Research in Astronomy and Astrophysics, 22:085014 (9pp), 2022 August https://doi.org/10.1088/1674-4527/ac777f

Predicting Supermassive Black Hole Mass with Machine Learning Methods

Yi He[1,2], Qi Guo[1,2], and Shi Shao[1]
[1] Key Laboratory for Computational Astrophysics, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; guoqi@nao.cas.cn
[2] School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China
Received 2022 April 24; revised 2022 May 18; accepted 2022 May 19; published 2022 July 15

Abstract

It is crucial to measure the mass of supermassive black holes (SMBHs) in understanding the co-evolution between the SMBHs and their host galaxies. Previous methods usually require spectral data which are expensive to obtain. We use the AGN catalog from the Sloan Digital Sky Survey project Data Release 7 (DR7) to investigate the correlations between SMBH mass and their host galaxy properties. We apply the machine learning algorithms, such as Lasso regression, to establish the correlation between the SMBH mass and various photometric properties of their host galaxies. We find an empirical formula that can predict the SMBH mass according to galaxy luminosity, colors, surface brightness, and concentration. The root-mean-square error is 0.5 dex, comparable to the intrinsic scatter in SMBH mass measurements. The 1σ scatter in the relation between the SMBH mass and the combined galaxy properties relation is 0.48 dex, smaller than the scatter in the SMBH mass versus galaxy stellar mass relation. This relation could be used to study the SMBH mass function and the AGN duty cycles in the future.

Key words: (galaxies:) quasars: supermassive black holes – galaxies: evolution – methods: data analysis

1. Introduction

Supermassive black holes (SMBHs) are prevalent at the centers of massive galaxies (e.g., Kormendy & Richstone 1995; Ferrarese & Ford 2005; Kormendy & Ho 2013). Recently, the SMBH in the elliptical galaxy M87 has been imaged by the Event Horizon Telescope (Event Horizon Telescope Collaboration 2019). The mass of the SMBHs is tightly related to the properties of the galaxies, such as bulge mass, velocity dispersion, and surface brightness (Ferrarese & Merritt 2000; Merritt & Ferrarese 2001; Häring & Rix 2004; Saglia et al. 2016), as well as the velocity dispersion of the entire elliptical galaxies (Liu et al. 2008; Graham & Scott 2013; Kormendy & Ho 2013). Such correlations present both in nearby and in high redshift galaxies (Wu et al. 2002; Shields et al. 2006; Shen & Kelly 2010; Schramm & Silverman 2013). The co-evolution between the SMBHs and their host galaxies invokes great interest, given their orders of magnitude differences in masses and sizes (Hopkins et al. 2008; Schawinski et al. 2010; Izumi et al. 2019; Pensabene et al. 2020). Recent studies suggest that massive black holes are formed in the central regions of galaxies as a result of nearby material feeding and SMBH mergers (Di Matteo et al. 2005, 2008; Alexander & Hickox 2012; Marasco et al. 2021). On the other hand, SMBHs could play an important role in shaping the formation and evolution of the host galaxies by releasing a vast amount of momentum and energy while accreting gas (Ciotti & Ostriker 2007; Sijacki et al. 2007; Hopkins et al. 2009).

It is crucial to measure the SMBH mass for understanding their formation and evolution, as well as their co-evolution with the host galaxies. However, accurate measurements require high spatial resolution within the gravitational regime of the SMBHs (Magorrian et al. 1998; Kormendy & Kennicutt 2004). The closest distance ever reached is via the stellar dynamics around the SMBH in the center of the Milky Way. The orbits of stars at ∼100 au from the center infer the mass of the SMBH to be $4 \times 10^6\, M_\odot$ (Ghez et al. 2008; Gillessen et al. 2009; Peißker et al. 2020). Such high-resolution data are only available in our Milky Way. For distant galaxies, gas dynamics at sub-kpc scales are usually adopted to estimate the SMBH mass. One of the most popular methods is reverberation mapping (Peterson 1993; Netzer & Peterson 1997; Kaspi et al. 2007; Grier et al. 2017), which uses the lag between broad emission-line flux and continuum flux to estimate the size of the broad-line region, and the width of the broad emission lines to estimate the velocity dispersion. Assuming an equilibrium status in the broad-line region, one can apply the viral theory to calculate the total mass enclosed as an approximation of the central SMBH mass. Recently Shen et al. (2019) collected 849 broad-line quasars from the Sloan Digital Sky Survey Reverberation Mapping (SDSS-RM) project, covering a redshift range of $0.1 < z < 4.5$.

Based on the reverberation mapping data, a correlation between the radius of the broad-line region and the continuum luminosity (R–L) can be derived (Bentz et al. 2006, 2009; Alvarez et al. 2020). It is much less expensive to obtain the continuum luminosity than to measure the lag between broad emission-line flux and continuum flux. Using the R–L relation, Liu et al. (2019) estimated the SMBH mass for a large, uniform and well-defined sample of 14,584 broad-line AGNs at $z < 0.35$.

Another way to estimate the SMBH mass is via the host galaxy properties, i.e., the SMBH mass is tightly related to the mass/velocity dispersion of the classical bulge or the same properties of elliptical galaxies. This method could extend the mass measurement to much larger SMBH samples. However, the scatter of estimated SMBH mass is very large for galaxies with pseudo-bulges or for spiral galaxies, ∼0.6 dex (Kormendy & Gebhardt 2001; Kormendy et al. 2011; Greene et al. 2016). It is therefore very important to find a robust relation between the SMBH mass and various properties of their host galaxies. Lin et al. (2021) used neural networks to model SMBH mass based on quasars' luminosity and colors. Their results have a small root-mean-square error (RMSE) value, 0.37, but the performance at high and low mass ends is not very good.

In this work, we use the SMBH mass estimated by Liu et al. (2019) with galaxy photometry from the Sloan Digital Sky Survey project Data Release 7 (SDSS DR7; York et al. 2000), and use the machine learning methods to investigate the correlation between the SMBHs and their host galaxies.

The paper is structured as follows. In Section 2, we present our sample galaxies and provide a brief introduction to the machine learning method, Lasso regression. We present the results in Section 3 and discussion in Section 4.

2. Data and Method

2.1. Data

In this work, we adopt the broad-line AGNs catalog from the SDSS DR7 presented by Liu et al. (2019). The SDSS conducts both imaging and spectroscopic surveys with the 2.5 m Sloan Foundation Telescope (Gunn et al. 2006) and the du Pont 2.5 m Telescope (Ahumada et al. 2020). Its imaging survey includes five photometric bands: u, g, r, i, and z, with the effective wavelength of 3550, 4770, 6230, 7620, and 9130 Å over 11,663 square degrees. The corresponding depth limit in each band is 22.0, 22.2, 22.2, 21.3, and 20.5 AB magnitudes, respectively (Abazajian et al. 2004). It contains 357 million unique objects. Among the five photometric bands, u-band has the largest uncertainty and the lowest sensitivity (Ivezić et al. 2004), we therefore exclude the u-band photometry in our analysis.

In SDSS, target galaxies are then selected from photometric data for spectroscopic observations. The spectroscopic survey consists of the main sample of bright galaxies with Petrosian r-band magnitude < 17.77, the luminous red galaxy sample (LRGs) with Petrosian r-band magnitude < 19.5 and the quasar candidate sample with point-spread function magnitude $i < 20.2$ up to $z < 5.5$. The spectroscopic catalog contains 930,000 galaxies and 120,000 quasars within 9380 square degrees.

Liu et al. (2019) compiled a comprehensive and uniform sample of broad-line AGNs catalog from the SDSS DR7 spectroscopic objects. It contains 14,584 well-defined broad-line AGNs over a redshift range of $0 < z < 0.35$, with a median redshift $z = 0.2$. Liu et al. (2019) properly removed the stellar continuum for each spectrum and carefully deblended the broad and narrow lines. Many observational studies (Hao et al. 2005; Pâris et al. 2012; Oh et al. 2015) use a conventionally broad-line criterion of FWHM $\sim 1000\,\mathrm{km\,s^{-1}}$, while Liu et al. (2019) extends the type 1 AGNs to the low-luminosity and low-BH-mass regimes with the minimum broad-line width down to 500 $\mathrm{km\,s^{-1}}$. This leads to a more complete (4 times larger in AGN numbers) type 1 AGN catalog than the previous quasar catalog based upon the SDSS DR7. The catalog spans the SMBH mass over a range of $10^{5.1} \sim 10^{10.3}\,M_{\odot}$ and the Eddington ratios from −3.3 to 1.3 in logarithmic scale. We remove all sources fall below the depth limits as indicated above for g, r, i and z bands, resulting in a final sample of 12,266 sources.

We briefly summarize the method used in Liu et al. (2019) to estimate the SMBH mass as follows. Liu et al. (2019) adopt the viral method to estimate the SMBH mass using the spectral measurements. The velocity is obtained directly by the width of the broad line, and the broad-line region (BLR) radius is estimated by adopting the empirical correlation with the broad-line luminosity (Kaspi et al. 2005; Wang et al. 2009). Thus, the SMBH mass, $M_{\mathrm{BH}}(\mathrm{H}\beta)$, can be derived by using the spectra of Hβ with the fitting formula as follows (Greene & Ho 2005; Ho & Kim 2015):

$$\log M_{\mathrm{BH}}(\mathrm{H}\beta) = \log\left[\left(\frac{\mathrm{FWHM}(\mathrm{H}\beta)}{1000\ \mathrm{km\ s^{-1}}}\right)^2\left(\frac{L_{5100}}{10^{44}\ \mathrm{erg\ s^{-1}}}\right)^{0.533}\right] + 6.91 \quad (1)$$

where L_{5100} is the rest frame continuum luminosity at 5100 Å ($L_{5100} \equiv \lambda L_{\lambda(5100\text{Å})}$). The mass can also be estimated using the broad line Hα with a similar fitting formula (Greene & Ho 2005) when data are available. SMBH mass, M_{BH}, is then defined as the average of $M_{\mathrm{BH}}(\mathrm{H}\beta)$ and $M_{\mathrm{BH}}(\mathrm{H}\alpha)$. The 1σ intrinsic scatter is about 0.35 dex.

We compare the M_{BH}–$M_{\star}$ relation in Liu's results to those found in the literature. To obtain the stellar mass of the host galaxy, we cross-match this catalog with the spectroscopic data products from the Max Planck Institute for Astrophysics and Johns Hopkins University DR7 catalog (MPA-JHU; Kauffmann et al. 2003; Brinchmann et al. 2004; Tremonti et al. 2004). Figure 1 presents the $M_{\mathrm{BH}} - M_{\star}$ relation. Red curves denote the median and 1σ scatter of 0.59 dex from Liu's AGN sample. The MPA-JHU catalog does not take into account the AGN contributions to the stellar mass. As a consequence, the stellar mass could be over-estimated, especially for those with luminous AGNs. This at least partly explains the flat feature at high masses. Häring & Rix (2004) measured the M_{BH}–M_{bulge} relation for 30 ellipticals and bulges with an observed scatter of <0.30 dex. Sahu et al. (2019) analyzed a sample of 84 early-type galaxies and their central SMBHs. The estimated scatter around

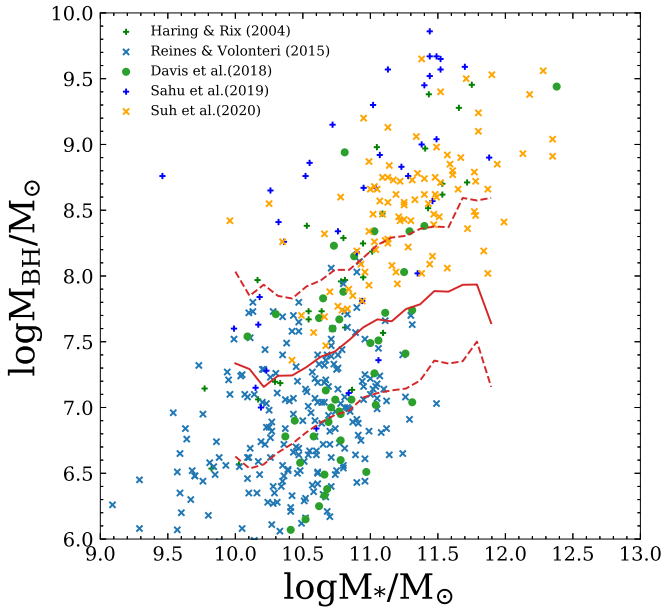

Figure 1. SMBH mass vs. galaxy stellar mass relation. The solid red curve and dashed curves denote the median value of Liu's AGN sample and the corresponding 1σ scatter. Measurements from the literature are presented using different symbols as indicated in the top left corner. Pluses are from Häring & Rix (2004) and Sahu et al. (2019), where the x-axis M_* represents stellar mass in bulge or in early-type galaxies. Green filled circles are taken from Davis et al. (2018) for local spirals. Blue crosses are results for nearby galaxies from Reines & Volonteri (2015). Orange crosses are for all galaxy types extending to high redshifts ($z < 2.5$) from Suh et al. (2020).

the $M_{\rm BH}$–$M_\star$ relation is about 0.52 dex. This relation is different from Liu's results because the former is based either on classical bulges or on elliptical galaxies, while the latter includes spiral galaxies and galaxy disks. Suh et al. (2020) conducted 100 X-ray-selected moderate-luminosity, broad-line AGNs up to $z \sim 2.5$ and estimated their masses based on the single-epoch virial method. Their observed scatter is ∼0.50 dex. Davis et al. (2018) used 40 local spiral galaxies with a regression scatter 0.66 dex. It is much larger than those for elliptical galaxies and for classical bulges, confirming a tighter relationship between the SMBH and the spheroidal component of galaxies. Reines & Volonteri (2015) analyzed 262 nearby broad-line galaxies with the same methods as described in Liu et al. (2019), finding a scatter of 0.55 dex. They did not distinguish the bulge component and morphology. These measurements are more in line with the $M_{\rm BH}$–$M_\star$ relation estimated using Liu's catalog.

As summarized in Table 1, the scatter between the SMBH mass and stellar mass for elliptical galaxies and classical bulges are smaller than those for all galaxies types and those for disk galaxies. Results from Liu's catalog include all galaxies types and have similar scatters to those reported in the literature.

Table 1
Summary of Previous Works on the $M_{\rm BH}$–$M_\star$ Relation

	Morphology	Redshift	Scatter
Häring & Rix (2004)	Bulge/Elliptical	<106 Mpc	0.30
Sahu et al. (2019)	Bulge/Elliptical	<158 Mpc	0.52
Davis et al. (2018)	Spiral	<258 Mpc	0.66
Reines & Volonteri (2015)	All	<0.055	0.55
Suh et al. (2020)	All	<2.5	0.50
This work	All	<0.35	0.59

Note. We present references (column 1), morphology of the target galaxies (column 2), distance/redshift range of the catalog (column 3), and scatter of the $M_{\rm BH}$–$M_\star$ relation (column 4).

2.2. Machine Learning Method

In this section, we use machine learning to investigate whether there is a tighter relation between the SMBH mass and galaxy properties other than stellar mass. The machine learning procedure is illustrated in Figure 2.

The linear regression analysis is widely employed in statistics. It uses least square method to evaluate the linear relationship between features $\boldsymbol{x}$ and their dependent variable y: $y = \boldsymbol{\omega x} + b$, where $\boldsymbol{\omega}$ denotes coefficients and b is a constant. The regression finds the optimal values of $\boldsymbol{\omega}$ by minimizing the loss function: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\omega x}_i - b)^2$ where i denotes a row of data and n is the number of the row. However, reducing the loss function as such could result in an over-fitting problem. As a consequence, the final formula would be too complex and easily perturbed.

Lasso (least absolute shrinkage and selection operator; Tibshirani 1996) is a regression analysis originally formulated for linear regression models. It is advanced in interpreting statistical models by performing both variable selection and regularization. Based on the loss function of linear regression, Lasso regression introduces a l^1 norm, i.e., the sum of $|\omega|$ less than a certain number. The loss function is modified as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\omega}_i\boldsymbol{x}_i - b)^2 + \lambda\sum|\omega|$. The linear limit pushes the absolute value of the coefficients to decrease, even down to zero. In practice, we use the Lasso algorithm from Scikit-Learn (Pedregosa et al. 2011), and utilize GridSearchCV from Scikit-Learn to find the best λ. We also test some non-linear regressions and several other popular machine learning algorithms, including Multilayer Perceptron, Xgboost (eXtreme Gradient Boosting), and Ridge regression. We find their efficiency is similar to the Lasso regression method, yet the latter has the advantage to provide a fitting formula to approximate the results given by the machine learning.

Here we use galaxy observables in the SDSS as input for the machine learning processes, which include the absolute magnitudes in g, r, i, and z band (M_g, M_r, M_i, M_z), $g-r$, $r-i$, $i-z$, and $g-z$ color, the r-band surface brightness within half-light radii R_{50}, $\Sigma_{\rm R50}$, concentration ($c = R_{90}/R_{50}$, where R_{90} is the radii

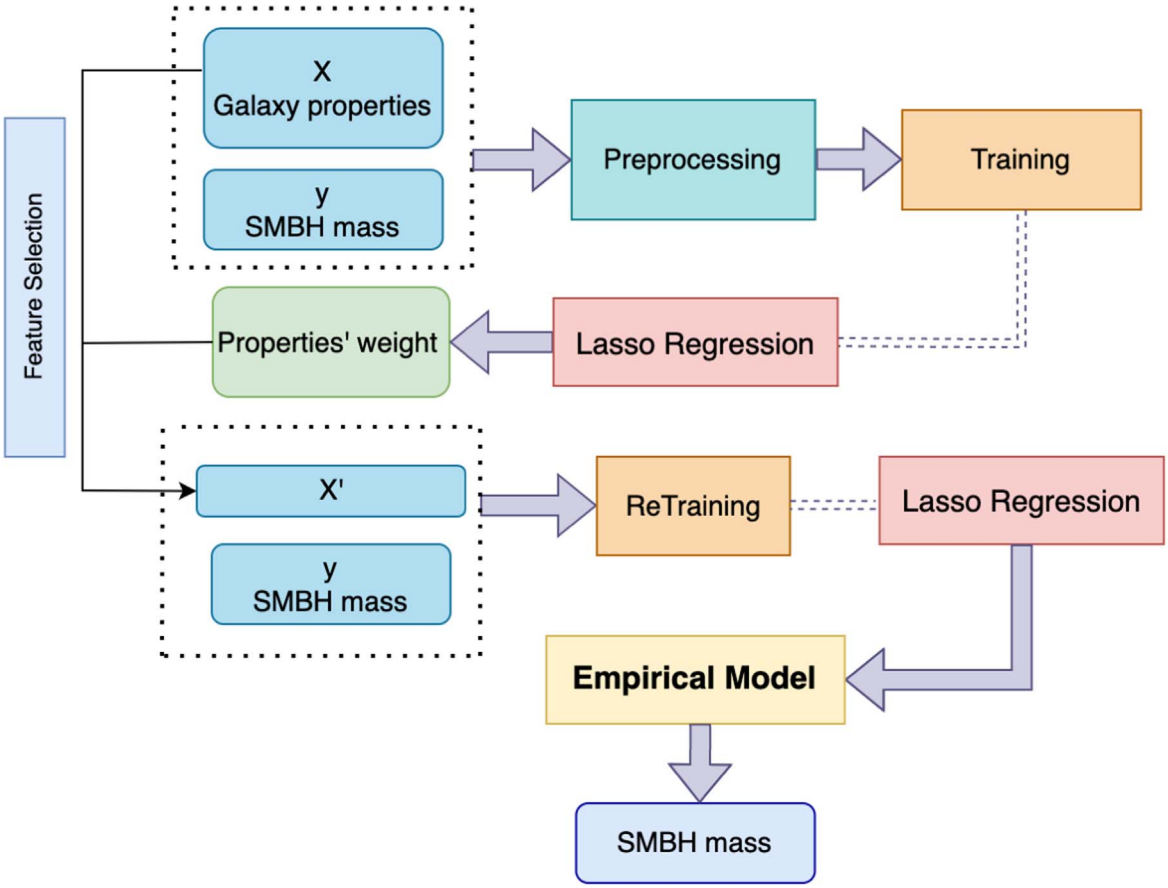

Figure 2. Methodology flowchart. The rounded rectangular boxes denote the data sets, and the rectangular boxes represent the operation performed. The arrows show the flow of data. We first perform feature selection according to the weights of the properties given by Lasso Regression, and then train the model based on the new feature set and present an empirical formula.

enclosing 90% Petrosian flux) and bulge fraction *fracDev _r* determined by the bulge-disk decomposition method using *r*-band. We do not apply attenuation corrections. Instead, we assume that the attenuation could be part of the constraint on the SMBH mass via their effects on luminosity and colors. We perform a test to correct intrinsic attenuation using CIGALE (Boquien et al. 2019; Yang et al. 2020) and find that the resulting RMSE is similar to what we find without taking into account the dust corrections. We ignore the dependence on redshift evolution because previous studies do not show strong evidence of evolution in the $M_{\rm BH}$–$M_{\star}$ relation with time (e.g., Cisternas et al. 2011; Suh et al. 2020). In addition, AGNs in Liu's catalog lie in a very narrow redshift range, $0 < z < 0.35$. The probability distributions of various properties are presented in Figure 3. Most properties have a well-spread distribution, except for the *fracDev_r* and $r-i$ whose distributions are rather concentrated. The bulge fraction could be higher and color could be redder in AGN host galaxies compared to non-AGN galaxies. This could partly explain the concentrated distributions of *fracDev_r* and $r-i$. In addition, AGN could also contribute to luminosity in the central regions and thus the *fracDev_r* is enlarged.

Samples in the data set are randomly shuffled and divided into two subsets: Subset A contains 75% of the population and Subset B contains the rest 25%. Subset B is regarded as the validation set. The distribution of the SMBH mass of Subset A is presented in Figure 4. It peaks at $10^{7.8}$ $M_{\odot}$ and drops both at high masses and low masses. The decline at high masses is mainly caused by the decreasing number of massive structures as predicted by the standard cosmology model. At low masses, it could either be caused by the low fraction of AGN in low mass systems or be limited by the detection ability. The cost function is to estimate the total deviations from the true values. The result thus could be biased by the most abundant population, i.e., both low mass and high mass SMBHs have a very low weighting in determining the machine learning results (see also Lin et al. 2021). To avoid such bias we generate the training sample by randomly selecting the same number of SMBHs in each bin from Subset A, i.e., 1000 per 0.2 dex. For those bins with fewer than 1000 sources, we duplicate the sample to have an even distribution in SMBH mass

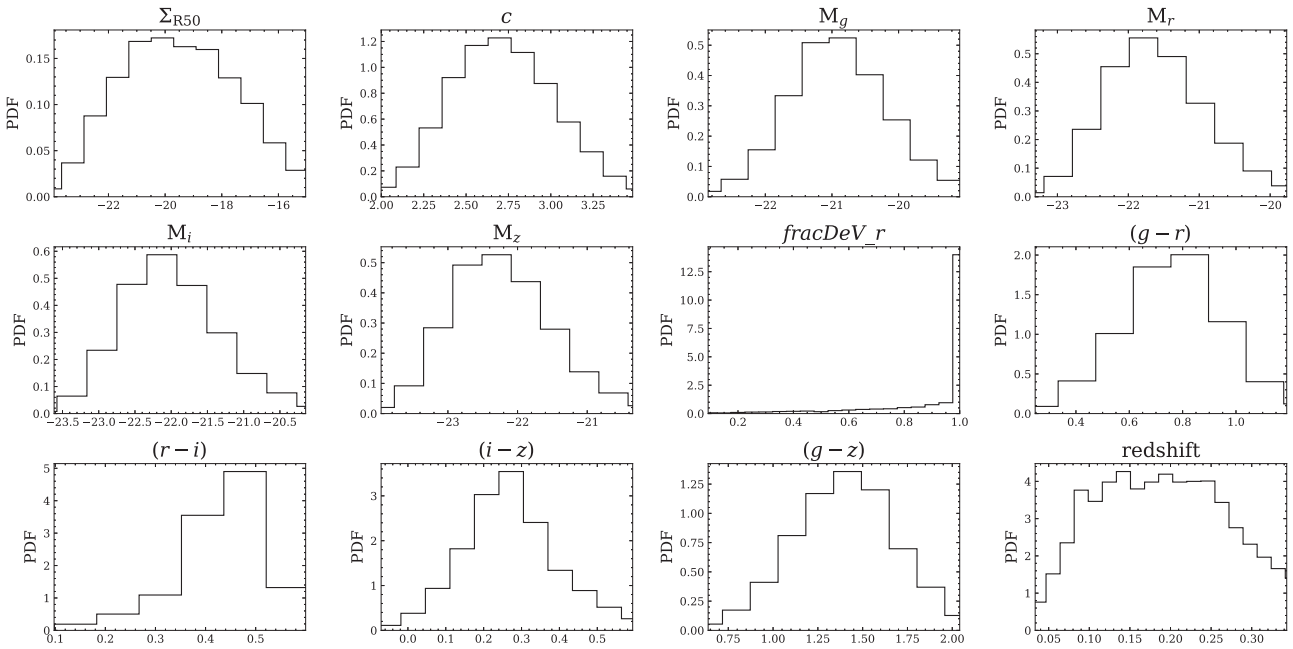

Figure 3. The probability distribution function (PDF) of galaxy properties used for the machine learning. Σ_{R50} denotes the surface brightness within *r*-band half light radii, and *c* means the ratio of the 90% *r*-band light radii and the half light radii. M_g, M_r, M_i, and M_z are the absolute Petrosian magnitude. *fracDeV_r* represents the de Vaucouleurs component weight in the bulge-disk decomposition model.

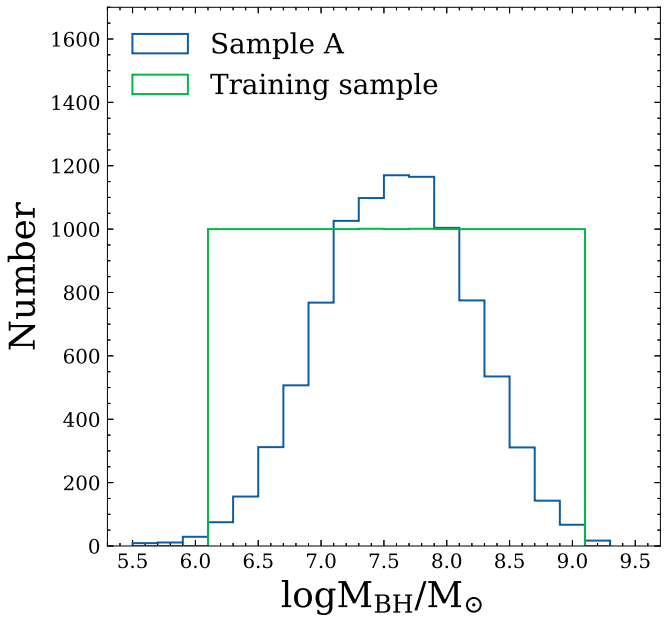

Figure 4. SMBH mass distributions. Blue and green histograms present the SMBH mass distributions of Subset A and the training sample, respectively. We transform the distribution of SMBH in Subset A into a uniform distribution of mass in the range from $10^{6.1}$ $M_\odot$ to $10^{9.1}$ $M_\odot$ to avoid the bias toward the most abundant population.

as indicated by the green histogram in Figure 4. We discard sources with SMBH mass less than $10^{6.1}M_\odot$ or larger than $10^{9.1}M_\odot$ to avoid shooting noises.

3. Results

3.1. Regression Results and Features Selection

We use the 11 variables as indicated in the last section to perform the machine learning and select a subset of the variables which have the highest contributions to reproduce the measured SMBH mass.

We apply the Scikit-Learning on all of the 11 features and present in Figure 5 the predicted SMBH mass (hereafter $M_{BH,pred}$) against the true SMBH mass (hereafter $M_{BH,true}$). It shows a clear positive relationship between the predicted SMBH mass and the true values both for the training sample and for the validation sample. Training results work better at intermediate regimes as expected. The median value of the SMBH mass is somehow overpredicted at low masses and underpredicted at high masses. It is at least partly because there is not enough intrinsic variation of the data points and the results are highly biased by a small number of statistics, especially at low masses. The RMSE of the validation sample is 0.50 dex, somehow lower than those in training samples, 0.55 dex. This is because the training samples have more

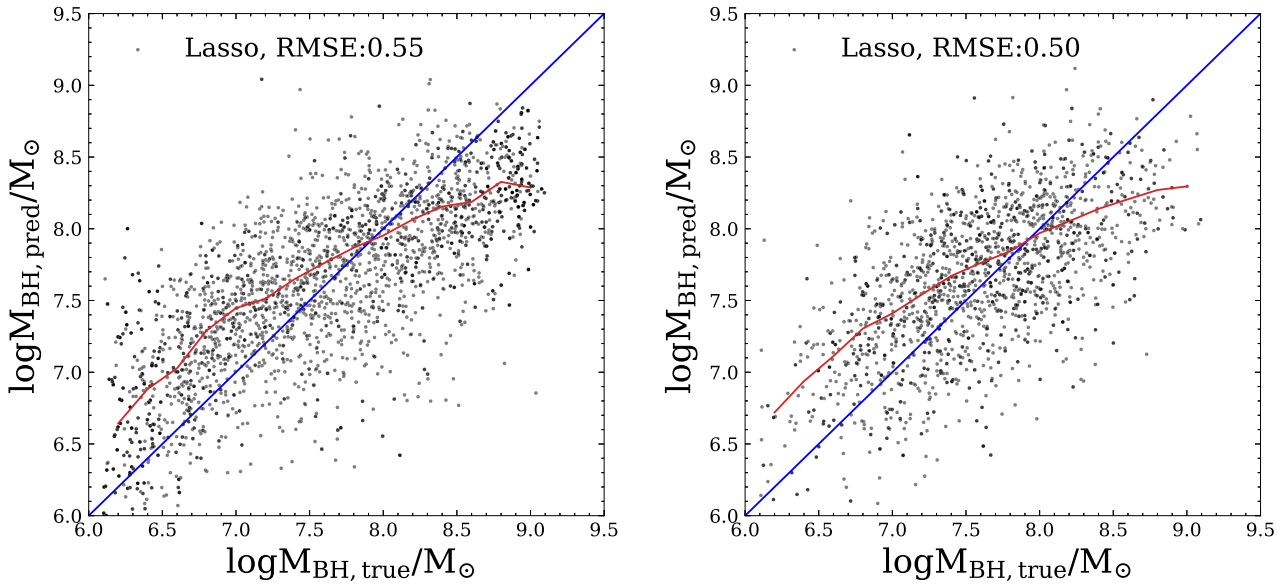

Figure 5. Predictions of Lasso regression based on all features in the training set (left) and the validation set (right). The solid red curves show the median values of the machine learning prediction, and the blue lines denote 1:1 ratio between the predicted masses and the true values. The corresponding RMSE is indicated in each panel.

weights at high and low masses where the training works less well. Given the fact that the intrinsic error in the SMBH is 0.35 dex, the accuracy is well enough.

We further explore the correlation of the 11 parameters in Figure 6. It shows that M_g, M_r, M_i and M_z are strongly correlated. It is because luminous galaxies are brighter in all bands, and vice versa. Interestingly, we find $g-r$ color and $g-z$ color are closely correlated. It could be due to the fact that $g-r$ and $g-z$ fall on the same side of big-blue-bump region (Shields 1978; Malkan & Sargent 1982).

We compare the contribution of each feature by ranking their coefficients in Figure 7. Since we use the normalized features, the coefficients are capable of indicating their contributions to the prediction. We notice that there is an obvious gap between the first six features and the rest of them. We thus keep the $i-z$, $r-i$, $g-r$ colors, M_g, $\Sigma_{\rm R50}$ and c in our optimized feature space.

We retrain the Lasso regression model utilizing the six selected features and present the result in Figure 8. Like those with the full features, it shows that the six selected features have a similar ability in reproducing the SMBH mass both for the training sample and the validation sample. Quantitatively, the RMSE is 0.50 dex for the validation set, similar to the validation results based on the full features. This demonstrates that the feature selection is reasonable.

In Lin et al. (2021), they used Neural Network to predict the SMBH mass of quasars based on photometric luminosities and colors. The RMSE of 0.37 dex in their work is lower than ours. They only consider quasars whose luminosity overweight the starlight and more closely related to the SMBH.

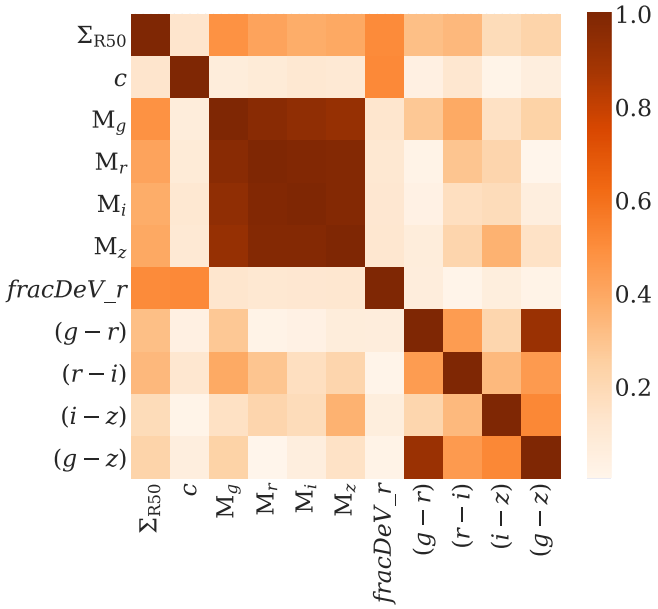

Figure 6. Pearson correlation coefficients of the galaxy properties adopted in machine learning. Color bar shows the absolute correlation strength between two features, with 1 for the strongest correlation.

3.2. Empirical Model

The Lasso linear regression provides the coefficients of the six features to predict the SMBH mass as follows:

$$\begin{aligned}\log M_{\rm BH,pred}/M_\odot = \; & 1.75(r-i) + 0.84(g-r) \\ & + 0.78(i-z) - 0.46M_g \\ & + 0.38c - 0.12\Sigma_{\rm R50} - 6.95 \end{aligned} \quad (2)$$

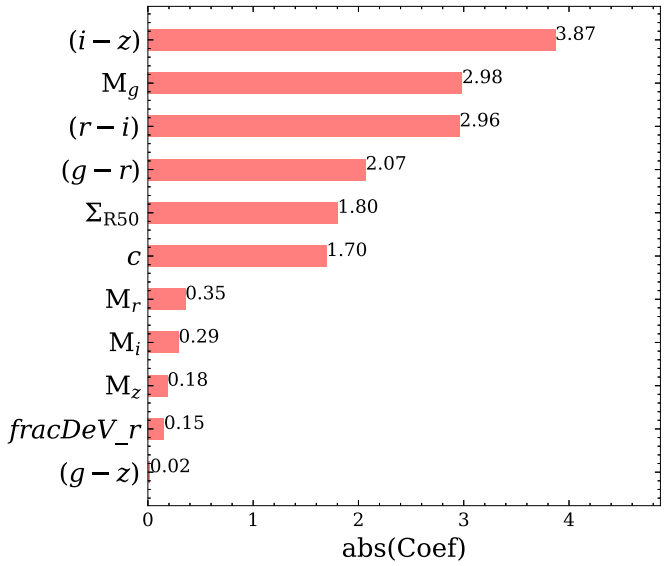

Figure 7. The absolute coefficients in the result of Lasso regression. The feature values are normalized in the training processes. The coefficients thus represent the contributions of different properties.

We compare $M_{\rm BH,pred}$ to $M_{\rm BH,true}$ for the full sample in Figure 9. Red curves denote the median value and the 1σ deviation. Interestingly, although the predicted SMBH masses deviate from the true values at high and low masses, when binned in the combined galaxy properties as indicated in the x-axis, the slope is close to one above $M_{\rm BH,pred} = 10^7\,M_\odot$, suggesting our model is appropriate for population studies. At low masses, the predicted mass is slightly higher, which could be due to the deficit of training samples at these masses.

The 1σ scatter around the median value is about 0.48 dex, much smaller than the scatter (0.59 dex) in the $M_{\rm BH}$–$M_\star$ relation in Liu's AGN catalog (Figure 1). The scatter is smaller than or comparable to those discovered by Davis et al. (2018), see also Sahu et al. (2019) and Suh et al. (2020), most of which performed more expensive dynamical measurements to obtain the SMBH mass. The scatter is somehow larger than those in Häring & Rix (2004). This is because they apply to the bulge and elliptical samples, while we include both spirals and elliptical galaxies.

3.3. Application on DR14 AGN catalog

We further quantify the performance of our fitting formula using type 1 AGNs reported in SDSS DR14 (Abolfathi et al. 2018) SPIDERS (SPectroscopic IDentification of eROSITA Sources, Coffey et al. 2019) which is an SDSS-IV (Blanton et al. 2017) X-ray selected AGN catalog, consisting of 7344 2RXS (Boller et al. 2016) and 1157 XMM-Newton (Dwelly et al. 2017) AGNs with masses measured using Mg II and Hβ emission lines (Coffey et al. 2019). Only those with broad-line width greater than 800 km s^{-1} AGNs are included in this catalog. We restrict our sample galaxies to have $z < 0.35$ to avoid the possible evolution effect. The final sample contains 2799 AGNs.

In Figure 10, we present the relation between the SMBH mass (Coffey et al. 2019) and the combined galaxy properties that we found. The RMSE is 0.50 dex, similar to the result based on Liu's AGN catalog. The fitting formula performs very well in reproducing the type 1 AGNs from SDSS DR14, except for those below $M_{\rm BH} = 10^{7.5}\,M_\odot$, where the training sample is too small. The data points are more concentrated with a smaller scatter of 0.42 dex.

4. Conclusion and Discussion

In the past few decades, numerous AGNs have been discovered, which allows us to establish the relationship between the SMBH growth and their host galaxy evolution. The mass of SMBHs is a crucial element in such studies. We use machine learning to extract the relation between the SMBH mass and their host galaxy properties using a comprehensive AGN catalog based on SDSS DR7.

We adopt a flexible and computationally efficient method, Lasso regression, which is powerful in variable selection. We find that colors, magnitude, surface density and concentration are most relevant in determining the SMBH mass. Based on the Lasso regression results, we provide an empirical formula to connect the SMBH mass and their host galaxy properties. The RMSE is 0.50 dex, comparable to the intrinsic uncertainty of 0.35 dex in the training data.

Interestingly, though the RMSE is not very small, when binned in galaxy properties, the predicted SMBH mass and the true SMBH mass follow the 1:1 ratio between $[10^7, 10^{8.5}]M_\odot$. The scatter is 0.48 dex, much lower than the scatter in the $M_{\rm BH}$–M_* relation. The machine learning results at high masses and low masses are somehow less accurate, which is mainly due to the poor intrinsic variance in the training samples at such masses.

In order to validate the performance of the Lasso regression algorithm, we tried several popular machine learning methods, including Multilayer Perceptron, Xgboost, and Ridge regression. The Multilayer Perceptron is a class of feedforward artificial neural networks, composed of several layers of nodes. Xgboost is a popular gradient boosted trees algorithm. Ridge, similar to the Lasso regression, introduces the l^2 norm instead of the l^1 norm. The RMSEs are 0.49, 0.52, 0.50 dex, respectively. The results based on different models are similar. We choose Lasso regression for it is capable of feature selection and could provide a fitting formula. We also tried to do the intrinsic dust correction by CIGALE (Boquien et al. 2019; Yang et al. 2020), and the RMSE is 0.53 dex.

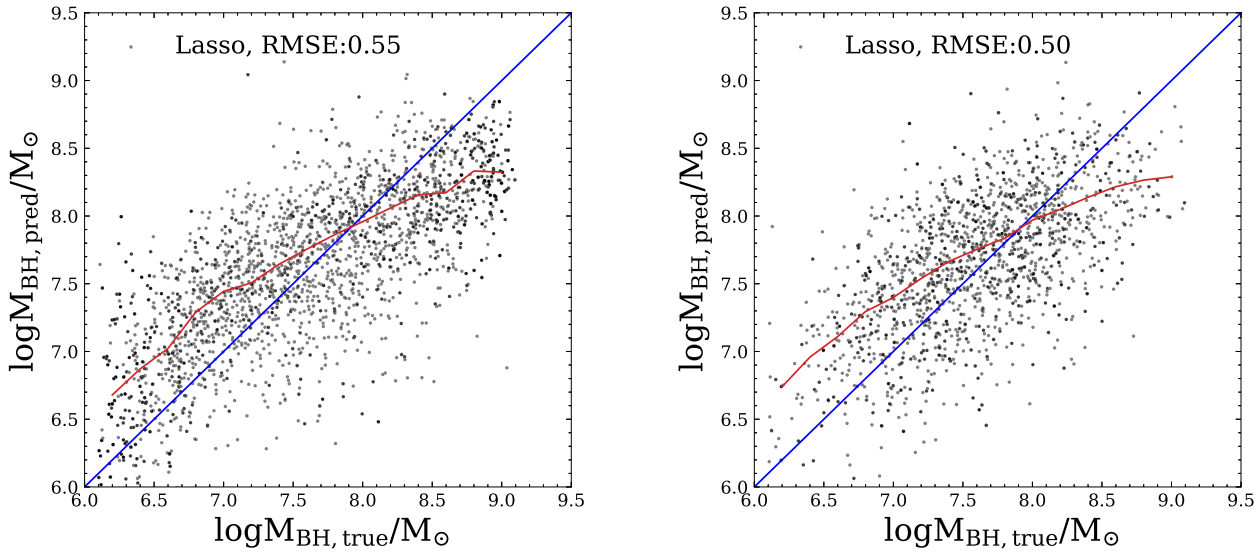

Figure 8. Predicted SMBH mass vs. true SMBH mass for the training set (left) and validation set (right) using the six selected features. Line types are the same as those in Figure 5.

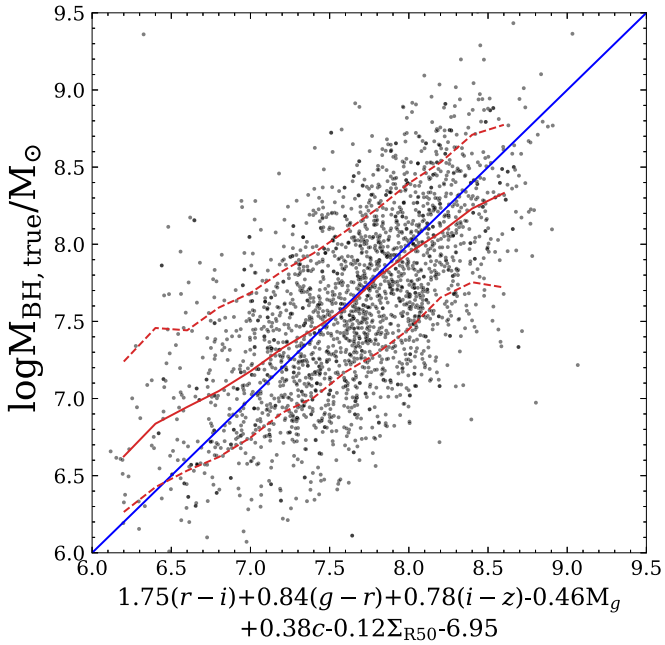

Figure 9. SMBH mass vs. the linear combination of the six selected galaxy properties provided by the machine learning for all samples in Liu's catalog. No duplication of samples is performed. Red solid curve and dashed curves show the median value and the 1σ scatter.

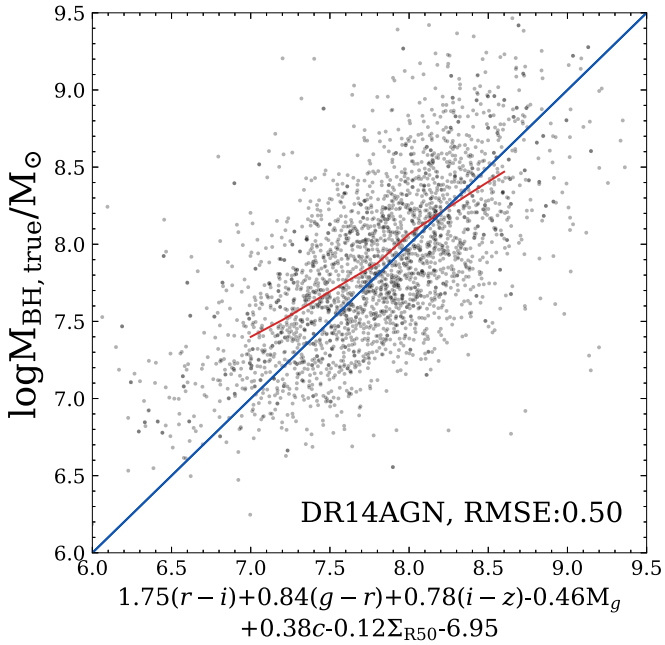

Figure 10. Same as Figure 9 but for the SDSS DR14 AGN catalog.

Shankar et al. (2008) derived the SMBH mass function by estimating the SMBH mass from the $M_{\rm BH}$–$M_\star$ relation. Using more galaxy properties, our formula could predict the SMBH mass more accurately. As a result, we could be able to provide more reliable SMBH mass functions. In combination with the AGN luminosity functions, it could also provide clues on the AGN duty cycles. In the future, we intend to collect more data at low and high masses. We will further divide samples into several subsamples according to their morphology which may improve the accuracy and reliability of the method. Using data from deeper surveys, we could also study the possible redshift evolution.

Acknowledgments

This work is supported by the National Key Research and Development of China (Grant No. 2018YFA0404503), NSFC (Grant Nos. 12033008 and 11988101), the K.C.Wong Education Foundation and the science research grants from the China·Manned Space Project with No. CMS-CSST-2021-A03.

References

Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2004, AJ, 128, 502
Abolfathi, B., Aguado, D., Aguilar, G., et al. 2018, ApJS, 235, 42
Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Alexander, D. M., & Hickox, R. C. 2012, NewAR, 56, 93
Alvarez, G. F., Trump, J. R., Homayouni, Y., et al. 2020, ApJ, 899, 73
Bentz, M. C., Peterson, B. M., Netzer, H., Pogge, R. W., & Vestergaard, M. 2009, ApJ, 697, 160
Bentz, M. C., Peterson, B. M., Pogge, R. W., Vestergaard, M., & Onken, C. A. 2006, ApJ, 644, 133
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Boller, T., Freyberg, M., Trümper, J., et al. 2016, A&A, 588, A103
Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103
Brinchmann, J., Charlot, S., White, S. D., et al. 2004, MNRAS, 351, 1151
Ciotti, L., & Ostriker, J. P. 2007, ApJ, 665, 1038
Cisternas, M., Jahnke, K., Bongiorno, A., et al. 2011, ApJL, 741, L11
Coffey, D., Salvato, M., Merloni, A., et al. 2019, A&A, 625, A123
Davis, B. L., Graham, A. W., & Cameron, E. 2018, ApJ, 869, 113
Di Matteo, T., Colberg, J., Springel, V., Hernquist, L., & Sijacki, D. 2008, ApJ, 676, 33
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Natur, 433, 604
Dwelly, T., Salvato, M., Merloni, A., et al. 2017, MNRAS, 469, 1065
Event Horizon Telescope Collaboration 2019, arXiv:1906.11238
Ferrarese, L., & Ford, H. 2005, SSRv, 116, 523
Ferrarese, L., & Merritt, D. 2000, ApJL, 539, L9
Ghez, A. M., Salim, S., Weinberg, N., et al. 2008, ApJ, 689, 1044
Gillessen, S., Eisenhauer, F., Trippe, S., et al. 2009, ApJ, 692, 1075
Graham, A. W., & Scott, N. 2013, ApJ, 764, 151
Greene, J. E., & Ho, L. C. 2005, ApJ, 630, 122
Greene, J. E., Seth, A., Kim, M., et al. 2016, ApJL, 826, L32
Grier, C., Trump, J. R., Shen, Y., et al. 2017, ApJ, 851, 21
Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332
Hao, L., Strauss, M. A., Tremonti, C. A., et al. 2005, AJ, 129, 1783
Häring, N., & Rix, H.-W. 2004, ApJL, 604, L89
Ho, L. C., & Kim, M. 2015, ApJ, 809, 123
Hopkins, P. F., Cox, T. J., Kereš, D., & Hernquist, L. 2008, ApJS, 175, 390
Hopkins, P. F., Murray, N., & Thompson, T. A. 2009, MNRAS, 398, 303
Izumi, T., Onoue, M., Matsuoka, Y., et al. 2019, PASJ, 71, 111
Kaspi, S., Brandt, W., Maoz, D., et al. 2007, ApJ, 659, 997
Kaspi, S., Maoz, D., Netzer, H., et al. 2005, ApJ, 629, 61
Kauffmann, G., Heckman, T. M., White, S. D., et al. 2003, MNRAS, 341, 33
Kormendy, J., Bender, R., & Cornell, M. 2011, Natur, 469, 374
Kormendy, J., & Gebhardt, K. 2001, Supermassive black holes in galactic nuclei, in AIP Conf. Pro. Vol 586, 20th Texas Symp. on Relativistic Astrophysics, 363
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Kormendy, J., & Kennicutt, R. C., Jr. 2004, ARA&A, 42, 603
Kormendy, J., & Richstone, D. 1995, ARA&A, 33, 581
Lin, J. Y.-Y., Pandya, S., Pratap, D., et al. 2021, arXiv:2108.07749
Liu, F., Xia, X., Mao, S., Wu, H., & Deng, Z. 2008, MNRAS, 385, 23
Liu, H.-Y., Liu, W.-J., Dong, X.-B., et al. 2019, ApJS, 243, 21
Ivezić, Ž., Lupton, R., Schlegel, D., et al. 2004, AN: Astronomical Notes, 325, 583
Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, AJ, 115, 2285
Malkan, M., & Sargent, W. 1982, ApJ, 254, 22
Marasco, A., Cresci, G., Posti, L., et al. 2021, MNRAS, 507, 4274
Merritt, D., & Ferrarese, L. 2001, MNRAS, 320, L30
Netzer, H., & Peterson, B. M. 1997, in Astronomical Time Series, 85
Oh, K., Sukyoung, K. Y., Schawinski, K., et al. 2015, ApJS, 219, 1
Pâris, I., Petitjean, P., Aubourg, É., et al. 2012, A&A, 548, A66
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Peißker, F., Eckart, A., Zajaček, M., Ali, B., & Parsa, M. 2020, ApJ, 899, 50
Pensabene, A., Carniani, S., Perna, M., et al. 2020, A&A, 637, A84
Peterson, B. M. 1993, PASP, 105, 247
Reines, A. E., & Volonteri, M. 2015, ApJ, 813, 82
Saglia, R., Opitsch, M., Erwin, P., et al. 2016, ApJ, 818, 47
Sahu, N., Graham, A. W., & Davis, B. L. 2019, ApJ, 876, 155
Schawinski, K., Urry, C. M., Virani, S., et al. 2010, ApJ, 711, 284
Schramm, M., & Silverman, J. D. 2013, ApJ, 767, 13
Shankar, F., Weinberg, D. H., & Miralda-Escudé, J. 2008, ApJ, 690, 20
Shen, Y., Hall, P. B., Horne, K., et al. 2019, ApJS, 241, 34
Shen, Y., & Kelly, B. C. 2010, ApJ, 713, 41
Shields, G. 1978, Natur, 272, 706
Shields, G., Menezes, K., Massart, C., & Bout, P. V. 2006, ApJ, 641, 683
Sijacki, D., Springel, V., Di Matteo, T., & Hernquist, L. 2007, MNRAS, 380, 877
Suh, H., Civano, F., Trakhtenbrot, B., et al. 2020, ApJ, 889, 32
Tibshirani, R. 1996, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
Wang, J.-G., Dong, X.-B., Wang, T.-G., et al. 2009, ApJ, 707, 1334
Wu, X.-B., Liu, F., & Zhang, T. 2002, A&A, 389, 742
Yang, G., Boquien, M., Buat, V., et al. 2020, MNRAS, 491, 740
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579