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


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/97997/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
The infrared-radio correlation of star-forming galaxies is strongly M∗-dependent but nearly redshift-invariant since z~4


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 (qIR), 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 qIR with redshift, finding a mild but significant decline, that is yet to be understood. For the first time, we calibrate qIR as a function of both stellar mass (M∗) and redshift, starting from an M∗-selected sample of >400,000 star-forming galaxies in the COSMOS field, identified via (NUV-r)µ and radio (rest 1.4 GHz) luminosities, with more massive galaxies displaying systematically lower qIR. A secondary, weaker dependence on redshift is also observed. The best-fit analytical expression is the following: qIR(M∗,z) = (2.646±0.024) x (1+z)^(-0.023±0.008) x (0.148±0.013) x (log[M∗/M⊙]-10). Adding the UV dust-uncorrected contribution to the IR as a proxy for the total SFR, would further steepen the qIR dependence on M∗. We interpret the apparent redshift decline reported in previous literature as due to low-M∗ 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 ratios seen in more massive galaxies are well described by their higher observed SFR surface densities. Our findings highlight that using radio-synchrotron emission as a proxy for SFR requires novel M∗-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∗ 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 (IR; rest-frame 8-1000 μm, i.e. $L_{\text{IR}}$) and radio (e.g. rest-frame 1.4 GHz, i.e. $L_{\text{GHz}}$) spectral luminosity arising from star formation, usually referred to as the “infrared-radio correlation” (IRRC; e.g. Helou et al. 1985; de Jong et al. 1985). This tight ($\sigma \sim 0.16$ dex, e.g. Molnár et al. 2020 submitted) correlation is often parametrized by the IR-to-radio luminosity ratio $q_{\text{IR}}$, defined as (e.g. Helou et al. 1985; Yun et al. 2001):

$$q_{\text{IR}} = \log \left( \frac{L_{\text{IR}}}{3.75 \times 10^{12} \text{Hz}} \right) - \log(L_{\text{GHz}}/\text{W Hz}^{-1})$$

where $3.75 \times 10^{12}$ Hz represents the central frequency over the far-infrared (FIR, rest-frame 42-122 μm) domain, usually scaled to IR in the recent literature. In the local Universe, the IRRC (or its parametrization $q_{\text{IR}}$) appears to hold over at least three orders of magnitude in both $L_{\text{IR}}$ and $L_{\text{GHz}}$ (e.g. Helou et al. 1985; Condon 1992; Yun et al. 2001). Broadly speaking, this because the infrared emission comes from dust heated by fairly massive ($> 5 M_\odot$) OB stars, while radio emission arises from relativistic cosmic ray electrons (CRe) accelerated by shock waves produced when massive stars ($> 8 M_\odot$) explode as supernovae. Nevertheless, CRe are also subject to different cooling processes, as they propagate throughout the galaxy, which are mainly caused by inverse Compton, bremsstrahlung and ionization losses (e.g. Murphy 2009).

Surprisingly enough, despite all such processes at play, infrared and radio emission are observed to correlate, both in local star-forming late-type galaxies and even in merging galaxies (e.g. Condon et al. 1993, 2002; Murphy 2013). 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; Garrett 2002; Appleton et al. 2004; Murphy et al. 2008; 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. 2020b).

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; $L_{\text{IR}} > 10^{12} L_\odot$; e.g. Yun et al. 2001) concluded that the IRRC remains linear across a wide range of $L_{\text{IR}}$. Conversely, other studies have argued that at low luminosities the IRRC may break down, consistent with a non-linear trend of the form $L_{\text{IR}} \propto L_{\text{GHz}}^{-0.75}$ (e.g. Bell 2003; Hodge et al. 2008; Davies et al. 2017; Gürkan et al. 2018), which might be partly induced by dust heating from old stellar populations (Bell 2003).

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 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 $M_\star \geq 10^{10} M_\odot$) SFGs, because of their increasing compactness (i.e. the size-mass relation $R \propto M_\star^{-1/3}$; 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 & Bica 1993; Niklas & Beck 1997; Bell 2003; Lacki et al. 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 IR 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 IR 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 $\sim 0.2–0.3$ dex). This relation holds from $z \sim 5$ down to the local Universe (e.g. Brinchmann et al. 2004; Noeske et al. 2007; Elbaz et al. 2011; Whittaker 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 $L_{\text{IR}}$, especially in massive SFGs (Kennicutt 1998), the existence of the MS gives an additional argument that studying $q_{\text{IR}}$ 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 $q_{\text{IR}} \propto (1+z)^{-0.2–0.1}$, although the physical explanation for such evolution is still uncertain. Somewhat different conclusions were reached by other works (e.g. Garrett et al. 2002; 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 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 $\geq 400,000$ 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; Dec=-2.20583333 (J2000)). Then we leverage the new deblended 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 $z \sim 4$. In addition, we exploit the deepest radio-continuum data taken from the VLA-COSMOS 3 GHz Large Project (Smolčić et al. 2017b). Individual detections will be combined with stacked flux densities of non-detections, at both IR and radio frequencies to assess the average $q_{IR}$ 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 $L_{opt}$ (Sect. 3.3). The average $q_{IR}$ 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 flux densities in Appendix A. A detailed comparison between radio stacking results is presented in Appendix B. In Appendix C we discuss how the final IRRC is sensitive to our AGN subtraction method. Finally, in Appendix D 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 LCDM cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, 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. The most recent collection of multi-wavelength 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 $Y$, $J$, $J$, $H$, 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 3.6–8$\mu$m data from the Infrared Array Camera (IRAC). We refer the reader to Laigle et al. (2016) for more details.

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

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.

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) ($3\sigma$ 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.

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 $\langle |\Delta z/(1+z_p)\rangle = 0.007$ 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 $\geq 5\%$ chance to be offset from the median by $>0.5\times(1+z_p)$). 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 quality flag $>3 \land |z_{p}-z_{p}| < 0.1\times(1+z_{p})$).

Infrared/sub-mm flux densities 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

![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.](image-url)
use of the rest-frame, dust-corrected \((NUV-r)\) and \((r-J)\) colours available in the parent catalogues (hereafter \(NUVrJ\)). As opposed to the widely used UVJ criterion, the \((NUV-r)\) colour is more sensitive to recent star formation \((10^6-10^8\ yr\ scales\) \cite{Salmietal2005,Arnouts2007,Davidzonetal2017}. Therefore, this criterion enables us to better distinguish between weakly star-forming galaxies (with specific-SFR, sSFR\(=\)SFR/M\(_\star\) \(\sim\)\(10^{-11}\)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. The fraction of catastrophic failures \(|lz-l\rho|>0.15\times(1+z)|\) is only 3.4\%.

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\(_\odot\)<\(10^{10}\) 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 the overall conclusions of this work are unchanged if we restrict ourselves to \(z<3\).

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\(_\nu\) and those extrapolated from the MS relation \cite{Schreiberetal2015} 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\(\sigma\) 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.

### 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. \cite{Jinetal2018}, based on the de-blending algorithm developed in Liu et al. \cite{Liuetal2018} for the GOODS-North field.

The dataset includes Spitzer-MIPS 24 \(\mu\)m data (P. D. Sanders; Le Floc’h et al. 2009); Herschel imaging from the PACS (100-160 \(\mu\)m, Poglitsch et al. 2010) and the SPIRE (250, 350, and 500 \(\mu\)m, Griffin et al. 2010) instruments, as part of the PEP \cite{Lutzetal2011} and HerMES \cite{Oliveretal2012} programs, respectively. In addition, JCMT/SCUBA2 (850 \(\mu\)m) images are taken from the S2CLS program \cite{Cowieetal2017,Geachtal2017}, the ASTE/AzTEC (1.1 mm) data are nested maps from Aretxaga et al. \cite{Aretxagetal2011} over a sub-area of 0.72 deg\(^2\). Finally, Jin et al. \cite{Jinetal2018} also included MAMBO data \cite{Bertoldietal2007} at 1.2 mm over an area of 0.11 deg\(^2\).

Briefly, Jin et al. \cite{Jinetal2018} used K\(_s\)-selected sources from the UltraVISTA survey \cite{Smolci etal2017b} as priors to perform PSF fitting of MIPS 24 \(\mu\)m, VLA-3 GHz \cite{Smolci etal2017b} and VLA-1.4 GHz \cite{Schinnereretal2010} images down to the 3\(\sigma\) level in each band. Within our final sample, this procedure identifies 67,114 MIPS 24 \(\mu\)m+VLA priors. Nevertheless, adopting a similar approach for extracting FIR/sub-mm flux densities 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 Jin et al. \cite{Jinetal2018}, only an M\(_\star\)-complete subset of K\(_s\) priors was added, which ultimately prioritizes IR brighter sources. This leads to a total of 136,584 K\(_s\)+MIPS 24 \(\mu\)m+VLA priors, that were used to de-blend and extract the Herschel, SCUBA2 and AzTEC flux densities \cite{Table1}. Within our final sample of 413,678 star-forming galaxies, 20,777 (5\%) have a combined S/N>3 over all FIR/sub-mm bands (10,285 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 flux densities, 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 \cite{Jinetal2018}.

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. Radio data in the COSMOS field

For our analysis we exploited data from the VLA-COSMOS 3 GHz Large Project \cite{Smolci etal2017b}, that is one of the largest and deepest radio survey ever conducted over a medium
Table 1: Main numbers of priors and detections that characterize our final sample of 413,678 star-forming galaxies. Note that subsets do not add up to make the final sample. (\(\ast\)): numbers reported also in Fig. 3.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Number of sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final sample (this work)</td>
<td>413,678 (\ast)</td>
</tr>
<tr>
<td>- MIPS 24 (\mu) m + VLA priors</td>
<td>67,114</td>
</tr>
<tr>
<td>- MIPS 24 (\mu) m + VLA + (K) priors</td>
<td>136,584</td>
</tr>
<tr>
<td>- (S/N_{IR}) &gt; 3</td>
<td>20,777 (\ast)</td>
</tr>
<tr>
<td>- (S/N_{3GHz}) &gt; 3</td>
<td>13,808 (\ast)</td>
</tr>
</tbody>
</table>

3. Stacking analysis

The aim of this paper is to investigate how the IRRC evolves with \(M_*\) 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 \(M_*\)-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 \(M_*\) and redshift, as shown in Fig. 3. For each bin, we also report the total number of \(M_*\)-selected SFGs (black), as well as the corresponding fractions having combined \(S/N_{IR}\) > 3 (red) and \(S/N_{3GHz}\) > 3 (blue). As can be seen, both fractions are a strong function of both \(M_*\) and redshift. Therefore, binning along both parameters enables us to account for the fact that galaxies of distinct \(M_*\) 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 D).

3.1. Infrared and sub-mm stacking

In this Section we estimate the average flux densities across eight infrared and sub-mm bands, namely MIPS 24 \(\mu\) m, PACS 100-160 \(\mu\) m, SPIRE 250-350-500 \(\mu\) m, SCUBA 850 \(\mu\) m and AzTEC 1100 \(\mu\) m. 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 flux densities 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. We stress that our procedure yields very consistent results with either median or mean stacking of detections and non-detections combined (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, \(M_*\) bin and redshift bin, we stack N\(\times\)N pixel cutouts from the residual images, each centred on the NIR position of the \(M_*\)-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\(\times\)FWHM, since a substantial fraction of the 24 \(\mu\) m flux is located in the first Airy ring. Since the AzTEC

\(^2\) https://www.ias.u-psud.fr/irgalaxies/downloads.php
The uncertainties on the stacked flux densities 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/sub-mm 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 flux densities 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).

$$S(x, y) = \varphi \times PSF(x, y) + \psi \times (PSF \odot w)(x, y) + \epsilon$$  \hspace{1cm} (2)

where $S(x, y)$ is the stacked image, $PSF$ the point spread function, and $w$ the auto-correlation function. The symbol $\odot$ represents the convolution. The parameters $\varphi, \psi,$ and $\epsilon$ 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. $\epsilon$) 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 $(1 - bias)$ for each band (see Table 2), drawn only from stacks with S/N $> 3$. For those images, we multiply the stacked flux by $(1 - bias)$ at the corresponding wavelength. Only MIPS 24 µm data are not shown, since their flux densities 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.

---

Fig. 4: Stacked cutouts of NUV/$J$–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 and AzTEC 1100 µm images are smoothed with a Gaussian kernel to ease the visualization. Each cutout size is $8\times$FWHM of the PSF, while for Spitzer-MIPS we choose $13\times$FWHM. Below each cutout we report the corresponding S/N ratio.
Table 2: 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.

<table>
<thead>
<tr>
<th>Wavelength</th>
<th>% Clustering signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PACS 100 μm</td>
<td>11.3±7.4</td>
</tr>
<tr>
<td>PACS 160 μm</td>
<td>10.2±16.5</td>
</tr>
<tr>
<td>SPIRE 250 μm</td>
<td>25.9±18.9</td>
</tr>
<tr>
<td>SPIRE 350 μm</td>
<td>31.3±20.8</td>
</tr>
<tr>
<td>SPIRE 500 μm</td>
<td>42.7±24.2</td>
</tr>
<tr>
<td>SCUBA 850 μm</td>
<td>19.2±10.7</td>
</tr>
<tr>
<td>AzTEC 1100 μm</td>
<td>20.1±12.9</td>
</tr>
</tbody>
</table>

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⋆ – 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 (U) per unit dust mass (M⋆), 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⋆. Therefore, (U) and the SED shape both vary as a function of redshift, for which Magdis et al. (2012) empirically found as (U)∝(1+z)1.15 up to z~2. More recently, Béthermin et al. (2015) revised the evolution of (U) with redshift out to z~4, using IR/sub-mm data in the COSMOS field, retrieving (U)∝(1+z)1.8. Here we adopt the set of 14 MS templates from Magdis et al. (2012), fit them to our data, and compare the (U)–z trend with Béthermin et al. (2015).

The SED-fitting routine performs a simultaneous fitting using AGN and dust emission models, looking for the best-fit solution via χ² minimization. In order to account for the typical photo-z uncertainty of the underlying galaxy population (at fixed M⋆, z), each template is fitted to the data across a range of ±0.05×(1+z) around the median redshift (z). The code keeps track of each SED solution and corresponding normalization, generating likelihood distributions and uncertainties on e.g. LIR, (U) 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 FIR dust features that cannot be captured by our limited set of templates (e.g., IR to rest-frame 8 μm ratio, IRS, 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 LIR 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⋆ (left to right, expressed in log(M⋆/M☉)) 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⋆ and redshift bin. Green dashed lines represent SEDs without FIR measurements and at z≥1.5, for which the integrated LIR is interpreted as 3σ upper limit (5/42 bins). Even though 24 μm has long been used as a proxy for LIR, this is only accurate at z≤1.5 (e.g. Elbaz et al. 2011; Lutz 2014). For this reason we still interpret as measurements the LIR obtained from SEDs without FIR data, but only at z≤1.5. That is the case for a few bins at the lowest M⋆, in which the SED reproduces a-posteriori the 24 μm data-point. Globally, our stacking analysis yields robust LIR estimates in 37/42 bins.

We find (U)∝(1+z)1.74±0.18, which is fully consistent with the revised (U)–z trend of Béthermin et al. (2015): (U)∝(1+z)1.8±0.4. This test is reassuring, since it confirms that one single z-dependent (or (U)–dependent) MS galaxy template is fully able to reproduce the observed SED across a wide M⋆ interval.
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 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_*, 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 calibrate the star-forming MS over an unprecedented M_* (down to M_*=10^{9.5} \text{M}_\odot) and redshift range (z \leq 4). Since we carried out a similar analysis, it is worth checking whether the SFR estimates...
I. Delvecchio et al.: The IRRC evolves primarily with $M_\star$

3.3. Radio stacking at 3 GHz

In this Section we describe the equivalent stacking analysis done with radio 3 GHz (Smolčić et al. 2017b) data, in order to derive average rest-frame 1.4 GHz spectral luminosities ($L_{1.4\GHz}$) in each $M_\star$ bin.

As done for IR stacking (Sect. 3.1), we treat detections and non-detections separately. Total flux densities 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 flux densities to those of the corresponding catalogues. The purpose of this approach is twofold: using the same published flux densities 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 $\mu$Jy), 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 minimal radio AGN contamination. This alternative approach will be tested in Appendix D.

Within the UltraVISTA area analyzed in this work, the 3 GHz rms ($2.3\, \mu$Jy 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 ($S_{\text{peak}}$) 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 brightness. The 1σ dispersion around this value is interpreted as the corresponding 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 flux densities ($S_{\text{tot}}$) are calculated by fitting a 2D elliptical Gaussian function to the median stacked image, using the IDL routine mmptr2dfun$^3$. 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 $A_{\text{gauss}}$. The integrated flux error was computed by

3 https://pages.physics.wisc.edu/~craigm/idl/fitqa.html
multiplying the peak flux error by \(\sqrt{A_{\text{beam}}/A_{\text{gauss}}/A_{\text{beam}}}\), where \(A_{\text{beam}}\) 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:

\[
\frac{S_{\text{tot}}}{S_{\text{peak}}} > 1 + (6 \times S/N_{\text{peak}})^{-1.44}
\]

where \(S/N_{\text{peak}}\) 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 flux densities are on average 1.8× higher than peak flux densities. 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 \(S/N_{\text{peak}}<3\). These are all among the 5 bins without \(L_{1.4}\) estimates from IR stacking (Sect. 3.2). Ano logically 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 flux densities 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 flux densities were first scaled to 1.4 GHz assuming of \(S_\nu \propto \nu^\alpha\), with spectral index \(\alpha=-0.75\pm 0.1\) (e.g. Condon 1992; Ibar et al. 2009,2010). This assumption is discussed in Appendix B.1. Lastly, 1.4 GHz flux densities were converted to rest-frame 1.4 GHz spectral luminosities \((L_{1.4\ \text{GHz}})\), again assuming \(\alpha=-0.75\). Formal \(L_{1.4\ \text{GHz}}\) errors were calculated by propagating the uncertainties on both combined flux and spectral index.

The various checks described in Appendix B prove our \(L_{1.4\ \text{GHz}}\) robust across the full range of \(M_\star\) and redshift analyzed in this work. We note that our \(L_{1.4\ \text{GHz}}\) 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 \(L_{1.4\ \text{GHz}}\). This issue will be addressed in Sect. 4.2.  

4. The IRRC and the contribution of radio AGN

Using the median \(q_{JR}\) and \(L_{1.4\ \text{GHz}}\) 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 \(q_{JR}\) errors. Among the 42 \(M_\star–z\) bins analyzed in this work, 37 yield robust estimates of \(L_{JR}\) and \(L_{1.4\ \text{GHz}}\), 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 \(10^9 <M_\star/M_\odot<10^{10.5}\) 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. \(q_{JR}\) before removing radio AGN

Fig. 8 shows the average \(q_{JR}\) 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 \(q_{JR}=2.64\pm 0.02\) (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 \(q_{JR}\)) 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: \(q_{JR}=(2.35\pm 0.08)\times (1+z)^{-0.12\pm 0.04}\), where the normalization is scaled to 2.63 in the \(q_{JR}\) 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 \(z>5\sigma\)) 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 \(z\sim 4\) as: \(q_{JR}=2.88\pm 0.03\times (1+z)^{-0.19\pm 0.01}\). 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: \(q_{JR}=2.83\pm 0.02\times (1+z)^{-0.15\pm 0.01}\), which becomes fully consistent with that of Magnelli et al. (2015).

When compared to the above literature, it is evident that our \(q_{JR}\) 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 \(q_{JR}\) estimates seem to display a clear \(M_\star\) stratification, with the most massive galaxies having typically lower \(q_{JR}\) 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 \(L_{1.4\ \text{GHz}}\), particularly at high \(M_\star\) (see e.g. Best & Heckman 2012) where radio AGN feedback is known to be prevalent. Our \(L_{JR}\) 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 \(q_{JR}\). 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. 8 should be taken as the AGN-uncorrected \(q_{JR}\). However, it is worth showing it to quantify how much \(q_{JR}\) 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 \(q_{JR}\) 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.

\footnote{The far-infrared luminosity used to compute \(q_{JR}\) was integrated between 42 and 122 \(\mu m\) rest-frame. This is quantified to be 1.91× smaller than the total \(L_{1.4}\) (Magnelli et al. 2015).}
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 median and median stacked L_{1.4 GHz} for non-detections (Fig. A.2, bottom panel). Indeed, if the contribution of radio-undetected AGN were substantial, 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. (2020c) argue that below 20 μJy (at 3 GHz), the fraction of radio-excess AGN is <10% (see also Smolić 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_{IR} distribution traced by individual 3 GHz detections as a function of M_{*} and redshift. First, we identify a subset of radio detections at M_{*} >10^{10.5} M_{\odot} that is representative of an M_{*}-selected sample. Then we decompose their q_{IR} 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_{*} >10^{10.5} M_{\odot} (Sect. 4.2.3). Later we extrapolate this calibration towards lower M_{*} bins (Sect. 4.2.4), where a similar in-depth analysis was not possible due to radio-detections being strongly incomplete in this M_{*} regime. Finally, the intrinsic (i.e. AGN-corrected) IRRC as a function of M_{*} and redshift is presented in Sect. 4.3.

4.2.1. The q_{IR} distribution of radio detections

In order to study the q_{IR} distribution of 3 GHz detections, we need to calculate their average L_{IR} as a function of M_{*} 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_{*}-z space. Out of 13,510 radio detections among our 37 bins, 8,762 (65%) have a combined S/N_{IR} >3, therefore reliable L_{IR} measurements from SED-fitting of FIR/sub-mm deblended 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_{*}-z bin. Stacked IR flux densities are corrected for clustering bias and converted to L_{IR} following the same procedure adopted for the prior M_{*} 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_{*} of that source (assuming the MS relation), in order to reduce the variance of the underlying sample within each M_{*}-z bin. We verified that our stacked L_{IR} are always systematically below the 3sigma 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 well placed to explore the full q_{IR} distribution of 3 GHz detections at different M_{*} and redshifts. Fig. 9 shows q_{IR} as a function of redshift, split in six M_{*} bins. Black dots mark individual 3 GHz detections, blue stars represent the q_{IR} obtained by combining detections and non-detections (same as in Fig. 8), 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_{*}, from 3.5% at 10^{9} <M_{*}/M_{\odot}<10^{9.5} to 78.3% at 10^{9.5} <M_{*}/M_{\odot}<10^{10.5}, which implies that at the lowest M_{*} nearly all q_{IR} 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_{*} 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_{*}-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_{*} and redshift. In order to extend this calibration to our full M_{*}-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_{*}-selected galaxies at a given redshift. To this end, within each bin we compare the q_{IR} of single radio detections against a specific threshold (q_{IR,lim}), corresponding to the 3 GHz survey limit at a given M_{*} and redshift (green upward arrows in Fig. 9). This threshold is proportional to the median stacked L_{IR} of the full SFGs sample, divided by the 3sigma luminosity limit at 1.4 GHz, scaled from 3 GHz by assuming a fixed spectral index α=-0.75 (Appendix. B.1). Specifically, q_{IR,lim} indicates the limiting q_{IR} at which a typical MS galaxy of a given M_{*}, z and L_{IR} drops below the 3 GHz detection limit, which translates into a lower q_{IR} limit. In other words, sources with q_{IR} <q_{IR,lim} lie within an M_{*} range that is virtually inaccessible by our 3 GHz survey. Therefore, any measurement above that threshold is not representative of an M_{*}-selected sample. Conversely, radio detections below that threshold would be seen also in an M_{*}-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 \( q_{IR, \text{lim}} \). This cutoff enables us to narrow down the position of the mode of the \( q_{IR} \) 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. 9.

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 \( q_{IR} \) 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 \( q_{IR} \) 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 \( q_{IR} \) distribution of radio detections, namely \( q_{IR, \text{peak}} \) (see red open circles in Fig. 9). We note that \( q_{IR, \text{peak}} \) 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 \( q_{IR, \text{peak}} \) with redshift using the IDL routine \texttt{mpfit2dpfun}, obtaining the best-fit expressions shown in Fig. 10.

### 4.2.2. Identifying radio AGN at high \( M_\star \)

After fitting the \( q_{IR, \text{peak}} \) trend with redshift for the two \( M_\star \) bins, we need to account for such redshift dependence while exploring the \( q_{IR} \) distributions of radio detections. To this end, we align the position of \( q_{IR, \text{peak}} \) 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 \( q_{IR} \) distribution is displayed in Fig. 10 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.

![Distributions of $q_{IR}$ as a function of redshift, split across increasing $M_\star$ bins.](image-url)

Fig. 9: Distribution of \( q_{IR} \) as a function of redshift, split across increasing \( M_\star \) bins. In each panel, we compare the \( q_{IR} \) estimates of individual radio detections (black dots) with the median stacked values of non-detections (yellow squares) and with the weighted-average \( q_{IR} \) of detections and non-detections together (from Eq. 3, blue stars). Green upward arrows indicate the corresponding threshold \( q_{IR,lim} \) 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 \( q_{IR} \) below the corresponding threshold. This criterion identifies 13 bins (in red square brackets). Within these bins, the peak of \( q_{IR} \) distribution (\( q_{IR, peak} \)) 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 \( S_{3GHz} > 3 \). See Sect. 4.2.1 for details.
Assuming that the peak of the distribution is populated by radio-detected SFGs, and that the intrinsic $q_{IR}$ distribution of SFGs is symmetric around the peak, we mirror the right-hand side of the observed $q_{IR}$ distribution to the left side. This symmetric function is interpreted as the intrinsic $q_{IR}$ distribution of 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 $q_{IR}$ 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 $q_{IRpeak}$ 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 C, 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. 10. 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, taking a single stacked $L_{IR}$ in each bin does not strongly impact the calibration of the SF locus.

Choosing the best dividing line between AGN and SF-dominated radio sources is a challenging, and somewhat arbitrary task. Moving the threshold to higher $q_{IR}$ 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 $q_{IR}$ distribution drawn by the two Gaussian fits (AGN in red, SF in blue), as shown at the bottom of Fig. 10, 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. 10), and $\sigma$ 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 3 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 on average 0.43 dex (that we deem robust in Sect. 4.2.3), which implies that our radio-excess AGN should have statistically at least 63% of their total radio emission arising from AGN activity.

### 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 $M_\star > 10^{10.5} M_\odot$. Then we combined the remaining radio-detected SFGs with...
Fig. 11: Same as Fig. 10, but normalizing the peak to the flatter $q_{IR}$--$z$ trend calibrated after removing AGN (Sect. 4.2.3). This two-fold approach slightly improves the $q_{IR}$ decomposition, as highlighted by the larger cumulative fraction of radio AGN that are rejected below the $q_{IR}$ threshold (red open circles, 81% against the previous 70%).

stacks of non-detections to compute the new $L_{1.4\,\text{GHz}}$ in those bins, which should be free from AGN contamination. We verified that the new $L_{1.4\,\text{GHz}}$ shifts the previously determined $q_{IR}$ (blue stars in Fig. 9) upward by a certain amount. In those complete bins, we fitted the AGN-corrected $q_{IR}$ with redshift, obtaining a significantly flatter relation than before, as shown in Fig. 11. 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 $q_{IR}$--$z$ trend, we again shift the $q_{IR}$ measurements of individual detections by the offset from such a trend at each $z$-bin, and perform a second $q_{IR}$ decomposition, as shown in Fig. 11. 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 $2\sigma$ threshold below the peak is also very similar: 0.42 and 0.44 dex in the two bins (therefore we use an average $\Delta q_{\text{AGN}}=0.43$ 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. 10 and 11), 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. 12 (at $M_\star>10^{10.5} M_\odot$), subtracting radio AGN (red dots) based on this latter locus shifts all the median $q_{IR}$ (blue stars) exactly on the fitted $q_{IR}$--$z$ trend (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 $q_{IR}$ coincide with the intrinsic 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 $-0.055\pm0.018$, i.e. not flat at a significance of $3\sigma$.

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 also at 2.5--$z<4.5$, finding a good agreement with the median $q_{IR}$ estimates.

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 ($\sigma\sim0.21--0.22$ dex), that we interpret as the intrinsic scatter of the IRCC 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. 11. 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 $L_{1.4\,\text{GHz}}$ 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 $q_{IR}$--$z$ trend might be also induced by a larger relative fraction of radio AGN with increasing redshift. As a sanity check, in both $M_\star$ bins we split and decomposed the $q_{IR}$ distribution of Fig. 11 separately at $z<1.2$ and $z>1.2$, examining the evolution of the relative fraction of radio AGN. Though we do find that radio AGN are slightly more prevalent at higher redshifts (i.e. on average from 12% at $z<1.2$ to 18% at $z>1.2$), we confirm that the dispersion of the SF population is redshift-invariant ($\sim0.20$ dex), both before and after removing radio AGN. This implies that the relative offset between the AGN and SF loci is unchanged, therefore our sample of $>2\sigma$ radio-excess AGN is globally preserved.

After removing those AGN, the flatter, yet declining $q_{IR}$ 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 since SFGs are intrinsically IR brighter, so the radio-excess contrast (at fixed $L_{1.4\,\text{GHz}}$) is less evident. Therefore, we argue that any further correction for mis-classified radio AGN would induce an even flatter $q_{IR}$ 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 4). 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).

Our empirical threshold 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 >5σ, 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_★ > 10^{11} M_☉). 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 incompletely inexact resolved subsample, the comparison with VLBA data further demonstrates the reliability of our radio AGN identification method.

### 4.2.4. Extrapolating the SF-vs-AGN loci at low M_★

We extrapolate the q_{IR}-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_★ < 10^{10.5} M_☉ are not representative of an M_★-selected sample. In particular, a galaxy of a given M_★ 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. 9. Radio detections at these masses are therefore quite peculiar relative to the overall galaxy population.

This is further suggested in Fig. 9 by the q_{IR} offset between median measurements (blue stars) and individual radio detections (black dots). The latter lie systematically below the median q_{IR}, deviating more and more at lower M_★. For these reasons, we refrain from calibrating the IRRC directly on those radio detections. We prefer to use the median q_{IR} values as benchmark, since they should be sensitive to a more representative sample of galaxies of that M_★.

We proceed as follows. Within each M_★ bin, the redshift trend of q_{IR} is extrapolated from that calibrated at higher M_★, in the form q_{IR} \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_{IR}. In other words, at M_★ < 10^{10.5} M_☉, we assume a constant q_{IR}-z slope. This approach is preferable to leaving also the slope as 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_{IR} trend as accurately as possible.
Following the iterative approach already tested at higher $M_\star$, the best-fitting trend of $q_{IR}$ with redshift enables us to identify radio AGN as sources lying $>0.3$ dex below the best-fit SF locus. After subtracting those radio AGN, we re-calculate the weighted-average $q_{IR}$ and search again for the best normalization that fits the new AGN-corrected $q_{IR}$ measurements with redshift. We repeat this procedure twice, i.e. until all median $q_{IR}$ are unchanged within the uncertainties, at each $M_\star$. This condition sets the end of our recursion.

The final, AGN-corrected $q_{IR}$ are shown in Fig. 12 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 $q_{IR}$ measurements nearly coincide with those obtained from stacking non-detections alone (yellow squares). These latter values delimit the highest $q_{IR}$ that could be reached if removing, by definition, all radio detections. The result of similarity between the two sets of $q_{IR}$ measurements is due to a heavy subtraction of radio AGN from the sample of radio detections. Within the sample of radio detections, the fraction of radio AGN identified at $M_\star<10^{10.5}$ $M_\odot$ increases with decreasing $M_\star$. From the first to the fourth $M_\star$ bin, these fractions are: 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 4). 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, according to our analysis the vast majority of radio-detected dwarf galaxies ($M_\star<10^{10.5}$ $M_\odot$, e.g. Mezcua 2017) in COSMOS are expected to be radio AGN.

Bearing this in mind, we note that the weighted average $q_{IR}$ (blue stars) are yet mostly driven by non-detections (yellow squares), which outnumber individual detections (dots) by a factor of $>100$ at $M_\star<10^{10.5}$ $M_\odot$. However, those few radio detections (mostly radio AGN, red dots) stand out from the stacks of non-detections (yellow squares) typically by over a factor of ten, up to one-hundred. As a consequence, after removing radio AGN at $M_\star<10^{10.5}$ $M_\odot$, the new average $q_{IR}$ still move upward by 0.2–0.3 dex.

### Table 4: Table summarizing the numbers and fractions of radio AGN and SFGs in different $M_\star$ bins, after fitting the AGN-corrected $q_{IR}$ with redshift (Sect. 4.2.4 and Fig. 12). Columns are sorted as follows: (1) $M_\star$ (M$_\odot$) bin; (2) best-fit normalization of the $q_{IR}$-$z$ trend, in the form $q_{IR}\propto(1+z)^\gamma$, by imposing $\gamma=-0.055\pm0.018$ 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$<z<2.5$).

<table>
<thead>
<tr>
<th>$M_\star$ (M$_\odot$) bin</th>
<th>$q_{IR}$-$z$ fit (normalization)</th>
<th>$#$ radio AGN</th>
<th>$#$ radio SFGs</th>
<th>$#$ $M_\star$ sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>$10^{10}-10^{9}$ M$_\odot$</td>
<td>2.83$\pm0.10^*$</td>
<td>482 (94.5%, 0.4%$)^*$</td>
<td>28 (5.5%, 0.0%)$^*$</td>
<td>129,658$^*$</td>
</tr>
<tr>
<td>$10^{9.5}-10^{9}$ M$_\odot$</td>
<td>2.78$\pm0.03^{(**)}$</td>
<td>489 (80.8%, 0.6%$)^{(**)}$</td>
<td>116 (19.2%, 0.1%)$^{(**)}$</td>
<td>78,563$^{(**)}$</td>
</tr>
<tr>
<td>$10^{10.5}-10^{10}$ M$_\odot$</td>
<td>2.75$\pm0.02$</td>
<td>802 (51.0%, 1.1%)</td>
<td>834 (49.0%, 1.2%)</td>
<td>17,122</td>
</tr>
<tr>
<td>$10^{10.5}-10^{10.5}$ M$_\odot$</td>
<td>2.65$\pm0.03$</td>
<td>608 (14.6%, 1.7%)</td>
<td>3,554 (85.4%, 9.6%)</td>
<td>36,838</td>
</tr>
<tr>
<td>$10^{11}-10^{12}$ M$_\odot$</td>
<td>2.58$\pm0.01$</td>
<td>359 (7.1%, 2.2%)</td>
<td>4,686 (92.9%, 28.4%)</td>
<td>16,489</td>
</tr>
<tr>
<td></td>
<td></td>
<td>182 (11.7%, 6.0%)</td>
<td>1,370 (88.3%, 44.8%)</td>
<td>3,060</td>
</tr>
</tbody>
</table>

Fig. 13: Intrinsic (i.e. AGN-corrected) $q_{IR}$ evolution as a function of redshift (x-axis) and $M_\star$ (colour bar). The $L_{IR}$ estimates are the same reported in Fig. 8, while $L_{IR}$ 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 2$\sigma$ outliers (triple dot-dashed lines).

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

After correcting our combined $L_{IR}$ 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. 13. 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. $-0.055\pm0.018$). 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 $q_{IR}$ with $M_\star$, with more massive galaxies showing systematically lower $q_{IR}$ values. This behaviour was already seen in Fig. 8 before re-
moving 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): \( q_{IR}(1+z) \sim 0.12 \pm 0.01 \) (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. 14 shows the equivalent of Fig. 13 but projected on \( M_\star \), with redshift bins in different colours. Error bars on each \( q_{IR} \) were re-scaled by a factor of \( \sqrt{\chi^2_{red}} \) in each \( M_\star \) bin, to bring the reduced \( \chi^2 \) of the corresponding \( q_{IR}-z \) fit to unity. It is quite evident that an \( M_\star \) dependence reduces substantially the scatter of the average \( q_{IR} \) around a single trend. To better quantify this, first we bootstrapped over the uncertainties of slope and normalization obtained from each \( q_{IR}-z \) trend (see Table 4). Then, at each \( M_\star \), we interpolated the full range of bootstrapped IRRCs at \( z=1 \), in correspondence of the 16\(^{\text{th}}\), 50\(^{\text{th}}\) and 84\(^{\text{th}}\) percentiles.

Interpolating at \( z=1 \), besides being at roughly the median redshift of our sample, reduces the increasing divergence of each \( q_{IR}-z \) fit at lower or higher redshifts. This leaves us with the interpolated median \( q_{IR}(1+z) \) 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: \( q(M_\star)_{z=1} = (2.586\pm0.011) - (0.124\pm0.015)\times\log(M_\star/M_\odot - 10) \). This function yields a \( \chi^2_{red} = 0.87 \), with an \( M_\star \) slope close to that commonly found when fitting \( q_{IR} \) 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 \( q_{IR}-M_\star-z \) space. This yields the following best-fit expression:

\[
q_{IR}(M_\star, z) = (2.646\pm0.024)\times10^{A(0.023\pm0.008)-B\times(0.148\pm0.013)} - 10
\]

where \( A = (1+z) \) and \( B = \log(M_\star/M_\odot - 10) \). The corresponding \( \chi^2_{red} = 0.90 \). 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.

The need for an additional \( M_\star \)-dependence of the IRRC (in the form of \( L_{\text{radio}}^{-}-\text{SFR} \)) has been also highlighted in previous low-z studies (Gürkan et al. 2018; Read et al. 2018) and recently extended out to \( z \sim 1 \) (Smith et al. 2020) using deep LOFAR-150 MHz data. A similar conclusion was independently reached in Molnár et al. (2020), when considering the dependence of the IRRC on galaxy spectral 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 \( L_{\text{1.4 GHz}} \), Molnár et al. (2020) report that \( q_{IR} \) decreases with increasing \( L_{\text{1.4 GHz}} \). 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 \( q_{IR} \) 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. 5). 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 across redshift from both an observational (e.g. Murphy 2009; 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 D 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 incidence of AGN activity is also presented (Sect. 5.3). Finally, we comment on the use of radio emission as a SFR tracer in the light of our results (Sect. 5.4).

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 $q_{IR}$ decreases by 0.25 dex (a factor of 1.8), at fixed redshift, and with high significance ($\sim 10\sigma$, see Eq. 5). This suggests that the dependence $L_{IR}/L_{AB}$ on $M_\star$ is steeper than the dependence $L_{IR}/M_\star$ (i.e. the MS). To translate this result into the corresponding IR-radio slope, we use our best $q_{LR}$-$M_\star$ relation (Eqs. 5) at fixed redshift, and assume for simplicity a linear MS between $M_\star$ and SFR (i.e. $L_{LR}$). This yields $L_{LR} \propto L_{AB}$; $M_\star$. In the past years, the deviation from the linear trend has been gaining increasing momentum, due to several studies finding a similar sub-linear behaviour in the local Universe ($L_{LR} \propto L_{AB}$; $M_\star$) (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. 5.4.

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

5.1.1. The role of the IMF

We quantify whether a deviation from a canonical IMF slope (e.g. Chabrier 2003; $n(M) \propto M^{-2.35}$ at $0.8 < M < 100 M_\odot$) could justify an $M_\star$-decreasing $q_{IR}$. 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 $q_{IR}$.

A top-heavy IMF has been directly constrained only in massive early-type galaxies at $z \sim 0$ (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 $q_{IR}$ 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 $L_{LR}/L_{AB}$; $M_\star$. 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) $q_{IR}$ variation across all $M_\star$. In line with the conclusions of Murphy (2009), we argue that a top-heavy IMF in the most massive galaxies proves insufficient to explain the evolving $q_{IR}$ with $M_\star$.

5.1.2. The role of the SFR surface density ($\Sigma_{SFR}$)

The model proposed by Schleicher & Beck (2013) postulates that the magnetic field strength scales with SFR, boosting radiative synchrotron emission during shocks or galaxy interactions (e.g. Donoskvi & Prodanović 2015; Tabatabaei et al. 2017). Because of the MS relation, this model implicitly predicts a net enhancement of radio emission with increasing $M_\star$, as well as an increase with redshift due to higher gas density in galaxies. Related to this, the single-zone galaxy model of Lacki & Thompson (2010), which includes a detailed CR description, suggests a variation of $q_{IR}$ as a function of SFR surface density ($\Sigma_{SFR}$), from “normal galaxies” ($\Sigma_{SFR} \lesssim 0.06 M_\odot$ yr$^{-1}$ kpc$^{-2}$) to “starbursts” ($\Sigma_{SFR} \gtrsim 2-4 M_\odot$ yr$^{-1}$ kpc$^{-2}$). In particular, this model argues that $q_{IR}$ slightly declines with $\Sigma_{SFR}$ (up to $\Sigma_{SFR} \sim 1 M_\odot$ yr$^{-1}$ kpc$^{-2}$) due to the escape of CRE, generating a radio dimming especially in lower $M_\star$ (or smaller size) galaxies. This effect is also expected to become more pronounced with redshift, due to IC scattering off the CMB that is expected to dominate over synchrotron cooling (Murphy 2009). Conversely, at $\Sigma_{SFR} \gtrsim 1 M_\odot$ yr$^{-1}$ kpc$^{-2}$, Lacki & Thompson (2010) invoke a “conspiracy” of ionization losses to balance spectral ageing, and additional synchrotron emission from secondary CRs, that together flatten $q_{IR}$ with $\Sigma_{SFR}$ and redshift.

Here we test the above models by relating $q_{IR}$ and average $\Sigma_{SFR}$ measured in this work. These estimates were obtained by using the total SFR as an input 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 $M_\star$ 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 $\theta_{FWHM}$ 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 $R_{ maj}$ can be calculated as $R_{ maj} = \theta_{FWHM} / 2.43$ (Murphy et al. 2017). Lastly, we take the circularized radius $R_e = R_{ maj} / \sqrt{A}$, where $A$ is the axial ratio.

Fig. 15 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 Appendix B). Our $R_e$ measurements are well consistent with VLA 3 GHz sizes independently 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 $q_{IR}$ in Fig. 16, colour-coded by redshift (left panel) and $M_\star$ (right panel). Empty symbols highlight $7/37$ bins with unresolved 3 GHz stacked maps, which translates into a lower limit in $\Sigma_{SFR}$. By fitting only the $q_{IR}$ and $\Sigma_{SFR}$ 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 ($q_{IR} \propto (-0.13\pm0.02) \times \log \Sigma_{SFR}$). From the $M_\star$ and redshift of each bin, we obtain a surrogate trend with rest-frame optical (5000Å) sizes estimated indirectly from the van der Wel et al. (2014) scaling relation for SFGs (grey dotted line). 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 $q_{IR}-\Sigma_{SFR}$ trend is linked to that with $M_\star$. Nevertheless, unlike the trend with optical sizes, our $\Sigma_{SFR}$ 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 $q_{IR}-\Sigma_{SFR}$ dependence is simply meant to be a more physical proxy for the observed $M_\star$ dependence. At fixed $M_\star$, the SFR surface density increases with redshift (left panel of Fig. 16), in qualitative agreement with our (weakly) decreasing $q_{IR}$ trend.

Both the slope and significance of the $q_{IR}-\Sigma_{SFR}$ relation are consistent with those found between $q_{IR}$ and $M_\star$ (Sect. 4.3). We argue that the declining $q_{IR}-\Sigma_{SFR}$ 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. 7), while $R_e$ only increases by a factor of 1.5–2.5 in the same interval. Thus 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.

The comparison with the model of Lacki & Thompson (2010) comes with a few caveats. First, we take a fixed spectral index $\alpha=0.75$ for all galaxies (which is supported by radio ancillary data in Appendix B), while Lacki & Thompson (2010) model a curved radio spectrum. Second, we label as “SBs” those galaxies that lie $>4\sigma$ above the MS (Sect. 5.2), while Lacki & Thompson (2010) identify them as $\Sigma_{SFR} > 2 \times 10^{-3} M_\odot yr^{-1} kpc^{-2}$. As we show in Fig. 16, the vast majority of our MS galaxies has $\Sigma_{SFR}$ below the “SB” threshold of Lacki & Thompson (2010).

That being said, their model predicts a decreasing $q_{IR}$ with $\Sigma_{SFR}$ (see their Fig. 1), that steepens with redshift, then followed by a flattening (or a reversal) at $\Sigma_{SFR} \approx 1 M_\odot yr^{-1} kpc^{-2}$. This behaviour is not clearly seen in our data, that instead display a smoothly declining $q_{IR}-\Sigma_{SFR}$ trend, and nearly redshift-invariant. Our trend is consistent with the low-$q$ values recently inferred by Algera et al. (2020b) in compact ($R_e \sim 1$ kpc) and massive ($M_\star > 10^{10} M_\odot$) sub-millimetre galaxies at 1.5$<z<3.5$. Indeed, their average $q_{IR} = 2.20\pm0.03$ lies close to the extrapolation of our best-fit $q_{IR}-\Sigma_{SFR}$ trend at $\Sigma_{SFR} \sim 100 M_\odot yr^{-1} kpc^{-2}$, thus further corroborating the relation between $q_{IR}$ and SFR per unit area in SFGs.

To explain low-$q_{IR}$ SBs, a further fine-tuning in the model of Lacki & Thompson (2010) is to invoke “puffy SBs” with larger disk scale height ($h=1$ kpc, i.e. SMC-like) than canonical “compact SBs” ($h=100$ pc, i.e. ULR-like). Indeed, in puffy SBs, CRe can travel longer distances before escaping the galaxy, creating secondary hadrons that induce an extra boost of radio emission. However, this process should globally steepen the observed radio spectra ($\alpha=0.9\pm1.0$), due to bremsstrahlung and ionization losses being weak with respect to synchrotron and IC losses. This prediction is again not confirmed by our data (Fig. B.2).

Another speculative hypothesis could be linked to the amplification of the magnetic field strength at higher SFRs, that boosts radio emission in more massive galaxies along the MS (e.g. Tabatabaei et al. 2017), though a fine-tuned balance between concomitant CRe losses and secondary CRe production is also required (Algera et al. 2020b).
In summary, our checks cannot firmly elucidate the main physical driver of the IRRC with $M_\star$, but they seem to support an empirical link between $q_{IR}$ and SFR surface density. Our findings do not seem to follow the $q_{IR}$ flattening or spectral index variations with $\Sigma_{SFR}$ predicted by models (e.g. Lacki & Thompson 2010; Schleicher & Beck 2013). Of course, our data do not have enough statistical power to discern all the underlying physical mechanisms and spectral variations that the model obviously addresses. We postulate that a more detailed data-vs-model comparison would require depth-matched observations at multiple radio frequencies of massive compact galaxies.

5.2. Does the IRRC evolve above the MS?

We investigate the behaviour of the average $q_{IR}$ above the MS. This is important to test whether radio emission follows a similar enhancement as $L_{IR}$ when moving above the MS, or instead $q_{IR}$ is not a good tracer of starburstiness (i.e. offset from the MS). This issue is still highly debated. For instance, Condon et al. (1991) found that the most extreme ULIRGs at $z=0$ have higher $q_{IR}$ and larger scatter compared to the MS population, which can be explained by flatter radio spectra due to free-free absorption (see also Murphy et al. 2013). On a different note, Helou et al. (1985) and Yun et al. (2001) do not report any significant deviation of $q_{IR}$ in local SB galaxies, though they also observed a larger scatter for this population. More recently, Magnelli et al. (2015) found a mild ($\pm0.2$ dex) enhancement of $q_{IR}$ above the MS, though not significant. Such apparent tension is probably also due to different definitions of ”starburst” galaxies and different sample selections.

For sake of consistency with Magnelli et al. (2015), in this Section we define ”SBs” as galaxies with SFR$>4\times$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 $>4x$ above the MS) lie 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 $q_{IR}$ 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. 17 shows the resulting $q_{IR}$ of SBs (circles) relative to the full SFG sample (MS+SB, stars) out to $z<2.5$, at $M_\star>10^{10.5} M_\odot$. For comparison, some previous IRRC trends are reported (black lines), as in Fig. 13. While some possible hints of $(-0.5$ dex) higher $q_{IR}$ in SBs could be present, these are consistent with MS analogues within 1$\sigma$ in all bins. Therefore, this test suggests that $q_{IR}$ evolves primarily with $M_\star$, irrespective of whether a galaxy is on or above the MS.

Though the (lack of) evolution of $q_{IR}$ above the MS is still debated, our decreasing $q_{IR}$-$\Sigma_{SFR}$ trend (Fig. 16) would predict lower $q_{IR}$ in SB than in MS galaxies, due to SBs being more compact. However, we note that our IR-detected SBs are both $>4x$ more star forming and smaller in size ($R_\star<1$ kpc at $z<2$, see Jiménez-Andrade et al. 2019) than MS analogues. Therefore, both parameters add to boost $\Sigma_{SFR}$. Since our SBs appear mostly unresolved in 3 GHz stacks, we can only place lower $\Sigma_{SFR}$ limits which prevent us from ruling out a possible flattening of $q_{IR}$ at the largest SFR surface densities. Higher resolution radio observations of these objects would be crucial to test such a behaviour.

On a side note, the sample of sub-millimetre galaxies for which Algera et al. (2020b) obtained an average $q_{IR}=2.20$ 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 $q_{IR}$, while our results predominantly reflect the behaviour of the MS population.

As mentioned in Sect. 5.1.2, Lacki & Thompson (2010) distinguish between ”compact SBs” ($h=100$ pc, ULIRG-like) similar to local merging galaxies, and ”puffy SBs” ($h=1$ kpc, SMG-like) with lower $q_{IR}$ values, that are more common at high-$z$ (Genzel et al. 2008). Nevertheless, a compact/puffy-SB transition above the MS should be reflected to a steepening of their radio spectra indices (Lacki & Thompson 2010), though we are unable to discern it from our data. In this respect, Magnelli et al. (2015) did not report any significant spectral index variation above the MS. Therefore, we caution that a simple dependence of $q_{IR}$ 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 or multiple mechanisms at play.
5.3. 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 ($M_{\star} < 10^{8.5} M_{\odot}$). As highlighted in Sect. 4.2.4 and Table 4, 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 differences suggest that radio-detected dwarves are a quite peculiar and not representative sub-sample 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$_{IR} > 4\times$ 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 $L_{1.4\,\text{GHz}}$ values than more massive counterparts, but still $100 \times$ 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$_{IR}$ 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 9 and 12). 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 & Deller 2012; Reines et al. 2014; 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 (Fakihouri 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).

5.4. 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$_{1.4\,\text{GHz}}$ is therefore not rigidly proportional to SFR.

For this reason, we express q$_{IR}$ 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 ($=\text{SFR}_{IR+UV}$). We thus define the parameter q$_{SFR}$ as:

$$q_{SFR} = \frac{q_{IR}}{L_{1.4\,\text{GHz}} [WHz^{-1}]}$$

where q$_{IR}$ is simply the SFR$_{IR+UV}$ scaled back to spectral luminosity units. This formalism enables us to keep similar units as for q$_{IR}$, while switching from luminosity to total SFR.

We repeated the analogous q$_{SFR}$ decomposition analysis at $M_{\star} > 10^{7.5} M_{\odot}$ 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 $-0.057 \pm 0.002$, which is strikingly similar to that inferred for q$_{IR}$ ($-0.055 \pm 0.018$, 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.

Using the same approach as for Eq. 5, the multi-parametric fitting in the q$_{SFR}$--$M_{\star}$--z plane yields the following expression:

$$q_{SFR}(M_{\star}, z) = (2.743 \pm 0.034) \times A^{(0.025 \pm 0.012)} - B \times (0.234 \pm 0.017)$$

(7)

where $A=(1+z)$ and $B=(\log M_{\star}/M_{\odot} - 10)$. Similarly to the fit in the q$_{IR}$ 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$_{1.4\,\text{GHz}}$, radio emission underestimates the total SFR by a larger factor as compared to the IR light. The sub-linear trend $L_{1.4\,\text{GHz}} \propto (1+z)^{0.9}$ 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 also Bell 2003; Hodge et al. 2008; Davies et al. 2017; Brown et al. 2017; Gürkan et al. 2018) becomes even steeper when further redshifts are considered, which could be possibly linked to stronger CRs scale heights (e.g. Helou & Bicay 1993; Lacki & Thompson 2010) or weaker magnetic fields (Donevski & Prodanović 2015; Tabatabaei et al. 2017) that are common in less dense SF environments.

I. Delvecchio et al.: The IRRC evolves primarily with $M_{\star}$
Moreover, our adopted $L_{IR}$-$SFR$ conversion does not account for the “cirrus” emission associated with cold dust heated by old (>A-type) stellar populations, which might lower the intrinsic SFR at fixed $L_{IR}$. However, this effect is expected to contribute in low-$sSFR$ galaxies, i.e. at high $M_*$ and low redshift (e.g. Yun et al. 2001). Hence, we expect it (if any) to further flatten the $q_{SFR}$-$z$ trend or to amplify the $M_*$ dependence of $q_{SFR}$.

In addition, we note that the lower efficiency in producing synchrotron emission in low-$SFR$, low-$M_*$ 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 $L_{1.4\,\mathrm{GHz}}$ and SFR functions. Therefore, our results reinforce the need for $M_*$-dependent, non-linear calibrations between radio-continuum emission and SFR, in order to develop successful observing strategies for targeting low-$M_*$ galaxies at radio wavelengths. These can be complemented with higher frequency observations that are more sensitive to thermal free-free emission as SFR tracer in high-redshift galaxies (e.g. Murphy et al. 2017, Penney et al. 2020, Van der Vlugt et al. 2020; Algera et al. 2020a).

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 workhouse 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+SFR 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 calibrate the IRRC of SFGs as a function of both $M_*$ and redshift, out to $z\sim4$. Starting from an $M_*$-selected sample of 413,678 galaxies SFGs selected via (NUV-$r$)/(r-$I$) colours in the COSMOS field, we leverage new de-blended IR/sub-mm data (Jiménez-Andrade et al. 2018), as well as deep radio images from the VLA COSMOS 3 GHz Large Project (Smołćich et al. 2017b).

In each $M_*$-$z$ 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 $q_{IR}$ as a function of $M_*$ and redshift (Sect. 4.1). We develop a recursive approach for identifying and then subtracting radio-excess AGN in different $M_*$ and redshift bins (Sect. 4.2). This technique is calibrated on a (>70%) $M_*$-complete subsample of 3 GHz detections at $M_*>10^{10.5} M_\odot$ 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 $M_*$, with more massive galaxies displaying systematically lower $q_{IR}$. A secondary, weaker dependence on redshift is also observed. The multi-parametric best-fitting expression is the following: $q_{IR}(M_*, z) = (2.646\pm0.024) \times (1+z)^{-0.023\pm0.008} \times (0.148\pm0.013) \times (\log M_*/M_\odot - 10)$. At fixed redshift, this trend translates into an IRRC of $L_{IR} \propto L_{1.4\,\mathrm{GHz}}^{-0.90}$, which corroborates the similar sub-linear behaviour reported in the literature (e.g. Bell 2003; Hodge et al. 2008; Güirkan et al. 2018). The typical scatter of the IRRC at $M_*>10^{10.5} M_\odot$ 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 $M_*$ and $z$.

2) Our recursive approach for removing radio AGN enables us to statistically decompose radio-detected SFGs and AGN (Figs. 10 and 11) as a function of $M_*$ and redshift. Removing radio AGN substantially flattens the observed $q_{IR}$-$z$ trend at $M_*>10^{10.5} M_\odot$ to a nearly flat slope. This correction nicely aligns the mode $q_{IR}$ of radio SFGs to the median stacked $q_{IR}$ of the full $M_*$ sample of non-AGN galaxies. Therefore, we interpret the resulting AGN-corrected $q_{IR}$ 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 $q_{IR}$ with redshift, and to induce an even steeper $M_*$ dependence, thus reinforcing our main findings.

3) The fraction of radio AGN identified within the full $M_*$ sample strongly increases with $M_*$, spanning from 0.4% to 6% across the full range (Table 4), in agreement with previous studies (e.g. Heckman & Best 2014). However, when limited to 3-GHz detected sources, about 90% of radio-detected dwarves ($M_*<10^{10.5} M_\odot$) are radio-excess AGN. We argue this is likely a selection effect induced by our 3 GHz-limited being biased towards the brightest radio sources in such low-$M_*$ systems. We test the reliability of our radio AGN identification owing to available VLBA data of radio AGN (Herrera Ruiz et al. 2017), confirming the AGN nature for 90% of them.

4) We examined the evolution of $q_{IR}$ as a function of SFR surface density ($\Sigma_{SFR}$), as a proxy for $M_*$, finding a very similar trend both in slope and statistical significance. In agreement with recent observations of high-redshift dusty SFGs (Algera et al. 2020b), our results support a decreasing $q_{IR}$ in MS galaxies towards higher $\Sigma_{SFR}$. Nevertheless, radio synchrotron models (e.g. Lacki & Thompson 2010; Schleicher & Beck 2013) predict a much stronger $q_{IR}$ evolution with redshift, and $\Sigma_{SFR}$ (i.e. $M_*$-) dependent radio spectral indices, neither of which are seen in our data. Another possibility links to magnetic field amplification in massive highly SFGs (Tabatabaei et al. 2017).

5) We compare the average $q_{IR}$ between MS galaxies and an $M_*$-complete subsample of SBS with SFRs $\sim4\times$ above the MS (Sect. 5.2). Despite SBS being more compact than MS analogues (Jiménez-Andrade et al. 2019), we do not observe a significant difference in $q_{IR}$, apparently at odds with our expectations. According to radio synchrotron models (Lacki & Thompson 2010), a “conspiracy” of different factors might induce a $q_{IR}$ flattening at $\Sigma_{SFR} \gtrsim 10 M_\odot \, \mathrm{yr}^{-1} \, \mathrm{kpc}^{-2}$. Our findings do not seem to support this prediction. However, our current data do not allow us to discriminate between various model scenarios in this $\Sigma_{SFR}$ regime. Alternatively, we postulate that SB galaxies might follow a different $q_{IR}$ relation with $\Sigma_{SFR}$ than MS analogues, in which multiple mechanisms 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 $q_{SFR}$-$M_*$ trend, leaving the evolution with redshift unchanged. These findings imply that using radio-synchrotron emission as a SFR tracer requires $M_*$-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. The authors are grateful to the referee for a detailed and constructive report that greatly helped us clarify the results and implications of this work. ID is supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 788679. ID thanks R. Gobat for useful discussions. MJJ acknowledges sup-
Appendix A: Testing total radio flux densities

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 flux densities 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 flux densities of each source, following the approach described in Sect. 3.3. Fig. A.1 displays the comparison between total flux densities (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 ~0.05 dex and dispersion of 0.12 dex. At 3<S/N<5, total flux densities 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 flux densities 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 $L_{1.4\,\text{GHz}}$ estimates. A comparison between median and mean $L_{1.4\,\text{GHz}}$ is presented in Fig. A.2. The top panel displays $L_{1.4\,\text{GHz}}$, 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 $M_*$ bins. Only stacks in which peak flux densities have S/N>3 are shown. No systematics is observed, at any $M_*$, between mean and median stacked $L_{1.4\,\text{GHz}}$. This is consistent with White et al. (2007), 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 $L_{1.4\,\text{GHz}}$ 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 flux densities, since these mean weighted-average flux densities could be biased towards fewer and brighter radio detections, which reduces the statistical weight of non-detections. Indeed, $L_{1.4\,\text{GHz}}$ of radio detections (top panel) are always $>3\times$ larger than $L_{1.4\,\text{GHz}}$ 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: Stacking ancillary VLA and MIGHTEE data

Here we perform a radio stacking analysis, as for 3 GHz data (Sect. 3.3), in order to check whether our 3 GHz based L_{1.4 GHz} are stable against different resolutions or spectral frequencies. We exploit VLA data from the 1.4 GHz Deep Project (Schinnerer et al. 2010) map. 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. In addition, we make use of 1.3 GHz MIGHTEE (Jarvis et al. 2016; I. Heywood et al. in prep.) data. MIGHTEE images formally reach 2.2 μJy beam^{-1} at 8.4"×8.6" 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).

Source flux densities in VLA 1.4 GHz and MIGHTEE 1.3 GHz maps were re-extracted, using K+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 flux densities 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 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. Stacked MIGHTEE flux densities are measured in the central pixel, that is assumed to trace the total flux. VLA 1.4 GHz peak flux densities were, instead, scaled to total flux densities 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_{peak}/S_{peak}^{3GHz} > 0.35^{-11}(S/N)_{peak}^{3GHz}. Because of the larger beamsize compared to 3 GHz, we find fewer resolved stacks (17/25). Total flux densities were converted to rest-frame L_{1.4 GHz}; assuming α=-0.75±0.1, that was propagated along with fluct errors to deliver reliable L_{1.4 GHz} uncertainties.

Upper limits at 3σ were assigned for S/N<3 stacks.

Fig. B.1 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⋆ (increasing from left to right). While stacking at 3 GHz delivers S/N>3 in 39/42 bins, only 25/42 and 29/42 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 flux densities 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). 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 0.1<z<0.5 (10^{11} < M⋆/M⊙ < 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 flux densities and comparing them with the previous measurements. The reason why this specific beamwidth was chosen is that beyond 3" it has already been shown that no significant missing flux is recorded at 3 GHz (see Table 2 in Delhaize et al. 2017). Of course, this convolution 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).

However, only the bin at the lowest z and highest M⋆ displays on average 0.3 dex larger total flux density, while the other bins show consistent estimates within the uncertainties. Since no extra flux is visible in the new residual image, we replaced the total flux density of that single bin with the 3" convolved value and used this value in the rest of our analysis. 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⋆ > 10^{11} M⊙ (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.

Appendix B.1: Considerations on the radio spectral index

We briefly discuss and test our assumption of taking a single spectral index α=-0.75 by comparing L_{1.4 GHz} estimates independently inferred from stacking the three above datasets. In Fig. B.2, 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⋆. 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⋆ range explored in this work. As a sanity check, the median spectral index 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 a M⋆-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⋆ or SFR, at least out to z~2. More recently, Calistro Rivera et al. (2017) ex-
Fig. B.2: Comparison between rest-frame 1.4 GHz spectral luminosity $L_{1.4 \text{ GHz}}$ 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 $\alpha = -0.75$ to scale flux densities from 3 GHz to 1.4 GHz. Colours indicate different $M_\ast$ ranges. Downward arrows with open symbols mark $3\sigma$ upper limits if S/N < 3. The broad agreement between the various datasets suggests that using a single $\alpha = -0.75$ is a reasonable assumption across the full $M_\ast$ range explored in this work.

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 in the observed range [150 MHz–1.4 GHz] (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 $L_{1.4 \text{ GHz}}$. Therefore, these studies provide mounting evidence that using a single power-law spectral index $\alpha = -0.75$ at our frequency is a reasonable assumption.

### Appendix C: 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 $q_\beta$ distribution ($q_{\beta, \text{peak}}$) 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-mm regime (e.g. Padovani et al. 2015; Smolčić et al. 2017b; Novak et al. 2018; Ceraj et al. 2018; Algera et al. 2020c), the contribution of radio-faint AGN to $q_{\beta, \text{peak}}$ 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 $M_\ast$-dependence of $q_\beta$, here we follow a different approach.

The observed $q_{\beta}$ distribution is fitted with two Gaussian functions simultaneously, which parameterize the contribution of SFGs and radio-excess AGN. Contrary to Sect. 4.2.2, we do not set the SF peak to $q_{\beta, \text{peak}}$, 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. C.1 for the two highest $M_\ast$ bins. As in Fig. 10, 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 $M_\ast$ bins. Secondly, the relative fraction of radio AGN that we mis-classify as SFGs (red tail at $q_{\beta, \text{peak}} - 2\sigma$) 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 $2\sigma$ threshold are about 80% at $10^{-5.5} < M_\ast / M_\odot < 10^{11}$ and 90% at $10^{11} < M_\ast / M_\odot < 10^{12}$.

Despite the lower level of purity of the SFG population, we emphasize 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 $q_{\beta, \text{peak}}$ and the dispersion reaching $-0.21$ dex. Therefore, it is reasonable to assume that the mode of the observed $q_{\beta}$ distribution is attributed to radio-detected SFGs. Related to this, the threshold $q_{\beta, \text{peak}} - 2\sigma$ 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
Appendix D: Differences compared to the literature

Our best-fit relation of $q_{IR}$ as a function of $M_*$ and redshift (Eq. 5 in Sect. 4.3) is fully consistent with the average $q_{IR}$ value measured in local SFGs (i.e. 2.64 in Bell 2003) for a typical galaxy with $M_*/M_\odot = 10^{10}$. At higher redshifts, instead, our average $q_{IR}$ measurements follow flatter evolutionary trends compared to previous studies (Fig. 13), while the best-fit normalization appears broadly consistent with the literature only at $M_*>10^{10.5} M_\odot$. 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 $M_*$ and redshift bin; (ii) exploiting an $M_*$-selected sample of SFGs; (iii) binning the derived $q_{IR}$ as a function of both $M_*$ and redshift. To test our results against different techniques, we expand on each of these aspects below.

Appendix D.1: Radio AGN subtraction

In Sect. 4.2, we performed a recursive subtraction of radio AGN as a function of $M_*$ and redshift, carefully calibrated on high-$M_*$ galaxies, and then extrapolated to lower $M_*$ 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 redshift $z\approx 2$. 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 $M_*$-selected sample, of which radio detections make a negligible fraction.

We compare our $q_{IR}$ with mock measurements obtained by following the stacking method of Magnelli et al. (2015), but applied to the sample used in our work. Fig. D.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 $L_{1.4\ GHz}$ 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 $q_{IR}$ trends. The colour bar highlights the average $M_*$ of each bin. Out of 37 bins analyzed in this work, 35 yield a S/N>3 from median $3\ GHz$ stacking (circles), while 3$\sigma$ 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 $\sim 0.1$ dex might be present at the highest $M_*$, 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 $M_*$-selected sample in that $M_*$ bin ($\sim 45\%$, Table 4). Therefore, this test proves our radio AGN subtraction broadly consistent with a totally independent approach.

Fig. D.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 $M_*$ ranges are colour-coded, while downward arrows mark 3$\sigma$ 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. D.2: Median $q_{IR}$ as a function of redshift obtained by analysing the SFG sample of Delhaize et al. (2017, stars). Black lines indicate the median $q_{IR}$ trend of Delhaize et al. before (dot-dashed) and after (triple dot-dashed) removing 2$\sigma$ outliers. The grey solid line marks the result of a quasar-$z$ trend of Delhaize et al. before (dot-dashed) and after (triple dot-dashed) removing 2$\sigma$ outliers. The grey solid line marks the resulting best-fit $q_{IR}$ 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 $M_*$ of the underlying sample.
Appendix D.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\textgreater{}5 in at least one PACS or SPIRE band) and radio (VLA 3 GHz with S/N\textgreater{}5; Smolčić et al. 2017b) sample of SFGs in the COSMOS field, out to $z\sim5$. By performing double-censored survival analysis to account for sources undetected at either radio or FIR wavelengths, they found an evolving $q_{\text{IR}} \propto (1+z)^{-0.19\pm0.01}$, which flattens to $q_{\text{IR}} \propto (1+z)^{-0.12\pm0.01}$ after removing 2$\sigma$ 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. D.1 that median radio stacking yields broadly consistent results (see Magnelli et al. 2015). Fig. D.2 displays the median $q_{\text{IR}}$ obtained by stacking the SFG sample of Delhaize et al. (2017) in different redshift bins (stars). This yields a best-fitting $q_{\text{IR}} \propto (1+z)^{-0.11\pm0.05}$, that is fully consistent with the flatter trend of Delhaize et al. (2017) after removing 2$\sigma$ outliers (triple dot-dashed line). This check proves our technique solid against different sample selections from the literature.

Fig. D.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 $M_\star \geq 10^{10.5} M_\odot$ (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 $q_{\text{IR}}$ 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.