 at S/N>5). These are displayed as red histograms in Fig. 2. The rest of the K_s sources are assumed to have negligible FIR/sub-mm fluxes, consistent with the background level in those bands. This is confirmed by the Gaussian-like behaviour of the noise (centered at zero) in the residual maps, after subtracting all S/N>3 sources in each band (Jin et al. 2018).

Throughout the rest of this paper, we interpret individual S/N>3 sources as detections, while S/N<3 sources will be stacked, as described in Sect. 3.1.

2.3. The radio legacy in the COSMOS field

For our analysis we make use of the deepest available radio surveys in the COSMOS field.

We exploit VLA data from the 1.4 GHz Deep Project (Schinnerer et al. 2010) and the 3 GHz Large Project (Smolčić et al.

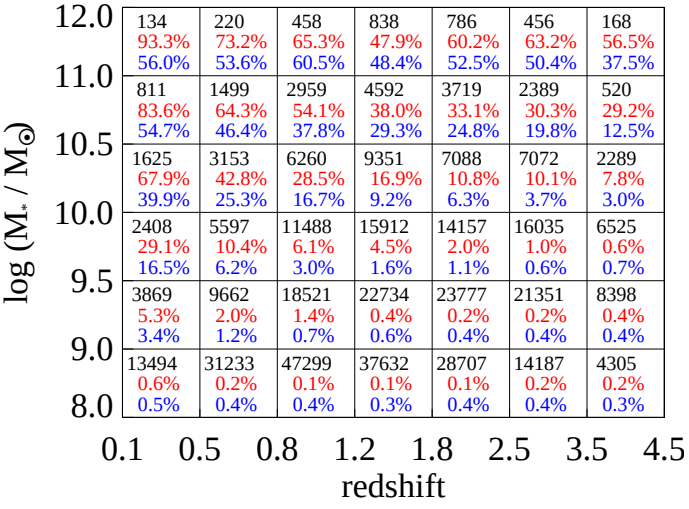

Fig. 3: Number of *NUVrJ*–based star-forming galaxies analyzed in this work, as a function of $M_\star$ and redshift (black). For convenience, in each bin we report the fraction of sources with stacked S/N_{IR} >3 across all FIR/sub-mm bands (red) and with $S/N_{3\ GHz}$ >3 (blue).

2017b). The 1.4 GHz Deep Project map was combined with the existing data from the VLA-COSMOS 1.4 GHz Large Project map (Schinnerer et al. 2010). It covers 1.7 deg^2 with an angular resolution of 2.5", reaching rms=12 μJy beam^{-1} in the central 50"×50". A total of 2,864 sources were blindly extracted down to S/N>5. The VLA-COSMOS 3 GHz Large Project (Smolčić et al. 2017b) is one of the largest and deepest radio survey ever conducted over a medium sky area like COSMOS. With 384h of observations, the final mosaic reaches a median rms=2.3 μJy beam^{-1} over 2.6 deg^2 at an angular resolution of 0.75", the highest among radio surveys in COSMOS. A total of 10,830 sources were blindly extracted down to S/N>5.

In addition, the COSMOS field boasts the current deepest radio-continuum data at 1.28 GHz from the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE, Jarvis et al. 2016) survey. With only 17h on-source over the central 1 deg^2 of COSMOS, the early science release of MIGHTEE reaches a thermal noise of 2.2 μJy beam^{-1}, though the effective noise is limited by confusion. This provides excellent sensitivity to large-scale radio emission. Nevertheless, given the relatively large beamsize (8.4"×6.8" FWHM), MIGHTEE fluxes were deblended using the same list of K_s+MIPS 24 μm+VLA priors applied on FIR/sub-mm images (Sect. 2.2). Similarly, VLA radio flux densities were re-extracted based on the same PSF-fitting technique down to S/N>3. However, for consistency with publicly available VLA catalogues, we adopt the 1.4 and 3 GHz radio flux densities of S/N>5 sources from Schinnerer et al. (2010) and Smolčić et al. (2017b), respectively. As a sanity check, we show in Appendix A that our procedure leads to fully consistent total 3 GHz fluxes.

Given its unparalleled depth and resolution over the full COSMOS area, we primarily use the VLA 3 GHz dataset for our radio analysis, which counts 13,808 S/N>3 sources out of 413,678 $M_\star$-selected star-forming galaxies (~3%). These radio detections are shown as blue histograms in Fig. 2. Fainter sources will be accounted for via stacking, as described in Sect. 3.3. Nevertheless, repeating the same stacking analysis with ancillary radio datasets at 1.3 GHz (MIGHTEE) and 1.4 GHz (VLA) is essential to validate our procedure against potential variations of radio spectral slope or different angular resolutions. For this reason, in Sect. 3.3.1 we compare our 3 GHz stacking results with the above ancillary datasets, while through the rest of the paper we will be using radio data only at 3 GHz.

3. Stacking analysis

The aim of this paper is to investigate how the IRRC evolves with $M_\star$ and redshift *simultaneously*. Contrary to studies in which galaxies were individually detected at IR and/or radio wavelengths, leading to complex selection functions and biased samples (see discussion in Sargent et al. 2010), we start from a well-defined $M_\star$-selected sample. As a consequence, our analysis makes use of IR (Sect. 3.1) and radio (Sect. 3.3) stacking. This includes a careful treatment of some common caveats concerning IR galaxy samples, such as clustering bias (Sect. 3.1.1) and spectral energy distribution (SED) fitting including AGN templates (Sect. 3.2). As for stacking radio data, special care is devoted to statistically removing radio AGN from our sample (Sect. 4.2).

In addition, our notably large star-forming galaxy sample allows us to bin as a function of *both* $M_\star$ and redshift, as shown in Fig. 3. For each bin, we also report the total number of $M_\star$-selected SFGs (black), as well as the corresponding fractions having combined S/N_{IR} >3 (red) and $S/N_{3\ GHz}$ >3 (blue). As can be seen, both fractions are a strong function of both $M_\star$ and redshift. Therefore, binning along both parameters enables us to account for the fact that galaxies of distinct $M_\star$ are detectable at IR and radio wavelengths over different redshift ranges. These aspects will be extensively discussed when comparing our results with previous literature (Appendix C).

3.1. Infrared and sub-mm stacking

In this Section we estimate the average fluxes across eight infrared and sub-mm bands, namely MIPS 24 μm, PACS 100-160 μm, SPIRE 250-350-500 μm, SCUBA 850 μm and AzTEC 1100 μm. Similarly to other studies, we perform *median* stacking on the residual maps from Jin et al. (2018), i.e. after subtracting all detected sources with S/N>3 in each band (see also Magnelli et al. 2009). Individual S/N>3 detections will be added to stacked fluxes a-posteriori through a weighted average (Eq. 3). Median stacking strongly mitigates contamination from bright neighbors and catastrophic outliers, and thus reduces the confusion noise for the faint sources. An exception is made for SCUBA maps: the heterogeneous sensitivity would not be captured via median stacking. Therefore, only in this case we apply mean stacking on SCUBA residual maps.

We stress that our procedure yields very consistent results with both median and mean stacking of both detections and non-detections (e.g. Magnelli et al. 2015; Schreiber et al. 2015), as shown in Sect. 3.2.

To produce stacked and rms images in each band, we used the publicly available IAS stacking library[2] (Bavouzet et al. 2008; Béthermin et al. 2010). For each band, $M_\star$ bin and redshift bin, we stack N×N pixel cutouts from the residual images, each centred on the NIR position of the $M_\star$-selected priors (Sect. 2). We choose the cutout size to be 8 times the full-width at half maximum (FWHM) of the PSF, while for *Spitzer*-MIPS we choose 13×FWHM, since a substantial fraction of the 24 μm flux is located in the first Airy ring. Since the AzTEC map covers only a central sub-area of 0.72 deg^2, at 1.1 mm we only stack within that region. We emphasize that the $M_\star$, z, and SFR distribution of the SFG population within the AzTEC region is fully

[2] https://www.ias.u-psud.fr/irgalaxies/downloads.php

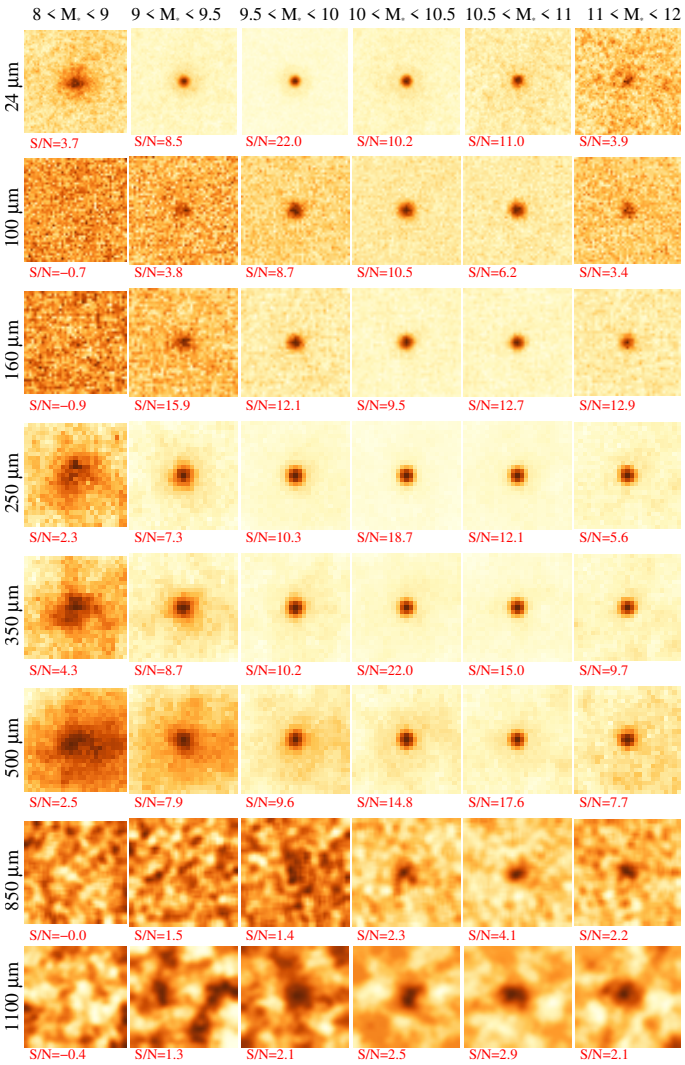

Fig. 4: Stacked cutouts of *NUVrJ*–based SFGs at 0.8<z<1.2, as a function of $M_\star$ (left to right, expressed in log $M_\odot$). Within each bin we stacked only sources with S/N<3 at a given band. SCUBA 850 μm (mean stacks) and AzTEC 1100 μm images are smoothed with a Gaussian kernel to ease the visualization. Each cutout size is 8×FWHM of the PSF, while for *Spitzer*-MIPS we choose 13×FWHM. Below each cutout we report the corresponding S/N ratio.

consistent with that derived in the rest of the COSMOS field, thus not biasing the resulting stacked fluxes. To measure total fluxes, we followed different techniques depending on the input map. For MIPS and PACS images, we used a PSF fitting technique (e.g. Magnelli et al. 2014). A correction of 12% is further applied to account for flux losses from the high-pass-filtering processing of PACS images (e.g. Popesso et al. 2012; Magnelli et al. 2013). For SPIRE images, the photometric uncertainties are not dominated by instrumental noise but by the confusion noise caused by neighboring sources (Dole et al. 2003; Nguyen et al. 2010). Since SPIRE, as well as SCUBA and AzTEC images are already beam-convolved, we first scale all maps to Jy beam^{-1} and then read the total flux from the central pixel. In the case of SCUBA maps, the mean stacked flux was computed by weighting each pixel with the inverse of the square of the error map. For SPIRE and AzTEC images, instead, the total flux was taken as the median of the input cube at the central pixel.

The uncertainties on the stacked fluxes are measured using a bootstrap technique (e.g. Béthermin et al. 2015). Within each $M_\star$–z bin, we run our stacking procedure 100 times, in all bands. For m non-detections at a given band, in each random realization we re-shuffle the input sample, preserving the same total m by allowing source duplication. We take the median of the resulting flux distribution as our formal stacked flux. The 1σ dispersion around this value is interpreted as the flux error. We propagate this uncertainty in quadrature with the standard deviation of the stacked map across 100 random positions within the cutout (after masking the central PSF). Though the latter component is typically sub-dominant relative to a bootstrapping dispersion, this conservative approach accounts for the strong fluctuations seen in low S/N stacked images, especially at low $M_\star$.

As an example, Fig. 4 shows stacked cutouts in all IR/submm bands at 0.8<z<1.2 (i.e. close to the median redshift of our sample) as a function of $M_\star$. As expected from the tight MS relation that links $M_\star$ and SFR in star-forming galaxies, stacks at low $M_\star$ display lower S/N, despite the larger numbers of input sources.

3.1.1. Correcting for clustering bias

The stacked fluxes calculated above can be biased high if the input sources are strongly clustered or very faint. This bias is caused by the greater probability of finding a source close to another one in the stacked sample compared to a random position. This generates an additional signal, as extensively discussed in the literature (e.g. Bavouzet et al. 2008; Béthermin et al. 2010, 2012; Kurczynski & Gawiser 2010; Bourne et al. 2012; Viero et al. 2013; Schreiber et al. 2015; Béthermin et al. 2015). Given the large number of stacked sources in each bin, the S/N is typically good enough to be able to correct for this effect, that becomes more prominent with increasing beam size (e.g. up to 50% for SPIRE images, see Béthermin et al. 2015). Here we briefly describe our approach, referring the reader to Appendix A.2 of Béthermin et al. (2015) for a detailed explanation.

We model the signal from stacking as the sum of three components: a central point source with the median flux of the underlying population, a clustering component convolved with the PSF, and a residual background term (Eq. 2). Following Béthermin et al. (2015), we attempt at separating these components via a simultaneous fit in the stacked images (Béthermin et al. 2012; Heinis et al. 2013, 2014; Welikala et al. 2016).

$$S(x,y) = \varphi \times PSF(x,y) + \psi \times (PSF \otimes w)(x,y) + \varepsilon \quad (2)$$

where $S(x,y)$ is the stacked image, PSF the point spread function, and w the auto-correlation function. The symbol $\otimes$ represents the convolution. The parameters φ, ψ, and ε are free normalizations of the source flux, clustering signal and background term, respectively.

We parametrize the "clustering bias" as bias=$\psi/(\varphi+\psi)$, once we have verified that residuals (i.e. ε) are always consistent with zero within the uncertainties. We do not see any obvious $M_\star$ or redshift dependence of the clustering bias. However, at fixed wavelength, this can fluctuate significantly depending on the S/N of the input stacked image. For these reasons, we prefer to use an *average* clustering correction $\langle 1 - bias \rangle$ for each band (see Table 1), drawn only from stacks with S/N>3. For those images, we multiply the stacked flux by $\langle 1 - bias \rangle$ at the corresponding wavelength. Only MIPS 24 μm data are not shown, since their fluxes will not be used for SED fitting in Sect. 3.2. Uncertainties on the clustering corrections were propagated quadratically with the stacked flux errors obtained in Sect. 3.1.

We stress that this method is suitable if the intrinsic source size is negligible compared to the PSF. This is especially true

Table 1: Average fraction of clustering signal at each FIR/sub-mm band. Uncertainties indicate the 1σ dispersion among all S/N>3 stacks at a given band.

Wavelength	% Clustering signal
PACS 100 μm	11.3±7.4
PACS 160 μm	10.2±16.5
SPIRE 250 μm	25.9±18.9
SPIRE 350 μm	31.3±20.8
SPIRE 500 μm	42.7±24.2
SCUBA 850 μm	23.1±13.9
AzTEC 1100 μm	20.1±12.9

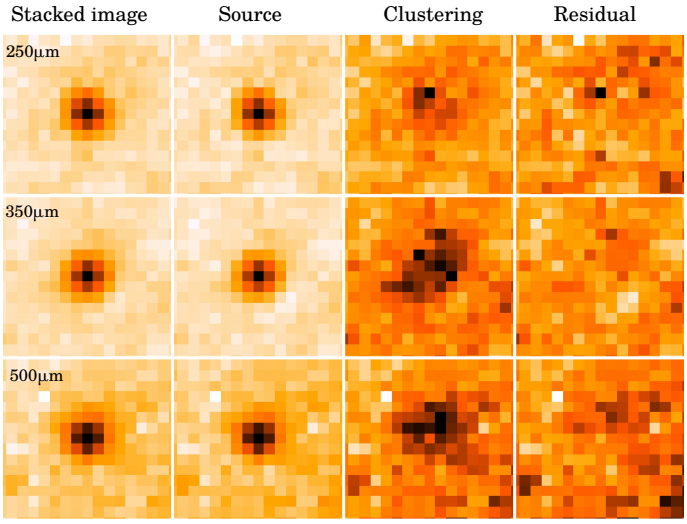

Fig. 5: Image decomposition of median stacks at 250, 350 and 500 μm, for a specific bin at 0.8<z<1.2 and 11<log($M_\star$/$M_\odot$)<12. From left to right, the stacked image is separated among a point source PSF, the clustering signal and a residual background term, respectively. The colour scale is normalized to the maximum in each cutout for visual purposes. See Sect. 3.1.1 for details.

in SPIRE images, of which we show an example in Fig. 5. This refers to a specific bin at 0.8<z<1.2 and 11<log($M_\star$/$M_\odot$)<12, for which all stacks give good enough S/N. Particularly for SPIRE images, the clustering bias can make up to 40% of the total flux, and it can be recognized as a more extended and diffuse emission. However, for consistency we extend this analysis to the full set of FIR/sub-mm data.

Lastly, the clustering-corrected source fluxes (S_{stack}) are combined with those of individual detections (S_i) with S/N>3 in each band. If (m, n) is the number of stacked and detected sources, respectively, the weighted-average flux S_{bin} in a given bin is derived as follows:

$$S_{bin} = \frac{m \times S_{stack} + \sum_{i=1}^{n} S_i}{n + m} \quad (3)$$

Flux uncertainties were propagated in quadrature. For stacks with S/N<3 in which we could not constrain the clustering correction, S_{stack} was set to the noise level of the stacked map (i.e. equal to its uncertainty). This way the weighted-average flux S_{bin} and its error are mainly driven by individual detections, for which the flux could be measured more accurately (Jin et al. 2018). If the combined flux S_{bin} has S/N<3, then it was set to 3× the noise and interpreted as 3σ upper limit.

3.2. Conversion to L_{IR} and SFR

This Section illustrates how we fit the observed FIR/sub-mm SEDs to determine the total (8-1000 μm rest-frame) IR luminosity within each $M_\star$–z bin. To this end, we use the two-component SED-fitting code developed by Jin et al. (2018) (see also Liu et al. 2018). Briefly, this includes: 3 mid-infrared AGN torus templates from Mullaney et al. (2011); 15 dust continuum emission models by Magdis et al. (2012), that were extracted from Draine & Li (2007) to best reproduce the average SEDs of MS (14) or SB (1) galaxies at various redshifts. While Draine & Li models were based on a number of physical parameters, the library of Magdis et al. (2012) depends exclusively on the mean radiation field $\langle U \rangle = L_{IR}$ per unit dust mass (M_d), and on whether the galaxy is on or above the MS. However, on the MS the average dust temperature strongly evolves with redshift (e.g. Magnelli et al. 2014) and directly enters M_d. Therefore, $\langle U \rangle$ and the SED shape both vary as a function of redshift, for which Magdis et al. (2012) empirically found as $\langle U \rangle \propto (1+z)^{1.15}$ up to z~2. More recently, Béthermin et al. (2015) revised the evolution of $\langle U \rangle$ with redshift out to z~4, using IR/sub-mm data in the COSMOS field, retrieving $\langle U \rangle \propto (1+z)^{1.8}$. Here we adopt the set of 14 MS templates from Magdis et al. (2012), fit them to our data, and we compare the $\langle U \rangle$–z trend with Béthermin et al. (2015) in Fig. 7.

The SED-fitting routine performs a simultaneous fitting using AGN and dust emission models, looking for the best-fit solution via χ^2 minimization. In order to account for the typical photo-z uncertainty of the underlying galaxy population (at fixed $M_\star$,z), each template is fitted to the data across a range of $\pm 0.05 \times (1+\langle z \rangle)$ around the median redshift $\langle z \rangle$. The code keeps track of each SED solution and corresponding normalization, generating likelihood distributions and uncertainties on e.g. L_{IR}, $\langle U \rangle$ and AGN luminosity, if any. We note that only FIR and sub-mm photometry (i.e. ignoring the MIPS 24 μm data-point) were used in the fitting procedure. This is to avoid internal variations of the MIR dust features that cannot be captured by our limited set of templates (e.g., IR to rest-frame 8 μm ratio, IR8, Elbaz et al. 2011), which might affect the global FIR/sub-mm SED fitting. This optimization clearly prioritizes the FIR/sub-mm part of the SED, while not impacting the final L_{IR} estimates (e.g. Liu et al. 2018).

Fig. 6 shows the best-fit star-forming galaxy template from the Magdis et al. (2012) library (green lines), as a function of $M_\star$ (left to right, expressed in log$M_\odot$) and redshift (top to bottom). Red circles indicate the IR/sub-mm photometry, while downward arrows mark 3σ upper limits. The red dotted line is the best-fit AGN template from Mullaney et al. (2011), shown if significant above 3σ. This is only found in the highest $M_\star$ and redshift bin. Green dashed lines represent SEDs without FIR measurements and at z$\gtrsim$1.5, for which the integrated L_{IR} is interpreted as 3σ upper limit (5/42 bins). Even though 24 μm has long been used as a proxy for L_{IR}, this is only accurate at z$\lesssim$1.5 (e.g. Elbaz et al. 2011; Lutz 2014). For this reason we still interpret as measurements the L_{IR} obtained from SEDs without FIR data, but only at z$\lesssim$1.5. That is the case for a few bins at the lowest $M_\star$, in which the SED reproduces a-posteriori the 24 μm data-point. Globally, our stacking analysis yields robust L_{IR} estimates in 37/42 bins.

In Fig. 7 we show the best-fit trend of $\langle U \rangle$ with redshift on our data. We find $\langle U \rangle = (2.4\pm0.5)\times(1+z)^{1.74\pm0.18}$, which is fully consistent with the revised $\langle U \rangle$–z trend of Béthermin et al. (2015): $\langle U \rangle = (3.0\pm1.1)\times(1+z)^{1.8\pm0.4}$. This test is reassuring, since it confirms that one single z-dependent (or $\langle U \rangle$–dependent)

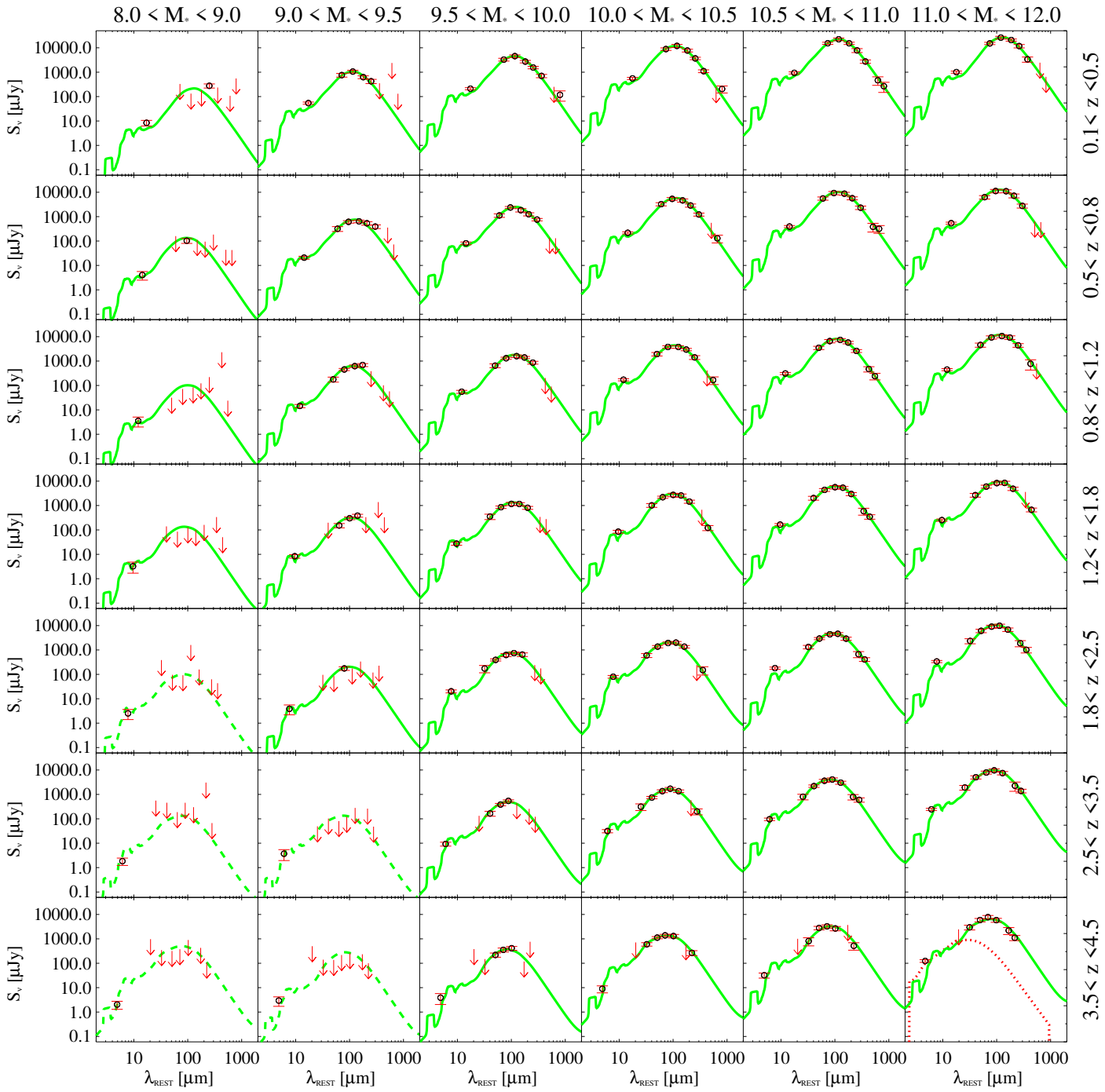

Fig. 6: Best-fit template obtained from SED-fitting decomposition (green lines), as a function of $M_\star$ (left to right, expressed in log $M_\odot$) and redshift (top to bottom). Red circles indicate the IR/sub-mm photometry (MIPS 24 μm, PACS 100-160 μm, SPIRE 250-350-500 μm, SCUBA 850 μm and AzTEC 1.1 mm), while downward arrows mark the corresponding 3σ upper limits. The red dotted line is the best-fit AGN template, shown in the only bin where its significance is above 3σ. Green dashed lines represent SEDs without FIR measurements and at z$\gtrsim$1.5, for which the integrated L_{IR} is interpreted as 3σ upper limit (5/42 bins). MIPS 24 μm fluxes are not used in the fitting.

MS galaxy template is fully able to reproduce the observed SED across a wide $M_\star$ interval.

Another caveat concerning the best-fit SEDs is that they are potentially broadened by the underlying distribution of redshifts and dust temperatures of the galaxies in the stacked samples (Magdis et al. 2012). To check for this potential issue, we convolved the observed z-distribution of each bin with a Gaussian kernel, whose dispersion is equal to the conservative photo-z uncertainty of 0.05×(1+z). Then we convolved the full MS library of Magdis et al. (2012) with such smoothed photo-z distribution, at each redshift bin. When shifted to the same rest-frame, the total SEDs always display a minimal broadening, mostly limited to PAH features, that affects the total IR luminosity by <<1%. This effect becomes observable only when stacking over much wider redshift bins. For this reason, we can safely ignore this effect.

Given the tight correlation between L_{IR} and SFR, IR data have been extensively used as proxy for SFR, assuming that most of galaxy star formation is obscured by dust (Kennicutt

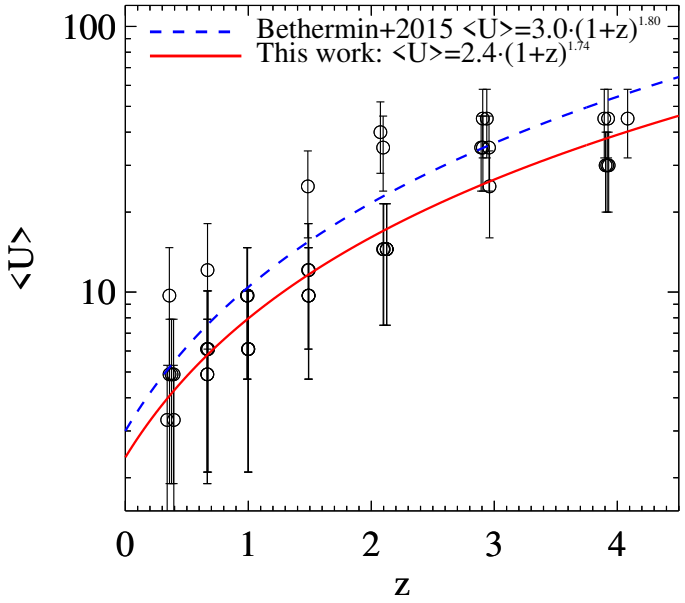

Fig. 7: Comparison of $\langle U \rangle$–z trends between Béthermin et al. (2015, blue dashed line) and our data (red solid line). Only the 37 bins with measured L_{IR} (thus $\langle U \rangle$) are shown.

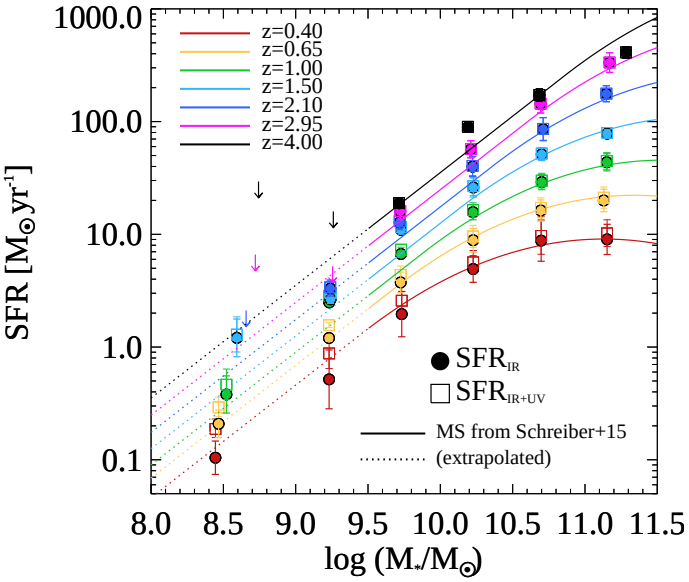

Fig. 8: SFR–$M_\star$ relation of the *NUVrJ* star-forming galaxies selected in this work, colour-coded by redshift over 0.1<z<4.5. At fixed $M_\star$ and redshift, SFR_{IR} measurements are converted from the L_{IR} obtained from IR/sub-mm SED-fitting (circles). Downward arrows indicate 3σ upper limits. For completeness, we also show the SFR_{IR+UV} estimates by combining SFR_{IR} with UV-uncorrected SFRs (open squares). Solid lines mark the evolving MS relation between SFR_{IR+UV} and $M_\star$ at different redshifts (Schreiber et al. 2015). We observe an excellent agreement with the MS, even below $M_\star \sim 10^{9.5} M_\odot$ that relies upon a linear extrapolation from Schreiber et al. (2015) not constrained by previous data.

1998; Kennicutt & Evans 2012). This is probably true inside the most massive star-forming galaxies (see Madau & Dickinson 2014 for a review). However, at decreasing $M_\star$, galaxies become metal poorer (e.g. Mannucci et al. 2010), thus less dusty and obscured. In these systems the ultraviolet (UV) domain provides a key complementary view on the unobscured star formation (Buat et al. 2012; Cucciati et al. 2012; Burgarella et al. 2013).

On this basis, the comprehensive work by Schreiber et al. (2015) exploited IR-based SFRs (i.e. SFR_{IR}) and UV-uncorrected SFRs, in the deepest CANDELS fields, to recalibrate the star-forming MS over an unprecedented $M_\star$ (down to $M_\star = 10^{9.5} M_\odot$) and redshift range ($z \lesssim 4$). Since we carried out a similar analysis, it is worth checking whether the SFR estimates based on our IR stacking reproduce or not the MS of Schreiber et al. (2015).

For consistency, we need to collect the UV-uncorrected SFRs for our input sample. For galaxies in the COSMOS2015 catalogue (Laigle et al. 2016), we use the rest-frame NUV luminosity L_{NUV} to estimate the UV-uncorrected SFR following Kennicutt & Evans (2012): $SFR_{UV}\ [M_\odot yr^{-1}] = 10^{-43.17} L_{NUV}/[erg\ s^{-1}]$. For the subset coming from the catalogue of Muzzin et al. (2013), we used the dust-uncorrected rest-frame UV luminosity at 2800 Å. In the latter case, the SFR_{UV} was calculated as $3.3 \times 10^{-10}\ L_{2800}/L_\odot$ (Kennicutt 1998) with Chabrier (2003) IMF.

Within each $M_\star$–z bin, we simply take the median SFR_{UV}, and we add it to the SFR_{IR} corresponding to the stacked L_{IR}, calculated as $SFR_{IR} = 10^{-10}\ L_{IR}/L_\odot$ (Kennicutt 1998, scaled to a Chabrier 2003 IMF). Fig. 8 displays our data in the SFR–$M_\star$ plane, colour-coded by redshift over 0.1<z<4.5. At fixed $M_\star$ and redshift, we show SFR_{IR} (circles) and the total SFR_{IR+UV} (open squares) for comparison. Downward arrows are 3σ upper limits scaled from L_{IR}. As can be seen, our data are in excellent agreement with the evolving MS relation at all redshifts (solid lines, Schreiber et al. 2015). While SFR_{UV} appears generally negligible compared to the total SFR, it becomes as high as SFR_{IR} towards low $M_\star$ and low redshift (e.g. Whitaker et al. 2012, 2017). Our median values agree with Schreiber et al. (2015) even below $M_\star \sim 10^{9.5} M_\odot$, at which we extrapolate the MS relation due to lack of previous data (dashed lines). This test compellingly demonstrates that our L_{IR} can be deemed robust over the full $M_\star$ and redshift interval explored in this work.

3.3. Radio stacking at 3 GHz

In this Section we describe the equivalent stacking analysis done with radio data, in order to derive average rest-frame 1.4 GHz luminosities ($L_{1.4\ GHz}$) in each $M_\star$–z bin. As mentioned in Sect. 2.3, the data used include VLA data at 1.4 GHz (Schinnerer et al. 2010) and 3 GHz (Smolčić et al. 2017b) over the full COSMOS field, as well as MIGHTEE data at 1.3 GHz over the central 1 deg^2 (Jarvis et al. 2016; I. Heywood et al. in prep.).

As done for IR stacking (Sect. 3.1), we treat detections and non-detections separately. Total fluxes of radio sources with 3<S/N<5 were taken from Jin et al. (2018) (see Sect. 2.3), while for brighter sources we matched their fluxes to those of the corresponding catalogues. The purpose of this approach is twofold: using the same published fluxes for S/N>5 detections for consistency and avoiding to deal with the effect of side-lobes from bright sources in stacked images, that might complicate total flux measurements (see Appendix A of Leslie et al. 2020 for a discussion). In addition, radio detections might contain a substantial fraction of AGN, that is expected to increase at higher $M_\star$ (e.g. Heckman & Best 2014). We will carefully deal with this issue in Sect. 4.2. At relatively faint flux densities (<100 μJy), most of radio emission is thought to arise from star formation (Bonzini et al. 2015; Padovani et al. 2015; Novak et al. 2017; Smolčić et al. 2017b), though some AGN-related radio emission might still be contributing (e.g. White et al. 2015; Jarvis et al. 2016). For this reason, median stacking of both detections and non-detections (e.g. Karim et al. 2011; Magnelli et al. 2015) in deep VLA-COSMOS 1.4 GHz images should result in mini-

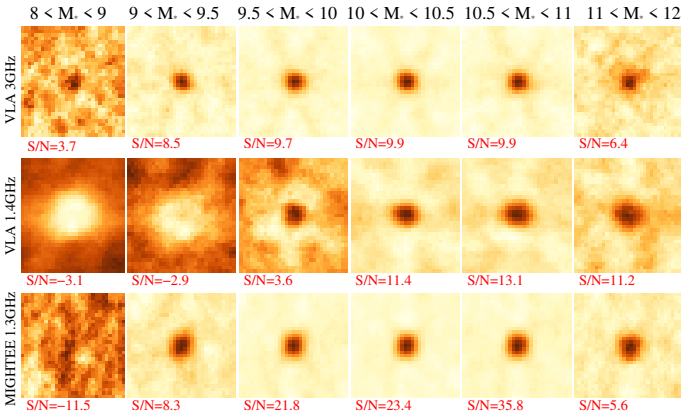

Fig. 9: Stacked cutouts of our sample at 0.8<z<1.2, as a function of $M_\star$ (left to right, expressed in log $M_\odot$). Only individual undetected sources (S/N<3) are stacked. The top, middle and bottom rows show VLA 3 GHz, VLA 1.4 GHz and MIGHTEE 1.3 GHz data, respectively. Each cutout size corresponds to 8×FWHM of the beam. Below each cutout we report the corresponding S/N of the total stacked flux.

mal radio AGN contamination. This alternative approach will be tested in Appendix C. Nevertheless, we note that median stacking yields meaningful results if the stacked image has homogeneous sensitivity, otherwise a mean rms-weighted stacking can better account for the position-dependent sensitivity.

Within the UltraVISTA area analyzed in this work, the 3 GHz rms (2.3 μJy beam^{-1}) fluctuates by less than 2% (Smolčić et al. 2017a). Indeed, we anticipate that no difference between median or rms-weighted mean stacking of non-detections is observed (see Appendix A and Leslie et al. 2020), as detailed below. For these reasons, we choose to perform median stacking of non-detections. Individual detections will be added a-posteriori via a simple mean weighted average, as done in Eq. 3.

Our stacking routine generates cutouts with size of 8×FWHM$_{3\ GHz}$ (i.e. 6" at 3 GHz), centered on the NIR position of each input galaxy. We acknowledge that an average offset of 0.1" was found between 3 GHz (Smolčić et al. 2017a) and UltraVISTA positions (Laigle et al. 2016), which is half the size of a pixel. To account for this systematic offset, our routine performs sub-pixel interpolation and searches for the peak flux (S_{peak}) within ±1 pixel from the center of the stacked image. The peak flux uncertainty is estimated via bootstrapping 100 times, as done in Sect. 3.1. We take the median of the resulting flux distribution as our formal peak flux. The 1σ dispersion around this value is interpreted as the flux error. We also measured the standard deviation across 100 random positions in the stack (masking the central beam of 0.75"). This gives less conservative errors compared to a bootstrap, but it is used to derive the formal rms of the stacked map.

Total fluxes (S_{tot}) are calculated by fitting a 2D elliptical Gaussian function to the median stacked image, using the IDL routine MPFIT2DFUN[3]. Given the typically high S/N (~10 on average) reached in the central pixel, we leave size, position angle and normalization of the 2D Gaussian as free parameters. We verified that adopting a circular Gaussian or forcing the normalization to the peak flux does not significantly affect any of our stacks. The total flux was calculated by integrating over the 2D Gaussian area A_{gauss}. The integrated flux error was computed by multiplying the peak flux error by $\sqrt{A_{gauss}/A_{beam}}$, where A_{beam} is the known beam area, and adding a known 5% flux calibration error in quadrature (Smolčić et al. 2017b). We remind the reader that the peak flux error already incorporates the variance of the stacked sample via bootstrapping.

In order to assess whether our sources are clearly resolved, we follow the same criterion applied to VLA 3 GHz detections (Smolčić et al. 2017b) to identify resolved sources:

$$\frac{S_{tot}}{S_{peak}} > 1 + (6 \times S/N_{peak})^{-1.44} \quad (4)$$

where S/N_{peak} is simply the peak flux divided by the rms of the image. This expression was obtained empirically to define an envelope containing 95% of unresolved sources, below such threshold. We find that 31 stacks out of 42 are resolved, according to Eq. 4. For these, total fluxes are on average 1.8× higher than peak fluxes. Similarly, Bondi et al. (2018) found 77% of VLA 3 GHz detected SFGs are resolved, and this fraction does not change significantly with $M_\star$ (Jiménez-Andrade et al. 2019). Of the 11 bins with unresolved emission, 3 have S/N_{peak} <3. These are all among the 5 bins without L_{IR} estimates from IR stacking (Sect. 3.2). Analogously to our treatment of the IR measurements, we discard all those 5 bins from the rest of our analysis.

For the stacks with resolved emission, we prefer to use their integrated flux from 2D Gaussian fitting as the most accurate estimate. Instead, for unresolved stacks we use the peak flux, consistent with the treatment of 3 GHz detections (Smolčić et al. 2017b). Fitting residuals are on average 3% of the total flux, and always consistent with zero within the uncertainties. We validate this approach by reproducing the total fluxes of 3 GHz detections presented in Smolčić et al. (2017b) at S/N>5 and in Jin et al. (2018) at 3<S/N<5, respectively (Appendix A).

Finally, we combined the radio stacked flux densities within each $M_\star$–z bin together with individual detections, following Eq. 3. The combined 3 GHz fluxes were first scaled to 1.4 GHz assuming of $S_\nu \propto \nu^\alpha$, with spectral index α=–0.75±0.1 (e.g. Condon 1992; Ibar et al. 2009,2010). This assumption is discussed in Sect. 3.3.2. Lastly, 1.4 GHz fluxes were converted to rest-frame 1.4 GHz luminosities ($L_{1.4\ GHz}$), again assuming α=–0.75. Formal $L_{1.4\ GHz}$ errors were calculated by propagating the uncertainties on both combined flux and spectral slope.

3.3.1. Stacking ancillary VLA and MIGHTEE data

As mentioned before, deep ancillary radio data are available in COSMOS, coming from the VLA-COSMOS (Schinnerer et al. 2010) survey at 1.4 GHz and the MIGHTEE (Jarvis et al. 2016) survey at 1.3 GHz. While VLA observations reach rms=12 μJy beam^{-1} with 2.5" resolution, MIGHTEE images formally reach 2.2 μJy beam^{-1} at 8.4"×6.8" resolution over 1 deg^2 in the MIGHTEE early science data, but the effective depth is limited by confusion (~5.5 μJy beam^{-1} in the central part).

Here we perform a radio stacking analysis, as for 3 GHz data, in order to check whether our 3 GHz based $L_{1.4\ GHz}$ are stable against different resolutions or spectral frequencies.

Source fluxes in VLA 1.4 GHz and MIGHTEE 1.3 GHz maps were re-extracted, using K_s+MIPS 24 positional priors. While the angular resolution at VLA 1.4 GHz is high enough to yield a negligible fraction of overlapping priors within the beam, MIGHTEE data suffer from blending issues. To this end, MIGHTEE fluxes were de-blended as in Jin et al. (2018) down to 3σ level. Then, individual S/N>3 detections were removed from the original image, and we used the residual map for stacking 1.3 GHz non-detections. Of course, only sources within

[3] https://pages.physics.wisc.edu/~craigm/idl/fitqa.html

the MIGHTEE area (central 1 deg^2) were stacked, containing roughly half of the sample size used for VLA stacking.

The stacking analysis follows the same reasoning and assumptions presented in Sect. 3.3. However, because of the heterogeneous depth of VLA 1.4 GHz and MIGHTEE 1.3 GHz observations, rms-weighted mean stacking can better account for the position-dependent sensitivity, and thus we use it instead of median stacking. We demonstrate that mean and median stacking at 3 GHz yield fully consistent results (Appendix A), given the minimal rms fluctuations across the VLA 3 GHz map.

Stacked MIGHTEE fluxes are measured in the central pixel, that is assumed to trace the total flux. VLA 1.4 GHz peak fluxes were, instead, scaled to total fluxes as done at 3 GHz. Nonetheless, a different, yet empirical relation was adopted to identify resolved sources at VLA 1.4 GHz, calibrated on 1.4 GHz detections (Schinnerer et al. 2010): $S_{tot}/S_{peak} > 0.35^{-11/(S/N^{1.45}_{peak})}$. Because of the larger beamsize compared to 3 GHz, we find fewer resolved stacks (3/20). Total fluxes were converted to rest-frame $L_{1.4\ GHz}$ assuming α=–0.75±0.1, that was propagated along with flux errors to deliver reliable $L_{1.4\ GHz}$ uncertainties. Upper limits at 3σ were assigned for S/N<3 stacks.

Fig. 9 shows stacked cutouts at 0.8<z<1.2 at VLA 3 GHz (top), VLA 1.4 GHz (middle) and MIGHTEE 1.3 GHz (bottom) data, as a function of $M_\star$ (increasing from left to right). While stacking at 3 GHz delivers S/N>3 in 39/42 bins, only 11/42 and 20/42 have S/N>3 in VLA 1.4 GHz and MIGHTEE 1.3 GHz stacked images, respectively. For VLA 1.4 GHz, the small number of bins is attributed to shallower than 3 GHz observations. For MIGHTEE, instead, this is probably induced by the confusion-limited signal in the stacks due to the larger MeerKAT primary beam at 1.3 GHz (e.g. Mauch et al. 2020). Nevertheless, because VLA 3 GHz data are much less sensitive than MeerKAT to large-scale radio emission, total radio fluxes might be underestimated at 3 GHz. This issue can be, however, especially relevant at low redshift (z<0.5) and for resolved/multi-component radio sources (e.g. Delhaize et al. 2020, submitted). In fact, a visual inspection of the median 3 GHz stacks of non-detections does not reveal clearly missing flux in the residual images at the scales of the MIGHTEE beam, except in the bin at z<0.5 and 10^{11} <$M_\star/M_\odot$ <10^{12}. To quantify this effect, we convolved all the original 3 GHz stacked cutouts with a Gaussian kernel of 3" FWHM, re-calculating the total fluxes and comparing them with the previous measurements. Of course, this drastically reduces the global S/N of the final stacks, leaving us with S/N>3 in only 16/42 bins (as opposed to 39/42 before). In those 16 bins, we confirm that the bin at the lowest z and highest $M_\star$ displays on average 0.3 dex larger total flux, while the remaining fluxes are fully consistent with our previous estimates. Given the lack of missing flux in the smoothed residual image, we replaced the total flux of that only bin with the convolved flux and treat it as unresolved. In any case, we stress that the final $L_{1.4\ GHz}$ obtained by combining both detections and non-detections is unchanged, since the fraction of radio detections is about 56% at z<0.5 and $M_\star$ >10^{11} $M_\odot$ (see Fig. 2), thus washing out the difference in the stacked flux. As a consequence, this effect has no impact on the rest of our analysis. In addition, we emphasize that any extra missing 3 GHz flux at low redshift would further strengthen our final redshift-invariant IRRC (Sect. 4.3). This motivates our choice of using primarily VLA 3 GHz images for our analysis.

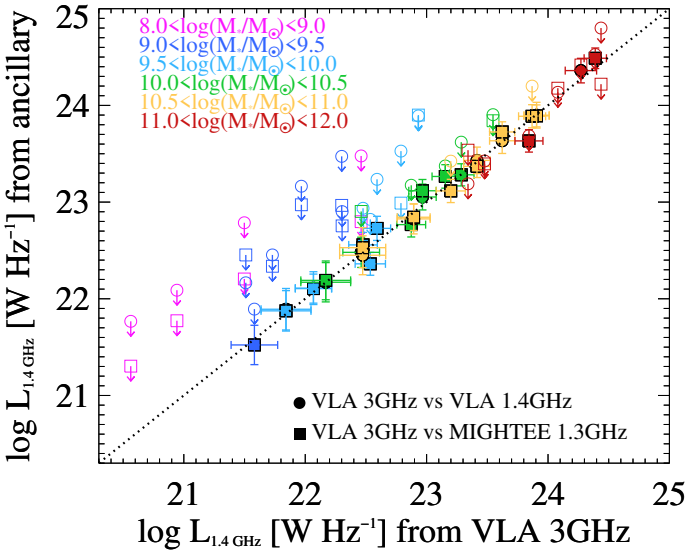

Fig. 10: Comparison between rest-frame 1.4 GHz luminosity $L_{1.4\ GHz}$ obtained from 3 GHz stacks (x-axis) and ancillary radio stacks (y-axis) using VLA (1.4 GHz, circles) and MIGHTEE (1.3 GHz, squares) data. We assumed a single spectral index α=–0.75 to scale fluxes from 3 GHz to 1.4 GHz. Colours indicate different $M_\star$ ranges. Downward arrows with open symbols mark 3σ upper limits if S/N<3. The broad agreement between the various datasets suggests that using a single α=–0.75 is a reasonable assumption across the full $M_\star$ range explored in this work.

3.3.2. Considerations on the radio spectral slope

We briefly discuss and test our assumption of taking a single spectral slope α=–0.75 by comparing $L_{1.4\ GHz}$ estimates independently inferred from stacking the three above datasets. In Fig. 10, we compare $L_{1.4\ GHz}$ obtained from 3 GHz stacks (x-axis) and ancillary radio stacks (y-axis), using either VLA (1.4 GHz, circles) or MIGHTEE (1.3 GHz, squares) data, colour-coded by $M_\star$. Downward arrows with open symbols mark 3σ upper limits where the stacked S/N<3. We find a good agreement between all the various datasets, suggesting that using a single α=–0.75 is a reasonable assumption across the full $M_\star$ range explored in this work. The global offset between 3 GHz and 1.4 GHz luminosities is –0.015 dex, and the scatter is 0.08 dex. We note again that radio stacks from VLA 1.4 GHz and MIGHTEE 1.3 GHz data are *mean*, while 3 GHz stacks are *median* stacks. However, the agreement would be unchanged if using VLA 3 GHz *mean* stacks, as shown in Fig. A.2. As a sanity check, the median spectral slope traced by VLA 3 GHz and MIGHTEE 1.3 GHz individual detections is –0.77, in agreement with our assumption. However, we prefer to adopt a fixed α=–0.75 in our calculation to treat both radio detections and non-detections in a self-consistent manner.

Magnelli et al. (2015) measured the average spectral index exploiting VLA 1.4 GHz and GMRT 610 MHz data for an $M_\star$-selected galaxy sample. They found that the observed 610 MHz–1.4 GHz slope, that probes closer to rest-frame 1.4 GHz than our 3 GHz data, does not seem to change with $M_\star$ or SFR, at least out to z~2. More recently, Calistro Rivera et al. (2017) exploited Low Frequency Array (LOFAR) data at 150 MHz in the Boötes field, out to z~2.5. Interestingly, they observed a spectral flattening of the radio SED of SFGs towards low frequency (observed range 150 MHz–1.4 GHz), probably due to synchrotron self-absorption (see also Read et al. 2018; Gürkan et al. 2018). However, they argue that this feature should not

affect the k-correction for the rest-frame 1.4 GHz luminosities $L_{1.4\,GHz}$. Therefore, these studies provide mounting evidence that using a single power-law spectral index α=–0.75 at our frequency is a reasonable assumption.

The various checks described above (see also Appendix A) prove our $L_{1.4\,GHz}$ robust across the full range of $M_\star$ and redshift analyzed in this work. We note that our $L_{1.4\,GHz}$ estimates do not necessarily trace radio emission from star formation. Indeed, radio AGN are not yet removed at this stage, and they might be potentially boosting the $L_{1.4\,GHz}$. This issue will be addressed in Sect. 4.2.

4. The IRRC and the contribution of radio AGN

Using the median L_{IR} and $L_{1.4\,GHz}$ luminosities obtained from stacking, we study the evolution of the IRRC as a function of $M_\star$ and redshift. Logarithmic uncertainties on each luminosity were propagated quadratically to get q_{TIR} errors. Among the 42 $M_\star$–z bins analyzed in this work, 37 yield robust estimates of L_{IR} and $L_{1.4\,GHz}$, while the remainder are discarded from the following analysis. Unsurprisingly, these latter 5 bins (three at 10^8 <$M_\star/M_\odot$ <10^9 and 1.8<z<4.5; two at 10^9 <$M_\star/M_\odot$ <$10^{9.5}$ and 2.5<z<4.5) are among the least complete in $M_\star$, as highlighted in Figs. 2 and 6. Therefore their exclusion partly mitigates the $M_\star$ incompleteness of the remaining sample.

4.1. q_{TIR} before removing radio AGN

Fig. 11 shows the average q_{TIR} as a function of redshift, colour-coded in $M_\star$ (stars). For comparison, other prescriptions of the evolution of the IRRC are overplotted (black lines). Bell (2003) inferred the average IRRC in local SFGs, finding q_{TIR}=2.64±0.02 (dotted line), with a scatter of 0.26 dex. Magnelli et al. (2015) studied an $M_\star$-selected sample at z≲2, and constrained the evolution of the far-infrared radio correlation (FIRC, parametrized via q_{FIR}[4]) across the SFR–$M_\star$ plane at $M_\star$ >10^{10} $M_\odot$. From stacking IR and radio images, they parametrized the evolution with redshift of the FIRC as: $q_{FIR}=(2.35\pm0.08)\times(1+z)^{-0.12\pm0.04}$, where the normalization is scaled to 2.63 in the q_{TIR} space. More recently, Delhaize et al. (2017) exploited a jointly-selected sample of IR (from *Herschel* PACS/SPIRE) or radio (from the VLA-COSMOS 3 GHz Large Project, Smolčić et al. 2017b) detected sources (at ≥5σ) in the COSMOS field. Through a survival analysis that accounts for non-detections in either IR or radio, they inferred the evolution of the IRRC with redshift out to z∼4 as: $q_{TIR}=(2.88\pm0.03)\times(1+z)^{-0.19\pm0.01}$. While this trend appears somewhat steeper than that of Magnelli et al., we note that Delhaize et al. (2017) did not formally remove objects with significant radio excess, while Magnelli et al. (2015) performed median radio stacking to mitigate the impact of potential outliers such as radio AGN. Nevertheless, Delhaize et al. (2017) argue that the IRRC trend with redshift would flatten if applying a 3σ-clipping: $q_{TIR}=(2.83\pm0.02)\times(1+z)^{-0.15\pm0.01}$, which becomes fully consistent with that of Magnelli et al. (2015).

When compared to the above literature, it is evident that our q_{TIR} values lie systematically below other studies at $M_\star$ >10^{11} $M_\odot$, while lower $M_\star$ galaxies lie closer or slightly above them. In other words, our q_{TIR} estimates seem to display

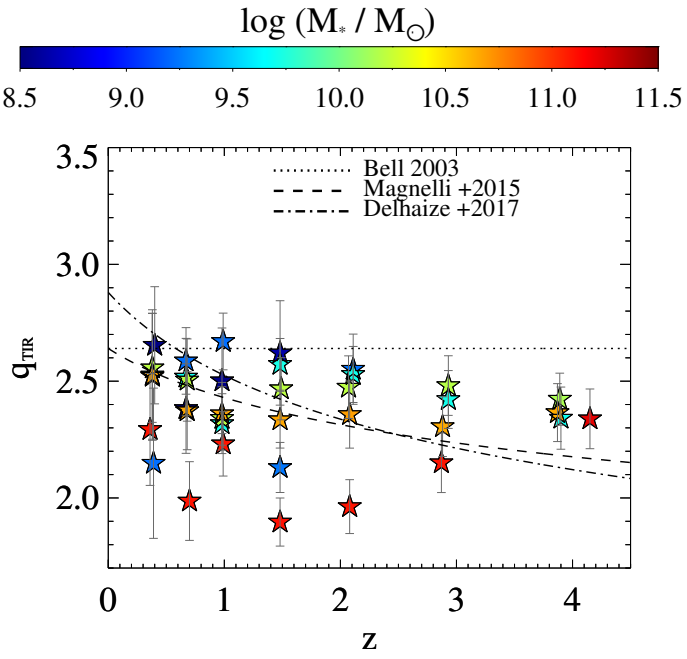

Fig. 11: AGN-uncorrected q_{TIR} evolution as a function of redshift (x-axis) and $M_\star$ (colour bar). Errors on q_{TIR} represent the 1σ scatter around the median value, estimated via bootstrapping over L_{IR} and $L_{1.4\,GHz}$. For comparison, other IRRC trends with redshift are taken from the literature (black lines): Bell (2003, dotted); Magnelli et al. (2015, dashed); Delhaize et al. (2017, dot-dashed). Our q_{TIR} values still include the contribution of radio AGN. See Sect. 4.1 for details.

a clear $M_\star$ stratification, with the most massive galaxies having typically lower q_{TIR} than less massive counterparts. As mentioned before, we remind the reader that our sample, at this point, contains some fraction of radio AGN, which might be boosting the $L_{1.4\,GHz}$, particularly at high $M_\star$ (see e.g. Best & Heckman 2012) where radio AGN feedback is known to be prevalent. Our L_{IR} estimates are, instead, corrected for a potential IR-AGN contribution (Sect. 3.2). Therefore, the net effect caused by including AGN is *lowering* the intrinsic q_{TIR}. Selecting typical SFGs on the MS should, however, reduce the incidence of powerful radio AGN expected in massive hosts, since most radio AGN at z<1 are found to reside in quiescent galaxies (e.g. Hickox et al. 2009; Goulding et al. 2014).

For these reasons, we caution that Fig. 11 should be taken as the AGN-uncorrected q_{TIR}. However, it is worth showing it to quantify how much q_{TIR} will change after removing radio AGN.

4.2. Searching for radio AGN candidates

In this Section we carry out a detailed study aimed at identifying potential radio AGN, removing them and ultimately deriving the intrinsic q_{TIR} trend purely driven by star formation.

In our radio analysis we have combined individual radio-detections (above S/N>3) with undetected sources via a weighted average (Eq. 3). Contrary to stacking detections and non-detections together, this formalism enables us to characterize the nature of individual radio detections, i.e. whether they show excess radio emission relative to star formation.

We make the underlying assumption that radio-undetected AGN do not significantly affect any of our radio stacks. This is supported by the excellent agreement between mean and median stacked $L_{1.4\,GHz}$ of non-detections (Fig. A.2, bottom panel). Indeed, if the contribution of radio-undetected AGN were sub-

[4] The far-infrared luminosity used to compute q_{FIR} was integrated between 42 and 122 μm rest-frame. This is quantified to be 1.91× smaller than the total L_{IR} (Magnelli et al. 2015).

stantial, the corresponding mean $L_{1.4\ GHz}$ would be significantly higher than the median $L_{1.4\ GHz}$ from stacking. This assumption is further supported by the fact that the fraction of identified radio AGN is a strong function of radio flux density, and the sources we stack are by construction faint in the radio. Algera et al. (2020b) argue that below 20 μJy (at 3 GHz), the fraction of radio-excess AGN is <10% (see also Smolčić et al. 2017a, Novak et al. 2018). We acknowledge that our assumption does not allow us to collect a *complete* sample of radio AGN, especially at high redshift where the fraction of radio detections notably drops (Fig. 3). Nevertheless, we will show that any residual AGN contribution does not change our conclusions.

We briefly summarize our next steps as follows. In Section 4.2.1 we explore the q_{TIR} distribution traced by individual 3 GHz detections as a function of $M_\star$ and redshift. First, we identify a subset of radio detections at $M_\star > 10^{10.5}$ $M_\odot$ that is representative of an $M_\star$-selected sample. Then we decompose their q_{TIR} distribution between AGN and star formation components (Sect. 4.2.2). This enables us to subtract potential radio AGN candidates, and calibrate the intrinsic best-fit IRRC with redshift at $M_\star > 10^{10.5}$ $M_\odot$ (Sect. 4.2.3). Later we extrapolate this calibration towards lower $M_\star$ bins (Sect. 4.2.4), where a similar in-depth analysis was not possible due to radio-detections being strongly incomplete in this $M_\star$ regime. Finally, the intrinsic (i.e. AGN-corrected) IRRC as a function of $M_\star$ and redshift is presented in Sect. 4.3.

4.2.1. The q_{TIR} distribution of radio detections

In order to study the q_{TIR} distribution of 3 GHz detections, we need to calculate their average L_{IR} as a function of $M_\star$ and redshift. For convenience, we refer the reader back to Fig. 2 (blue histograms) for visualizing the distribution of 3 GHz detections in the $M_\star$–z space. Out of 13,510 radio detections among our 37 bins, 8762 (65%) have a combined $S/N_{IR} > 3$, therefore reliable L_{IR} measurements from SED-fitting of FIR/sub-mm de-blended photometry (Jin et al. 2018). For the remainder of the sample, we stack again their IR/sub-mm images in all bands in each $M_\star$–z bin. Stacked IR fluxes are corrected for clustering bias and converted to L_{IR} following the same procedure adopted for the prior $M_\star$ sample (Sect. 3.1). Median stacked L_{IR} are retrieved for the same 37/42 bins of the full parent sample, since a stacked S/N>3 flux was obtained in at least one FIR/sub-mm band. Then, for each source we re-scaled its median stacked L_{IR} to the redshift and $M_\star$ of that source (assuming the MS relation), in order to reduce the variance of the underlying sample within each $M_\star$–z bin. We verified that our stacked L_{IR} are always systematically below the 3σ L_{IR} upper limits inferred from FIR/sub-mm SED-fitting (Jin et al. 2018). This ensures that our stacking analysis provides more stringent constraints on the L_{IR} of individual non-detections.

From this analysis, we are in a sweet spot to explore the full q_{TIR} distribution of 3 GHz detections at different $M_\star$ and redshifts. Fig. 12 shows q_{TIR} as a function of redshift, split in six $M_\star$ bins. Black dots mark individual 3 GHz detections, blue stars represent the q_{TIR} obtained by combining detections and non-detections (same as in Fig. 11), while yellow squares are the stacks of non-detections only. In each panel we report the number of 3 GHz detected sources, and the fraction of them with combined $S/N_{IR} > 3$. This fraction strongly increases with $M_\star$, from 3.5% at $10^8 < M_\star/M_\odot < 10^9$ to 78.3% at $10^{11} < M_\star/M_\odot < 10^{12}$, which implies that at the lowest $M_\star$ nearly all q_{TIR} estimates of radio detections rely upon IR stacking. This is because the 3 GHz detection limit sets a rough threshold in SFR (if radio emission primarily arises from star formation), therefore is biased towards high-$M_\star$ galaxies because of the MS relation. Because of these potential biases, it is essential to identify the bins in which radio detections give us access to a representative sample of $M_\star$-selected galaxies.

Indeed, our purpose is to use single radio detections to calibrate a threshold that best distinguishes radio AGN from radio SFGs, as a function of $M_\star$ and redshift. In order to extend this calibration to our full $M_\star$-selected sample, we need to make sure that our derived trends are not affected by selection biases, i.e. that the radio-detected sources we rely upon are fully representative of $M_\star$-selected galaxies at a given redshift. To this end, within each bin we compare the q_{TIR} of single radio detections against a specific threshold ($q_{TIR,lim}$), corresponding to the 3 GHz survey limit at a given $M_\star$ and redshift (green upward arrows in Fig. 12). This threshold is proportional to the median stacked L_{IR} of the full SFGs sample, divided by the 3σ luminosity limit at 1.4 GHz, scaled from 3 GHz by assuming a fixed spectral index α=-0.75 (Sect. 3.3.2). Specifically, $q_{TIR,lim}$ indicates the limiting q_{TIR} at which a typical MS galaxy of a given $M_\star$, z and L_{IR} drops below the 3 GHz detection limit, which translates into a *lower* q_{TIR} limit. In other words, sources with $q_{TIR} > q_{TIR,lim}$ lie within an $M_\star$ range that is virtually inaccessible by our 3 GHz survey. Therefore, any measurement above that threshold is not representative of an $M_\star$-selected sample. Conversely, radio detections below that threshold would be seen also in an $M_\star$-selected sample of SFGs.

In this framework, we consider as *complete* only those $M_\star$–z bins in which at least 70% of radio detections are below $q_{TIR,lim}$. This cutoff enables us to narrow down the position of the mode of the q_{TIR} distribution, leaving us with a total of 13 *complete* bins (at >70% level) across the full sample. Unsurprisingly, they are preferentially located in high-$M_\star$ galaxies, and/or at low redshift. These are delimited by a red segment in Fig. 12.

It is quite evident that the locus populated by radio detections tends to decline with redshift, at each $M_\star$. However, this behaviour is far more pronounced at low $M_\star$, and likely driven by selection effects. In fact, by definition the $L_{1.4\ GHz}$ of radio detections increases with redshift at all $M_\star$, because 3 GHz sources are drawn from a flux-limited sample. On the contrary, the L_{IR} of radio detections behaves differently with $M_\star$: at higher $M_\star$ it is mostly based on IR-detected sources, while at lower $M_\star$ it comes predominantly from IR stacking. At higher $M_\star$, L_{IR} increases with redshift similarly to $L_{1.4\ GHz}$, giving rise to a nearly flat q_{TIR} locus. At lower $M_\star$, instead, L_{IR} stands typically below the IR detection limit, thus not bound to a monotonic redshift increase. This effect causes an apparent *decrease* of q_{TIR} with redshift, that is mostly driven by the radio detection limit. Indeed, a similar trend can be seen in the green arrows, that move down in redshift at each $M_\star$.

Since the single complete z-bin found at $10^{10} < M_\star/M_\odot < 10^{10.5}$ is insufficient for us to constrain a redshift trend, we only consider the remaining 12 complete bins placed at $M_\star > 10^{10.5}$. For each of them, we identify the peak of the corresponding q_{TIR} distribution of radio detections, namely $q_{TIR,peak}$ (see red open circles in Fig. 12). We note that $q_{TIR,peak}$ represents the *mode* of radio detections, rather than the *average* that is, instead, potentially affected by underlying radio AGN (Sect. 4.2.2). Then we fitted a power law trend of $q_{TIR,peak}$ with redshift using the IDL routine MPFIT2DFUN, obtaining the following best-fit expressions: $q_{TIR}=(2.78\pm0.05)\times(1+z)^{-0.16\pm0.03}$ at $10^{10.5} < M_\star/M_\odot < 10^{11}$, and $q_{TIR}=(2.66\pm0.04)\times(1+z)^{-0.08\pm0.01}$ at $10^{11} < M_\star/M_\odot < 10^{12}$.

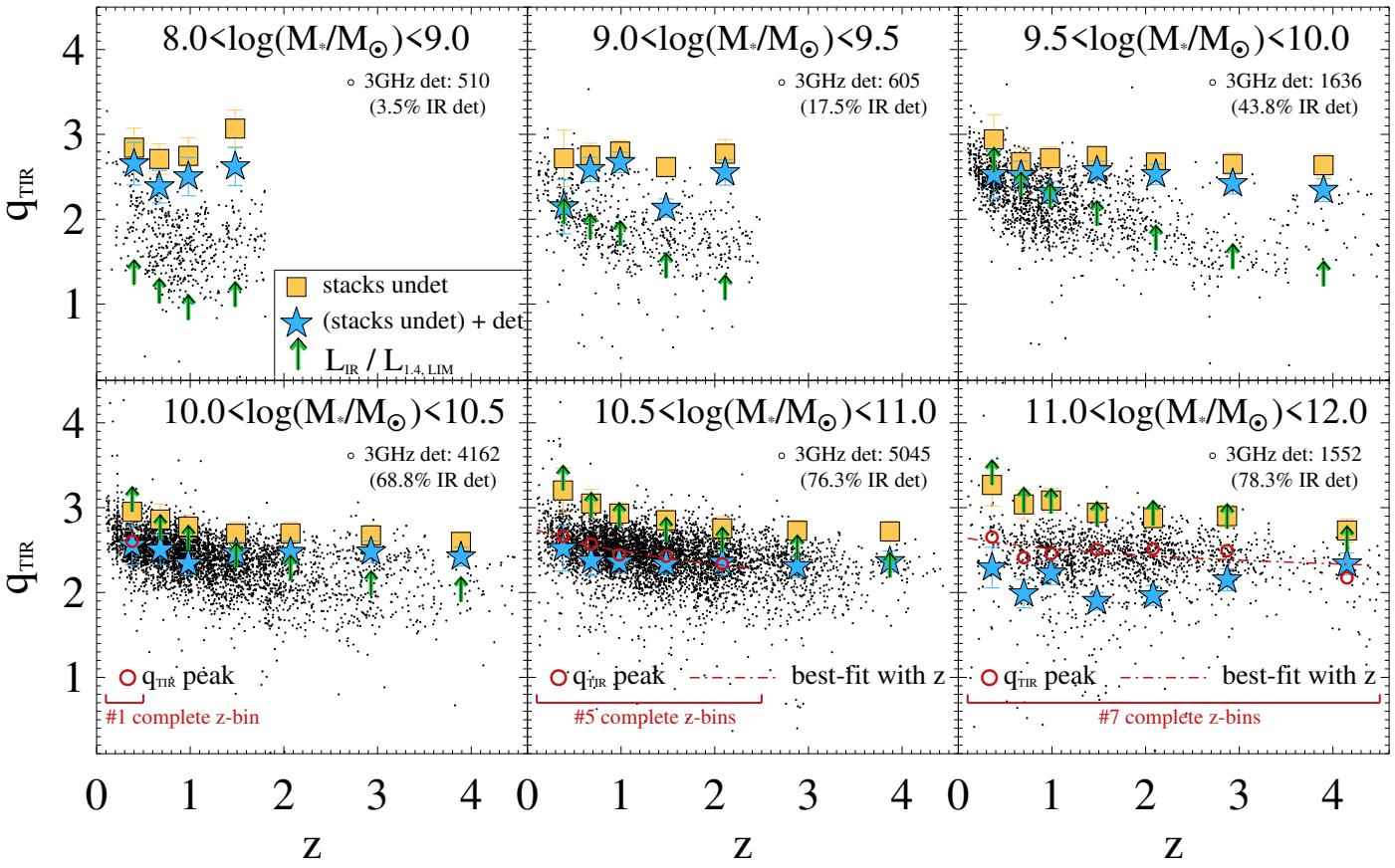

Fig. 12: Distribution of q_{TIR} as a function of redshift, split across increasing $M_\star$ bins. In each panel, we compare the q_{TIR} estimates of individual radio detections (black dots) with the median stacked values of non-detections (yellow squares) and with the weighted-average q_{TIR} of detections and non-detections together (from Eq. 3, blue stars). Green upward arrows indicate the corresponding threshold $q_{TIR,lim}$ above which radio detections become inaccessible from an $M_\star$-selected sample. We select relatively *complete* $M_\star$–z bins, in which at least 70% of radio detections have q_{TIR} below the corresponding threshold. This criterion identifies 13 bins (in red square brackets). Within these bins, the peak of q_{TIR} distribution ($q_{TIR,peak}$) is indicated with red open circles. In the two highest $M_\star$ bins, the best fitting trends with redshift are shown by the red dashed lines. Each panel reports the number of individual 3 GHz sources and their fraction with S/N_{IR} >3. See Sect. 4.2.1 for details.

4.2.2. Identifying radio AGN at high $M_\star$

After fitting the $q_{TIR,peak}$ trend with redshift for the two $M_\star$ bins, we need to account for such redshift dependence while exploring the q_{TIR} distributions of radio detections. To this end, we align the position of $q_{TIR,peak}$ in each redshift bin to match the best-fit redshift trend. This allows us to marginalize over the internal redshift trend, and merge all radio detection homogeneously within the same $M_\star$ bin. The resulting redshift-corrected q_{TIR} distribution is displayed in Fig. 13 for the two highest $M_\star$ bins (left and right panel). Each total histogram (black) includes the contribution from both galaxy- and AGN-dominated radio sources. We proceed to dissecting into the two components as follows , leaving the discussion on how radio AGN affect the redshift trend in Sect. 4.2.3.

Assuming that the peak of the distribution is populated by radio-detected SFGs, and that the intrinsic q_{TIR} distribution of SFGs is symmetric around the peak, we mirror the right-hand side of the observed q_{TIR} distribution to the left side. This symmetric function is interpreted as the intrinsic q_{TIR} distribution of normal SFGs (blue histogram). We fitted it with a Gaussian function, leaving normalization and dispersion free to vary. The Gaussian fit yields a dispersion of 0.20 and 0.23 dex at $10^{10.5}$ $<M_\star/M_\odot<10^{11}$ and 10^{11} $<M_\star/M_\odot<10^{12}$, respectively (blue dot-dashed lines). The residual histogram (total-SF) is then fitted with a second Gaussian function (red dot-dashed lines), that parametrizes the additional radio-excess population ascribed to AGN. We attempted to fit the AGN population with other non-Gaussian functions, since the lowest q_{TIR} tail is not perfectly reproduced with a Gaussian shape. However, we stress that our purpose is separating the two populations statistically and prioritize a clean identification of SFGs, while a proper characterization of the *shape* of the AGN population is beyond the scope of this paper.

We note that our fitting approach relies on the assumption that $q_{TIR,peak}$ is entirely attributed to SF. Therefore, by mirroring and fitting the SF Gaussian first, it is possible that we might be underestimating the intrinsic relative fraction of radio AGN. We discuss this potential issue in Appendix B, though we anticipate that our main findings could only be reinforced if addressing this effect.

Another possible caveat of our approach lies in the assumption that IR-undetected sources are represented by a single stacked L_{IR}, though rescaled to the $M_\star$ and redshift of each object based on the MS relation. However, we checked that the distribution of radio detections that are also IR detected displays an average scatter of 0.22 dex, as for the full radio-detected sample shown in Fig. 13. This is because the vast majority of radio sources at $M_\star$ $>10^{10.5}$ $M_\odot$ is also individually detected at IR wavelengths (see Fig. 3). Therefore, our assumption does not strongly impact the calibration of the SF locus.

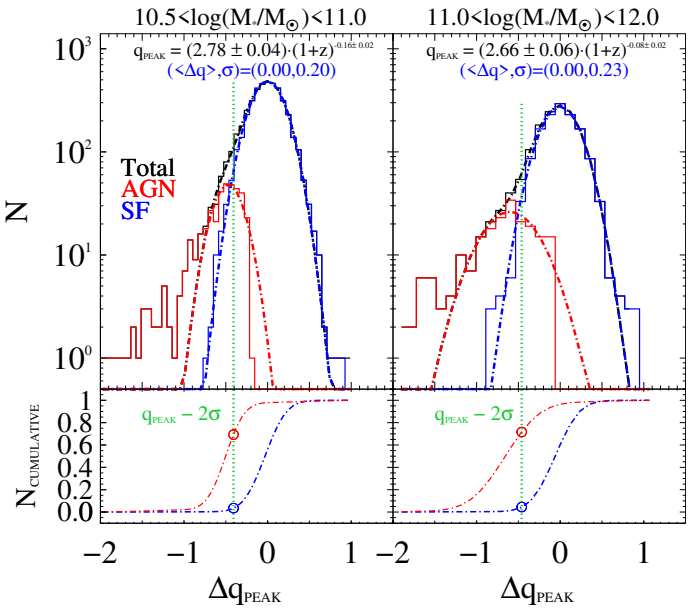

Fig. 13: q_{TIR} distribution of 3 GHz detections in the two highest $M_\star$ bins (left and right panels), after correcting for the internal q_{TIR}–redshift dependence. Only sources within complete z-bins were considered at each $M_\star$. The total distribution (black histogram) was dissected among SF-dominated (blue) and AGN-dominated (red) radio sources, and fitted with two separate Gaussian functions (dot-dashed lines). While the SF population was fitted first, the AGN part was fitted in a second step from the total–SF residuals. The 1σ dispersion attributed to SF equals 0.20 and 0.23 dex at $10^{10.5}$ $<M_\star/M_\odot<10^{11}$ and 10^{11} $<M_\star/M_\odot<10^{12}$, respectively. The bottom panels display the corresponding cumulative Gaussian fits, both normalized to unity. The vertical green dotted line marks the threshold that best separates between the SF and AGN populations (see Table 2), which rejects about 70% AGN and only 3–4% SFGs (open circles).

Choosing the best dividing line between AGN and SF-dominated radio sources is a challenging, and somewhat arbitrary task. Moving the threshold to higher q_{TIR} increases the purity of SFGs to the detriment of completeness, and vice-versa for a lower threshold. Here we attempt to reach low levels of cross-contamination between the SF and AGN populations, while keeping a high completeness of the SF population. For this reason, we checked the cumulative q_{TIR} distribution drawn by the two Gaussian fits (AGN in red, SF in blue), as shown at the bottom of Fig. 13, each normalized to unity.

Four different thresholds (q_{thres}) were examined: (1) $q_{thres}=q_{peak}-1\sigma$; (2) $q_{thres}=q_{peak}-2\sigma$; (3) $q_{thres}=q_{peak}-3\sigma$; (4) $q_{thres}=q_{cross,AGN=SF}$. In this formalism, q_{peak} is still the peak of the SF population (blue Gaussian fit in Fig. 13), and σ its dispersion, while $q_{cross,AGN=SF}$ represents the cross-over value at which the numbers of radio AGN and radio SFGs match each other. For each threshold, in Table 2 we report the cumulative fractions of SF and AGN populations lying below it. Qualitatively speaking, an ideal compromise consists of a low fraction of SF galaxies and a high fraction of AGN below the threshold.

This comparison highlights that the best trade-off between cross-contamination and completeness is given by the threshold $q_{thres}=q_{peak}-2\sigma$, in both $M_\star$ bins. This method rejects about 70% of potential radio-excess AGN, and only 3–4% of SFGs, that we believe is quite acceptable. The offset from the corresponding q_{peak} value is about 0.40 and 0.46 dex at $10^{10.5}$ $<M_\star/M_\odot<10^{11}$ and 10^{11} $<M_\star/M_\odot<10^{12}$, respectively. Given the similarity between these values, for simplicity we set an average threshold of 0.43 dex (that we deem robust in Sect. 4.2.3). With Δq_{AGN} being the offset from the SF peak, the minimum AGN fraction f_{AGN} (e.g. Ceraj et al. 2018) is calculated as:

$$f_{AGN} = (1 - 10^{-\Delta q_{AGN}}) \quad (5)$$

By inputing Δq_{AGN}=0.43 dex, Eq. 5 implies that, if SFGs lie exactly on the IRRC, our radio-excess AGN have statistically at least 63% of their total radio emission arising from AGN activity.

Table 2: Comparison between fractions of radio-SFGs and radio-excess AGN below some threshold q_{thres}, for the two highest $M_\star$ bins. Four different thresholds are examined. The best trade-off between cross-contamination and completeness is given by q_{thres} = q_{peak} -2σ (green dotted line in Fig. 13), which we apply in the following analysis. See Sect. 4.2.2 for details.

$M_\star$ ($M_\odot$) bin	q_{thres}	% SF ($q<q_{thres}$)	% AGN ($q<q_{thres}$)
$10^{10.5}$–10^{11} $M_\odot$	$q_{peak}-1\sigma$	20.1%	93.6%
	$q_{peak}-2\sigma$	3.5%	69.2%
	$q_{peak}-3\sigma$	0.4%	27.3%
	$q_{cross,AGN=SF}$	2.6%	64.7%
10^{11}–10^{12} $M_\odot$	$q_{peak}-1\sigma$	23.0%	89.2%
	$q_{peak}-2\sigma$	4.5%	71.5%
	$q_{peak}-3\sigma$	0.7%	45.2%
	$q_{cross,AGN=SF}$	3.7%	70.2%

4.2.3. Re-calibrating the radio AGN threshold

According to the threshold defined above, we removed radio-excess AGN from our 12 complete z-bins at $M_\star$ $>10^{10.5}$ $M_\star$. Then we combined the remaining radio-detected SFGs with stacks of radio non-detections to compute the new $L_{1.4\ GHz}$ in those bins, which should be free from AGN contamination. We verified that the new $L_{1.4\ GHz}$ shifts the previously determined q_{TIR} (blue stars in Fig. 12) upward by a certain amount. In those complete bins, we fitted the AGN-corrected q_{TIR} with redshift, obtaining a significantly flatter relation than before: $q_{TIR}=(2.59\pm0.05)\times(1+z)^{-0.055\pm0.017}$ at $10^{10.5}$ $<M_\star/M_\odot<10^{11}$, and $q_{TIR}=(2.58\pm0.09)\times(1+z)^{-0.054\pm0.032}$ at 10^{11} $<M_\star/M_\odot<10^{12}$. This suggests that the steeper redshift trend seen before (Sect. 4.2.1) might be driven by radio AGN contamination, while the intrinsic redshift trend is significantly flatter, and possibly $M_\star$ invariant.

To test the robustness of the newly derived q_{TIR}–z trend, we again shift the q_{TIR} measurements of individual detections by the offset from such a trend at each z-bin, and perform a second q_{TIR} decomposition, as shown in Fig. 14. The Gaussian fit that parametrizes star formation is nearly unchanged, with a dispersion of 0.21-0.22 dex in the two highest $M_\star$ bins. The 2σ threshold below the peak is also very similar: 0.42 and 0.44 dex in the two bins (therefore we use an average Δq_{AGN}=0.43 dex). Moreover, the cumulative histograms (bottom panels) underline that this latter decomposition rejects about 81% of radio AGN below the threshold, as opposed to 70% estimated in the first step (see red open circles in Figs. 13 and 14), while missing a comparable 3–4% of SFGs. This confirms the effective improvement led by our re-calibration of the SF locus in removing radio AGN.

As shown in the updated Fig. 15 (at $M_\star$ $>10^{10.5}$ $M_\odot$), subtracting radio AGN (red dots) based on this latter locus shifts all the median q_{TIR} (blue stars) exactly on the fitted q_{TIR}–z trend

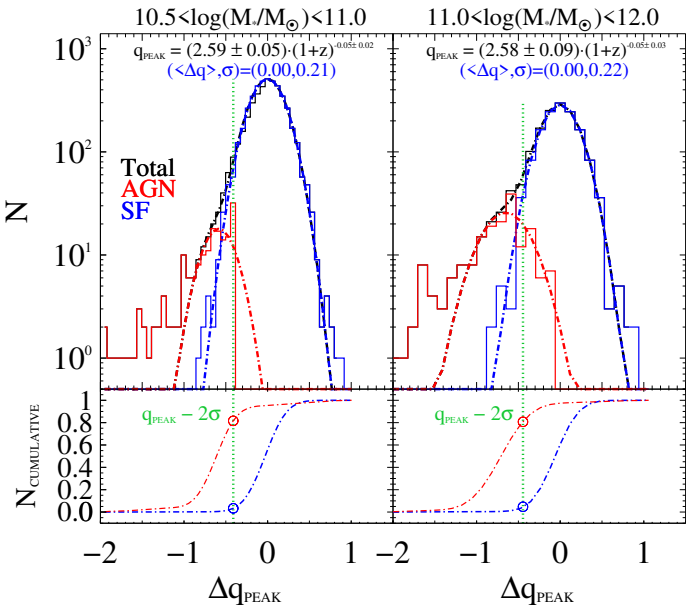

Fig. 14: Same as Fig. 13, but normalizing the peak to the flatter q_{TIR}–z trend re-calibrated after removing AGN (Sect. 4.2.3). This two-fold approach slightly improves the q_{TIR} decomposition, as highlighted by the better agreement between SF Gaussian dispersions, and by the larger cumulative fraction of radio AGN that are rejected below the q_{TIR} threshold (red open circles, 81% against the previous 70%).

(blue solid lines). This agreement suggests that no further AGN subtraction is needed in those complete bins. Therefore, we can confidently assume that the new median q_{TIR} coincide with the instrinsic peak of the SF population. Given the robustness of our analysis, we compute a weighted-average redshift slope among the two highest $M_\star$ bins, by simply weighing each slope by the inverse square of its uncertainty. This way, we obtain an average slope of -0.055±0.018, i.e. not flat at a significance of 3σ.

While at 10^{11} <$M_\star/M_\odot$ <10^{12} all z-bins (0.1<z<4.5) were used to constrain this trend, at $10^{10.5}$ <$M_\star/M_\odot$ <10^{11} we only used the first 5/7 z-bins (0.1<z<2.5). We now extrapolate the same relation towards the two highest z-bins, finding a good agreement with the median q_{TIR} estimates. For this reason, radio AGN were subtracted also in those incomplete bins, as done for the rest of radio detections at $10^{10.5}$ <$M_\star/M_\odot$ <10^{11}.

The resulting fractions of radio AGN identified in the two highest $M_\star$ bins should be quite representative of the overall incidence of radio AGN in these galaxies. This is suggested by the tightness of the SF Gaussian fit (σ~0.21–0.22 dex), that we interpret as the instrinsic scatter of the IRRC in these galaxies. Therefore, radio-undetected AGN that are not captured in our analysis, if any, are expected to be mostly composite (AGN+SF) radio sources whose total emission is predominantly arising from star formation processes.

It is worth noting that about 20% radio AGN still lie within our clean sample of SFGs, as shown in Fig. 14. As highlighted in Molnár et al. (2020), while this high-q tail of AGN is SF-dominated in the radio, it could add to the intrinsic scatter of the underlying pure SFG sample. Therefore, our inferred scatter of 0.21–0.22 dex could be slightly overestimated (see e.g. 0.16 dex in Molnár et al. 2020 for local SFGs), also due to larger uncertainties on $L_{1.4\ GHz}$ and L_{IR} than in the local Universe.

The fact that in both $M_\star$ bins the subtraction of radio-excess AGN leads to a *flattening* of the q_{TIR}–z trend might be also induced by an increasing incidence of radio AGN towards higher redshifts. If true, this would suggest that we have been combining multiple q_{TIR} distributions that are not self-similar at all redshifts, that is, the relative fractions of radio AGN and pure SFGs are not redshift invariant. As a sanity check, in both $M_\star$ bins we split and decomposed the q_{TIR} distribution of Fig. 14 separately at z<1.2 and z>1.2, examining the evolution of the relative fraction of radio AGN. As expected from the flattening of the q_{TIR}–z trend after removing them, radio AGN appear to be slightly more prevalent relative to SFGs at higher redshifts (i.e. on average from 12% at z<1.2 to 18% at z>1.2). However, we confirm that the dispersion of the SF population remains nearly unchanged (about ~0.20 dex) with redshift, both before and after removing radio AGN. This implies that the relative offset between the AGN and SF loci is redshift-invariant, therefore our sample of >2σ radio-excess AGN is globally preserved irrespective of their changing fraction with redshift. After removing those AGN, the flatter, yet declining q_{TIR} evolutionary trend could be explained by residual radio AGN activity *within* the SF locus. We estimate the overall fraction of "pure" SFGs to be 95% at z<1.2 and 90% at z>1.2. Such minimal AGN contamination is probably more important at higher redshifts because SFGs are intrinsically IR brighter, so the radio-excess contrast (at fixed $L_{1.4\ GHz}$) is less evident. Therefore, we argue that any further correction for mis-classified radio AGN would induce an even flatter q_{TIR} trend with redshift.

Finally, our approach leads to the following fractions of radio-excess AGN. At $10^{10.5}$ <$M_\star/M_\odot$ <10^{11}, radio AGN are 7.1% of all radio-detections and 2.2% of the full $M_\star$ sample of SFGs. At 10^{11} <$M_\star/M_\odot$ <10^{12}, radio AGN are 11.7% of all radio-detections and 6.0% of the full $M_\star$ sample of SFGs (see Table 3). These numbers are consistent with the known prevalence of radio AGN in the most massive galaxies (e.g. Heckman & Best 2014; Hardcastle & Croston 2020). An increasing incidence of (X-ray) AGN activity with $M_\star$ has also been reported in recent studies (Aird et al. 2019; Delvecchio et al. 2020; Carraro et al. 2020), and possibly driven by the ability of the dark matter halo mass to regulate the amount of cold gas that trickles to the central black hole (Delvecchio et al. 2019).

4.2.4. Extrapolating the SF-vs-AGN loci at low $M_\star$

We extrapolate the q_{TIR}–z trend of non-AGN galaxies calibrated in the previous Section towards less massive counterparts. As mentioned in Sect. 4.2.1, 3 GHz detections placed at $M_\star$ <$10^{10.5}$ $M_\odot$ are not representative of an $M_\star$-selected sample. In particular, a galaxy of a given $M_\star$ and redshift, with infrared luminosity L_{IR} of a typical MS galaxy would likely fall below the 3 GHz detection limit, as indicated by the green arrows in Fig. 12. Radio detections at these masses are therefore quite peculiar relative to the overall galaxy population.

This is further suggested by the discrepancy in q_{TIR} between the median measurements (blue stars) and individual radio detections (black dots). The latter lie systematically below the median q_{TIR}, deviating more and more at lower $M_\star$. For these reasons, we refrain from calibrating the IRRC directly on those radio detections. We prefer to use the median q_{TIR} values as benchmark, since they should be sensitive to a more representative sample of galaxies of that $M_\star$.

We proceed as follows. Within each $M_\star$ bin, the redshift trend of q_{TIR} is extrapolated from that calibrated at higher $M_\star$, in the form $q_{TIR} \propto (1+z)^{-0.055\pm0.018}$ (Sect. 4.2.3). Only the normalization is left free to vary, in order to best fit the median q_{TIR}. In other words, at $M_\star$ <$10^{10.5}$ $M_\odot$, we *assume* a constant q_{TIR}–z slope. This approach is preferable to leaving also the slope as

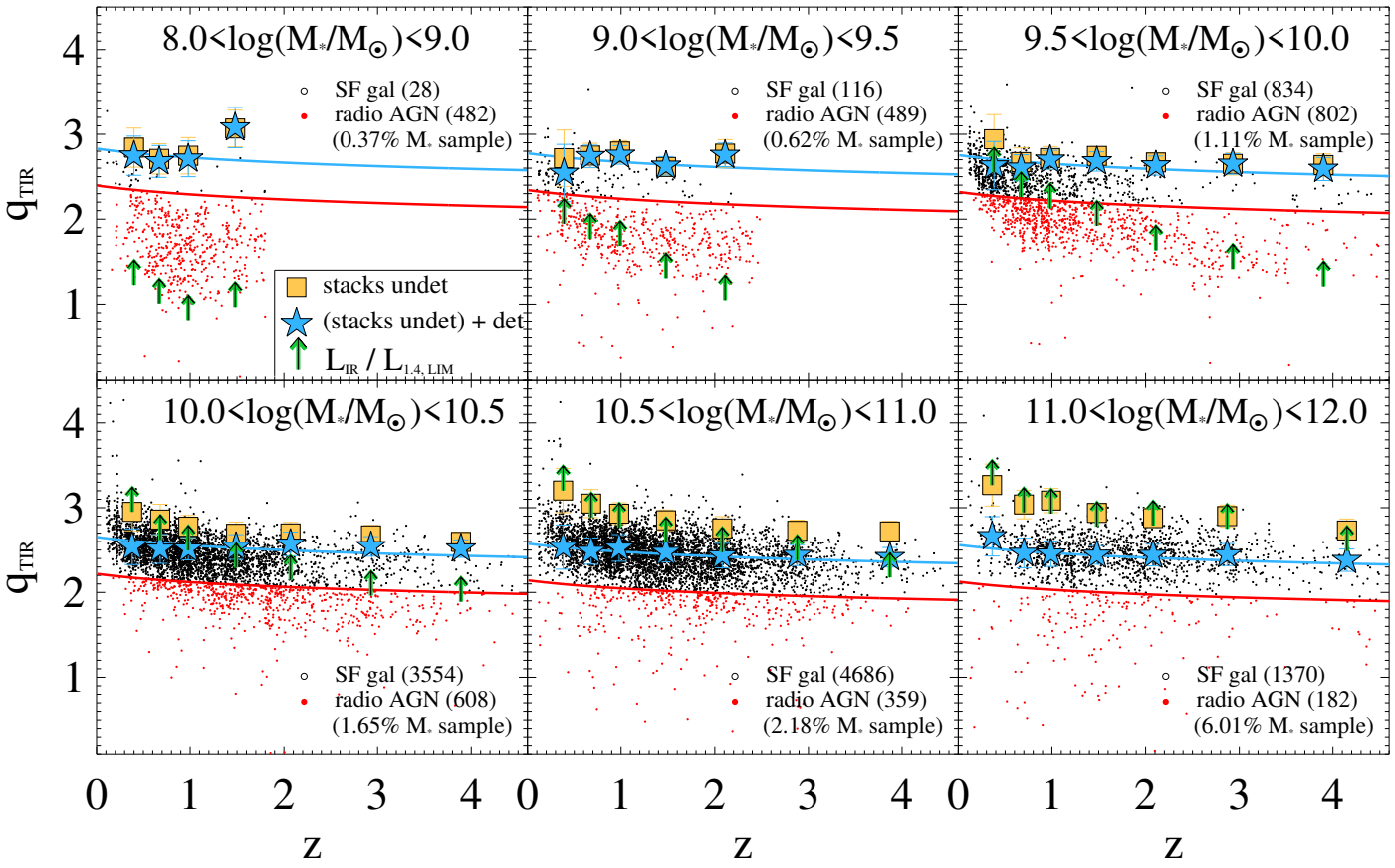

Fig. 15: Distribution of q_{TIR} as a function of redshift and $M_\star$, after removing radio AGN (red dots). Symbols are the same of Fig. 12, except for the median q_{TIR} estimates (blue stars), which are here re-calculated from the combination of stacks and the subset of radio-detected SFGs. Fractions of radio-detected AGN and SFGs are reported in each $M_\star$-bin, as well as the fraction of AGN relative to the full $M_\star$ sample analyzed in this work (in brackets). The blue and red solid lines denote the intrinsic IRRC of SFGs and the locus below which we classify radio sources as AGN (0.43 dex below the IRRC), respectively.

Table 3: Table summarizing the numbers and fractions of radio AGN and SFGs in different $M_\star$ bins, after fitting the AGN-corrected q_{TIR} with redshift (Sect. 4.2.4 and Fig. 15). Columns are sorted as follows> (1) $M_\star$ range; (2) best-fit normalization of the q_{TIR}–z trend, in the form $q_{TIR} \propto (1+z)^\gamma$, by imposing γ=-0.055±0.018 as found in the two highest $M_\star$ bins (Sect. 4.2.3); (3,4) number of identified radio AGN and radio SFGs, respectively. In brackets we report their fractions relative to the radio-detected sample, and relative to the full $M_\star$ sample; (5) Number of $M_\star$-selected SFGs analyzed in this work. $^{(*)}$: calculated over four redshift bins (0.1<z<1.8). $^{(**)}$: calculated over five redshift bins (0.1<z<2.5).

$M_\star$ ($M_\odot$) bin	q_{TIR}–z fit (normalization)	# radio AGN (% radio-det, % $M_\star$ sample)	# radio SFGs (% radio-det, % $M_\star$ sample)	# $M_\star$ sample
(1)	(2)	(3)	(4)	(5)
10^8–10^9 $M_\odot$	2.83±0.10$^{(*)}$	482 (94.5%, 0.4%)$^{(*)}$	28 (5.5%, 0.0%)$^{(*)}$	129,658$^{(*)}$
10^9–$10^{9.5}$ $M_\odot$	2.78±0.03$^{(**)}$	489 (80.8%, 0.6%)$^{(**)}$	116 (19.2%, 0.1%)$^{(**)}$	78,563$^{(**)}$
$10^{9.5}$–10^{10} $M_\odot$	2.75±0.02	802 (51.0%, 1.1%)	834 (49.0%, 1.2%)	72,122
10^{10}–$10^{10.5}$ $M_\odot$	2.65±0.03	608 (14.6%, 1.7%)	3,554 (85.4%, 9.6%)	36,838
$10^{10.5}$–10^{11} $M_\odot$	2.58±0.01	359 (7.1%, 2.2%)	4,686 (92.9%, 28.4%)	16,489
10^{11}–10^{12} $M_\odot$	2.56±0.02	182 (11.7%, 6.0%)	1,370 (88.3%, 44.8%)	3,060

a free parameter, since the small number of bins is insufficient for us to constrain the redshift trend as accurately as previously done with single detections. However, we stress that if we leave the slope free when fitting the q_{TIR} in each $M_\star$ bin, we always obtain slopes consistent between zero and -0.055, within 1σ uncertainties.

Following the iterative approach already tested at higher $M_\star$, the best-fitting trend of q_{TIR} with redshift enables us to identify radio AGN as sources lying at least 0.43 dex below the best-fit SF locus. After subtracting those radio AGN, we re-calculate the weighted-average q_{TIR} and search again for the best normalization that fits the new AGN-corrected q_{TIR} measurements with redshift. We repeat this procedure twice, i.e. until all median q_{TIR} are unchanged within the uncertainties, at each $M_\star$. This condition sets the end of our recursion.

The final, AGN-corrected q_{TIR} are shown in Fig. 15 for all $M_\star$ bins (blue stars). This plot highlights the sample of radio-detected AGN that was removed (red dots) and the final SF locus (blue solid lines) that we eventually inferred after subtracting

those AGN. The numbers of radio-detected AGN and SFGs are reported in each panel for convenience.

In most bins at $M_\star$ $<10^{10.5}$ $M_\odot$, the AGN-corrected q_{TIR} measurements nearly coincide with those obtained from stacking non-detections alone (yellow squares). These latter values delimit the highest q_{TIR} that could be reached if removing, by definition, all radio detections. The result of similarity between the two sets of q_{TIR} measurements can be attributed to a heavy subtraction of radio AGN from the sample of radio detections. The fraction of radio AGN identified at $M_\star$ $<10^{10.5}$ $M_\odot$ *increases* with *decreasing* $M_\star$, relative to the sample of radio detections. From the first to the fourth $M_\star$ bin, these fractions are as high as 94.5%, 80.8%, 51.0% and 14.6%, respectively. However, when compared to the size of our full $M_\star$ sample in each bin, they drop to (in the same order): 0.4%, 0.6%, 1.1% and 1.7%, respectively (see Table 3).

These latter numbers are consistent with a decreasing incidence of radio AGN towards lower $M_\star$ systems, following the trend constrained at $M_\star$ $>10^{10.5}$ $M_\odot$ (Sect. 4.2.3). Nevertheless, it is important to acknowledge that our analysis predicts that the vast majority of radio-detected dwarf galaxies ($M_\star$ $<10^{9.5}$ $M_\odot$, e.g. Mezcua 2017) in COSMOS are radio AGN.

Bearing this in mind, we note that the weighted average q_{TIR} (blue stars) are yet mostly driven by non-detections (yellow squares), which outnumber individual detections (circles) by a factor of >100 at $M_\star$ $<10^{9.5}$ $M_\odot$. However, when removing radio AGN, those average q_{TIR} still move upward by 0.2–0.3 dex, despite the negligible fraction of radio AGN. This is because radio detections, and in particular radio AGN, stand out from the stacks of non-detections (yellow squares) typically by over a factor of ten, up to one-hundred. Therefore, the linear weighted average computed after removing those outliers can change significantly, as we see at $M_\star$ $<10^{9.5}$ $M_\odot$.

4.3. The intrinsic IRRC evolves primarily with $M_\star$

After correcting our combined $L_{1.4\ GHz}$ measurements for radio AGN contamination, we are finally able to examine the evolution of the intrinsic IRRC as a function of $M_\star$ and redshift, as presented in Fig. 16. For each $M_\star$ bin, we show the best-fit power-law trend, whose slope was directly inferred in Sect. 4.2.3 in the two highest $M_\star$ bins (i.e. -0.055±0.018). We verified that our median L_{IR} estimates are, instead, totally unchanged after removing radio-excess AGN, as expected given their minimal fraction relative to the parent $M_\star$-selected SFG sample.

The colour bar highlights a clear stratification of q_{TIR} with $M_\star$, with more massive galaxies showing systematically lower q_{TIR} values. This behaviour was already seen in Fig. 11 before removing radio AGN, but here it suggests that some additional mechanisms unrelated to AGN activity might be boosting (reducing) radio emission in more (less) massive systems, relative to the IR.

For comparison, other IRRC trends with redshift are reported from Bell (2003, dotted line), Magnelli et al. (2015, dashed line) and Delhaize et al. (2017, dot-dashed line). Since Delhaize et al. (2017) did not remove radio-excess AGN, we also show their AGN-corrected relation by removing 2σ outliers (as reported in Delvecchio et al. 2018): $q_{TIR} = 2.80\pm0.02\times(1+z)^{-0.12\pm0.01}$ (triple dot-dashed line). This trend is flatter than the previous one, more consistent with that of Magnelli et al. (2015) and more appropriate for a comparison with our approach.

In the following, we examine the significance of the $M_\star$ dependence at fixed redshift, and we provide a multi-parametric fit as a function of both parameters.

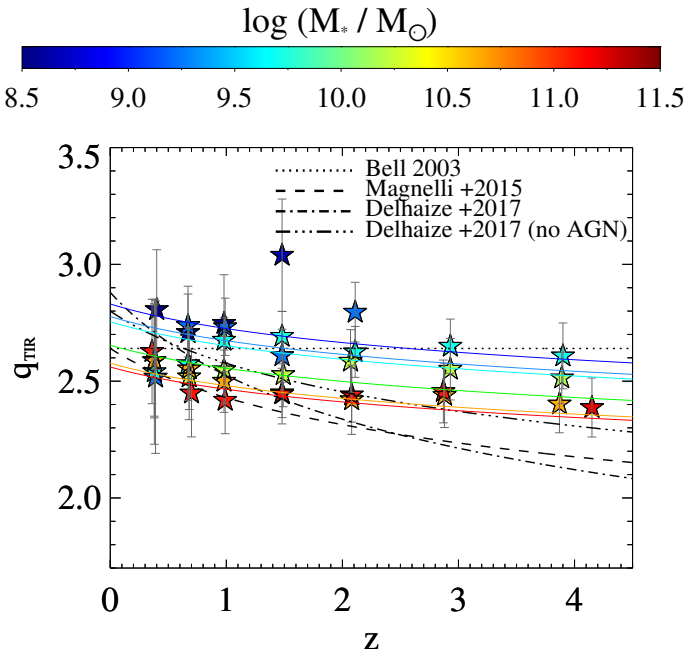

Fig. 16: Intrinsic (i.e. AGN-corrected) q_{TIR} evolution as a function of redshift (x-axis) and $M_\star$ (colour bar). The L_{IR} estimates are the same reported in Fig. 11, while $L_{1.4\ GHz}$ measurements have been re-calculated after excluding radio-detected AGN (Sect. 4.2). For comparison, other IRRC trends with redshift are taken from the literature (black lines): Bell (2003, dotted); Magnelli et al. (2015, dashed); Delhaize et al. (2017, dot-dashed) and their AGN-corrected version after removing 2σ outliers (triple dot-dashed lines).

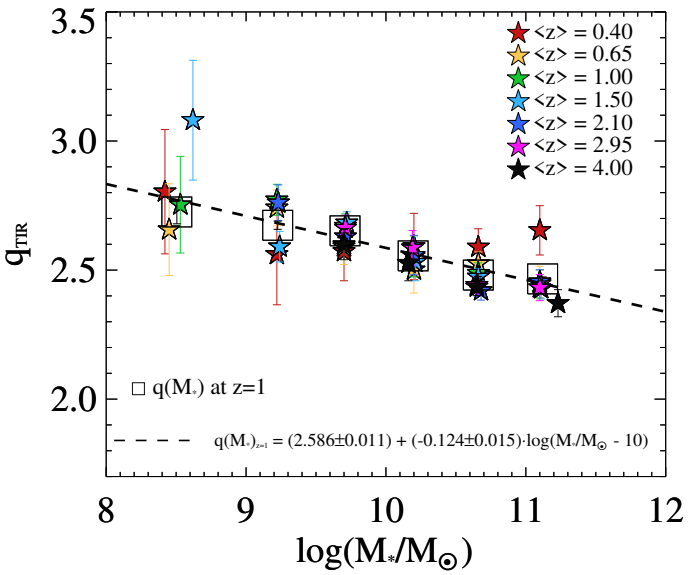

Fig. 17: Distribution of AGN-corrected q_{TIR} as a function $M_\star$, colour-coded by redshift (stars). At each $M_\star$, open squares indicate the median q_{TIR} values at z=1, obtained after propagating the uncertainties of slope and normalization of the corresponding q_{TIR}–z fit and interpolating each at z=1. These values were fitted with a linear function in log–log space (black dashed line).

Fig. 17 shows the equivalent of Fig. 16 but projected on $M_\star$, with redshift bins in different colours. Error bars on each q_{TIR} were re-scaled by a factor of $\sqrt{\chi^2_{red}}$ in each $M_\star$ bin, to bring the reduced χ^2 of the corresponding q_{TIR}–z fit to unity. It is quite evident that an $M_\star$ dependence reduces substantially the scatter

of the average q_{TIR} around a single trend. To better quantify this, first we bootstrapped over the uncertainties of slope and normalization obtained from each q_{TIR}–z trend (see Table 3). Then, at each $M_\star$, we interpolated the full range of bootstrapped IRRCs at z=1, in correspondence of the 16^{th}, 50^{th} and 84^{th} percentiles. Interpolating at z=1, besides being at roughly the median redshift of our sample, reduces the increasing divergence of each q_{TIR}–z fit at lower or higher redshifts. This leaves us with the interpolated median q_{TIR}(z=1) as a function of $M_\star$ (black open squares). Error bars indicate the uncertainty on the median value. The black dashed line marks the corresponding linear best-fit trend: $q(M_\star)_{z=1}$=(2.586±0.011) + (-0.124±0.015)×log($M_\star/M_\odot$-10). This function yields a χ^2_{red}=0.87, with an $M_\star$ slope close to that commonly found when fitting q_{TIR} as a function of redshift (e.g. Magnelli et al. 2015), and significant at over 8σ. Though the interpolated fit at z=1 is purely indicative, this check suggests that $M_\star$ might be the primary driver of the evolution of the IRRC across redshift.

Moreover, in order to incorporate the dependence of the IRRC on both $M_\star$ and redshift *simultaneously*, we performed a multi-parametric fit in the 3-dimensional q_{TIR}–$M_\star$–z space. This yields the following best-fit expression:

$$q_{TIR}(M_\star, z) = (2.646\pm0.024)\times A^{(-0.023\pm0.008)} - B\times(0.148\pm0.013) \quad (6)$$

where A=(1+z) and B=(log $M_\star/M_\odot$ − 10). The corresponding χ^2_{red}=0.90. The best-fit slopes with redshift and $M_\star$ are significant at 2.9σ and 11σ levels, respectively. This further strengthens the need for a primary $M_\star$ dependence, followed by a weaker and less significant redshift dependence. These numbers and confidence levels refer to the median trend. However, we acknowledge that, if assuming a constant IRRC scatter of 0.21–0.22 dex across all $M_\star$ galaxies, the weak co-dependence on redshift could be easily washed out. This dilution might also hide a mildly increasing redshift trend, which could be expected by Inverse Compton cooling of cosmic ray electrons (Murphy 2009). Nevertheless, the main argument of our analysis is to demonstrate how previously reported *best-fitting* IRRC trends with redshift are likely a red herring, whereas the $M_\star$ (or a related proxy) is a better predictor of the average IRRC in SFGs.

A similar conclusion was recently reached by Molnár et al. (2020), when considering the dependence of the IRRC on galaxy radio luminosity. To mitigate selection effects, they exploited a depth-matched sample of SFGs at z<0.2. After performing a radio decomposition analysis in different bins of $L_{1.4\ GHz}$, Molnár et al. (2020) report that q_{TIR} decreases with increasing $L_{1.4\ GHz}$. Assuming that radio emission comes predominantly from star formation, this is in line with our inferred $M_\star$ dependence, since more massive SFGs are also brighter in radio (Leslie et al. 2020). This further corroborates the idea that the IRRC varies across different types of galaxies, at fixed redshift (but see e.g. Pannella et al. 2015 for an alternative interpretation). Therefore, we conclude that our results are in qualitative agreement with Molnár et al. (2020), who also demonstrate the implications of such a non-linearity for decreasing q_{TIR} vs. z trends in the literature.

5. Discussion

The main result of this work is the finding that the IRRC primarily evolves with $M_\star$, and only weakly with redshift (Eq. 6). While the $M_\star$ dependence has not been explored in detail so far, except in the local Universe (e.g. Gürkan et al. 2018, see Sect. 5.1), for several years much effort has been devoted in understanding the mild, but significant decline of the IRRC with redshift from both an observational (e.g. Ivison et al. 2010a; Sargent et al. 2010; Magnelli et al. 2015; Delhaize et al. 2017; Calistro Rivera et al. 2017; Molnár et al. 2018) and a theoretical (Lacki & Thompson 2010; Schleicher & Beck 2013; Schober et al. 2016; Bonaldi et al. 2019) perspective. In Appendix C we expand on the role played by various assumptions in deriving different IRRC trends presented in the literature. In this Section, instead, we interpret our results and discuss the many implications of our findings in the context of the origin and evolution of the IRRC. In particular, we split the discussion in several sections, each focusing on a specific issue. First, we explore some physical interpretations of the origin of an $M_\star$ and redshift-dependent IRRC (Sect. 5.1). We further investigate the possible evolution of the IRRC above the MS (Sect. 5.2). A discussion on the reliability and incidence of AGN activity is also presented (Sect. 5.3 and 5.4). Finally, we comment on the use of radio emission as a SFR tracer in the light of our results (Sect. 5.5).

5.1. What drives the primary $M_\star$ dependence?

Our main finding is that the IRRC decreases primarily with $M_\star$, and only weakly with redshift. In particular, within the range 10^9 <$M_\star/M_\odot$ <10^{12}, the median q_{TIR} decreases by 0.25 dex (a factor of 1.8), at fixed redshift, and with high significance (∼10σ, see Eq. 6). To quantify the corresponding IR-radio slope, we take our best q_{TIR}–$M_\star$ relation (Eq. 6) at fixed redshift, and assume for simplicity a linear MS between $M_\star$ and SFR (i.e. L_{IR}). This yields $L_{IR} \propto L^{0.90}_{1.4\ GHz}$. In the past years, the deviation from a linear trend has been gaining increasing momentum, due to several studies finding a similar sub-linear behaviour in the local Universe ($L_{IR} \propto L^{0.75-0.90}_{1.4\ GHz}$, Bell 2003; Hodge et al. 2008; Davies et al. 2017; Brown et al. 2017; Gürkan et al. 2018; Molnár et al. 2020). This might challenge the idea of calibrating radio emission as a universal SFR tracer, as we discuss later in Sect. 5.5.

Here we explore some physical parameters behind this non-linearity, that might induce an $M_\star$-evolving q_{TIR} similar to our findings. To test some radio synchrotron models (e.g. Lacki & Thompson 2010), we study the relation between q_{TIR} and SFR surface density, and then we also discuss the possible role of a top-heavy IMF in driving the IRRC.

5.1.1. The role of the SFR surface density (Σ_{SFR})

The model proposed by Schleicher & Beck (2013) postulates that the magnetic field strength scales with SFR, boosting radio synchrotron emission during shocks or galaxy interactions (e.g. Donevski & Prodanović 2015; Tabatabaei et al. 2017). Because of the MS relation, this model predicts a net enhancement of radio emission with increasing $M_\star$. Related to this, the semi-analytical model of Schober et al. (2016, 2017) also predicts enhanced radio synchrotron emission in more massive systems, due to increasing gas density and cosmic ray scale height. Similarly, the non-calorimetric model (Lacki & Thompson 2010) suggests that increasing Σ_{SFR} might boost radio emission relative to the IR, due to larger cosmic ray scale height, generating a lower q_{TIR} in more compact SF regions. It is known that SB galaxies above the MS exhibit systematically more compact SF than $M_\star$- and z-matched MS analogues (e.g. Elbaz et al. 2011). Later, we will explore whether q_{TIR} evolves above the MS relation (Sect. 5.2). Here we focus on our sample of typical MS galaxies, and test the above models by relating q_{TIR} and average

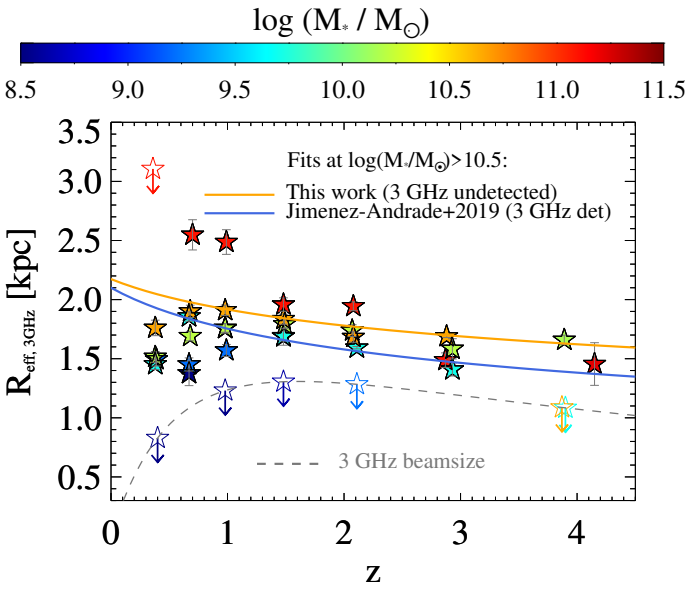

Fig. 18: Distribution of 3 GHz effective radius (in kpc) as a function of redshift and colour-coded by $M_\star$ (stars). Size measurements are taken from median stacked 3 GHz images of non-detections. Upper limits are given for unresolved stacks and correspond to the angular 3 GHz beamsize (0.75", grey dashed line), except for the highest $M_\star$ bin at z<0.5 that was convolved with a Gaussian kernel of 3" FWHM. We observe a clear increase of R_e with galaxy $M_\star$. The bins at $M_\star > 10^{10.5}$ $M_\odot$ with resolved emission are fitted with a power-law redshift trend, which yields $R_e \propto (1+z)^{-0.18\pm0.07}$ (orange solid line). A comparison study by Jiménez-Andrade et al. (2019) is shown (blue solid line) for 3 GHz detected SFGs at similar $M_\star$ in COSMOS, obtaining $R_e \propto (1+z)^{-0.26\pm0.08}$.

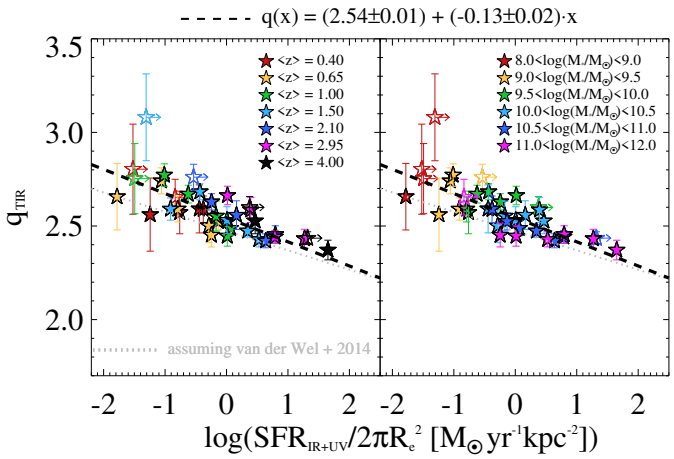

Fig. 19: Evolution of q_{TIR} as a function of SFR surface density (Σ_{SFR}=$SFR_{IR+UV}/2\pi R_e^2$), colour coded by redshift (left panel) and by $M_\star$ (right panel). The SFR_{IR+UV} estimates are taken from Sect. 3.2, while the effective radius R_e is measured from stacked 3 GHz images via 2D Gaussian fitting. This plot shows a significant anti-correlation similar in slope to that observed with $M_\star$ (Sect. 4.3), marked by the black dashed line ($q_{TIR} \propto \Sigma_{SFR}^{-0.13\pm0.02}$). For comparison, the best-fit trend with rest-frame optical (5000Å) sizes estimated from van der Wel et al. (2014) scaling relation is shown (grey dotted line).

Σ_{SFR} estimates derived in this work. These latter were obtained by using the total SFR_{IR+UV} calculated from IR stacking and adding the dust-uncorrected UV contribution (Sect. 3.2). Galaxy sizes are drawn from median radio stacking of non-detections, carried out in Sect. 3.3 at each $M_\star$–z bin via 2D elliptical Gaussian fitting. Though these measurements do not include the contribution of single 3 GHz detections, they come from about 97% of all $M_\star$-selected galaxies in our sample, hence they should be statistically representative of their average radio properties. This approach implicitly assumes that radio emission encloses the total star formation of the host, that is quite plausible especially in high-$M_\star$ galaxies, where the dominant obscured SF traced by IR is also seen in the radio (e.g. Jiménez-Andrade et al. 2019). To scale angular sizes θ_{FWHM} into effective radius (R_e, enclosing half of the total flux density), we assume that our galaxies follow a disk-like surface brightness profile (Sérsic index n=1), as found for MS galaxies (e.g., Nelson et al. 2016). Under this assumption, the major-axis $R_{e,maj}$ can be calculated as $R_{e,maj} = \theta_{FWHM}$ / 2.43 (Murphy et al. 2017). Lastly, we take the circularized radius $R_e = R_{e,maj} / \sqrt{A_r}$, where A_r is the axial ratio.

Fig. 18 displays our median stacked 3 GHz size measurements (or upper limits) as a function of redshift and $M_\star$. Error bars are obtained from the IDL routine MPFIT2DFUN. Upper limits are shown for unresolved stacks and correspond to the angular 3 GHz beamsize (0.75", grey dashed line), except for the highest $M_\star$ bin at z<0.5 that was convolved with a Gaussian kernel of 3" FWHM (see Sect. 3.3.1). Our 3 GHz sizes are about 1.5–2.5 kpc and slightly increase with $M_\star$, qualitatively similar to the behaviour of optical sizes (e.g. van der Wel et al. 2014, though a factor of two smaller). Our R_e measurements are well consistent with VLA 3 GHz sizes indipendently derived in the recent study of Jiménez-Andrade et al. (2019). The authors used the same VLA 3 GHz COSMOS images to construct a $M_\star$-complete sample of 3,184 radio-detected SFGs with $M_\star > 10^{10.5}$ $M_\odot$, most of which lie around the MS relation (Schreiber et al. 2015). The best-fitting R_e trend with redshift reported by Jiménez-Andrade et al. (2019) for MS galaxies (blue solid line, $R_e \propto (1+z)^{-0.26\pm0.08}$) is broadly consistent with our evolutionary trend based on median 3 GHz stacks (orange solid line, $R_e \propto (1+z)^{-0.18\pm0.07}$). Our slightly larger size measurements are likely due to radio-detected SFGs (Jiménez-Andrade et al. 2019) having a more centrally peaked surface brightness compared to our stacks (Bondi et al. 2018).

We calculate $\Sigma_{SFR} = SFR_{IR+UV}/2\pi R_e^2$ (see e.g. Jiménez-Andrade et al. 2019) and show its relation with q_{TIR} in Fig. 19, colour-coded by redshift (left panel) and $M_\star$ (right panel). Empty symbols highlight 7/37 bins with unresolved 3 GHz stacked emission, which translates into a lower limit in Σ_{SFR}. By fitting only the 30 q_{TIR} and Σ_{SFR} measurements, we obtain a significant anti-correlation similar in slope to that observed with $M_\star$ (Sect. 4.3), marked by the black dashed line ($q_{TIR} \propto \Sigma_{SFR}^{-0.13\pm0.02}$). For comparison, we also show the best-fit trend with rest-frame optical (5000Å) sizes estimated indirectly from the van der Wel et al. (2014) scaling relation for SFGs, at each $M_\star$ and z (grey dotted line). Though this latter trend is a surrogate of the trend with $M_\star$, the consistent slope of (-0.10±0.02) suggests that q_{TIR} significantly decreases in more compact SFGs, irrespective of using optical or radio size estimates. We note that the mild difference between the two latter trends originates from the 2× smaller radio sizes compared to rest-frame optical sizes (Bondi et al. 2018).

Since more massive galaxies are characterized by more compact star formation (Elbaz et al. 2011), the decreasing q_{TIR}–Σ_{SFR} trend is linked to that with $M_\star$. Nevertheless, unlike the trend with optical sizes, our Σ_{SFR} measurements are not bound to $M_\star$ *by construction*, but rather measured from independent tracers (IR+UV and 3 GHz data). We thus stress that our proposed q_{TIR}–Σ_{SFR} dependence is meant to be a proxy for the observed $M_\star$ de-

pendence, simply interpreted on more physical grounds. At fixed $M_\star$, the SFR surface density increases with redshift (left panel of Fig. 19), in qualitative agreement with our (weakly) decreasing q_{TIR} trend.

Both the slope and significance of the q_{TIR}–Σ_{SFR} relation are consistent to those found between q_{TIR} and $M_\star$ (Sect. 4.3). We argue that the declining q_{TIR}–Σ_{SFR} slope is primarily driven by the SFR, and only weakly by radio sizes. Indeed, at fixed redshift, the SFR(IR+UV) increases along the MS by a factor >30 from 10^9 to 10^{11} $M_\odot$ (Fig. 8), while R_e^2 only increases by a factor of 1.5–2.5 in the same interval. Though this is not a conclusive evidence, our analysis seems to suggest that the larger SFR per unit area in more massive (and higher-z) galaxies might be driving the sub-linear behaviour of the IRRC. This finding is in qualitative agreement with the predictions of the non-calorimetric model (Lacki & Thompson 2010), and also consistent with the low-q values recently inferred by Algera et al. (2020a) in compact (R_e ~1 kpc) and massive ($M_\star$ >10^{10} $M_\odot$) sub-millimetre galaxies at 1.5<z<3.5. Indeed, their average q_{TIR}=2.20±0.03 lies close to the extrapolation of our best-fit q_{TIR}–Σ_{SFR} trend at Σ_{SFR} ~100 $M_\odot$ yr^{-1} kpc^{-2}, thus further corroborating the relation between q_{TIR} and SFR per unit area in SFGs.

5.1.2. The role of the IMF

We quantify whether a deviation from a canonical IMF slope (e.g. Chabrier 2003; n(M)∝$M^{-2.35}$ at 0.8<M<100 $M_\odot$) could justify an $M_\star$-decreasing q_{TIR}. In particular, we note that reprocessed IR light comes predominantly from stars with M>5$M_\odot$, while radio synchrotron emission comes from more massive stars with M>8$M_\odot$. We check whether a systematically flatter IMF in more massive galaxies could explain the observed decreasing q_{TIR}. A top-heavy IMF has been directly constrained only in massive early-type galaxies at z~0 (Cappellari et al. 2012) from the comparison between dynamical masses and optical light, but only proposed or indirectly inferred otherwise (e.g. Baugh et al. 2005; Hopkins & Beacom 2006; Davé 2008; van Dokkum 2008; Dabringhausen et al. 2009). To quantify the change of q_{TIR} as a function of IMF slope, we integrate the IMF over the ranges 5-100 $M_\odot$ and 8-100 $M_\odot$, with varying IMF slope. The ratio between the two integrals is somewhat proportional to $L_{IR}/L_{1.4\ GHz}$. However, we find only 8% variation of the integral ratio across the full range of slopes [-2.35, 0] , as compared to 80% (i.e. 0.25 dex) q_{TIR} variation across all $M_\star$. Therefore, a top-heavy IMF in the most massive galaxies proves insufficient to explain the evolving q_{TIR} with $M_\star$.

Our checks cannot firmly elucidate the main physical driver of the IRRC with $M_\star$, but they seem to support a link between q_{TIR} and compactness of star formation. More massive and more distant galaxies display higher Σ_{SFR}, which corresponds to a larger cosmic-ray scale height, according to the non-calorimetric model (Lacki & Thompson 2010). This model predicts a slight (0.3 dex) decrease of q_{TIR} with redshift. When binning q_{TIR} as a function of both $M_\star$ and redshift, we find a similar q_{TIR} drop (0.25 dex) across the $M_\star$ range of 10^9 <$M_\star/M_\odot$ <10^{12}. This underpins Σ_{SFR} as a good predictor of the average q_{TIR} in SFGs. We acknowledge that also other mechanisms, or a combination of them, could also be at play.

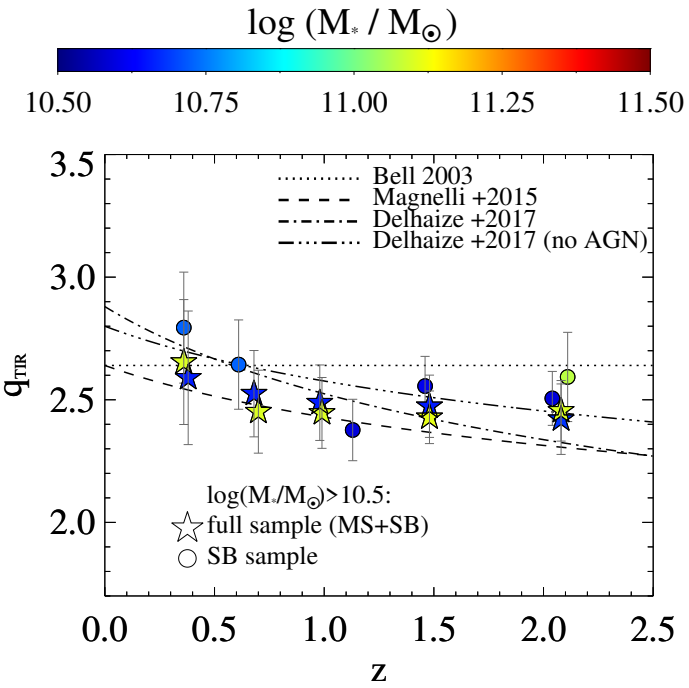

Fig. 20: Comparison of q_{TIR} between our full SFG sample (MS+SB, stars) and the SB subsample (circles), as a function of redshift. To mitigate the incompleteness of an IR-based selection of SBs, we only show bins with $M_\star$ >10^{10} $M_\odot$. Black lines highlight best-fit IRRC trends from the literature for comparison. This test suggests that q_{TIR} evolves irrespective of whether a galaxy is on or above the MS.

5.2. Does the IRRC evolve above the MS?

We investigate the behaviour of the average q_{TIR} above the MS. This is important to test whether radio emission follows a similar enhancement as L_{IR} when moving to the starburst region, or instead q_{TIR} is not a good tracer of starburstiness (i.e. offset from the MS). This issue is still highly debated. On the one hand, Condon et al. (1991) found that the most extreme ULIRGs at z~0 have higher q_{TIR} and larger scatter compared to the MS population. On the other hand, Helou et al. (1985) and Yun et al. (2001) do not report any significant deviation of q_{TIR} in local SB galaxies, though they also observed a larger scatter for this population. More recently, Magnelli et al. (2015) found a mild (+0.2 dex) enhancement of q_{FIR} in SBs relative to MS galaxies, though not significant. Such apparent tension is probably also due to different definitions of "starburst" galaxies and different sample selections.

Here we define SBs as galaxies with SFR>4×SFR_{MS} (e.g. Rodighiero et al. 2011), where SFR_{MS} corresponds to the SFR predicted by the MS (Schreiber et al. 2015), at each $M_\star$ and redshift. Our measured SFR estimates come from IR+UV, as described in Sect. 3.2. However, following Carraro et al. (2020), we select as SBs only *individually* IR-detected galaxies (S/N_{IR} >3) that meet the above criterion. This is because our stacked SFR_{IR} estimates are mostly dominated by MS galaxies, while the SB subsample is likely washed out in all median stacks. Especially at low $M_\star$ and high-redshift, this approach yields an incomplete SB sample due to galaxies being IR fainter. In order to mitigate possible selection biases, we only focus on SB galaxies with $M_\star$ >$10^{10.5}$ $M_\odot$ and z≲2.5. This interval is set to ensure that all SB galaxies (i.e. lying >4× above the MS) stand above the limiting L_{IR} of *Herschel* PACS+SPIRE data in COSMOS (Béthermin et al. 2015), and thus are IR detected. We further remove radio-

excess AGN (pre-identified in Sect. 4.2) from the SB subsample of radio detections, in order to consider only bona-fide SFGs and fairly compare the AGN-corrected q_{TIR} between the SB and MS populations. This leaves us with a sample of 554 SBs. As done for the full SFG sample, we performed median stacking at 3 GHz and combined the stacked signal with radio-detected SBs.

Fig. 20 shows the resulting q_{TIR} of SBs (circles) relative to the full SFG sample (MS+SB, stars) out to $z \lesssim 2.5$, at $M_\star > 10^{10.5}$ $M_\odot$. For comparison, some previous IRRC trends are reported (black lines), as in Fig. 16. While some possible hints of (~0.05 dex) higher q_{TIR} in SBs could be present, these are consistent with MS analogues within 1σ in all bins. Therefore, this test suggests that q_{TIR} evolves primarily with $M_\star$, irrespective of whether a galaxy is on or above the MS.

Though the (lack of) evolution of q_{TIR} above the MS is still debated, our decreasing q_{TIR}–Σ_{SFR} trend (Fig. 19) would predict lower q_{TIR} in SB than in MS galaxies, due to SBs being more compact. However, we note that our IR-detected SBs are *both* >4× more star forming and smaller in size ($R_e \lesssim 1$ kpc at $z<2$, see Jiménez-Andrade et al. 2019) than MS analogues. Therefore, both parameters add to boost Σ_{SFR}.

On a side note, the sample of sub-millimetre galaxies for which Algera et al. (2020a) obtained an average q_{TIR}=2.20 includes SFGs within a factor of three from the MS relation, thus not formally SBs. It might be possible that SB galaxies follow a different regime of q_{TIR}, while our results predominantly reflect the behaviour of the MS population. For instance, Lacki & Thompson (2010) distinguish SBs depending on the disk scale height h, between "compact SBs" (h=100 pc) similar to local merging galaxies, and "puffy SBs" (h=1 kpc) that are more common at high-z (Genzel et al. 2008). Since the latter are characterized by more efficient Inverse Compton losses than the former, puffy SBs lie generally at higher q_{TIR} than compact SBs, thus closer to the MS population (Lacki & Thompson 2010). It is possible that our non-local SB sample resembles that class of objects, which might explain the similarity in q_{TIR} with MS analogues (Fig. 20). Therefore, we caution that a simple dependence of q_{TIR} on the SF compactness might not be suitable for unveiling the physics behind the IRRC in SBs, which might be also connected to the geometry of the SF regions.

5.3. Reliability of radio AGN identification

In Sect. 4.2 we carried out an $M_\star$ and redshift dependent radio decomposition to separate radio SFGs from radio-excess AGN. Our empirical threshold (Eq. 5) identifies as radio-excess AGN sources with at least 63% of the total radio emission arising from AGN activity. Therefore, radio sources with lower, yet substantial AGN contribution could still be mis-classified as radio-SFGs (e.g. White et al. 2015, Wong et al. 2016; White et al. 2017). We attempt at quantifying this fraction by comparing our classification against ancillary VLBA data in the COSMOS field (Herrera Ruiz et al. 2017, 2018). This excellent dataset contains 468 VLBA sources detected at $>5\sigma$, targeted from a pre-selected sample of VLA-COSMOS 1.4 GHz sources at $S/N_{1.4} > 5.5$ (Schinnerer et al. 2010, 2,864 sources). Since the brightness temperature reached by VLBA observations at about 0.01" resolution exceeds 10^6 K, detections are most likely to be radio AGN (Herrera Ruiz et al. 2017). Therefore, this sample provides an unambiguous method to test our source classification, though for a very tiny fraction of our sample with 1.4 GHz flux $S_{1.4} > 55$ μJy, typically hosted in massive galaxies ($M_\star > 10^{10}$ $M_\odot$). Out of 13,510 3 GHz radio detections among our 37 bins, we found only 189 VLBA counterparts within 0.5" search radius. A fraction as high as 90% (170/189) were identified as "radio-excess AGN" based on our recursive approach. The remaining 10% AGN mis-classified as SFGs from our approach are all IR-detected sources with typically high SFRs, which clearly reduces the apparent contrast between AGN- and SF-driven radio emission at arcsec scales. Although limited to a relatively bright and highly incomplete subsample, the comparison with VLBA data further demonstrates the reliability of our radio AGN identification method.

5.4. Is there widespread AGN activity in radio-detected dwarves?

A noteworthy implication raised from our radio AGN subtraction is the possibly widespread AGN activity within radio-detected dwarf galaxies ($M_\star < 10^{9.5}$ $M_\odot$). As highlighted in Sect. 4.2.4 and Table 3, about 90% of radio-detected dwarves are classified as radio AGN. This fraction drops down to only ~0.5% relative to the full $M_\star$ sample of dwarves. Such huge difference suggests that radio-detected dwarves are a quite peculiar and not representative subsample of these low-$M_\star$ galaxies.

From an IR perspective, nearly all radio-detected dwarves (>99%) are completely undetected ($S/N_{IR} < 3$) at any IR/sub-mm band (Fig. 3). This is likely a natural effect due to the increasing incompleteness of IR selection towards low $M_\star$ galaxies. From IR/sub-mm stacking, however, we obtain SFR_{IR} >4× higher than the MS relation, placing these sources in the SB region (e.g. Rodighiero et al. 2011; Sargent et al. 2012). This might apparently support a SF-driven origin of radio emission in dwarves.

Nevertheless, on the radio side, these sources display on average lower $L_{1.4\ GHz}$ values than more massive counterparts, but still 100× larger than those obtained from median radio stacking of non-detections. This effect fully counter-balances the high starburstiness seen in the IR, causing an overall drop of q_{TIR} in radio-detected dwarves by over a factor of 10, with respect to the stacked population (see black dots relative to yellow squares in Figs 12 and 15). These arguments let us suppose that radio-detected dwarves are consistent with being AGN-dominated in the radio.

While there is broad consensus on the prevalence of radio AGN within massive galaxies (e.g. Heckman & Best 2014), in which AGN-driven feedback could hamper star formation, little is known about its incidence and impact in dwarves. These systems are thought to host the pristine relics of the first black hole seeds, whose growth has been long believed to be disfavoured by SNa-driven feedback (e.g. Reines et al. 2013; Dubois et al. 2015; Mezcua et al. 2016; Marleau et al. 2017). However, there is mounting evidence that AGN feedback may also play a role at the low-mass end of the galaxy population.

From a theoretical perspective, cosmological simulations find that starbursting dwarf galaxies triggered by major mergers can be very frequent (Fakhouri et al. 2010; Deason et al. 2014). These events can induce widespread AGN feedback at low-$M_\star$ regimes, that could help solve the so-called "too-big-to-fail" problem, whereby simulated dwarves outnumber by several factors their observed counterparts (Garrison-Kimmel et al. 2013; Kaviraj et al. 2017). This excess number cannot be suppressed via SNa feedback alone, but through additional AGN feedback (Keller et al. 2016; Silk 2017; Koudmani et al. 2020).

To search for observational AGN signatures in dwarf galaxies, spatially-resolved emission line diagnostics (Mezcua & Domínguez Sánchez 2020), deep X-ray and high angular resolution radio observations have been used (e.g. Reines et al. 2011,

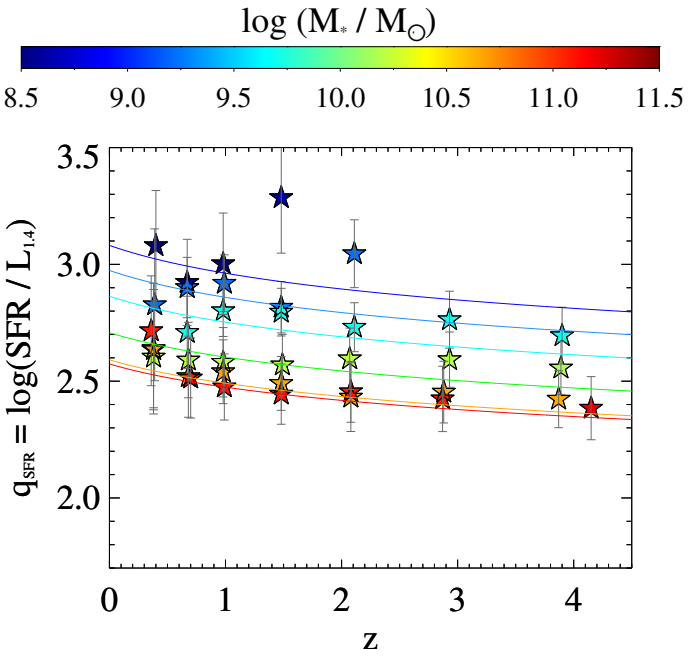

Fig. 21: Distribution of $q_{SFR} \propto \log(SFR/L_{1.4\ GHz})$ as a function of redshift and $M_\star$, similarly to what shown in Fig. 16. An equivalent analysis was carried out to separate SFGs and radio-excess AGN within radio detections, and remove AGN contaminants to derive the present q_{SFR} measurements. While the redshift dependence is roughly unchanged from the q_{TIR} space, the increasing UV contribution at low-$M_\star$ amplifies and steepens the final $M_\star$ dependence. See Sect. 5.5 for details.

Reines & Deller 2012; Reines et al. 2014; Mezcua et al. 2019). In the local Universe, these campaigns led to the confirmation of on-going AGN activity in starbursting dwarf galaxies (Reines & Deller 2012). At higher redshifts, Mezcua et al. 2019 performed a statistical study of radio-detected dwarf galaxies at z<3.4 using deep VLA-COSMOS 3 GHz data (Smolčić et al. 2017b). They isolated a sample of 35 bona-fide dwarf galaxies, which displayed radio jets powers and efficiencies as high as those of more massive galaxies. These studies argue that AGN feedback may be more common than previously thought, and potentially impactful for regulating galaxy star formation (Kaviraj et al. 2019). Our findings that most radio-detected dwarves stand above the MS and display excess radio emission are therefore not surprising, and in broad agreement with the above literature.

As an independent check, we investigated the possible AGN nature of radio-detected dwarves by stacking deep *Chandra* images from the COSMOS-Legacy survey (Civano et al. 2016; Marchesi et al. 2016) with the online tool CSTACK[5]. We stacked our combined radio-AGN sample identified at $M_\star < 10^{9.5}$ $M_\odot$ in different redshift bins, finding no X-ray detection in any of them. We then converted the 3σ upper limit fluxes to 2-10 keV (rest-frame) X-ray luminosity, by assuming a photon index Γ=1.4 (e.g. Gilli et al. 2007) and compared those L_X limits with the level of X-ray emission predicted by star formation, using the most recent prescription of Lehmer et al. (2016). We find that our L_X limits are still 5× higher than the X-ray emission expected at the average $M_\star$, redshift and SFR of our sources, which does not rule out their AGN nature.

[5] Developed by T. Miyaji, it is accessible at http://lambic.astrosen.unam.mx/cstack/

5.5. Is radio emission a good SFR tracer in all galaxies?

In this Section we discuss the link between the IRRC and SFR in galaxies. As mentioned in Sect. 3.2, the conversion from L_{IR} to SFR is quite accurate in massive galaxies, while towards less massive and less obscured systems the UV may contribute as much as the IR to the global SFR. The observed correlation between L_{IR} and $L_{1.4\ GHz}$ is therefore not rigidly proportional to SFR.

For this reason, we express q_{TIR} through a slightly different formalism that accounts for the addition of dust-uncorrected UV emission, in order to study the connection between radio emission and total SFR (=SFR_{IR+UV}). We thus define the parameter q_{SFR} as:

$$q_{SFR} = \log\left(\frac{L_{SFR}\ [W]}{3.75\times10^{12}\ Hz}\right) - \log(L_{1.4\ GHz}\ [WHz^{-1}]) \quad (7)$$

where L_{SFR} is simply the SFR_{IR+UV} scaled back to luminosity units [W], by multiplying it by a factor 10^3. This incorporates the L_{IR}–SFR conversion factor (×10^{10}, Kennicutt 1998) and the conversion from erg s^{-1} to W (×10^{-7}). This formalism enables us to keep similar units as for q_{TIR}, while switching from luminosity to total SFR.

We repeated the analogous q_{SFR} decomposition analysis at $M_\star > 10^{10.5}$ $M_\odot$ to calibrate the AGN-vs-SF locus of radio detections (Sect. 4.2). Within the two highest $M_\star$ bins, the best-fitting trend of q_{SFR} with redshift has slope -0.057±0.002 (see fits in Fig. 21). This value is strikingly similar to that inferred for q_{TIR} (-0.055±0.018, Sect. 4.2.3). Then we extrapolated such trend at lower $M_\star$ bins (Sect. 4.2.4) and recursively removed radio AGN to derive the AGN-corrected IRRC. Fig. 21 displays the final q_{SFR} as a function of $M_\star$ and redshift. While the normalizations at $M_\star > 10^{10.5}$ $M_\odot$ are fully consistent with those derived in the q_{TIR} space, they progressively increase up by a factor of two at the lowest $M_\star$, due to mounting dust-uncorrected UV contribution. This further amplifies the overall $M_\star$ stratification. Using the same approach as for Eq. 6, the multi-parametric fitting in the q_{SFR}–$M_\star$–z plane yields the following expression:

$$q_{SFR}(M_\star, z) = (2.743\pm0.034)\times A^{(-0.025\pm0.012)} - B\times(0.234\pm0.017) \quad (8)$$

where A=(1+z) and B=($\log M_\star/M_\odot - 10$). Similarly to the fit in the q_{TIR} space, the redshift dependence is weaker and less significant than the $M_\star$ dependence, which is unsurprisingly steeper than before. This suggests that radio emission drops considerably more than SFR in low-$M_\star$ non-AGN galaxies. Reversing the argument, at fixed L_{IR}, radio emission underestimates the total SFR by a larger factor as compared to the IR light. The sub-linear trend $L_{IR} \propto L_{1.4\ GHz}^{0.90}$ that we inferred in our analysis (see also Bell 2003; Hodge et al. 2008; Davies et al. 2017; Brown et al. 2017; Gürkan et al. 2018) becomes even steeper when adding the UV contribution to L_{IR}, i.e. $SFR_{IR+UV} \propto L_{1.4\ GHz}^{0.81}$. Such a radio deficit in the dwarf-galaxy regime could be possibly linked to shorter CRe scale height (Lacki & Thompson 2010) or weaker magnetic fields (Donevski & Prodanović 2015; Tabatabaei et al. 2017) that are common in less dense SF enviroments.

In addition, we note that the lower efficiency in producing synchrotron emission in low-SFR, low-$M_\star$ galaxies is already factored in recent synchrotron emission models of SFGs (e.g. Massardi et al. 2010; Mancuso et al. 2015; Bonaldi et al. 2019) based on empirical matching between local $L_{1.4\ GHz}$ and

SFR functions. Therefore, our results reinforce the need for $M_\star$-dependent, non-linear calibrations between radio-continuum emission and SFR, in order to develop successful observing strategies for targeting low-$M_\star$ galaxies at radio wavelengths.

These considerations are relevant in the context of the forthcoming SKA. In particular, the SKA mid-frequency receivers will be equipped with five bands, of which the SKA Band2 (0.95–1.76 GHz) will be the workhorse for radio-continuum based SFR measurements. Even the faintest and least massive galaxies in our sample will be routinely observed by SKA, probing diverse populations of SFGs (and composite AGN+SF objects). Our findings highlight that a detailed understanding of the physics behind the relation between radio synchrotron emission and SFR is fundamental for fully exploiting the unique SKA capabilities in terms of depth and angular resolution.

6. Summary and conclusions

In this manuscript we re-calibrate the IRRC of SFGs for the first time as a function of *both* $M_\star$ and redshift, out to z~4. Starting from an $M_\star$-selected sample of 413,678 galaxies SFGs selected via (NUV-r)/(r-J) colours in the COSMOS field, we leverage new de-blended IR/sub-mm data (Jin et al. 2018), as well as deep radio images from the VLA COSMOS 3 GHz Large Project (Smolčić et al. 2017b). Ancillary radio data in COSMOS from VLA 1.4GHz (Schinnerer et al. 2010) and MIGHTEE 1.3 GHz (Jarvis et al. 2016) surveys are also used to validate our stacking analysis and assumptions on the radio spectral slope (Sect. 3.3.1).

In each $M_\star$–z bin, we performed stacking of undetected sources at both IR (Sect. 3.1) and radio (Sect. 3.3) frequencies, and combined the stacked signal with individual detections a-posteriori to infer the average q_{TIR} as a function of $M_\star$ and redshift (Sect. 4.1). We develop a recursive approach for identifying and then subtracting radio-excess AGN in different $M_\star$ and redshift bins (Sect. 4.2). This technique is calibrated on a (>70%) $M_\star$-complete subsample of 3 GHz detections at $M_\star > 10^{10.5}$ $M_\odot$ and extrapolated to the rest of the sample to infer the AGN-corrected IRRC (Sect. 4.3). Finally, we interpret our findings in the context of existing IRRC studies, from both models and observations. The main results of this work are listed below.

1) The IRRC evolves primarily with $M_\star$, with more massive galaxies displaying systematically lower q_{TIR}. A secondary, weaker dependence on redshift is also observed. The multi-parametric best-fitting expression is the following: $q_{TIR}(M_\star,z) = (2.646\pm0.024) \times (1+z)^{(-0.023\pm0.008)} - (0.148\pm0.013) \times (\log M_\star/M_\odot - 10)$. At fixed redshift, this trend translates into an IRRC of $L_{IR} \propto L_{1.4\,GHz}^{0.90}$, which corroborates the similar sub-linear behaviour reported in the literature (e.g. Bell 2003; Hodge et al. 2008; Gürkan et al. 2018). The typical scatter of the IRRC at $M_\star > 10^{10.5}$ $M_\odot$ is around 0.21–0.22 dex (a factor of 1.7), consistent with other studies (Yun et al. 2001; Bell 2003; Molnár et al. 2020) and roughly constant with $M_\star$ and z.

2) Our recursive approach for removing radio AGN enables us to statistically decompose radio-detected SFGs and AGN (Figs. 13 and 14) as a function of $M_\star$ and redshift. Removing radio AGN substantially flattens the observed q_{TIR}–z trend at $M_\star > 10^{10.5}$ $M_\odot$ to a nearly flat slope. This correction nicely aligns the mode q_{TIR} of radio SFGs to the median stacked q_{TIR} of the full $M_\star$ sample of non-AGN galaxies. Therefore, we interpret the resulting AGN-corrected q_{TIR} measurements as robust against further AGN removal. We acknowledge that residual radio AGN activity within radio-detected SFGs (10–20%) could be possible. Nevertheless, we expect this effect, if any, to further flatten out the evolution of q_{TIR} with redshift, and to induce an even steeper $M_\star$ dependence, thus reinforcing our main findings.

3) The fraction of radio AGN identified within the full $M_\star$ sample strongly increases with $M_\star$, spanning from 0.4% to 6% across the full range (Table 3), in agreement with previous studies (e.g. Heckman & Best 2014). However, when limited to 3-GHz detected sources, radio-detected dwarves ($M_\star < 10^{9.5} M_\odot$) are biased towards AGN-dominated sources, as our AGN/SF classification identifies ~90% of them as radio-excess AGN, yet they are a rare population occupying only ~0.4% of all SFGs in this $M_\star$ regime. We test the reliability of our radio AGN identification against available VLBA data of radio AGN (Herrera Ruiz et al. 2017), finding that 90% of them were also classified as radio-excess AGN from our analysis.

4) We examined the evolution of q_{TIR} as a function of SFR surface density (Σ_{SFR}), as a proxy for $M_\star$, finding a very similar trend both in slope and statistical significance. In qualitative agreement with models (e.g. Lacki & Thompson 2010) and recent observations (Algera et al. 2020a), our results support a decreasing q_{TIR} in more compact star-forming enviroments in MS galaxies, due to increasing cosmic ray scale height that boosts radio synchrotron emission.

5) We compare the average q_{TIR} between MS galaxies and an $M_\star$-complete subsample of SBs detected at IR wavelengths (Sect. 5.2). Despite SBs being more compact than MS analogues (Jiménez-Andrade et al. 2019), we do not observe a significant offset in q_{TIR}, apparently at odds with our expectations. However, as already foreseen by radio synchrotron models (Lacki & Thompson 2010), we postulate that SB galaxies might follow a different q_{TIR} relation with Σ_{SFR}, in which other factors (i.e. "puffy" vs "compact" geometry) could play a role.

6) We verified that adding the UV dust-uncorrected contribution to the IR, as a proxy for the total SFR, would further steepen the q_{SFR}–$M_\star$ trend, leaving the evolution with redshift unchanged. These findings imply that using radio emission as a SFR tracer requires $M_\star$-dependent conversion factors. Finally, our results can be useful to make accurate calibrations for future radio-continuum surveys as SFR machines down to dwarf galaxy regimes, especially in the upcoming SKA era.

Acknowledgements. ID is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 788679. MJJ acknowledges support from the UK Science and Technology Facilities Council [ST/N000919/1], the Oxford Hintze Centre for Astrophysical Surveys which is funded through generous support from the Hintze Family Charitable Foundation and a visiting Professorship from SARAO. SJ acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities (MICIU) under grant AYA2017-84061-P, co-financed by FEDER (European Regional Development Funds). DL acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 694343). IHW acknowledges support from the Oxford Hintze Centre for Astrophysical Surveys which is funded through generous support from the Hintze Family Charitable Foundation. RC acknowledges financial support from CONICYT Doctorado Nacional N° 21161487 and the Max-Planck Society through a Partner Group grant with MPA. JD acknowledges the financial assistance of SARAO. MN acknowledges support from the ERC Advanced Grant 740246 (Cosmic Gas). IP acknowledges financial support from the Italian Ministry of Foreign Affairs and International Cooperation (MAECI Grant Number ZA18GR02) and the South African Department of Science and Technology's National Research Foundation (DST-NRF Grant Number 113121) as part of the ISARP RADIOSKY2020 Joint Research Scheme. SMR hereby acknowledged the financial assistance of the National Research Foundation (NRF) towards this research. JS acknowledges the funding from the Swiss National Science Foundation under Grant No. 185863. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. We acknowledge use of the Inter-University Institute for Data Intensive Astronomy (IDIA) data intensive research cloud for data processing. IDIA is a

South African university partnership involving the University of Cape Town, the University of Pretoria and the University of the Western Cape. The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

References

Aird, J., Coil, A. L., & Georgakakis, A. 2019, MNRAS, 484, 4360
Algera, H. S. B., Smail, I., Dudzevičiūtė, U., et al. 2020a, arXiv e-prints, arXiv:2009.06647
Algera, H. S. B., Van der Vlugt, D., Hodge, J. A., et al. 2020b, arXiv e-prints, arXiv:2009.13531
Appleton, P. N., Fadda, D. T., Marleau, F. R., et al. 2004, ApJS, 154, 147
Aretxaga, I., Wilson, G. W., Aguilar, E., et al. 2011, MNRAS, 415, 3831
Arnouts, S., Walcher, C. J., Le Fèvre, O., et al. 2007, A&A, 476, 137
Baugh, C. M., Lacey, C. G., Frenk, C. S., et al. 2005, MNRAS, 356, 1191
Bavouzet, N., Dole, H., Le Floc'h, E., et al. 2008, A&A, 479, 83
Bell, E. F. 2003, ApJ, 586, 794
Bertoldi, F., Carilli, C., Aravena, M., et al. 2007, ApJS, 172, 132
Best, P. N. & Heckman, T. M. 2012, MNRAS, 421, 1569
Béthermin, M., Daddi, E., Magdis, G., et al. 2015, A&A, 573, A113
Béthermin, M., Dole, H., Beelen, A., & Aussel, H. 2010, A&A, 512, A78
Béthermin, M., Le Floc'h, E., Ilbert, O., et al. 2012, A&A, 542, A58
Bonaldi, A., Bonato, M., Galluzzi, V., et al. 2019, MNRAS, 482, 2
Bondi, M., Zamorani, G., Ciliegi, P., et al. 2018, A&A, 618, L8
Bonzini, M., Mainieri, V., Padovani, P., et al. 2015, MNRAS, 453, 1079
Bourne, N., Dunne, L., Ivison, R. J., et al. 2011, MNRAS, 410, 1155
Bourne, N., Maddox, S. J., Dunne, L., et al. 2012, MNRAS, 421, 3027
Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
Brown, M. J. I., Moustakas, J., Kennicutt, R. C., et al. 2017, ApJ, 847, 136
Bruzual, G. & Charlot, S. 2003, MNRAS, 344, 1000
Buat, V., Noll, S., Burgarella, D., et al. 2012, A&A, 545, A141
Burgarella, D., Buat, V., Gruppioni, C., et al. 2013, A&A, 554, A70
Calistro Rivera, G., Williams, W. L., Hardcastle, M. J., et al. 2017, MNRAS, 469, 3468
Capak, P., Aussel, H., Ajiki, M., et al. 2007, ApJS, 172, 99
Cappellari, M., McDermid, R. M., Alatalo, K., et al. 2012, Nature, 484, 485
Carraro, R., Rodighiero, G., Cassata, P., et al. 2020, arXiv e-prints, arXiv:2007.11002
Ceraj, L., Smolčić, V., Delvecchio, I., et al. 2018, A&A, 620, A192
Chabrier, G. 2003, PASP, 115, 763
Civano, F., Marchesi, S., Comastri, A., et al. 2016, ApJ, 819, 62
Condon, J. J. 1992, ARA&A, 30, 575
Condon, J. J., Huang, Z. P., Yin, Q. F., & Thuan, T. X. 1991, ApJ, 378, 65
Cowie, L. L., Barger, A. J., Hsu, L. Y., et al. 2017, ApJ, 837, 139
Cucciati, O., Tresse, L., Ilbert, O., et al. 2012, A&A, 539, A31
Dabringhausen, J., Kroupa, P., & Baumgardt, H. 2009, MNRAS, 394, 1529
Davé, R. 2008, MNRAS, 385, 147
Davidzon, I., Ilbert, O., Laigle, C., et al. 2017, A&A, 605, A70
Davies, L. J. M., Huynh, M. T., Hopkins, A. M., et al. 2017, MNRAS, 466, 2312
de Jong, T., Klein, U., Wielebinski, R., & Wunderlich, E. 1985, A&A, 147, L6
Deason, A., Wetzel, A., & Garrison-Kimmel, S. 2014, ApJ, 794, 115
Del Moro, A., Alexander, D. M., Mullaney, J. R., et al. 2013, A&A, 549, A59
Delhaize, J., Smolčić, V., Delvecchio, I., et al. 2017, A&A, 602, A4
Delhaize et al., s. 2020, MNRAS
Delvecchio, I., Daddi, E., Aird, J., et al. 2020, ApJ, 892, 17
Delvecchio, I., Daddi, E., Shankar, F., et al. 2019, ApJ, 885, L36
Delvecchio, I., Smolčić, V., Zamorani, G., et al. 2017, A&A, 602, A3
Delvecchio, I., Smolčić, V., Zamorani, G., et al. 2018, MNRAS, 481, 4971
Dole, H., Lagache, G., & Puget, J. L. 2003, ApJ, 585, 617
Donevski, D. & Prodanović, T. 2015, MNRAS, 453, 638
Donley, J. L., Rieke, G. H., Rigby, J. R., & Pérez-González, P. G. 2005, ApJ, 634, 169
Draine, B. T. & Li, A. 2007, ApJ, 657, 810
Dubois, Y., Volonteri, M., Silk, J., et al. 2015, MNRAS, 452, 1502
Elbaz, D., Dickinson, M., Hwang, H. S., et al. 2011, A&A, 533, A119
Fakhouri, O., Ma, C.-P., & Boylan-Kolchin, M. 2010, MNRAS, 406, 2267
Franco, M., Elbaz, D., Zhou, L., et al. 2020, arXiv e-prints, arXiv:2005.03040
Garrison-Kimmel, S., Rocha, M., Boylan-Kolchin, M., Bullock, J. S., & Lally, J. 2013, MNRAS, 433, 3539
Geach, J. E., Dunlop, J. S., Halpern, M., et al. 2017, MNRAS, 465, 1789
Genzel, R., Burkert, A., Bouché, N., et al. 2008, ApJ, 687, 59
Gilli, R., Comastri, A., & Hasinger, G. 2007, A&A, 463, 79
Goulding, A. D., Forman, W. R., Hickox, R. C., et al. 2014, ApJ, 783, 40
Griffin, M. J., Abergel, A., Abreu, A., et al. 2010, A&A, 518, L3
Gürkan, G., Hardcastle, M. J., Smith, D. J. B., et al. 2018, MNRAS, 475, 3010
Hales, C. A., Murphy, T., Curran, J. R., et al. 2012, MNRAS, 425, 979
Hales, C. A., Norris, R. P., Gaensler, B. M., et al. 2014, MNRAS, 441, 2555
Hardcastle, M. J. & Croston, J. H. 2020, New A Rev., 88, 101539
Harwit, M. & Pacini, F. 1975, ApJ, 200, L127
Heckman, T. M. & Best, P. N. 2014, ARA&A, 52, 589
Heinis, S., Buat, V., Béthermin, M., et al. 2013, MNRAS, 429, 1113
Heinis, S., Buat, V., Béthermin, M., et al. 2014, MNRAS, 437, 1268
Helou, G. & Bicay, M. D. 1993, ApJ, 415, 93
Helou, G., Khan, I. R., Malek, L., & Boehmer, L. 1988, ApJS, 68, 151
Helou, G., Soifer, B. T., & Rowan-Robinson, M. 1985, ApJ, 298, L7
Herrera Ruiz, N., Middelberg, E., Deller, A., et al. 2017, A&A, 607, A132
Herrera Ruiz, N., Middelberg, E., Deller, A., et al. 2018, A&A, 616, A128
Hickox, R. C., Jones, C., Forman, W. R., et al. 2009, ApJ, 696, 891
Hodge, J. A., Becker, R. H., White, R. L., & de Vries, W. H. 2008, AJ, 136, 1097
Hopkins, A. M. & Beacom, J. F. 2006, ApJ, 651, 142
Hummel, E., Davies, R. D., Wolstencroft, R. D., van der Hulst, J. M., & Pedlar, A. 1988, A&A, 199, 91
Ibar, E., Cirasuolo, M., Ivison, R., et al. 2008, MNRAS, 386, 953
Ibar, E., Ivison, R. J., Best, P. N., et al. 2010, MNRAS, 401, L53
Ibar, E., Ivison, R. J., Biggs, A. D., et al. 2009, MNRAS, 397, 281
Ivison, R. J., Magnelli, B., Ibar, E., et al. 2010a, A&A, 518, L31
Ivison, R. J., Swinbank, A. M., Swinyard, B., et al. 2010b, A&A, 518, L35
Jarvis, M., Taylor, R., Agudo, I., et al. 2016, in MeerKAT Science: On the Pathway to the SKA, 6
Jarvis, M. J., Smith, D. J. B., Bonfield, D. G., et al. 2010, MNRAS, 409, 92
Jiménez-Andrade, E. F., Magnelli, B., Karim, A., et al. 2019, A&A, 625, A114
Jin, S., Daddi, E., Liu, D., et al. 2018, ApJ, 864, 56
Karim, A., Schinnerer, E., Martínez-Sansigre, A., et al. 2011, ApJ, 730, 61
Kaviraj, S., Laigle, C., Kimm, T., et al. 2017, MNRAS, 467, 4739
Kaviraj, S., Martin, G., & Silk, J. 2019, MNRAS, 489, L12
Keller, B. W., Wadsley, J., & Couchman, H. M. P. 2016, MNRAS, 463, 1431
Kennicutt, R. C. & Evans, N. J. 2012, ARA&A, 50, 531
Kennicutt, Jr., R. C. 1998, ApJ, 498, 541
Koudmani, S., Henden, N. A., & Sijacki, D. 2020, arXiv e-prints, arXiv:2007.10342
Kurczynski, P. & Gawiser, E. 2010, AJ, 139, 1592
Lacki, B. C. & Thompson, T. A. 2010, ApJ, 717, 196
Lacki, B. C., Thompson, T. A., & Quataert, E. 2010, ApJ, 717, 1
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Le Floc'h, E., Aussel, H., Ilbert, O., et al. 2009, ApJ, 703, 222
Lee, N., Sanders, D. B., Casey, C. M., et al. 2015, ApJ, 801, 80
Lehmer, B. D., Basu-Zych, A. R., Mineo, S., et al. 2016, ApJ, 825, 7
Leslie, S. K., Schinnerer, E., Liu, D., et al. 2020, ApJ, 899, 58
Liu, D., Daddi, E., Dickinson, M., et al. 2018, ApJ, 853, 172
Lutz, D. 2014, ARA&A, 52, 373
Lutz, D., Poglitsch, A., Altieri, B., et al. 2011, A&A, 532, A90
Madau, P. & Dickinson, M. 2014, ARA&A, 52, 415
Magdis, G. E., Daddi, E., Béthermin, M., et al. 2012, ApJ, 760, 6
Magnelli, B., Elbaz, D., Chary, R. R., et al. 2009, A&A, 496, 57
Magnelli, B., Ivison, R. J., Lutz, D., et al. 2015, A&A, 573, A45
Magnelli, B., Lutz, D., Saintonge, A., et al. 2014, A&A, 561, A86
Magnelli, B., Popesso, P., Berta, S., et al. 2013, A&A, 553, A132
Mancuso, C., Lapi, A., Cai, Z.-Y., et al. 2015, ApJ, 810, 72
Mannucci, F., Cresci, G., Maiolino, R., Marconi, A., & Gnerucci, A. 2010, MNRAS, 408, 2115
Marchesi, S., Civano, F., Elvis, M., et al. 2016, ApJ, 817, 34
Marleau, F. R., Clancy, D., Habas, R., & Bianconi, M. 2017, A&A, 602, A28
Massardi, M., Bonaldi, A., Negrello, M., et al. 2010, MNRAS, 404, 532
Mauch, T., Cotton, W. D., Condon, J. J., et al. 2020, ApJ, 888, 61
McCracken, H. J., Le Fèvre, O., Brodwin, M., et al. 2001, A&A, 376, 756
McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, A&A, 544, A156
Mezcua, M. 2017, International Journal of Modern Physics D, 26, 1730021
Mezcua, M., Civano, F., Fabbiano, G., Miyaji, T., & Marchesi, S. 2016, ApJ, 817, 20
Mezcua, M. & Domínguez Sánchez, H. 2020, ApJ, 898, L30
Mezcua, M., Suh, H., & Civano, F. 2019, MNRAS, 488, 685
Molnár, D. C., Sargent, M. T., Delhaize, J., et al. 2018, MNRAS, 475, 827
Molnár et al., s. 2020, MNRAS
Mullaney, J. R., Alexander, D. M., Goulding, A. D., & Hickox, R. C. 2011, MNRAS, 414, 1082
Murphy, E. J. 2009, ApJ, 706, 482
Murphy, E. J., Bremseth, J., Mason, B. S., et al. 2012, ApJ, 761, 97
Murphy, E. J., Condon, J. J., Schinnerer, E., et al. 2011, ApJ, 737, 67
Murphy, E. J., Momjian, E., Condon, J. J., et al. 2017, ApJ, 839, 35
Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJS, 206, 8
Nelson, E. J., van Dokkum, P. G., Förster Schreiber, N. M., et al. 2016, ApJ, 828, 27
Nguyen, H. T., Schulz, B., Levenson, L., et al. 2010, A&A, 518, L5
Niklas, S. & Beck, R. 1997, A&A, 320, 54
Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJ, 660, L43
Novak, M., Smolčić, V., Delhaize, J., et al. 2017, A&A, 602, A5
Novak, M., Smolčić, V., Schinnerer, E., et al. 2018, A&A, 614, A47
Oke, J. B. 1974, ApJS, 27, 21

Oliver, S. J., Bock, J., Altieri, B., et al. 2012, MNRAS, 424, 1614
Padovani, P., Bonzini, M., Kellermann, K. I., et al. 2015, MNRAS, 452, 1263
Pannella, M., Elbaz, D., Daddi, E., et al. 2015, ApJ, 807, 141
Poglitsch, A., Waelkens, C., Geis, N., et al. 2010, A&A, 518, L2
Popesso, P., Magnelli, B., Buttiglione, S., et al. 2012, arXiv e-prints, arXiv:1211.4257
Read, S. C., Smith, D. J. B., Gürkan, G., et al. 2018, MNRAS, 480, 5625
Reines, A. E. & Deller, A. T. 2012, ApJ, 750, L24
Reines, A. E., Greene, J. E., & Geha, M. 2013, ApJ, 775, 116
Reines, A. E., Plotkin, R. M., Russell, T. D., et al. 2014, ApJ, 787, L30
Reines, A. E., Sivakoff, G. R., Johnson, K. E., & Brogan, C. L. 2011, Nature, 470, 66
Rickard, L. J. & Harvey, P. M. 1984, AJ, 89, 1520
Rodighiero, G., Daddi, E., Baronchelli, I., et al. 2011, ApJ, 739, L40
Salim, S., Charlot, S., Rich, R. M., et al. 2005, ApJ, 619, L39
Sargent, M. T., Béthermin, M., Daddi, E., & Elbaz, D. 2012, ApJ, 747, L31
Sargent, M. T., Schinnerer, E., Murphy, E., et al. 2010, ApJ, 714, L190
Schinnerer, E., Sargent, M. T., Bondi, M., et al. 2010, ApJS, 188, 384
Schleicher, D. R. G. & Beck, R. 2013, A&A, 556, A142
Schober, J., Schleicher, D. R. G., & Klessen, R. S. 2016, ApJ, 827, 109
Schober, J., Schleicher, D. R. G., & Klessen, R. S. 2017, MNRAS, 468, 946
Schreiber, C., Pannella, M., Elbaz, D., et al. 2015, A&A, 575, A74
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Silk, J. 2017, ApJ, 839, L13
Smith, D. J. B., Jarvis, M. J., Hardcastle, M. J., et al. 2014, MNRAS, 445, 2232
Smolčić, V., Delvecchio, I., Zamorani, G., et al. 2017a, A&A, 602, A2
Smolčić, V., Novak, M., Bondi, M., et al. 2017b, A&A, 602, A1
Speagle, J. S., Steinhardt, C. L., Capak, P. L., & Silverman, J. D. 2014, ApJS, 214, 15
Spergel, D. N., Verde, L., Peiris, H. V., et al. 2003, ApJS, 148, 175
Steinhardt, C. L., Speagle, J. S., Capak, P., et al. 2014, ApJ, 791, L25
Tabatabaei, F. S., Schinnerer, E., Krause, M., et al. 2017, ApJ, 836, 185
van der Kruit, P. C. 1971, A&A, 15, 110
van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, ApJ, 788, 28
van Dokkum, P. G. 2008, ApJ, 674, 29
Viero, M. P., Wang, L., Zemcov, M., et al. 2013, ApJ, 772, 77
Voelk, H. J. 1989, A&A, 218, 67
Welikala, N., Béthermin, M., Guery, D., et al. 2016, MNRAS, 455, 1629
Whitaker, K. E., Pope, A., Cybulski, R., et al. 2017, ApJ, 850, 208
Whitaker, K. E., van Dokkum, P. G., Brammer, G., & Franx, M. 2012, ApJ, 754, L29
White, R. L., Helfand, D. J., Becker, R. H., Glikman, E., & de Vries, W. 2007, ApJ, 654, 99
White, S. V., Jarvis, M. J., Häußler, B., & Maddox, N. 2015, MNRAS, 448, 2665
White, S. V., Jarvis, M. J., Kalfountzou, E., et al. 2017, MNRAS, 468, 217
Wong, O. I., Koss, M. J., Schawinski, K., et al. 2016, MNRAS, 460, 1588
Wu, Y., Charmandaris, V., Houck, J. R., et al. 2008, ApJ, 676, 970
Yun, M. S., Reddy, N. A., & Condon, J. J. 2001, ApJ, 554, 803

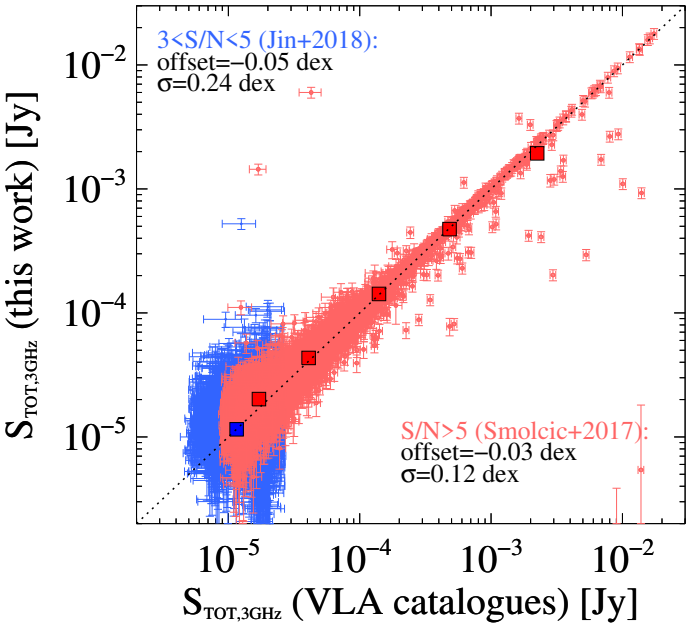

Fig. A.1: Comparison of total fluxes of 3 GHz detections between our procedure and catalogue fluxes, both at S/N>5 (Smolčić et al., red dots) and at 3<S/N<5 (Jin et al. 2018; blue dots). Squares highlight the median ratio at various intervals. The global offset and dispersion suggest a good agreement within the uncertainties down to S/N~3.

Appendix A: Testing total radio fluxes

We validate our total flux estimation against individual detections taken from published VLA catalogues at 3 GHz. At S/N>5 we used the catalogue of Smolčić et al. (2017b), while total fluxes at 3<S/N<5 were taken from Jin et al. (2018). After excluding the 67/10830 multi-component sources identified in Smolčić et al. (2017b), we calculate peak and total fluxes of each source, following the approach described in Sect. 3.3. Fig. A.1 displays the comparison between total fluxes (dots), highlighting the corresponding median ratio at various intervals (squares). It is worth noting that Smolčić et al. (2017b, red) used the software BLOBCAT (Hales et al. 2012; 2014) to sum over all blobs identified in the 3 GHz image above a certain S/N cut. Therefore, it is best suited for non-Gaussian shapes. On the other hand, our approach described in Sect. 3.3 assumes a 2D elliptical Gaussian, with size, angle and normalization being free to vary. Despite the different techniques, we find a good agreement at S/N>5, with a logarithmic offset of -0.028 dex and dispersion of 0.12 dex. At 3<S/N<5, total fluxes from Jin et al. (2018, blue) were computed via Gaussian PSF fitting, using a circular beamsize of 0.75". Despite the low S/N regime, we also observe a fair agreement, with an offset of -0.05 dex and dispersion of 0.24 dex. This check proves our total fluxes fully consistent with the published values of Smolčić et al. (2017b) and Jin et al. (2018) for individual 3 GHz detections down to S/N~3.

We further demonstrate that our choice of performing median stacking at 3 GHz, rather than rms-weighted mean stacking, does not impact our final $L_{1.4\ GHz}$ estimates. A comparison between median and mean $L_{1.4\ GHz}$ is presented in Fig. A.2. The top panel displays $L_{1.4\ GHz}$ from the combined flux of detections and non-detections (see Eq. 3), while the bottom panel refers to the case of purely undetected sources. Colours indicate different $M_\star$ bins. Only stacks in which peak fluxes have S/N>3 are shown. No systematics is observed, at any $M_\star$, between mean and median stacked $L_{1.4\ GHz}$. This is consistent with White et al. (2007),

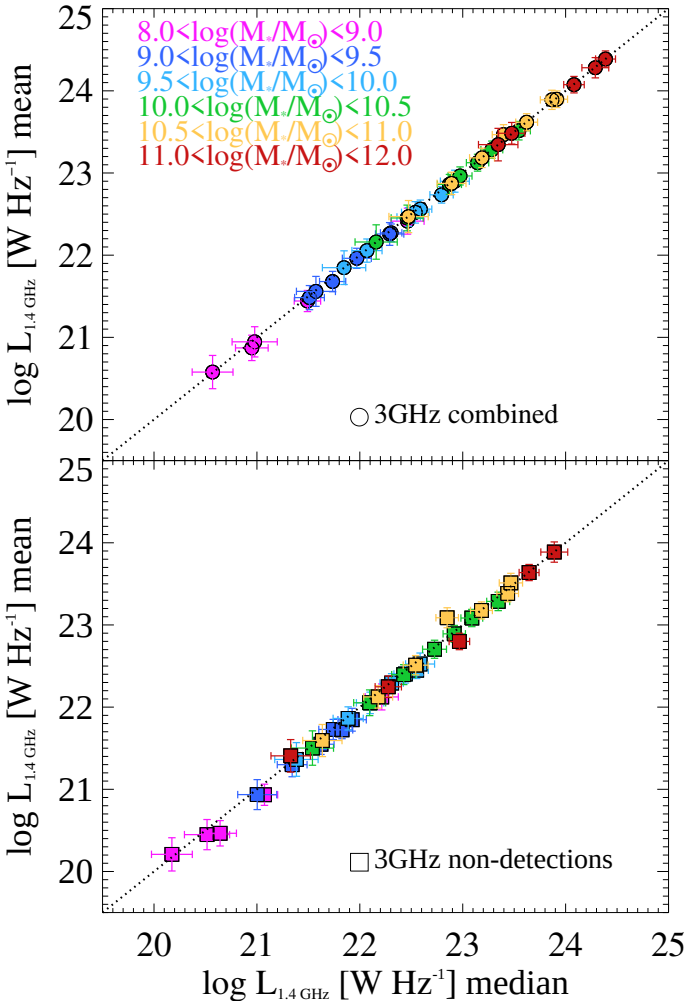

Fig. A.2: Top panel: comparison between median $L_{1.4\ GHz}$ (x-axis) and rms-weighted mean $L_{1.4\ GHz}$ (y-axis) for combined 3 GHz detections and non-detections, following Eq. 3. Colours indicate various $M_\star$ bins. Bottom panel: same comparison, but referred to 3 GHz undetected sources only.

who showed that, in the noise-dominated regime, the stacked median traces the population mean. Moreover, such excellent agreement confirms that the uniform 3 GHz sensitivity across the full map ensures that either stacking method can reliably recover the average flux of the underlying galaxy population.

The fact that non-detections (bottom panel) display consistent $L_{1.4\ GHz}$ between mean and median stacking suggests that, if any, radio AGN do not dominate the total radio emission in our stacks. The same argument cannot be implicitly extended to the combined fluxes, since these mean weighted-average fluxes could be biased towards fewer and brighter radio detections, which reduces the statistical weight of non-detections. Indeed, $L_{1.4\ GHz}$ of radio detections (top panel) are always >3× larger than $L_{1.4\ GHz}$ of non-detections (bottom panel), despite the smaller numbers. This partly smooths over the initial fluctuations between mean and median stacking, thus delivering an even tighter agreement, as we observe.

Appendix B: Impact of a different radio AGN-vs-SFG fitting approach

We discuss a potential caveat related to our AGN-vs-SF decomposition presented in Sect. 4.2.2. Specifically, our procedure relies on the assumption that the mode of the observed q_{TIR} distribution ($q_{TIR,peak}$) of radio detections is entirely attributed to SF. Though this is supported by a number of previous studies arguing that radio AGN are a sub-dominant population in the sub-mJy regime (e.g. Padovani et al. 2015; Smolčić et al. 2017b; Novak et al. 2018; Ceraj et al. 2018; Algera et al. 2020b), the contribution of radio-faint AGN to $q_{TIR,peak}$ might not be negligible. If this is the case, by mirroring and fitting the SF Gaussian first, it is possible that we are underestimating the true fraction of radio AGN relative to SFGs. To quantify this potential issue and test how much it would affect our final $M_\star$-dependence of q_{TIR}, here we follow a different approach.

The observed q_{TIR} distribution is fitted with two Gaussian functions *simultaneously*, which parametrize the contribution of SFGs and radio-excess AGN. Contrary to Sect. 4.2.2, we do not set the SF peak to $q_{TIR,peak}$, but we leave it free to vary along with the dispersion and normalization for both functions. In this simultaneous fitting we give equal input weights to all bins, regardless of the number of sources in each. This approach is thus expected to return a rather conservative AGN contribution relative to SFGs.

The results are shown in Fig. B.1 for the two highest $M_\star$ bins. As in Fig. 13, the best-fit SF (blue) and AGN (red) Gaussians head up to reproduce the total distribution (black). However, we clearly notice two main differences compared to the previous approach. Firstly, the AGN distribution is far broader than the SF distribution in both $M_\star$ bins. Secondly, the relative fraction of radio AGN that we mis-classify as SFGs (red tail at $>q_{peak}–2\sigma$) is as high as 40–70%, hence much higher than the 30% obtained in Sect. 4.2.2 when fitting and mirroring the SF part first. This is clearly displayed by the cumulative AGN fraction in the bottom panels. Instead, the relative fractions of "pure" SFGs above the 2σ threshold are about 80% at $10^{10.5}<M_\star/M_\odot<10^{11}$ and 90% at at $10^{11}<M_\star/M_\odot<10^{12}$.

Despite the lower level of purity of the SFG population, we emphaize that the main results of this paper are quite robust against the AGN-vs-SF fitting procedure. Indeed, both peak and dispersion (~0.21 dex) of the SF population are essentially unchanged, the peak being identical to $q_{TIR,peak}$ and the dispersion reaching ~0.21 dex. Therefore, it is reasonable to *assume* that the mode of the observed q_{TIR} distribution is attributed to radio-detected SFGs. Related to this, the threshold $q_{peak}–2\sigma$ is still equal to 0.42 dex, implying that roughly the same exact sources as in Sect. 4.2.2 would be identified as radio-excess AGN. This agreement demonstrates that our recursive radio AGN removal would lead to the same final IRRC, regardless of the assumed shape of the AGN distribution.

If we were able to statistically remove the underlying radio AGN contribution within the SF population (though impossible with the present data), this would systematically increase q_{TIR} by a larger amount towards *lower $M_\star$ galaxies*. Indeed, at $10^{11}<M_\star/M_\odot<10^{12}$ the radio AGN distribution is clearly broader, but far more offset than at $10^{10.5}<M_\star/M_\odot<10^{11}$, thus at higher $M_\star$ the two populations are more distinguishable. As a consequence, we argue that a proper correction for such an effect would further amplify the $M_\star$ stratification of q_{TIR} reported in this work.

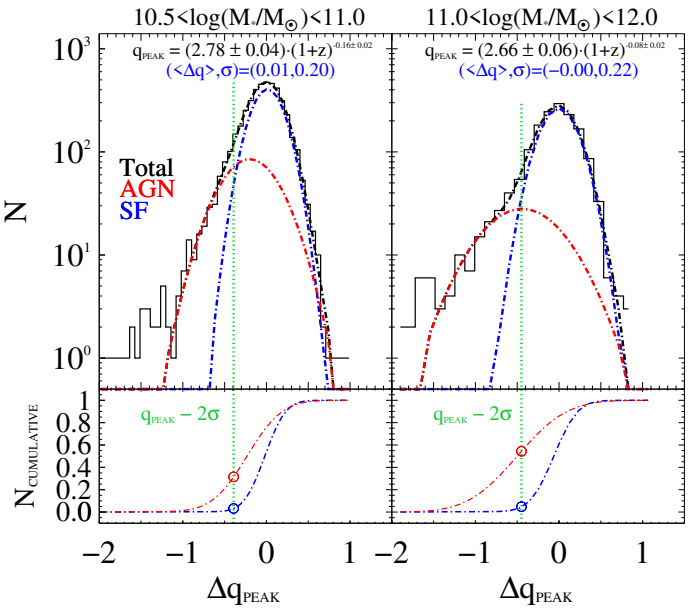

Fig. B.1: Same as Fig. 13, but fitting the total q_{TIR} distribution of 3 GHz detections (black histogram) simultaneously with a SF Gaussian (blue) and an AGN Gaussian (red dashed). The 1σ dispersion has remained unchanged to about 0.21 dex. The bottom panels display the corresponding cumulative Gaussian fits, both normalized to unity. The vertical green dotted line marks the 2σ threshold of 0.42 dex below which we consider a source as radio-excess AGN. As opposed to Fig. 13, this cutoff removes only 30-50% of the total radio AGN population. However, we estimate this effect to be more prevalent in lower $M_\star$ galaxies. This implies that accounting for such mis-classified radio AGN would likely strengthen our final $M_\star$-dependent q_{TIR}.

Appendix C: Differences compared to the literature

Our best-fit relation of q_{TIR} as a function of $M_\star$ and redshift (Eq. 6 in Sect. 4.3) is fully consistent with the average q_{TIR} value measured in local SFGs (i.e. 2.64 in Bell 2003) for a typical galaxy with $M_\star\sim10^{10}$ $M_\odot$. At higher redshifts, instead, our average q_{TIR} measurements follow flatter evolutionary trends compared to previous studies (Fig. 16), while the best-fit normalization appears broadly consistent with the literature only at $M_\star>10^{10.5}$ $M_\odot$. In order to interpret these differences in a quantitative fashion, we identify three key points that combined differentiate our approach from that adopted in the previous literature: (i) removing radio AGN via a recursive approach in each $M_\star$ and redshift bin; (ii) exploiting an $M_\star$-selected sample of SFGs; (iii) binning the derived q_{TIR} as a function of both $M_\star$ and redshift. To test our results against different techniques, we expand on each of these aspects below.

Appendix C.1: Radio AGN subtraction

In Sect. 4.2, we performed a recursive subtraction of radio AGN as a function of $M_\star$ and redshift, carefully calibrated on high-$M_\star$ galaxies, and then extrapolated to lower $M_\star$ analogues. However, other studies followed alternative approaches to discard radio AGN when deriving the intrinsic IRRC. For instance, Magnelli et al. (2015) performed median stacking of both radio detections and non-detections out to z~2. This method strongly reduces the contribution of a few bright outliers, assuming that the bulk radio population is made of SFGs. This assumption is quite reasonable, since Magnelli et al. (2015) started from an

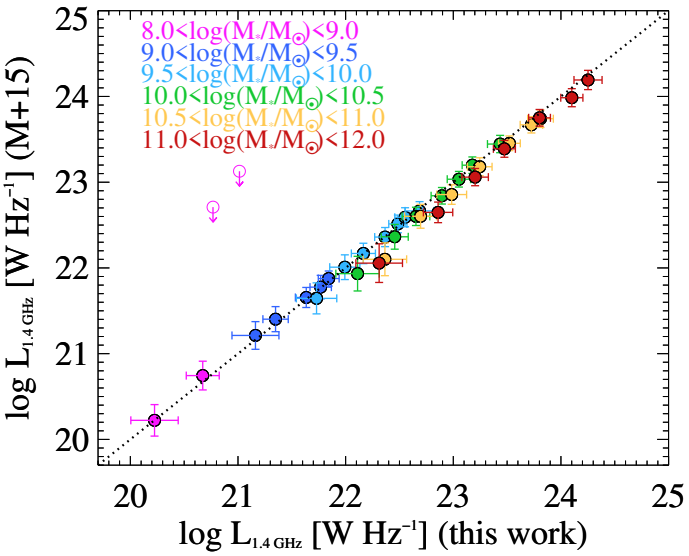

Fig. C.1: Comparison between AGN-corrected $L_{1.4\ GHz}$ from this work (x-axis) and median $L_{1.4\ GHz}$ obtained from stacking detections and non-detections together (Magnelli et al. 2015, y-axis). Different $M_\star$ ranges are colour-coded, while downward arrows mark 3σ upper limits for 2/37 bins. Despite these different approaches, we notice a very good agreement in all bins, that strengthens the reliability of our recursive AGN subtraction.

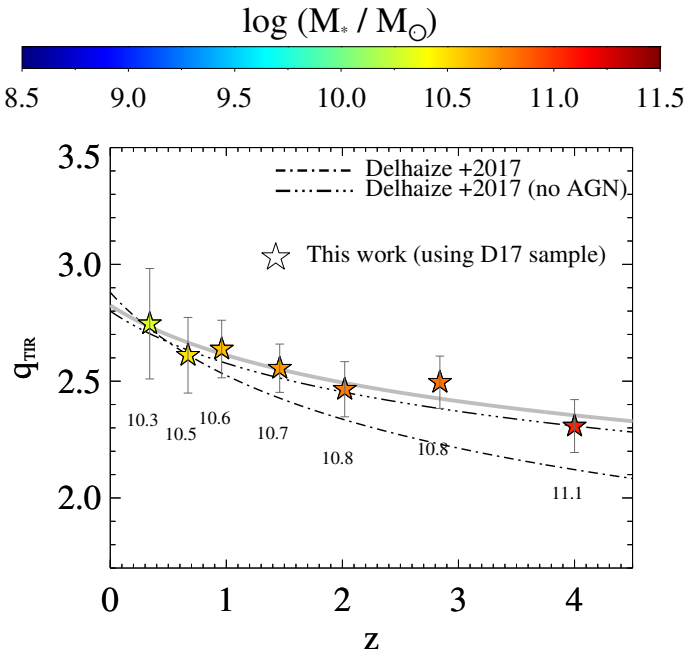

Fig. C.2: Median q_{TIR} as a function of redshift obtained by analysing the SFG sample of Delhaize et al. (2017, stars). Black lines indicate the median q_{TIR}–z trend of Delhaize et al. before (dot-dashed) and after (triple dot-dashed) removing 2σ outliers. The grey solid line marks the resulting best-fit q_{TIR} trend with redshift, that is highly consistent with that of Delhaize et al. (2017) after removing radio AGN. Numbers below each star denote the median $M_\star$ of the underlying sample.

$M_\star$-selected sample, of which radio detections make a negligible fraction.

We compare our q_{TIR} with mock measurements obtained by following the stacking method of Magnelli et al. (2015), but applied to the sample used in our work. Fig. C.1 displays the final $L_{1.4\ GHz}$ estimates that we obtained after removing radio AGN (x-axis) against those derived from median radio stacking (Magnelli et al. 2015, y-axis). We note that our L_{IR} estimates and Magnelli et al.'s were instead calculated through a fully consistent approach, therefore only a difference in $L_{1.4\ GHz}$ might lead to systematics in the final q_{TIR} trends. The colour bar highlights the average $M_\star$ of each bin. Out of 37 bins analyzed in this work, 35 yield a S/N>3 from median 3 GHz stacking (circles), while 3σ upper limits are shown for the remaining bins (downward arrows). This comparison clearly reveals a very good agreement between final 1.4 GHz luminosities, with all measurements being consistent within the uncertainties. Despite the different approaches, the agreement extends down to dwarf galaxies, supporting the AGN nature of most radio-detected sources (Sect. 4.2.4). A possible (though not significant) deviation of ~0.1 dex might be present at the highest $M_\star$, with our measurements returning slightly higher $L_{1.4\ GHz}$ measurements than those of Magnelli et al.. This might be ascribed to the contribution of radio-detected SFGs to our weighted average $L_{1.4\ GHz}$, since they make a substantial fraction of the $M_\star$-selected sample in that $M_\star$ bin (~45%, Table 3). Therefore, this test proves our radio AGN subtraction broadly consistent with a totally independent approach.

Appendix C.2: Sample selection and binning

An additional aspect worth testing is whether different sample selections lead to distinct IRRC trends. We started from an $M_\star$-selected sample of SFGs based on K_s-band priors, that typically reaches much deeper than any infrared or radio survey, compared to an average galaxy SED. A rare exception is represented by very high-redshift (z>4) or heavily dust-obscured systems, which are visible only in IRAC (e.g. Davidzon et al. 2017) or deep ALMA imaging (e.g. Franco et al. 2020). For this reason, studies that derived the IRRC based on exclusive or joint samples of radio/IR detections, are partly biased against low-$M_\star$ galaxies. For instance, the work of Delhaize et al. (2017) was based on a jointly-selected infrared (from *Herschel*, with S/N≥5 in at least one PACS or SPIRE band) and radio (VLA 3 GHz with S/N≥5; Smolčić et al. 2017b) sample of SFGs in the COSMOS field, out to z~5. By performing double-censored survival analysis to account for sources undetected at either radio or FIR wavelengths, they found an evolving $q_{TIR}=(2.88\pm0.03)\times(1+z)^{-0.19\pm0.01}$, which flattens to $q_{TIR}=(2.80\pm0.02)\times(1+z)^{-0.12\pm0.01}$ after removing 2σ outliers (as reported in Delvecchio et al. 2018), particularly radio-excess AGN. We repeat our IR and radio stacking analysis using the same sample of SFGs from Delhaize et al. (2017) (9,575 sources), to demonstrate that our analysis leads to consistent results when matching the input sample.

We split the sample of Delhaize et al. (2017) among the same seven redshift bins analyzed in this work. For each, we perform median stacking of 3 GHz and IR images in all bands, combining both detections and non-detections. This approach should be comparable to the search for the median value carried out via survival analysis (Delhaize et al. 2017). Although we do not formally remove radio AGN in this check, we showed in Sect. C.1 that median radio stacking yields broadly consistent results (see Magnelli et al. 2015). Fig. C.2 displays the median q_{TIR} obtained by stacking the SFG sample of Delhaize et al. (2017) in different redshift bins (stars). This yields a best-fitting trend $q_{TIR}=(2.82\pm0.15)\times(1+z)^{-0.11\pm0.05}$, that is fully consistent with the flatter trend of Delhaize et al. (2017) after removing 2σ outliers (triple dot-dashed line). This check proves our technique solid against different sample selections from the literature.

Fig. C.2 also highlights the important role played by the binning grid in driving a declining IRRC with redshift. In particular, the colour-coded $M_\star$ clearly indicates how a joint IR and radio selection is sensitive to increasing galaxy $M_\star$ with redshift. Moreover, the scatter of the IRRC reported by Delhaize et al. (2017) is around 0.35 dex, while the dispersion that we measured at $M_\star > 10^{10.5}$ $M_\odot$ (Sect. 4.2.2) is only 0.21–0.22 dex. This is similar to the value reported by Bell (2003) (i.e. 0.26 dex) for nearby galaxies, recently narrowed down to 0.16 dex in Molnár et al. (2020). A possible reason for the smaller than 0.35 dex dispersion in our study might be that we are splitting SFGs among different $M_\star$, each carrying an intrinsically smaller dispersion compared to the full SFG sample. Because of the decreasing q_{TIR} with $M_\star$, binning only as a function of redshift leads to a mixture of different galaxy $M_\star$ that results into a larger global dispersion.