# Galaxy-AGN co-evolution

GIANFRANCO DE ZOTTI INAF – OSSERVATORIO ASTRONOMICO DI PADOVA

### Layout

- Introduction to AGNs
- AGN host galaxies
- BH mass estimates
- BH growth. The Soltan argument
- How can AGNs affect the host galaxy?
- AGN impact on galaxy evolution. Feedback
- Do stellar and BH masses really grow in parallel?
- Conclusions

#### The quasar discovery

Maarten Schmidt (1963, Nature): 3C273

The stellar object is the nuclear region of a galaxy with a cosmological redshift of 0.158, corresponding to an apparent velocity of 47,400 km/s. The distance would be around 500 megaparsecs, and the diameter of the nuclear region would have to be less than 1 kiloparsec. This nuclear region would be about 100 times brighter than the luminous galaxies which have been identified with radio sources so far

3C273







## The luminosity of active nuclei is due to accretion onto black holes.

#### *Ya. B. Zeldovich (1963)*

- The total energy output from a quasar is at least the energy stored in its radio halo ( $\approx 10^{54} \text{ J}=10^{61} \text{ erg}$ ); via E=mc<sup>2</sup> this corresponds to 10<sup>7</sup> M<sub>sun</sub>.
- Nuclear reactions have at best an efficiency of 0.7% (H burning). So the waste mass left behind in powering a quasar is >  $10^9 M_{sun}$ .
- Rapid brighness variations show that a typical quasar is no bigger than a few light-hours. But the gravitational energy of 10<sup>9</sup> M<sub>sun</sub> compressed within this size is  $\approx 10^{55}$  J, i.e. 10 times larger than the fusion energy.

"Evidently, although our aim was to produce a model based on nuclear fuel, we have ended up with a model which has produced more than enough energy by gravitational contraction. The nuclear fuel has ended as an irrelevance."

Donald Lynden-Bell (1969)

The idea that active galactic nuclei (AGNs) are powered by accretion onto supermassive BHs was elaborated by several authors (Hoyle & Fowler 1963; Salpeter 1964; Zel'dovich 1964; Lynden-Bell 1969, 1978; Lynden-Bell & Rees 1971) and quickly gained acceptance.



Although it was clear from the beginning that quasars are located in the nuclei of galaxies, their presence was considered for decades just as an incidental diversion, an ornament irrelevant for galaxy formation and evolution:

- The physical scale of the AGNs is incomparably smaller than that of galaxies: typical radii of the stellar distribution of galaxies are of several kpcs, to be compared with the Schwarzschild radius  $r_s=2GM_{BH}/c^2\approx 9.56\times 10^{-6} (M_{BH}/10^8 M_{sun})$  pc; i.e.  $r_{BH}\sim 10^{-9} r_{gal}$ .
- The radius of the "sphere of influence" of the BH (the distance at which its potential signifcantly affects the motion of the stars or of the ISM) is also small:  $r_{inf}=GM_{BH}/\sigma_*^2\approx 11 (M_{BH}/10^8 M_{sun})/(\sigma_*/200 \text{ km/s})^2 \text{ pc.}$  Hence even large BH have a negligible impact on the global stellar and ISM dynamics. Even in nearby galaxies the angular scale associated to  $r_{inf}$  is small, ~ 0.2 ( $M_{BH}/10^8 M_{sun}$ ) ( $\sigma_*/200 \text{ km/s}$ )<sup>-2</sup> (D/10 Mpc)<sup>-1</sup> arcsec.
- Quasars are much rarer than galaxies at low redshifts, where galaxies were most extensively studied, and become much more numerous at  $z \ge 2$  (evolution much stronger than that of galaxies).

#### **BH demographyc studies**

- On the other hand, the BHs powering the quasars at high z do not disappear. After they stop accreting, they should live essentially forever as dark remnants. So dead quasar engines should hide in many nearby galaxies (Schmidt 1978).
- However, the required BH masses are a tiny fraction of the stellar mass (let alone the total mass!) of galaxies (~0.1% M<sub>\*</sub>) and their radius of influence is very small. Thus dynamical evidence for BHs is hard to find; the first stellar dynamical BH detections followed in the mid- to late-1980s, when CCDs became available on spectrographs and required the excellent seeing of observatories like Palomar and Mauna Kea (see Kormendy & Richstone 1995 for a review).
- The Hubble Space Telescope (HST), by delivering five-times-better resolution than ground-based, optical spectroscopy, made it possible to find BHs in many more galaxies.
- This led to the convincing conclusion that **BHs are present in essentially every** galaxy that has a bulge component.

#### AGN host galaxies



Kormendy J, Ho LC. 2013.

Annu. Rev. Astron. Astrophys. 51:511–653

Optical *R*-band Hubble Space Telescope images of *z* ~0.1 Seyfert 1 galaxies and quasars. Some of these nearby AGNs have elliptical hosts, but most have hosts that are early-type spirals, which have a substantial bulge component.



R Kormendy J, Ho LC. 2013. Annu. Rev. Astron. Astrophys. 51:511–653





Host galaxy morphologies of 94 intermediate-redshift  $(0.2 \le z < 1.2)$  AGNs, selected using Chandra Xray data in the Extended Groth Strip. Using the second-order moment of the brightest 20% of a galaxy flux (M20), sensitive to the light concentration, and the Gini coefficient, which measures the distribution of flux among a galaxy's pixels, as morphology measures, Pierce et al. (2007) found that X-ray-selected AGNs mostly reside in E/So/Sa galaxies (53%), i.e. in galaxies with prominent -2.5 bulges.

#### **BH mass estimates**



- Because of the smallness of the BH sphere of influence, resolving it is possible only for relatively nearby, very massive BHs.
- Dynamical estimates based on line widths may not be reliable because of contributions to the mass from other components (dense star clusters, dark matter, ...).
- Completely reliable estimates would require resolved proper motions of surrounding stars but so far this could be achieved only for the Milky Way.

Image of the Milky Way Center (Genzel et al. 2010





Mass distribution in the Galactic center, assuming an 8 kpc distance. The filled circles at the shortest projected distances denote the masses derived from the orbits of S2 and S12. The solid curve shows the overall best-fit model to all data. It is the sum of a  $2.87 \times 10^6 \,\mathrm{M_{\odot}}$ point mass and a stellar cluster of central density  $3.6 \times 10^6 \,\mathrm{M_{\odot}}$ pc<sup>-3</sup>, core radius 0.34 pc, and power-law index  $\alpha = 1.8$ . The long-dash-short-dashed curve shows the same stellar cluster separately. The dashed curve shows the sum of the visible cluster and a Plummer model of a hypothetical very compact (core radius ~0.00019 pc) dark cluster of central density  $2.2 \times$  $10^{17} \,\mathrm{M_{\odot}} \,\mathrm{pc}^{-3}$ .



Innermost radius (normalized to the Schwarzschild radius) probed by measurements which led to the detection of super-massive BHs in nearby galactic nuclei, plotted against the inferred mass density within that region. The detection in the Milky Way is shown as the solid triangle, the water maser detection of the Seyfert 2 galaxy NGC 4258 is shown as a solid circle, detections based on stellar and gas dynamics are shown as open circles and triangles, respectively. The solid curves show the maximum lifetime of a dark cluster against collision and evaporation, using the prescription in Maoz (1998), from 10<sup>4</sup> to 10<sup>13</sup> years. The thick solid line represents a lifetime of 15 billion years.

### BH required?

Apart from the Milky Way, the presence of a super-massive BH is required for:

- A handful of objects with VLBI measurements (milliarcsec resolution, i.e. ~100 times better than the HST) of the  $H_2O$  mega-maser at 22 GHz from circum-BH molecular gas disks. Observations of NGC 4258 (Miyoshi et al. 1995) established this as one of the most powerful techniques of measuring BH masses. Useful maser disks are however rare, not least because they must be edge-on, and the orientation of the host galaxy gives no clue about when this is the case. Also in several cases the disk mass was found to be comparable to the BH mass, a major complication for BH mass measurements. However, progress on this subject has been accelerating in recent years (Kuo et al. 2011, Gao et al. 2017).
- Asymmetric K $\alpha$  iron line. X-ray observations have shown that AGNs possess a strong, broad Fe K $\alpha$  line at 6.4 keV (rest-frame). Variability shows that the line is produced in the innermost part of the accretion disk that may extend to the innermost stable circular orbit. The line than experiences general-relativistic effects, such as gravitational redshift. If the BH is rotating its angular momentum manifests through Lense-Thirring precession which occurs only in the innermost part of the accretion disk where the spacetime becomes twisted in the same direction that the BH is rotating. The BH spin is measured by the parameter a=cJ /GM<sup>2</sup>. Negative spin values represent retrograde configurations in which the BH spins in the opposite direction to the disk, positive values denote prograde spin configurations, and a = 0 implies a non-spinning BH.



Rotation curve traced by the  $H_2O$ mega-maser clouds in NGC 4258. Units are milliarcsec, with 1 mas corresponding to 0.035 pc at the distance of the galaxy. The masers have almost exactly a Keplerian velocity (V $\propto$  r<sup>-1/2</sup>), so it is likely that the disk mass is negligible and that  $M_{BH} = V^2 r/G.$ 

Velocity



Estimates of the half-light radii of the X-ray and optical emitting regions of the doubly-imaged lensed quasar Q 0158-4325 obtained from optical and X-ray monitoring of its rapid microlensing variability (Morgan et al 2012). Most of the X-ray emission from luminous accreting BHs comes from within 20 gravitational radii. The effective emission radius is several times smaller if the BH is rapidly spinning. Large spacetime curvature causes strong light bending and large gravitational redshifts. The hard X-ray, power-law-emitting corona irradiates the accretion disc generating an X-ray reflection component. Atomic features in the reflection spectrum allow gravitational redshifts to be measured (Fabian 2013).



Change in the shape of the Fe K $\alpha$  line as a function of BH spin. The black solid line represents a =J/M= -0.998, the red dashed line shows a=0.0 and the blue dotted line shows a=+0.998. Note the enhancement in the breadth of the red wing of the line as the black hole spin increases.

Suzaku (black) and XMM-Newton (red) data showing the broad Fe K $\alpha$  line in MCG-6-30-15 ratioed against a power-law continuum. A lower limit to the spin parameter  $a \ge 0.97$  was derived (Brenneman 2013).



Most AGN BH mass estimates are based on single-epoch spectra. This assumes that the broad-line region is virialized, the continuum luminosity is used as a proxy for the BLR radius, and the broad line width (FWHM) is used as a proxy for the virial velocity. The figure compares virial masses between two different lines. The left panels are one-to-one plots, where the contours are for a grid size of  $\Delta = 0.1$  on both axes. The right panels show the distribution of mass ratios between two lines, and the mean and  $1\sigma$  from a Gaussian fit to the distribution are indicated in the top-left corner. From Shen, Y. et al. 2011 ApJS 194

### Results from BH demographic studies

- Most important is the agreement between the global volume density of BH mass with that predicted by the **Soltan argument**. Soltan (1982) pointed out that the luminosity function of QSOs as a function of redshift traces the accretion history of these BHs.
- For an assumed mass-to-energy conversion efficiency, the luminosity function at any given redshift translates into an accreted mass density at that redshift. Integrating such mass density over redshift, gives a present day accreted mass density, which is a lower limit to the present day BH mass density since the BH mass can have also increased non-radiatively (e.g. via BH mergers).
- Comparison of the accreted mass density with the local BH mass density provides considerable insight into the formation and growth of massive BHs.

### The "Soltan" approach - 1

If a BH is accreting at a rate  $\dot{M}$ , its emitted luminosity is

 $L = \epsilon \dot{M}_{\rm acc} c^2$ 

where  $\epsilon$  is the radiation efficiency, i.e. the fraction of the accreted mass which is converted into radiation and thus escapes the BH.

The growth rate of the BH,  $\dot{M}$ , is thus given by

 $\dot{M} = (1 - \epsilon) \dot{M}_{\rm acc}.$ 

Let's neglect any process which, at time t, might 'create' or 'destroy' a BH with mass M. In particular, **this means neglecting BH merging**. Indeed, in the merging process of two BH's,  $M_1 + M_2 \rightarrow M_{12}$ , means that BH's with  $M_1$  and  $M_2$  are destroyed while a BH with  $M_{12}$  is created.

#### The "Soltan" approach - 2

Then, if  $\epsilon$  is constant we have:

$$\rho_{\rm BH} = \frac{1-\epsilon}{\epsilon c^2} U_T$$

where  $U_T$  is the total *comoving* energy density from AGNs (not to be confused with the total *observed* energy density), given by

$$U_T = \int_0^{z_s} dz \frac{dt}{dz} \int_{L_1}^{L_2} L\phi(L, z) \, dL \, .$$

Here  $\phi(L, z) dL$  is the **bolometric** AGN luminosity function. Note the factor  $(1 - \epsilon)$  which is needed to account for the part of the accreting matter which is radiated away during the accretion process. If BH mergers, which don't yield electromagnetic radiation but only gravitational waves, are important, the derived  $\rho_{\rm BH}$  is a lower limit.





Optical (g-band) AGN luminosity function (Palanque-Delabrouille 2013)





Bolometric correction (Marconi et al. 2004)



**Bolometric** corrections for B band, mid-IR, and soft and hard X-ray bands, determined from a number of observations as a function of luminosity. The shaded areas show the dispersion in the distribution of bolometric corrections at fixed L.

Marconi et al. (2004) used luminosity functions in the optical B-band, soft X-ray (0.5–2 keV) and hard X-ray (2–10 keV) band (Ueda et al. 2003) transformed into a bolometric luminosity function using bolometric corrections obtained from template spectral energy distributions (SEDs). In addition, they constrained the redshift-dependent X-ray luminosity function to reproduce the hard X-ray background, which can be considered an integral constraint on the total mass accreted over the cosmic time and locked in super-massive BHs (SMBHs). They obtained:

$$\rho_{\rm BH} = (4.7 - 10.6) \left[ \frac{(1 - \epsilon)}{9\epsilon} \right] \times 10^5 \,\,\mathrm{M_{\odot} \, Mpc^{-3}},$$

consistent with their own estimate from the local BH mass function,  $\rho_{\rm BH} = (3.2 - 6.5) \times 10^5 \, {\rm M}_{\odot} \, {\rm Mpc}^{-3}$ , for the 'canonical' value  $\epsilon = 0.1$ .

Similar conclusion reached by Ueda et al. (2014) using updated X-ray luminosity functions, still constrained to reproduce the X-ray background, and comparing with the local SMBH mass density by Vika et al. (2009),  $\rho_{\rm BH} = (4.9 \pm 0.7) \times 10^5 \, {\rm M}_{\odot} \, {\rm Mpc}^{-3}$ , from the empirical relation between SMBH mass and host-spheroid luminosity (or mass). However, recent analyses of BH mass measurements and scaling relations concluded that the black-hole-to-bulge mass ratio, shows a mass dependence and varies from 0.1–0.2% at  $M_{\rm bulge} \simeq 10^9 M_{\odot}$  to  $\simeq 0.5\%$  at  $M_{\rm bulge} \simeq 10^{11} M_{\odot}$  (Graham & Scott 2013; Kormendy & Ho 2013). A similarly large median  $M_{\rm BH}/M_{\rm bulge}$  ratio for early type galaxies was found by Savorgnan et al. (2016) who however reported a substantial decrease of the ratio with decreasing stellar mass of the bulges of late-type galaxies.

- The revised normalization is a factor of 2 to 5 larger than previous estimates ranging from ≈ 0.10% (Merrit & Ferrarese 2001; McLure & Dunlop 2002; Sani et al. 2011) to ≈ 0.23% (Marconi & Hunt 2003), therefore resulting in an overall increase in the effective ratio, which is dominated by massive bulges.
- This would imply either a lower mean radiative efficiency or an important nonradiative (e.g. merging) contribution to the BH growth. Radiatively inefficient processes (i.e. slim accretion disks) have been independently advocated to explain the fast growth of SMBH in the early Universe (e.g. Madau et al. 2014).
- On the other hand, it was pointed out the bulge-disk decomposition can lead to a considerable underestimate of the spheroid luminosity and stellar mass (Savorgnan & Graham 2016) and the selection bias can lead to an overestimate of  $M_{BH}$  (Shankar et al. 2016). Both effects lead to an overestimate of the  $M_{BH}/M_*$  ratio. According to Shankar et al. (2016) the selection bias leads to an estimate of  $M_{BH}$  by at least a factor of 3.

- Based on a comprehensive analysis of the co-evolution of galaxies and SMBHs throughout the history of the universe by a statistical approach using the continuity equation and the abundance matching technique Aversa et al. (2015) found a mean  $M_{BH} M_*$  relation systematically lower by a factor  $\approx 2.5$  than that proposed by Kormendy & Ho (2013).
- The BH-to-stellar mass ratio was found to evolve mildly at least up to z ≤ 3, indicating that the BH and stellar mass growth occurs in parallel by in situ accretion and star formation processes, with dry mergers playing a marginal role at least for the stellar and BH mass ranges for which the observations are more secure.
- In conclusion, although issue is still debated, the emerging picture of the BH-galaxy co-evolution indicates that the AGN radiative efficiency,  $\varepsilon$ , varies with the system age, although how this happens is not yet clear.
- The global consistency between the BH mass density inferred from the "Sołtan" approach and from the local BH mass function constrains predictions on the gravitational wave signals expected from black hole mergers.

- In the case of thin-disk accretion,  $\varepsilon$  may range from 0.057 for a nonrotating BH to 0.32 for a rotating Kerr BH with spin parameter of 0.998 (Thorne 1974). During a coherent disk accretion, the BH is expected to spin up very rapidly, and correspondingly the efficiency is expected to increase up to  $\approx$  0.3
- On the other hand, when the mass is flowing towards the BH at high rates (super-Eddington accretion), the matter accumulates in the vicinity of the BH and the accretion may happen via the radiatively-inefficient `slim-disk' solution (Abramowicz et al. 1988, Begelman 2012, Madau et al. 2014, Volonteri et al. 2015) that speeds up the growth of SMBHs. This would relieve the challenge set by the existence of billion-solar-mass black holes at the end of the reionization epoch. The most distant quasar discovered to date, ULAS J1120+0641 at a redshift *z* = 7.084 is believed to host a black hole with a mass of 2.0(+1.5, -0.7)×10<sup>9</sup> M<sub>sun</sub>, 0.78 Gyr after the big bang (Mortlock et al. 2011; see, e.g., Haiman 2013 for a review).
- The need of a radiatively inefficient phase is however debated. For example, Trakhtenbrot et al.(2017) argue that "the available luminosities and masses for the highest-redshift quasars can be explained self-consistently within the thin, radiatively efficient accretion disk paradigm".

### BH – host galaxy correlations

- A tight correlation betwee  $M_{BH}$  and the galaxy velocity dispersion,  $\sigma$ , was reported, independently, by Ferrarese & Merritt (2000) and by Gebhardt et al. (2000).
- Both papers claimed that the scatter was only 0.30 dex over almost 3 orders of magnitude in  $M_{\rm BH}$  and no larger than expected on the basis of measurement error alone. This suggested that the most fundamental relationship between BHs and host galaxies had been found and that it implies that BH growth and bulge formation are closely linked. Many papers have expanded on this result with bigger samples (see Kormendy & Ho 2013 for a review).
- The  $M_{BH} \sigma$  correlation has been confirmed to be the strongest, with the lowest measured and intrinsic scatter (Saglia et al. 2016). The correlations with the bulge mass,  $M_{Bu}$ , is also strong, except for the pseudo-bulge subsample (while bulge properties are indistinguishable from those of elliptical galaxies, except that they are embedded in disks, pseudo-bulges have more disk-like properties and are though to be made by slow ("secular") evolution internal to isolated galaxy disks).





The ellipses show the 10 errors. The labels name particularly deviant galaxies. The solid lines indicate the best-fit relations the ellipticals and classical bulges. The dotted lines indicate the estimated intrinsic scatter. Arrows describe the effect of an equal-mass dry merger (red), of a sequence of minor mergers doubling the bulge mass (orange), an equal-mass, gas-rich merger of two spiral galaxies with 20% bulge mass with bulge-scales ratio 3 (blue) or 0.5 (dashed blue), and doubling the BH mass through accretion or BH merging (black).
## Correlations with classical bulges

• Saglia et al. (2016) find, for ellipticals and classical bulges:

 $\log(M_{BH}/M_{sun}) = (4.868 \pm 0.32)\log(\sigma/km \text{ s}^{-1}) - (2.827 \pm 0.75)$ 

 $\log(M_{BH}/M_{sun}) = (0.846 \pm 0.064)\log(M_{bulge}/M_{sun}) - (0.713 \pm 0.697)$ 

- Both relationships agree, within the errors, with those derived by Kormendy & Ho (2013); however, while the latter find a mean  $M_{BH}/M_{bulge}$  ratio slowly decreasing with increasing  $M_{bulge}$ , Kormendy & Ho find a slowly increasing trend.
- The normalization of the  $M_{BH} M_{bulge}$  relation is currently debated. The main issues are the difficulties of removing the disk component to derive  $M_{bulge}$  and, even more, the bias on  $M_{BH}$  estimate. We will come back to that later.
- An inspection of the figures shows that: i) core ellipticals have more massive BHs than other classical bulges, at a given  $\sigma$  or bulge mass; moreover, the smallest intrinsic and measured scatter of the  $M_{\rm BH}-\sigma$  and  $M_{\rm BH}-M_{\rm bulge}$  relations are measured for the sample of core ellipticals; ii) power-law early-type galaxies and classical bulges follow similar  $M_{\rm BH}-\sigma$  and  $M_{\rm BH}-M_{\rm bulge}$  relations; iii) pseudo-bulges have smaller BH masses than the rest of the sample at a given  $\sigma$  or

M<sub>bulge</sub>.



BH mass, M<sub>•</sub>, versus the K-band absolute magnitude of the host galaxy bulge and versus  $\sigma$ , of the host bulge averaged inside the radius that contains one-half of the bulge light. Black dots: elliptical galaxies Red dots: classical bulges Blue dots: pseudo-bulges Green points are for galaxies that contain neither a classical bulge nor a large pseudo-bulge but only a nuclear star cluster. Open symbols represent galaxies for which we have only upper limits on M.

**Disks do not correlate with M. and it is unclear whether pseudo-bulges do**: BH masses do not "know about" galaxy disks (Kormendy & Bender 2011).



Evolution of the AGN bolometric luminosity density compared with the evolution of the luminosity density due to star formation. The red band has been computed from a compilation of X-ray luminosity functions and assuming the Marconi et al. (2004) bolometric correction.The blue solid line is the Aird et al. (2015) determination.

The cyan band is the average luminosity density due to star formation based on a compilation from Santini et al. (2009), Gruppioni et al. (2015), Bouwens et al. (2011, 2015). The black solid line is the Madau & Dickinson (2014) determination.



The growth histories of the stellar mass (left panel) and of the AGN luminosity (hence of BH mass) share further similarities: the most massive galaxies form in intense starbursts at high redshift while less massive galaxies have more extended star formation histories that peak later with decreasing mass. This 'anti-hierarchical' nature (*downsizing*) is mirrored in BH growth: the most massive BHs likely grow in intense quasar phases which peak in the early universe, while less massive BH have more extended, less intense growth histories that peak at lower redshift.

### How can AGNs affect the host galaxy? - 1

As we have seen, the smallness of the radius of influence means that the BH's gravity has a completely negligible effect on its host galaxy. On the other hand, the energy released by the AGN

$$E_{\rm BH} \simeq \epsilon M c^2 \sim 2 \times 10^{61} \frac{\epsilon}{0.1} \frac{M_{\rm BH}}{10^8 M_{\odot}} \,\mathrm{erg},$$

where  $\epsilon$  is the mass to radiation conversion efficiency, is far larger than the gas binding energy. Setting  $M_{\text{gas}} = f M_{\text{bulge}}$ , with f < 1, we have

$$E_{\rm gas} \sim \frac{3}{2} f M_{\rm bulge} \sigma^2 \sim 1.2 \times 10^{58} f \frac{M_{\rm bulge}/M_{\rm BH}}{10^3} \frac{M_{\rm BH}}{10^8 \, M_{\odot}} \left(\frac{\sigma}{200 \, \rm km/s}\right)^{-2} \, \rm erg,$$

where  $\sigma$  is the line-of-sight velocity dispersion (the corresponding 3D velocity is  $v = \sqrt{3}\sigma$ ). This means that only a few percent of the BH energy output may have a strong influence on the gas in the host galaxy, potentially expelling it and, at the same time, limiting its own growth.

### How can AGNs affect the host galaxy? - 2

There are two main ways that the SMBH energy release can potentially affect its surroundings.

- By far the stronger one (in principle) is through direct radiation. This is particularly effective during heavily dust-obscured phases of AGN evolution (Fabian, Wilman & Crawford 2002).
- The second form of coupling SMBH binding energy to a host bulge is mechanical. The huge SMBH accretion luminosity may drive powerful gas flows into the host, impacting into its interstellar medium (ISM).
- A well know form of flow often is jets—highly collimated flows driven from the immediate vicinity of the SMBH (radio-mode feedback). However to affect most of the bulge requires a way of making the interaction relatively isotropic, perhaps with changes of the jet direction over time. Moreover, radio observations show that the jet energy is dissipated on scales from several kpc to Mpc, i.e. on scales larger than that of a galaxy. Therefore they are more relevant for heating the intergalactic medium (IGM) in galaxy clusters.
- A form of mechanical interaction that has automatically the right property are nearisotropic winds carrying large momentum fluxes. Such winds are indeed observed in many AGNs, as we will see.

Radio-optical image of Centaurus A. The jets transport energy from the nucleus to the radio lobes well beyond the boundary of the galaxy, before being stopped by the intergalactic medium surrounding the galaxy

From Alan Bridle's (NRAO) Images of Radio Galaxies and Quasars

## Energy-driven flows - 1

Silk & Rees (1998) pointed out that an AGN at the Eddington limit can prevent accretion into a galaxy at the maximum possible rate provided that

$$M_{\rm BH} \sim \frac{f\sigma^5 \sigma_{\rm T}}{4\pi G^2 m_{\rm p} c},$$

where  $\sigma_{\rm T}$  is the Thomson cross section for electron scattering. The galaxy bulge is assumed to be isothermal with radius r, so that its mass is  $M_{\rm bulge} = 2\sigma^2 r/G$ . The gas mass can be written as  $M_{\rm gas} = f M_{\rm bulge}$  with f < 1. The maximum collapse rate of the gas is  $\dot{M}_{\rm gas} = M_{\rm gas}/t_{\rm free-fall}$  with  $t_{\rm free-fall} = r/\sigma$ . The corresponding power is

$$\dot{E}_{\rm gas} = 1/2 (M_{\rm gas}/t_{\rm free-fall}) v^2 = 3f\sigma^5/G,$$

 $(v = \sqrt{3}\sigma).$ 

### Energy-driven flows - 2

The relation  $M_{\rm BH}-\sigma$  then follows equating  $\dot{E}_{\rm gas}$  to the Eddington luminosity

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} m_p c}{\sigma_T},$$

where G,  $m_p$  and c are the gravitational constant, the proton mass and the speed of light. Plugging in the numbers and considering that only a fraction,  $f_{\rm Edd}$ , of the Eddington luminosity can be used to prevent accretion we get:

$$M_{\rm BH} = \frac{3.6 \times 10^5}{f_{\rm Edd}} \left(\frac{\sigma}{100 \,\rm km/s}\right)^5 \,M_{\odot}.$$

A comparison with the empirical  $M_{\rm BH}$ – $\sigma$  relation shows that  $f_{\rm Edd}$  at the few/several percent level is enough to stop the accretion and to expel the gas from the host galaxy.

## Momentum-driven flows - 1

Alternatively, we may have momentum or force (instead of energy) balance (Fabian 1999, Fabian, Wilman & Crawford 2002, King 2003, 2005, Murray, Quataert & Thompson 2005). Balancing the outward radiation force with the inward one due to gravity gives

$$\frac{4\pi G M_{\rm BH} m_{\rm p}}{\sigma_{\rm T}} = \frac{L_{\rm Edd}}{c} = \frac{G M_{\rm gal} M_{\rm gas}}{r^2} = \frac{f G M_{\rm gal}^2}{r^2} = \frac{f G}{r^2} \left(\frac{2\sigma^2 r}{G}\right)^2$$
  
i.e.  
$$\frac{4\pi G M_{\rm BH} m_{\rm p}}{\sigma_{\rm T}} = \frac{4f\sigma^4}{G},$$

from which we get

$$M_{\rm BH} = \frac{f\sigma^4 \sigma_{\rm T}}{\pi G^2 m_{\rm p}} = 1.43 \times 10^9 f \left(\frac{\sigma}{100 \,\rm km/s}\right)^4 \,M_{\odot}.$$

## Momentum-driven flows - 2

- Thus in the case of momentum-driven flows, the BH mass required to stop the accretion is a factor  $\sim c/\sigma$  larger than in the energy-driven case.
- A comparison with the empirical  $M_{BH}-\sigma$  relation shows that in this case, even the full Eddington luminosity is not enough to unbind the gas unless the gas fraction f < 0.1 or the AGN has a strongly super-Eddington luminosity.
- That the full AGN luminosity goes into radiation pressure is expected in the case of `Compton-thick' objects, most of whose radiation is absorbed by the gas.
- In the radiation-driven case it is implicitly assumed that the cooling of the BH wind is negligible. Under what condition this applies?

## Properties of the winds - 1

We can crudely model the outflows as quasi-spherical winds from SMBHs accreting at about the Eddington rate

$$\dot{M}_w \simeq \dot{M}_{\rm Edd} = \frac{L_{\rm Edd}}{\epsilon c^2} \simeq 0.22 \frac{M_{\rm BH}}{10^8} M_{\odot} \,\mathrm{yr}^{-1}.$$

Winds like this have electron scattering optical depth  $\tau \sim 1$ , measured inward from infinity to a distance of order the Schwarzschild radius  $R_{\rm S} = 2GM/c^2$  (King & Pounds 2015). So on average every photon emitted by the AGN scatters about once before escaping to infinity.

## Properties of the winds - 2

Because electron scattering is front-back symmetric, each photon on average gives up all its momentum to the wind, and so the total (scalar) wind momentum should be of order the photon momentum, or

$$\dot{M}_w v \sim \frac{L_{\rm Edd}}{c} \simeq \dot{M}_{\rm Edd} \epsilon c,$$

where v is the winds terminal velocity. Since  $M_w \simeq M_{\rm Edd}$  we get (King & Pounds 2015).

$$v \simeq 0.1 \frac{\epsilon}{0.1} c$$
.

The instantaneous wind mechanical luminosity is then

$$L_{\rm BHwind} = \frac{1}{2} v^2 \dot{M}_w \simeq \frac{1}{2} \frac{v}{c} L_{\rm Edd} \simeq 0.05 \frac{\epsilon}{0.1} L_{\rm Edd},$$

in good agreement with the earlier estimate for the energy-driven winds.

## The wind at work - 1

A self-consistent model is described by King & Pounds (2015).

- The black hole wind is abruptly slowed in an inner (within the BH sphere of influence) shock, in which the temperature approaches  $\sim 10^{11}$  K.
- The shocked wind gas acts like a piston, sweeping up the host ISM at a contact discontinuity moving ahead of it. Because this swept-up gas moves supersonically into the ambient ISM, it drives an outer (forward) shock into it. The dominant interaction here is the reverse shock slowing the black hole wind, which injects energy into the host ISM.
- The nature of this shock differs sharply depending on whether some form of cooling (typically radiation) removes significant energy from the hot shocked gas on a timescale shorter than its flow time.
- If the cooling is strong (momentum-driven flow), most of the pre-shock kinetic energy is lost (usually to radiation). As momentum must be conserved, the post-shock gas transmits just its ram pressure to the host ISM. As we have seen, this amounts to transfer of only a fraction  $\sim \sigma/c \sim 10^{-3}$  of the mechanical luminosity  $L_{\rm BH \, wind} \cong 0.05 L_{\rm Edd}$  to the ISM. In other words, in the momentum-driven limit, only energy  $\sim 10\%$  of the gas binding energy for BHs close to the  $M_{\rm BH}-\sigma$  relation is injected into the bulge ISM, which is therefore stable.

## The wind at work - 2

- In the opposite limit in which cooling is negligible, the post-shock gas retains all the mechanical luminosity and expands adiabatically into the ISM. The post-shock gas is now geometrically extended.
- The mechanical energy, thermalized in the shock and released to the ISM, now equals the gas binding energy. The energy-driven flow is much more violent than momentum-driven flow and can unbind the bulge.
- Note that if the BH and galaxy evolve strictly in parallel, preserving the  $M_{\rm BH}-\sigma$  relation, a BH in an energy-driven environment is unlikely to reach observed SMBH masses because it would stop its gas accretion. More later.



Schematic view (not to scale) of the shock pattern resulting from the impact of a black hole wind (blue) on the interstellar gas (red ) of the host galaxy. The accreting SMBH drives a fast wind with v~0.1c whose ionization state makes it observable in X-ray absorption lines. It collides with the ambient gas in the host galaxy and is slowed in a strong shock. The inverse Compton effect from the AGN's radiation field rapidly cools the shocked gas, strongly compressing and slowing it. In the most compressed gas cooling becomes important, and the flow rapidly cools and slows over an even narrower region ("cooling clumps"). The cooled gas exerts the pre-shock ram pressure on the galaxy's interstellar gas and sweeps it up into a dense shell. The shell's motion then drives an outward shock.



Schematic picture of momentum- and energy-driven outflows. In both cases a fast wind (velocity ~0.1c) impacts the interstellar gas of the host galaxy, producing an inner reverse shock that slows the wind and an outer forward shock that accelerates the swept-up gas. In the momentum-driven case, the shocks are very narrow and rapidly cool to become effecttively isothermal; only the ram pressure is communicated to the outflow, leading to very low kinetic energy  $\sim (\sigma/c) L_{Edd}$ . In an energy-driven outflow, the shocked regions are much wider and do not cool; they expand adiabatically, transferring most of the kinetic energy of the wind to the outflow.



Wind kinetic power as a function of the AGN bolometric luminosity. Solid, dashed and dotted line represent the correlations  $\dot{E}_{kin} = 1, 0.1, 0.01L_{bol}$ . Blue symbols: AGNs for which molecular winds have been reported (mostly local ULIRGs and Sevfert galaxies). Green symbols: ionised outflows; filled squares mark z>1 AGNs; filled triangles mark z=0.1-0.2 AGN; open triangles mark z=0.4-0.6 type 2 AGN; pentagons mark z=2-3 radiogalaxies; filled circles mark hyper-luminous z=2-3 QSOs. BAL winds are shown with black stars. The black open pentagon highlights the [CII] wind in J1148+5251 at z=6.4. Red symbols mark X-ray outflows.



SFR

The mass-loading factor  $\eta =$  $(dM/dt)_{OF}/SFR$  as a function of AGN bolometric luminosity. The mass-loading factor of molecular winds is > 1 in most cases, and > 10in about half the cases. The median mass-loading factor of ionised winds is  $\approx$  1, with a rather large distribution between 0.001 and 100.  $\eta$  is not correlated with the AGN bolometric luminosity. The AGN driven wind mass-loading factors are systematically larger than those of starburst driven winds in local star-forming galaxies estimated by Heckman et al. (2015). From Fiore et al. (2017).



Without some form of feedback, neither current semi-analytic models nor numerical simulations can successfully reproduce the properties of massive galaxies. The naïve assumption that stellar mass follows halo mass, leads to too many small galaxies, too many big galaxies in the nearby universe, too few massive galaxies at high redshift, and too many baryons within the galaxy halos. Large amount of energy can be injected into the interstellar medium by supernovae (SNe) and by AGNs.

## Observed wind properties

- Several fast (v<sub>OF</sub> of the order of 1000 km/s), massive outflows of ionised, neutral and molecular gas, extended on kpc scales, have been discovered.
- The mass outflow rate is correlated with the AGN bolometric luminosity.
- The fraction of outflowing gas in the ionised phase increases with the bolometric luminosity.
- The wind kinetic energy rate  $\dot{E}_{kin}$  (kinetic power) is correlated with  $L_{bol}$  for both molecular and ionized outflows. We have  $\dot{E}_{kin}/L_{bol} \sim 1 10\%$  for molecular winds,  $\dot{E}_{kin}/L_{bol} \sim 0.1-10\%$  for ionised winds. About half X-ray absorbers and broad absorption line (BAL) winds have  $\dot{E}_{kin}/L_{bol} \sim 0.1-1\%$  with another half having  $\dot{E}_{kin}/L_{bol} \sim 1 10\%$ .
- Most molecular winds and the majority of ionised winds have kinetic power in excess to what would be predicted if they were driven by SNe, based on the SFR measured in the AGN host galaxies. The straightforward conclusion is that most powerful winds are AGN driven.
- The average AGN wind mass-loading factor, <  $\eta$  > is between 0.2 and 0.3 for the full galaxy population while <  $\eta$  >> 1 for massive galaxies at z≤ 2. Tentative conclusion: AGN winds are, on average, powerful enough to clean galaxies from their molecular gas (either expelling it from the galaxy or by destroying the molecules) in massive systems only, and at z≤ 2.
- What happens at z> 2 is still unclear.

# Feedback and galaxy evolution-1

- There are also structural problems:
  - ≻ massive galaxies with thin disks and/or without bulges are missing;
  - ≻the concentration and cuspiness of cold dark matter is found to be excessive in barred galaxies and in dwarfs.
- The solution to all of these difficulties must lie in feedback.
- There are various flavors of feedback:
  >reionization,
  >supernova (SN) explosions,
  >tidal stripping,
  >input from AGNs.

## Feedback and galaxy evolution- 2

- Although the total *energy* released by SNe, integrated over the galaxy lifetime, may be large, the mean *power* (energy released per unit time) is enough only to cause gas disruption and dispersal in intermediate mass and massive dwarfs (halo mass~ $10^8-10^{10}M_{\odot}$ ). SNe expel the remaining baryons in systems of halo mass up to ~  $10^8 M_{\odot}$ , leaving behind dim remnants of dwarf galaxies (Dekel & Silk 1986). Presumably the luminous dwarfs accrete gas at later epochs.
- In very low-mass halos gas cannot even fall in, because its specific entropy is too high (Rees 1986). This entropy barrier amounts to a temperature barrier since the gas density, which to first order is proportional to the total mass density, is the same in different halos at a given epoch. Only halos of mass > 10<sup>5</sup> M<sub>☉</sub> trap baryons that are able to undergo early H<sub>2</sub> cooling and eventually form stars.
- Heating by re-ionization increases this mass limit. The abrupt increase of the sound speed to ~10 20 km s<sup>-1</sup> at z ~ 8 means that dwarfs of halo mass ~ 10<sup>6</sup> 10<sup>7</sup> M<sub>☉</sub>, which have not yet collapsed and fragmented into stars, will be disrupted. However massive dwarfs are unaffected, as are the high  $\sigma$  peaks that develop into early collapsing, but rare, low mass dwarfs.

## Feedback and galaxy evolution: the galaxy bimodality - 1



Galaxy colors illustrate the bimodality of SFRs. The contours are the density of SDSS galaxies in color-luminosity space, after correction for selection effects (Baldry et al. 2004). Elliptical and lenticular galaxies (early-type) are red, spirals (late-type) are blue. The bimodality is real: most galaxies lie in either the Red Sequence or the Blue Cloud. This suggests a different star formation history of the two populations: in late-type galaxies star-formation is ongoing while in early-type galaxies it was quenched several Gyr ago. The small fraction of intermediate population, Green Valley galaxies, suggests that some galaxies have experienced a recent quenching of their star formation.

#### (c) Interaction/"Merger"

now within one halo, galaxies interact &

- lose angular momentum SFR starts to increase
- stellar winds dominate feedback
- rarely excite QSOs (only special orbits)

#### (b) "Small Group"



 halo accretes similar-mass companion(s) can occur over a wide mass range

- Music still similar to before: dynamical friction merges the subhalos efficiently

(a) Isolated Disk



 halo & disk grow, most stars formed - secular growth builds bars & pseudobulges "Seyfert" fueling (AGN with Me>-23) cannot redden to the red sequence

#### (d) Coalescence/(U)LIRG



 galaxies coalesce: violent relaxation in core. gas inflows to center;

-1

Time (Relative to Merger) [Gyr]

Hopkins et al. (2008)

starburst & buried (X-ray) AGN starburst dominates luminosity/feedback. but, total stellar mass formed is small

1000

100

0.3 (

3

No.

10

82.80

(e) "Blowout"



 BH grows rapidly: briefly dominates luminosity/feedback remaining dustigas expelled - get reddened (but not Type II) QSO: recent/ongoing SF in host high Eddingtion ratios morger signatures still visible.



(f) Quasar

 dust removed now a "traditional" Q5O host morphology difficult to observe: tidal features fade rapidly - characteristically blue/young spheroid

#### (g) Decay/K+A



 tidal features visible only with very deep observations. - remnant reddens rapidly (E+A/K+A) "hot halo" from feedback sets up quasi-static cooling

#### (h) "Dead" Elliptical

46 W

- star formation terminated large BH/spheroid - efficient feedback halo grows to "large group" scales; mergers become inefficient. growth by "dry" mergers.

**Do SFR and BH** accretion rate simply track each other? In the picture outlined in the figure, galactic disks grow mainly in quiescence until the onset of a major merger. During the early stages of the merger, tidal torques excite some enhanced star formation and BH accretion. During the final coalescence of the galaxies, massive inflows of gas trigger strong starbursts. The high gas densities feed rapid BH growth.







HST image of NGC 4676 (the Mice)



The merger scenario by Hopkins et al. (2008) does not work! The figure shows the average  $\log(L_{SE,IR})$ in bins of  $log(L_x)$  in five redshift bins: blue for 0.1 < z < 0.4, cyan for 0.4 < z < 0.8, green for 0.8 < z < 1.2, red for 1.2 < z < 2 and magenta for 2 < z < 4. The short dashed line is the correlation derived in Mullaney et al. (2011) for a pure AGN SED. At any redshift there is no correlation between AGN activity and SF over several orders of magnitudes in luminosity.



On the other hand, computing average  $L_X$  in  $L_{SF,IR}$  bins (in log scale), from the same bivariate distribution, gives different results. At all redshifts the average  $L_x$  correlates with the L<sub>SF IR</sub> and the binned points are close to the SFR/BHAR~ 500 ratio found 0.4 has a very flat slope, possibly due to the small volume sampled, the 0.4  $\leq z < 0.8$  interval shows a correlation with slope consistent with 1 at  $\sim 1\sigma$ . At higher redshifts, i.e. as we approach the epoch at which both the SFR and the BH accretion rate peak, the slope becomes increasingly flat and the correlation weakens.

The simplest interpretation of these data goes as follows. At high-z the gas is very abundant in galactic halos and therefore both the star-formation and the BH accretion can proceed vigorously. However the timescales are widely different. In the case of star-formation, the *minimum* timescale is the dynamical time, which is  $\sim 0.1$  Gyr for massive galaxies. But, as argued by several authors (Granato et al. 2004, Thomas et al. 2010; Lapi et al. 2014) is more likely of at least 0.5–0.7 Gyr for massive spheroidal galaxies and larger for less massive galaxies.

The AGN accretion rate likely occurs at about the Eddington limit:

$$L_{\rm AGN} = \epsilon c^2 \dot{M}_{\rm BH} = \lambda L_{\rm Edd}$$

where  $\epsilon \simeq 0.1$  is the matter to radiation conversion efficiency of the accretion,  $\lambda$  is the Eddington ratio and

$$L_{\rm Edd} = \frac{4\pi c G M_{\rm BH} \mu_e}{\sigma_T} = 1.51 \times 10^{38} \frac{M_{\rm BH}}{M_{\odot}} \,\,{\rm erg\,s^{-1}}$$

 $\mu_e \simeq 1.2 m_p$  being the mass per unit electron and  $\sigma_T$  the Thomson cross-section.

If  $L_{\text{AGN}} = L_{\text{Edd}}$ , i.e.  $\lambda = 1$ ,  $M_{\text{BH}}$  grows exponentially:

$$M_{\rm BH} = M_{\rm seed} e^{t/\tau_{\rm S}}$$

where  $\tau_{\rm S}$  is the Salpeter time

$$\tau_{\rm S} = \frac{\epsilon c \sigma_T}{4\pi G \mu_e} = 3.7 \times 10^7 \frac{\epsilon}{0.1} \, {\rm yr}.$$

Thus the accretion timescale is much shorter than the star-formation timescale. Until we are *in the Eddington-limited regime* the star-formation and the AGN luminosities cannot be proportional to each other.

On the other hand, at later times, the accretion is strongly sub-Eddington and the accretion timescale can match the star-formation timescale, as in the case of re-activation of star-formation and nuclear activity ("rejuvenation") by interactions or mergers.



Evolution of the bolometric AGN luminosity due to accretion (blue lines) and of the FIR luminosity due to star formation (red lines) in galaxies with halo mass  $M_{\rm H}$  = 2 × 10<sup>12</sup>  $M_{\odot}$ (solid lines) and  $M_{\rm H} = 6 \times$  $10^{12} \,\mathrm{M}_{\odot}$  (dashed lines) at redshift z = 2. The light curves are plotted as a function of the galactic age in units of 10<sup>8</sup> yr (lower scale) and of the Salpeter time (upper scale).

## Conclusions - 1

- BH masses correlate tightly only with classical bulges and ellipticals. In contrast, they do not correlate with disk properties at all. Since spheroidal components of galaxies possess the older stellar populations, this suggests that the BH growth happens in the context of dissipative baryon collapse at substantial redshifts.
- The evolution of *luminosity densities* due to star-formation in spheroidal galaxies and to AGN activity proceed in parallel, implying a mutual relationship (but SFR and BH accretion rate do **not** simply track each other).
- Most growth of large BHs happens by radiatively efficient gas accretion (Soltan argument).
- The energy radiated by a BH (~0.1M<sub>BH</sub>• $c^2$ , assuming a radiative efficiency of 10%) is much larger than the binding energy ~ M<sub>bulge</sub> $\sigma^2$  of its host bulge. If only ~5% of the AGN energy output couples to gas in the forming galaxy, then all of the gas can be blown. Thus, BH growth may be self-limiting, and AGNs may quench star formation.
- The substantial stellar masses and star-formation rates of sub-millimeter galaxies (SMGs) and the evidence for subdominant AGN activity and moderate BH masses imply that most gas is converted into stars before BH reach large masses. This is easily understood since the early BH growth is Eddington limited.

# Conclusions - 2

- When the BHs reach a critical threshold, their "quasar-mode energy feedback" balances outward radiation or mechanical pressure against gravity. Then the AGNs blow away the interstellar gas, quenching star formation and leaving the galaxies red and dead. AGNs become visible and continue to shine for a few Salpeter times, accreting the `reservoir' (torus) mass.
- This convincingly solves some problems of galaxy formation, such as expelling a large fraction of initial gas to account for the present day baryon to dark matter ratio in galaxies and to prevent excessive star formation (the stellar mass function of galaxies sinks down, at large masses, much faster that the halo mass function).
- However, observations of winds capable of removing the interstellar gas from z>2 galaxies are still missing. And a large fraction of massive early type galaxies formed most of their stars at these redshifts. To remove the gas we need a kinetic power of ~5% of the bolometric luminosity, but the scanty observations of high z winds generally indicate lower kinetic powers.

# Conclusions - 3

- Two kinds of AGN feedback have been considered. Radio-mode feedback (powerful jets of radio sources) is a well known phenomenon but can hardly substantially affect the galaxy evolution: it is very hard to confine at least some effects of well-collimated jets within their galaxies. As Kormendy & Ho (2013) put it: "Firing a rifle in a room does not much heat the air in the room". It is much more plausible that the jet energy is dissipated by interactions with the intergalactic gas, especially in galaxy clusters.
- To efficiently operate on galaxy scales, the feedback must be relatively isotropic (quasarmodel feedback). But the physics is not well understood. The correlations between  $M_{BH}$  and the galaxy velocity dispersions are consistent with both energy- and momentum-driven feedback. But only energy-driven feedback can efficiently quench star formation.
- Outflows with kinetic power sufficient to clean the galaxy of cold gas (kinetic power ~5% of the AGN bolometric emission) were found for only a few high-z galaxies. They are more common at low z. But most local AGNs accrete at very sub-Eddington rates. Very few galaxies are still growing their BHs at a significant level. Rapid BH growth by radiatively efficient accretion took place mostly in more massive galaxies that are largely quenched today. That is, the era of BH growth by radiatively efficient accretion is now mostly over: co-evolution happened at high z.
## Thanks for your attention!