Towards Better Selection and Characterisation Criteria for High-Redshift Radio Galaxies Using Machine-Assisted Pattern Recognition

Rodrigo Carvajal

Supervised by
Dr J. Afonso
Dr I. Matute
Dr H. Messias

Ciências ULisboa
instituto de astrofísica e ciências do espaço

PD F FCT PhD PROGRAMMES
COFINANCIAMENTO / COFINANCING
fct Fundação para a Ciência e a Tecnologia
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

Towards Better Selection and Characterisation Criteria for High-Redshift Radio Galaxies Using Machine-Assisted Pattern Recognition

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) through the grant UIDP/04434/2020, UIDB/04434/2020, and the PhD Fellowship PD/BD/150455/2019 (PhD::SPACE Doctoral Network PD/00040/2012) and POCH/FSE (EC).

Stargazing LIVE is a BBC and Open University co-production. Credit: Photography sourced from NASA.

Stargazing LIVE is a BBC and Open University co-production. Credit: Photography sourced from NASA.

Need to understand role of SMBH in galaxy evolution!

Need to understand role of SMBH in galaxy evolution!

AGN + SF

Need to understand role of SMBH in galaxy evolution!

AGN + SF

Redshifts

Outline

Separating AGN from SFGs

Our approach

Results

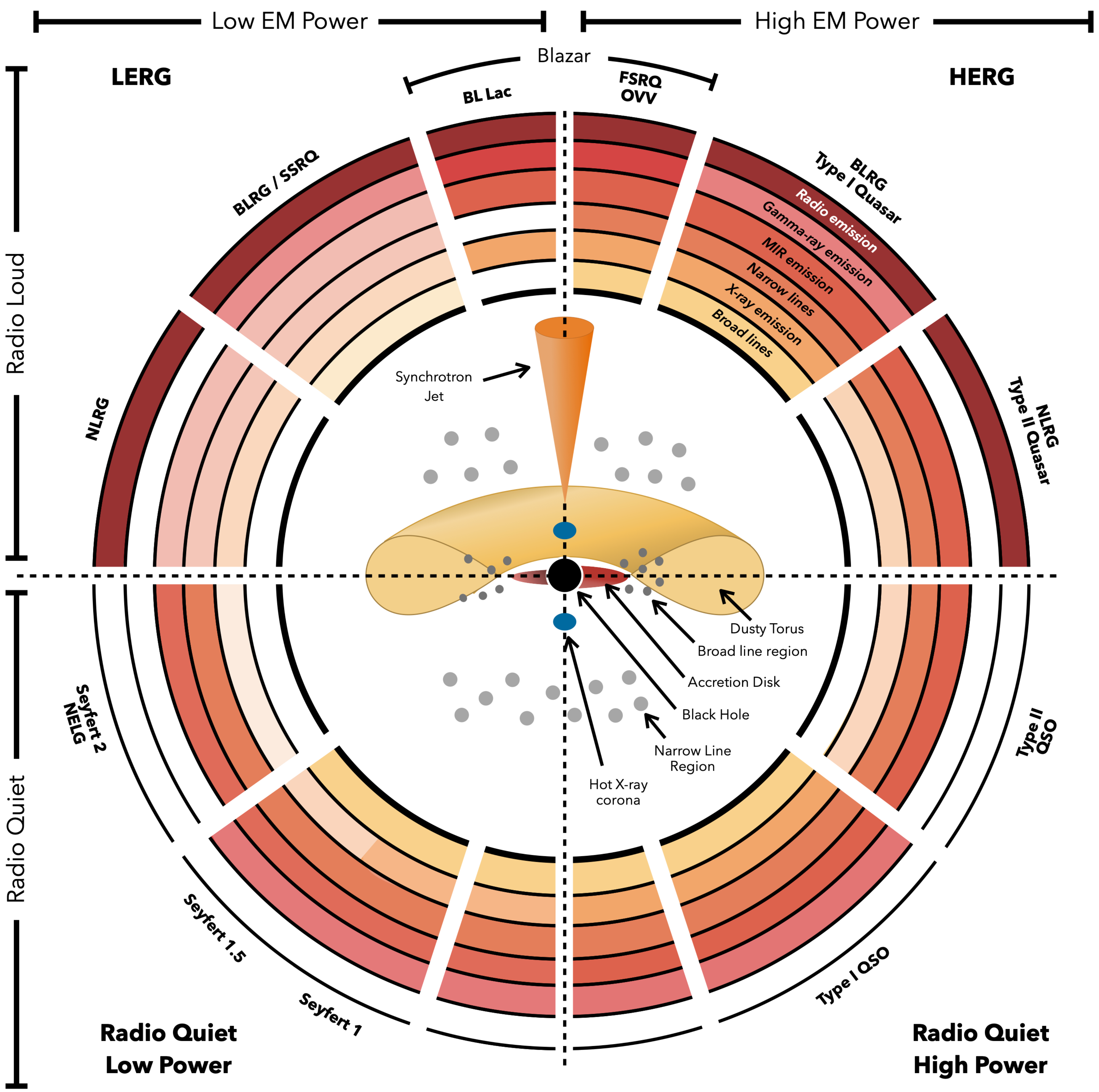
Analysing our tool

Conclusions and future work

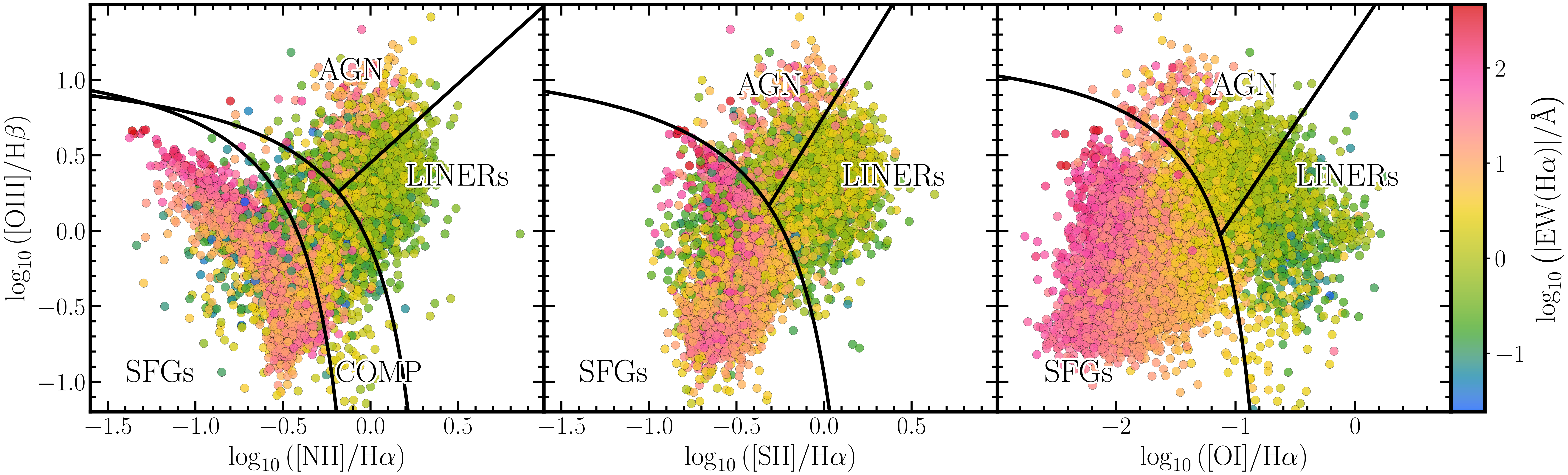
Separating AGN from SFG (non-AGN)

DETECT AGN THROUGH DIFFERENT WAVELENGTHS

ACCESSING VARIOUS PHYSICAL PROCESSES AND PROPERTIES

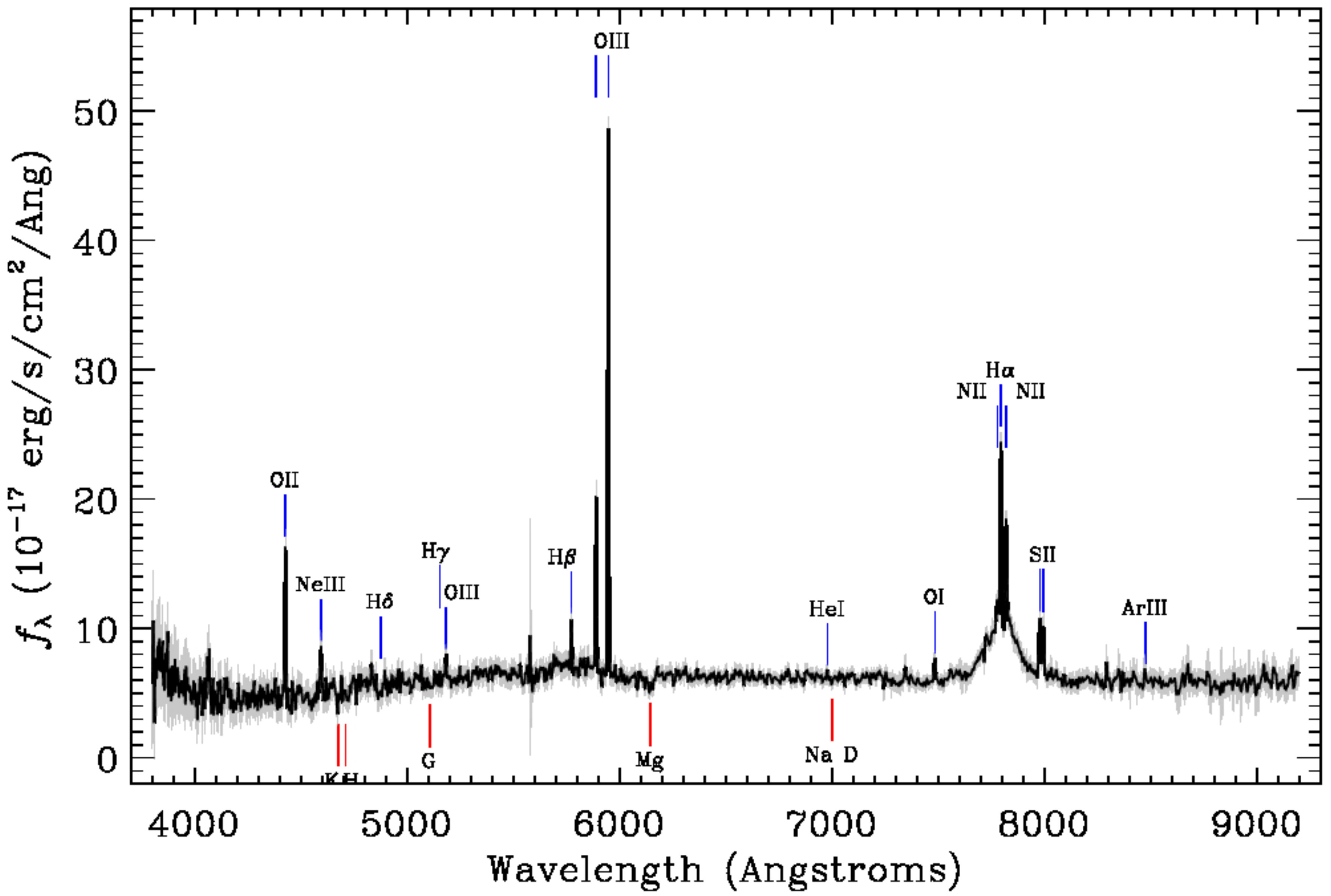


J. E. Thorne

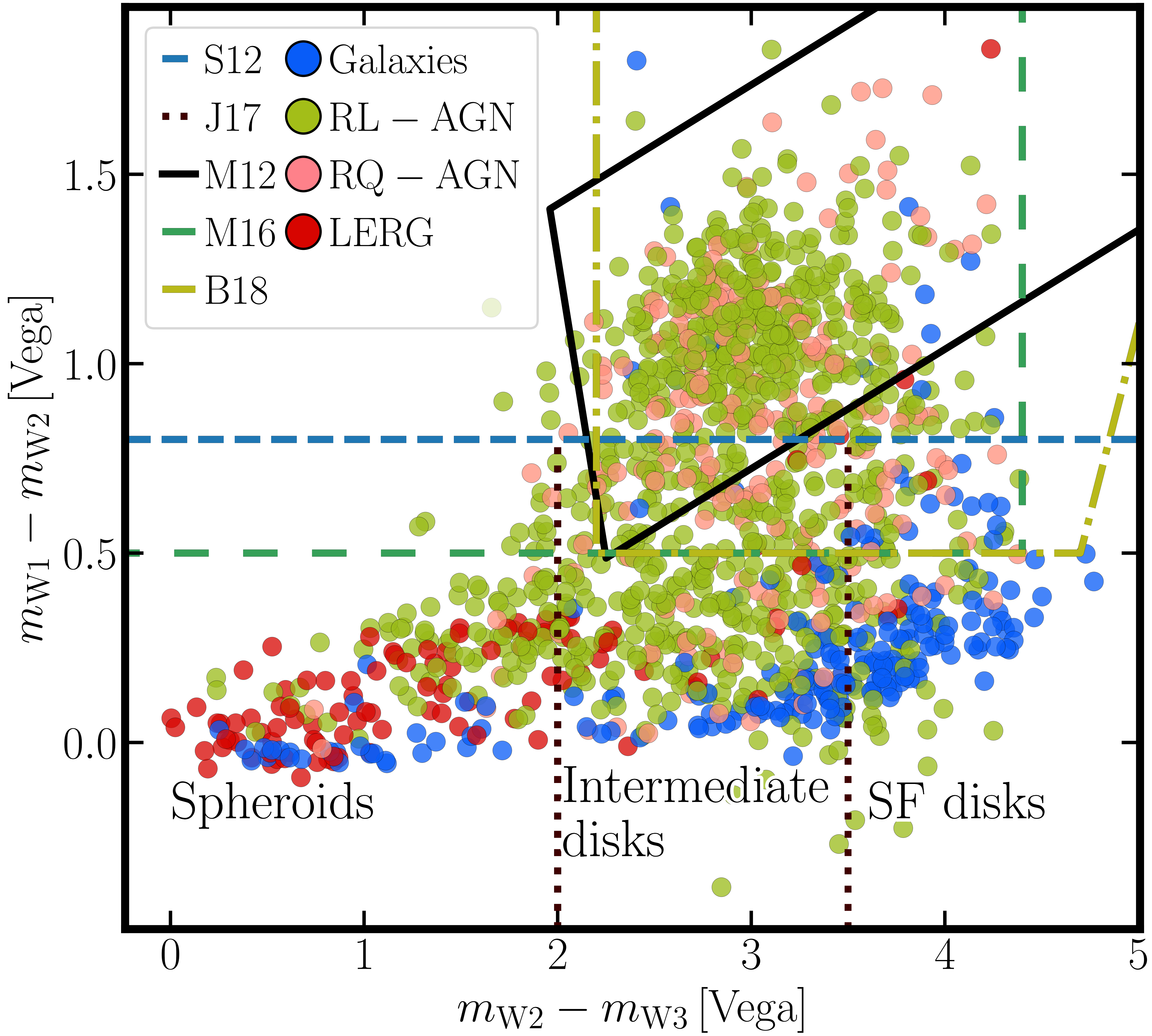


BPT Diagram

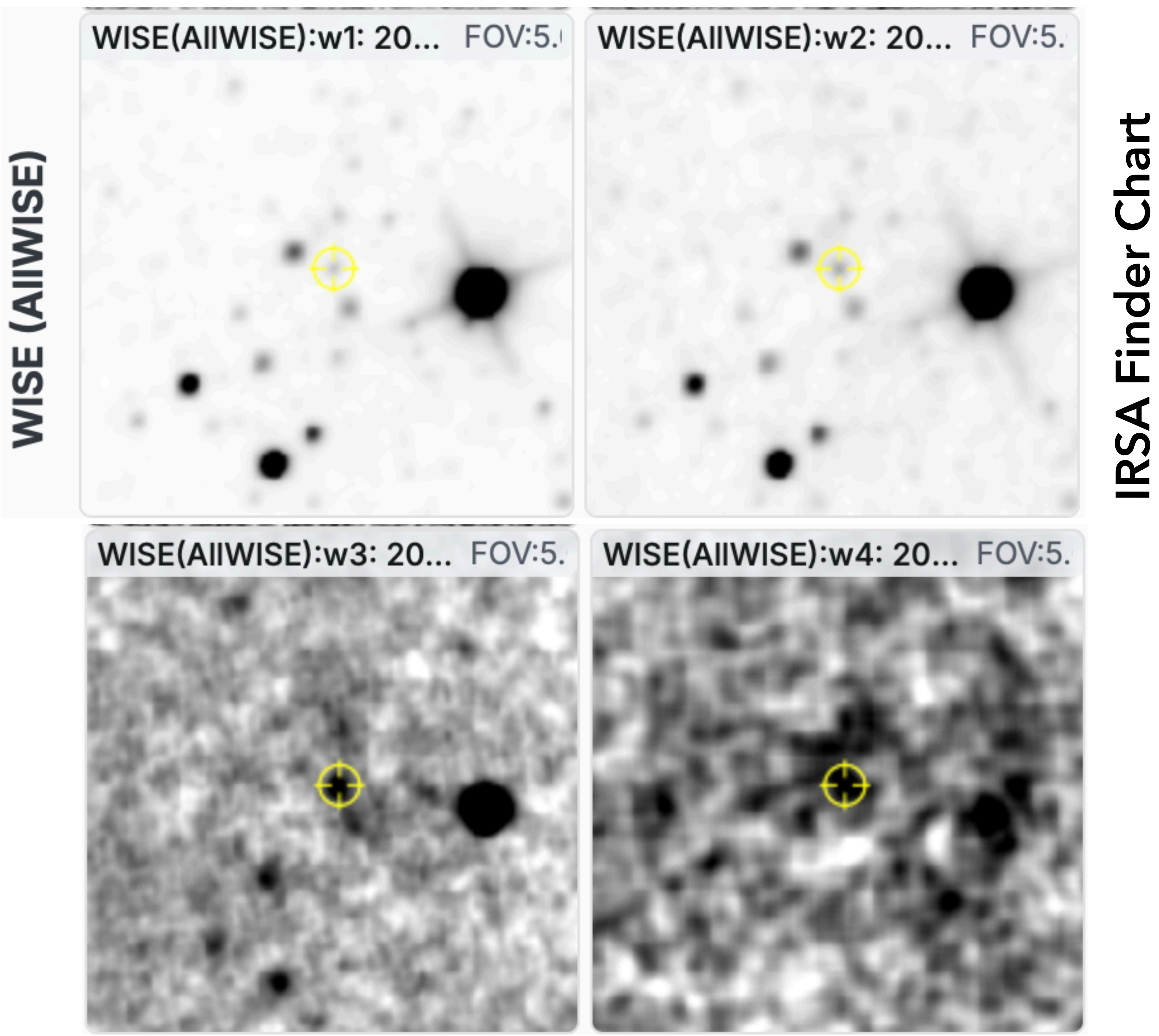
SPECTROSCOPY



SDSS DR18 Quick look



PHOTOMETRY



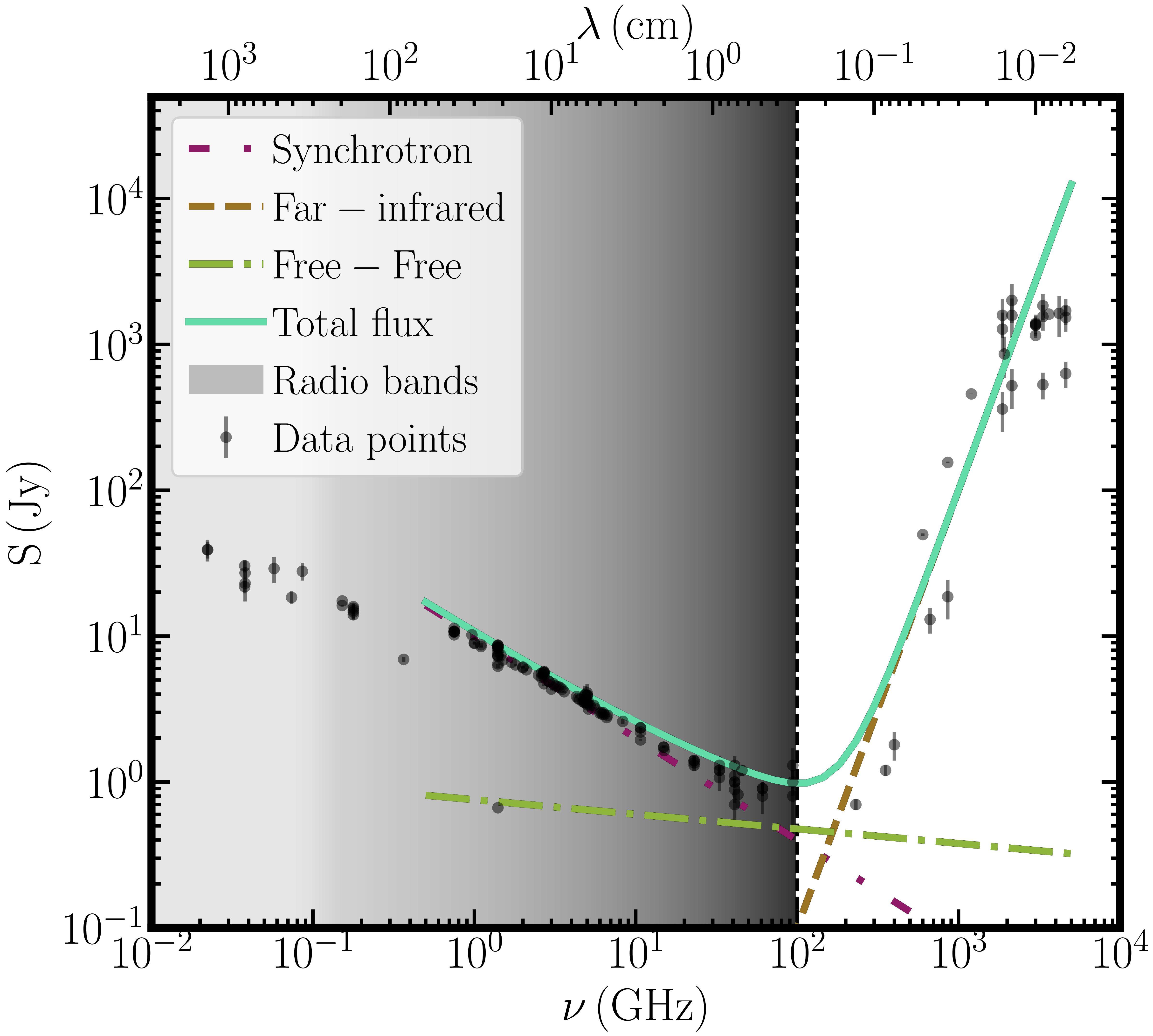
For ex. WISE Colours

ALMOST DIRECT OBSERVATION OF AGN

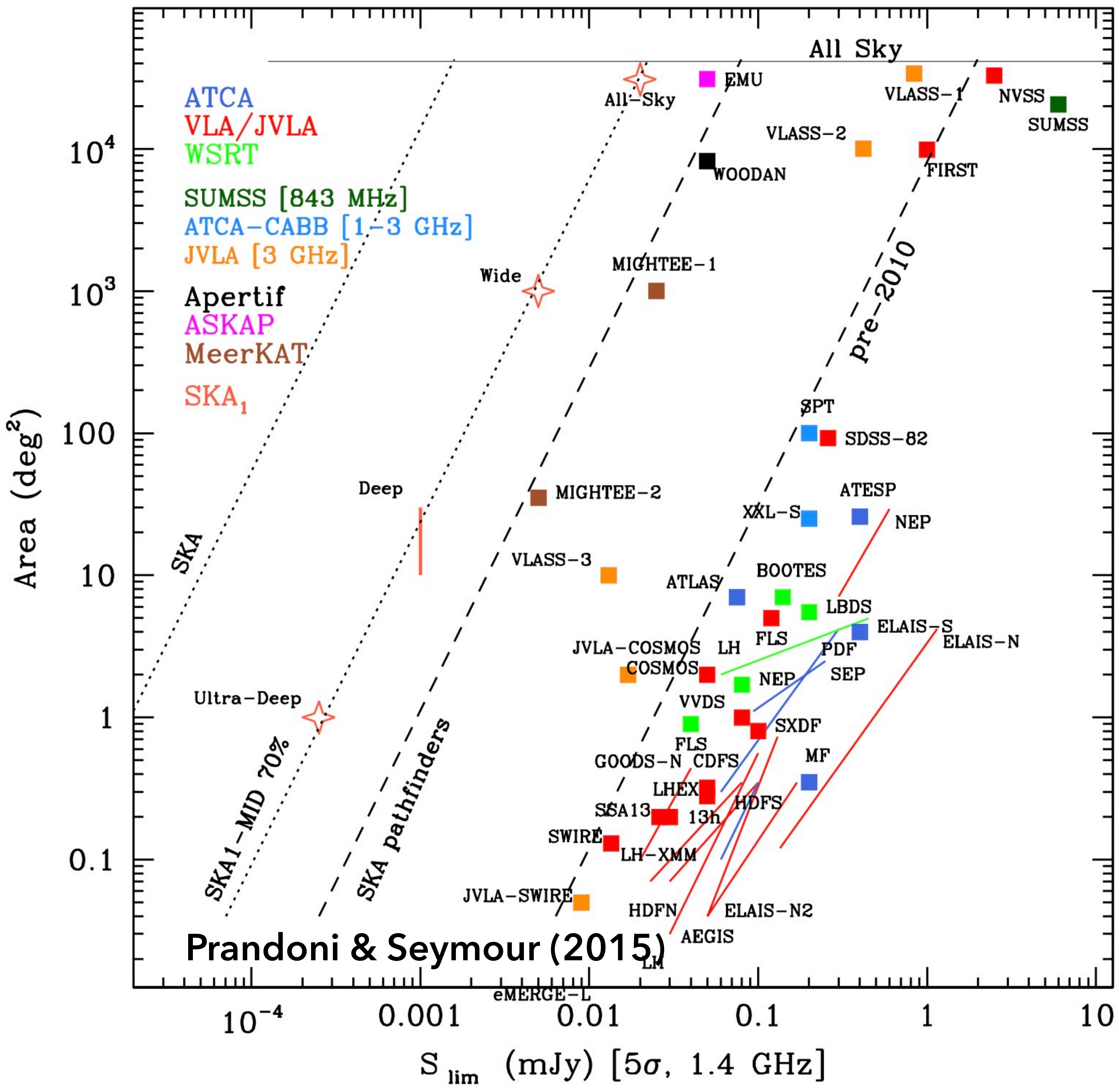
SKA + PRECURSORS

LARGE DATASETS

RADIO BANDS



BIG-DATA ERA



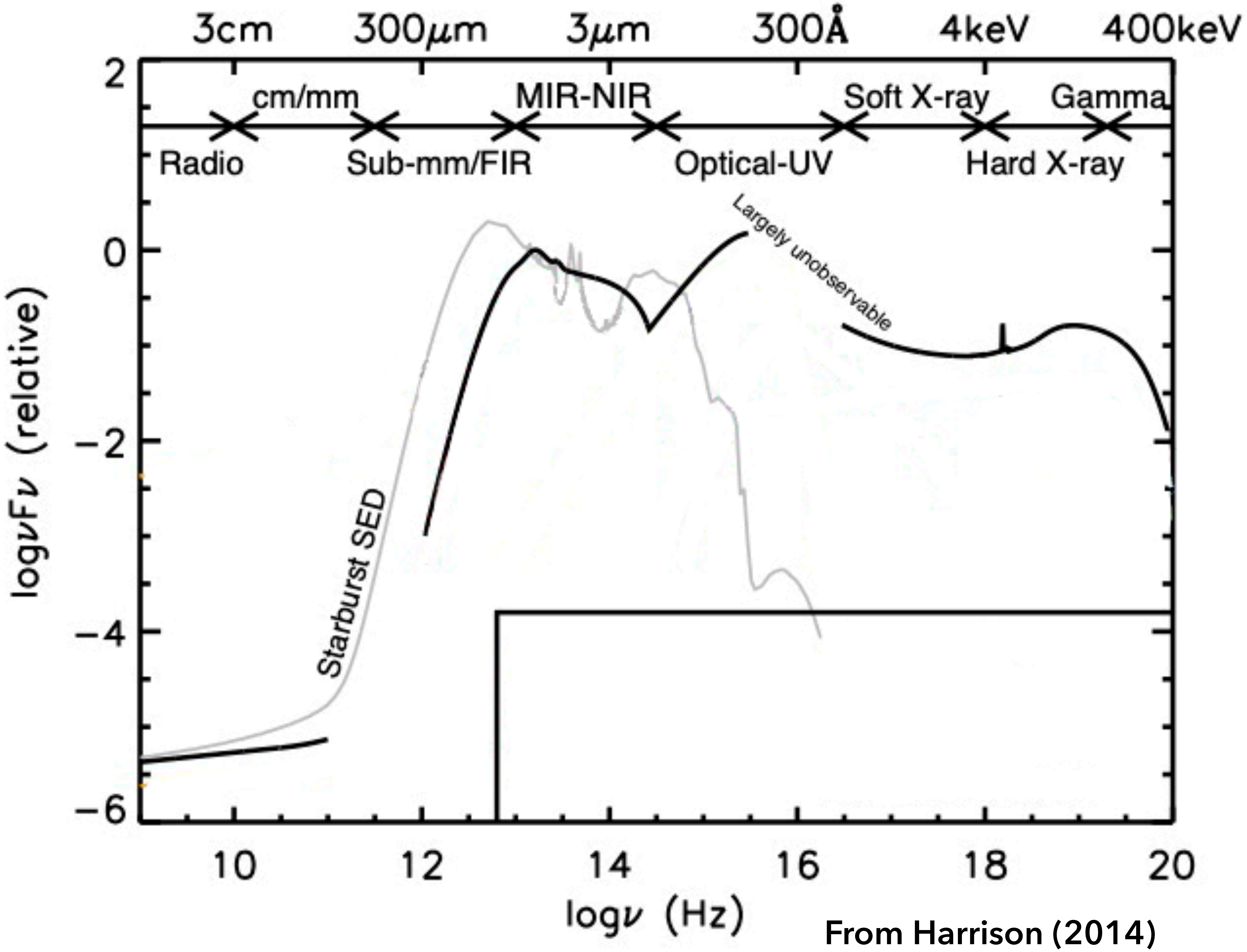
All Sky
ATCA
VLA/JVLA
WSRT
SUMSS [843 MHz]
ATCA-CABB [1-3 GHz]
JVLA [3 GHz]
Apertif
ASKAP
MeerKAT
SKA$_1$
EMU
All-Sky
VLASS-1
NVSS
SUMSS
VLASS-2
FIRST
WOODAN
MIGHTEE-1
Wide
pre-2010
SPT
SDSS-82
Deep
MIGHTEE-2
ATESP
XXL-S
NEP
SKA
VLASS-3
ATLAS
BOOTES
LBDS
FLS
ELAIS-S
ELAIS-N
JVLA-COSMOS
LH
PDF
COSMOS
NEP
SEP
VVDS
Ultra-Deep
SXDF
SKA1-MID 70%
SKA pathfinders
FLS
GOODS-N
CDFS
MF
LHEX
SSA13
13h
HDFS
SWIRE
LH-XMM
JVLA-SWIRE
HDFN
ELAIS-N2
AEGIS
LH
eMERGE-L
Prandoni & Seymour (2015)
Area (deg^2)
S_{lim} (mJy) [5σ, 1.4 GHz]

We need a multi-wavelength approach!

Is it possible to incorporate several indicators into one tool?

SED FITTING

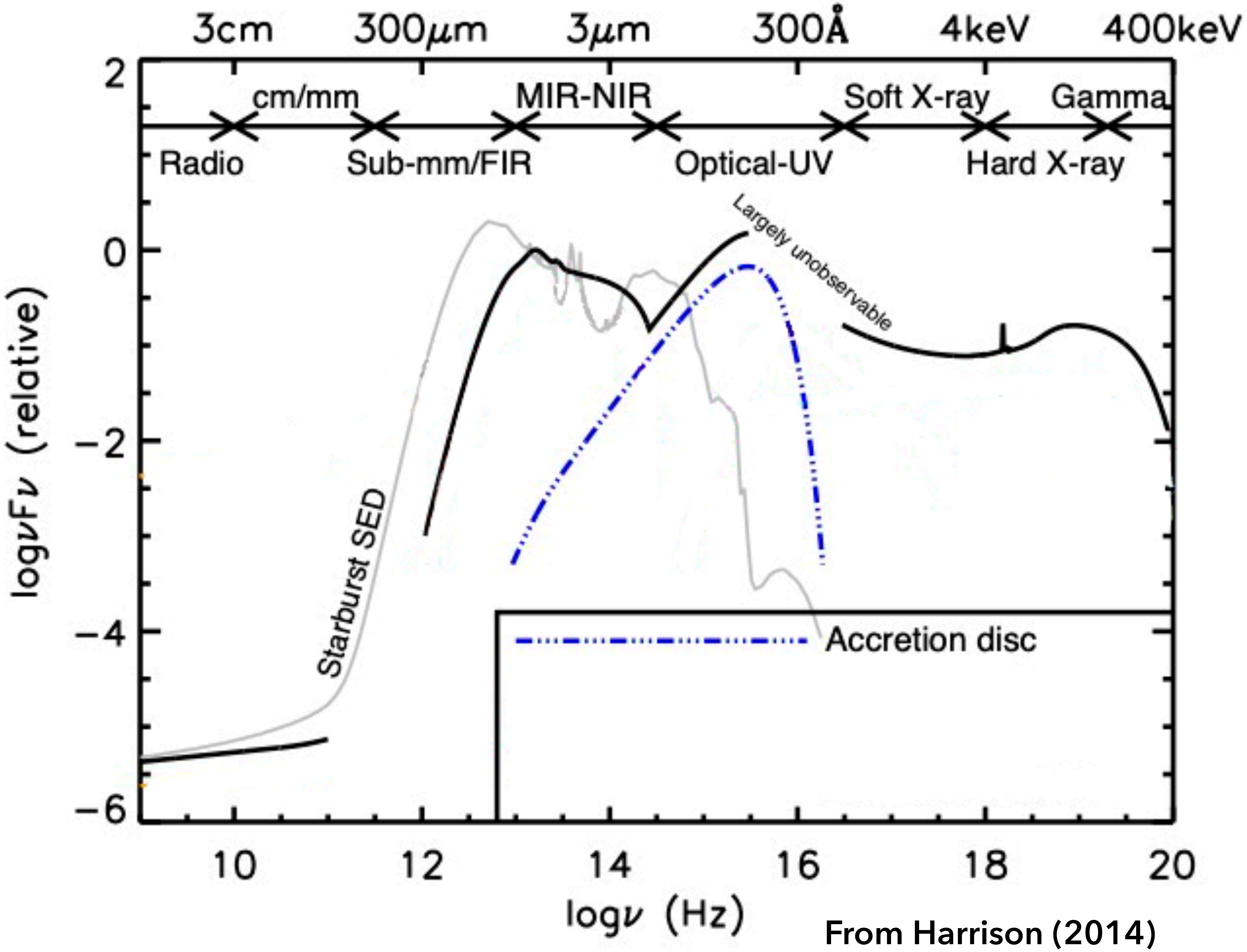
ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



From Harrison (2014)

SED Fitting

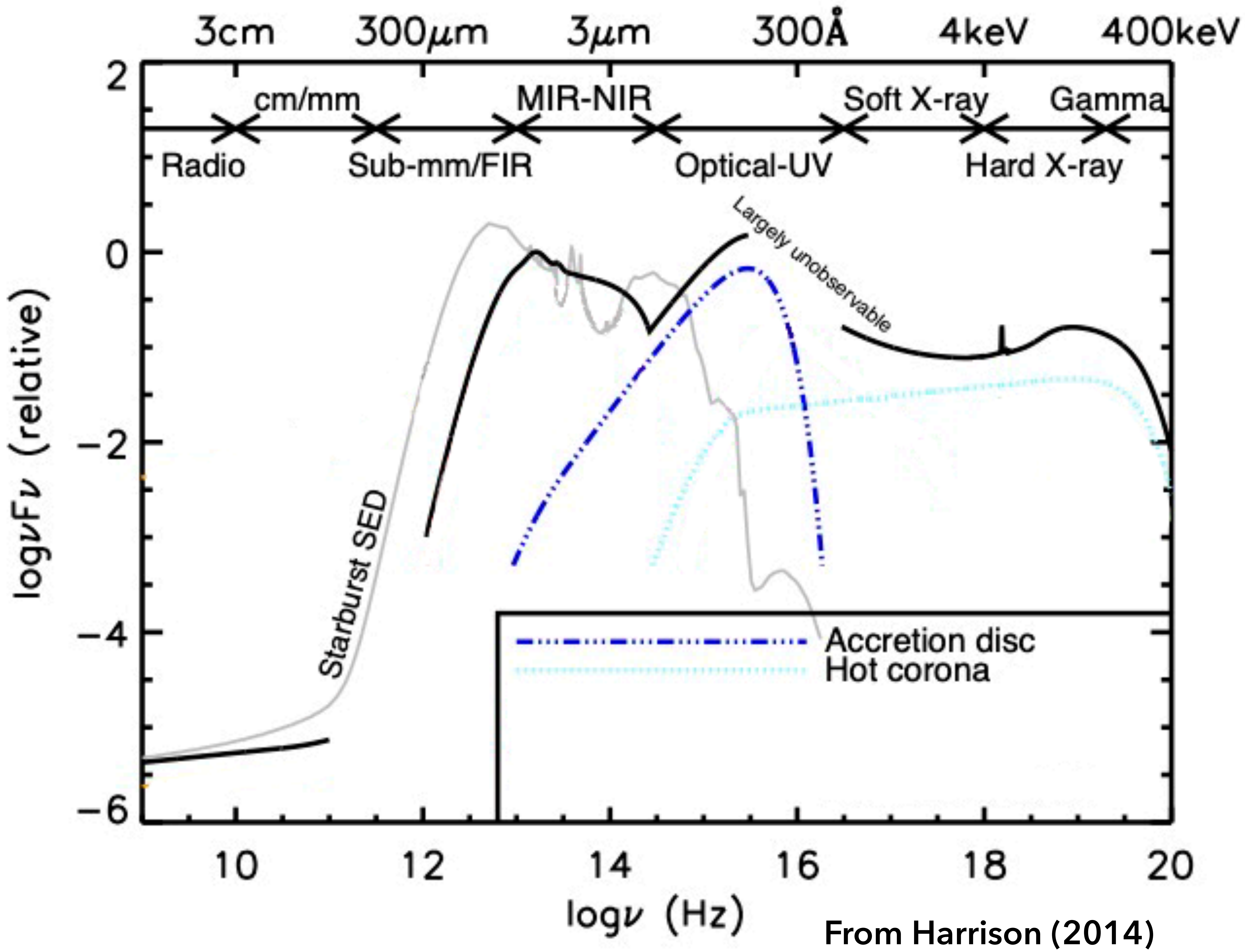
One way to do multi-wavelength analysis



From Harrison (2014)

SED FITTING

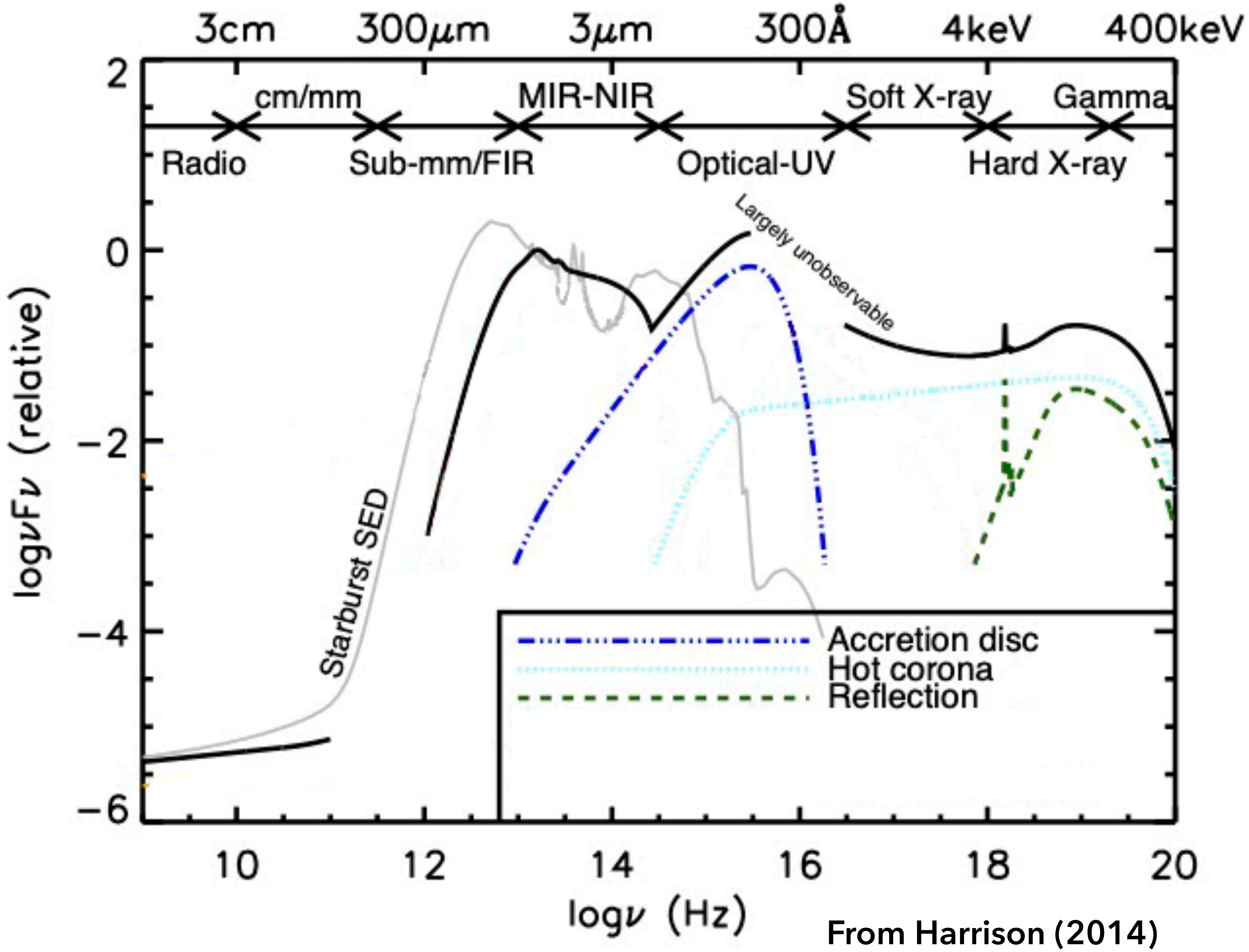
ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



From Harrison (2014)

SED FITTING

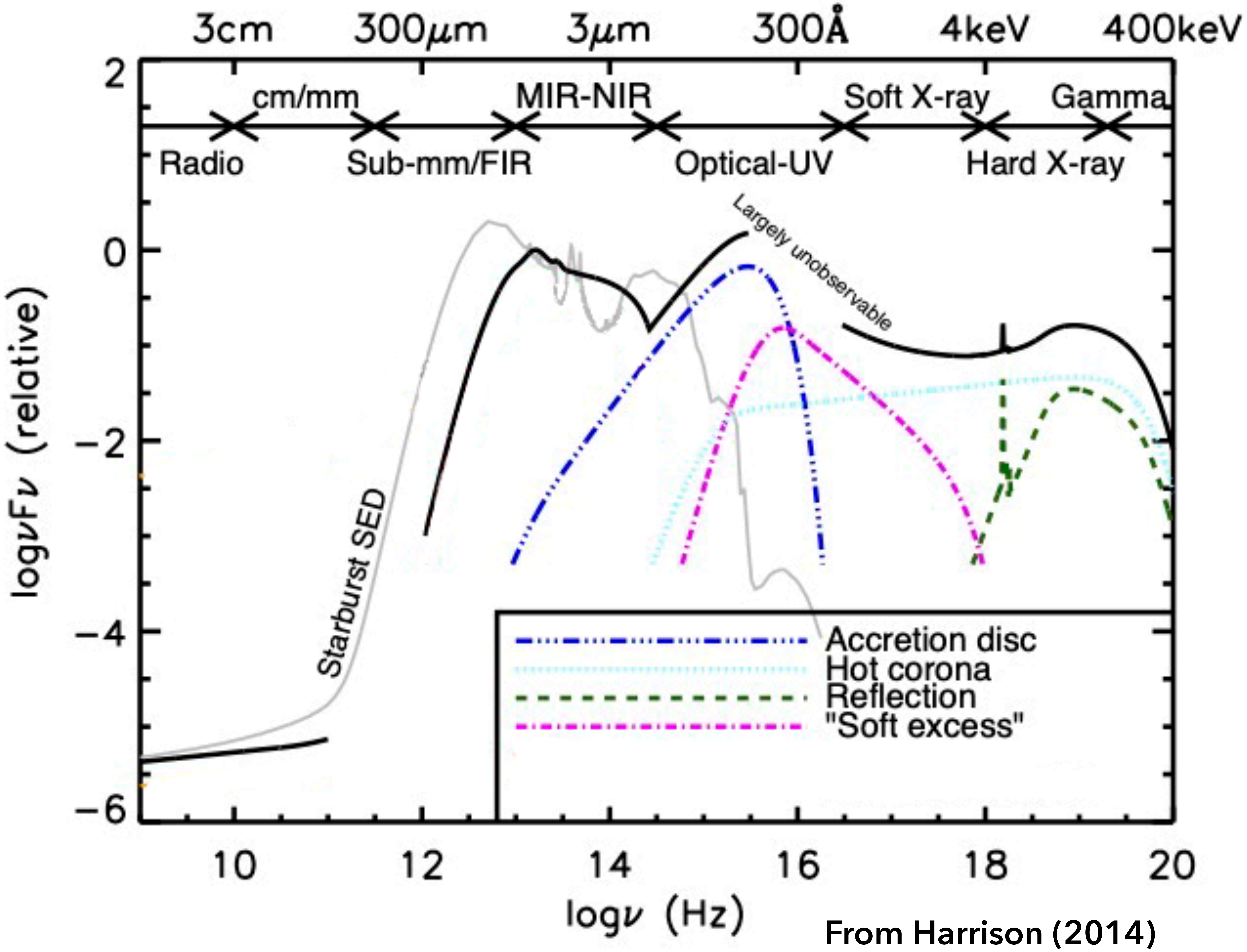
ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



From Harrison (2014)

SED Fitting

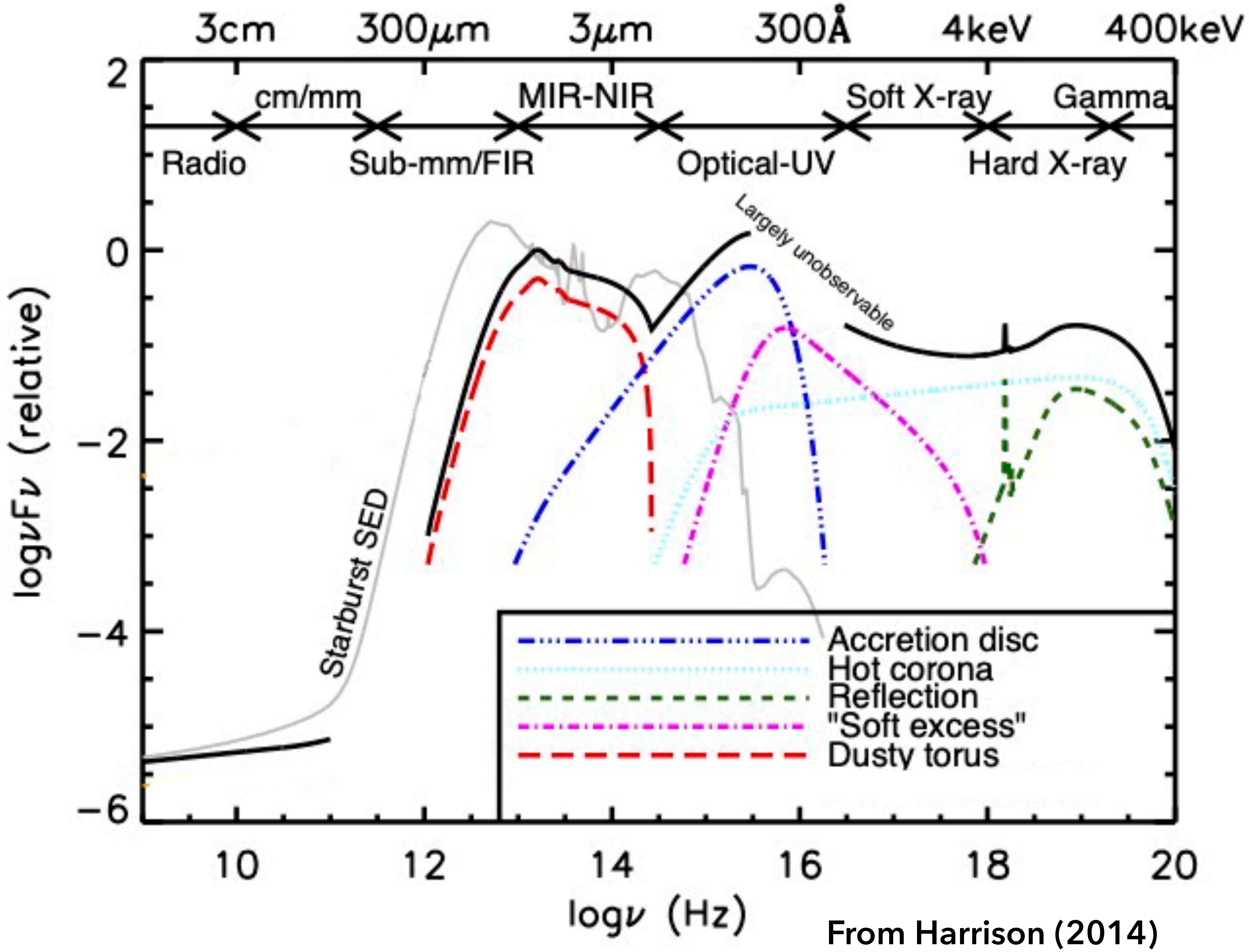
One way to do multi-wavelength analysis



From Harrison (2014)

SED FITTING

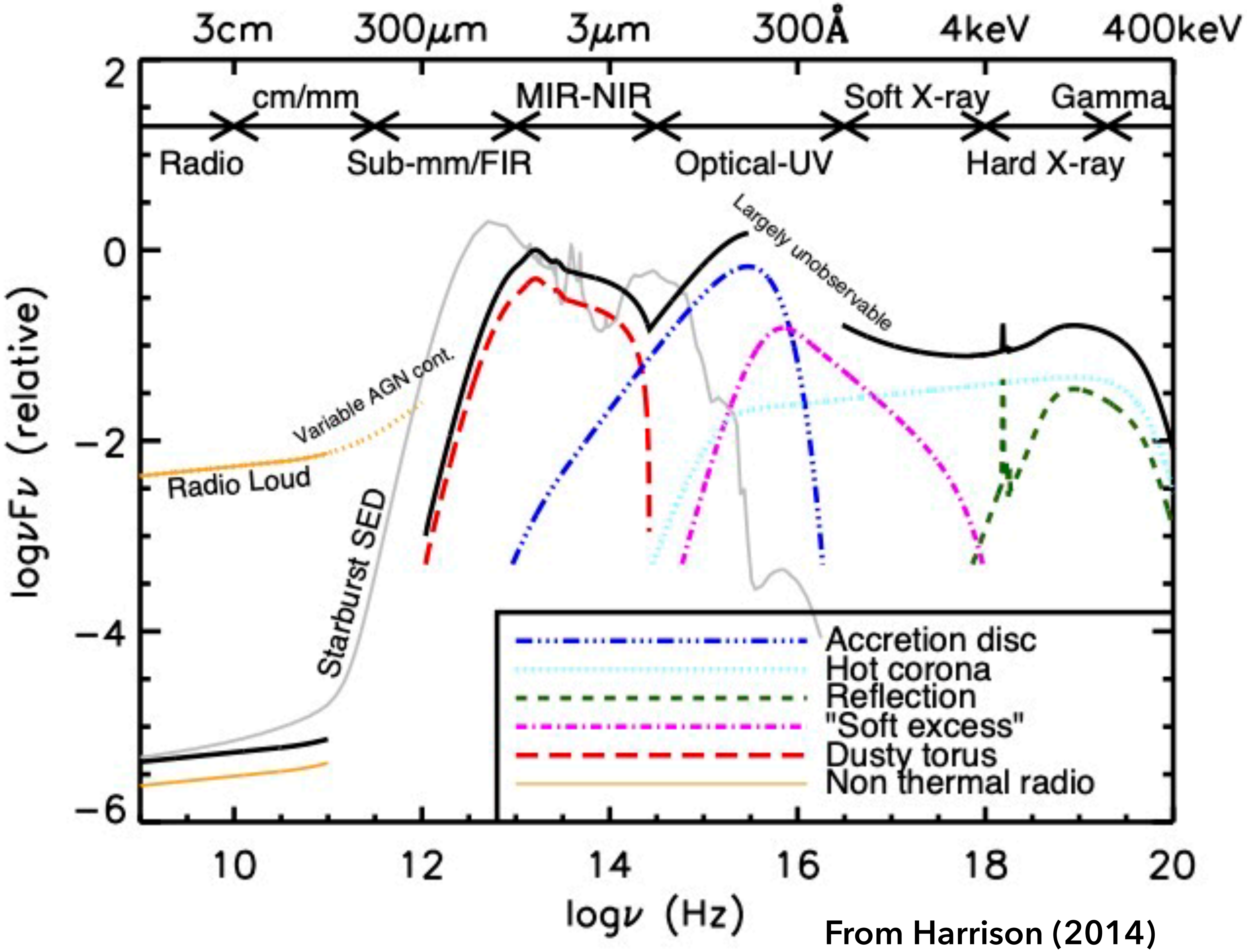
ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



From Harrison (2014)

SED FITTING

ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



From Harrison (2014)

Machine Learning can help!

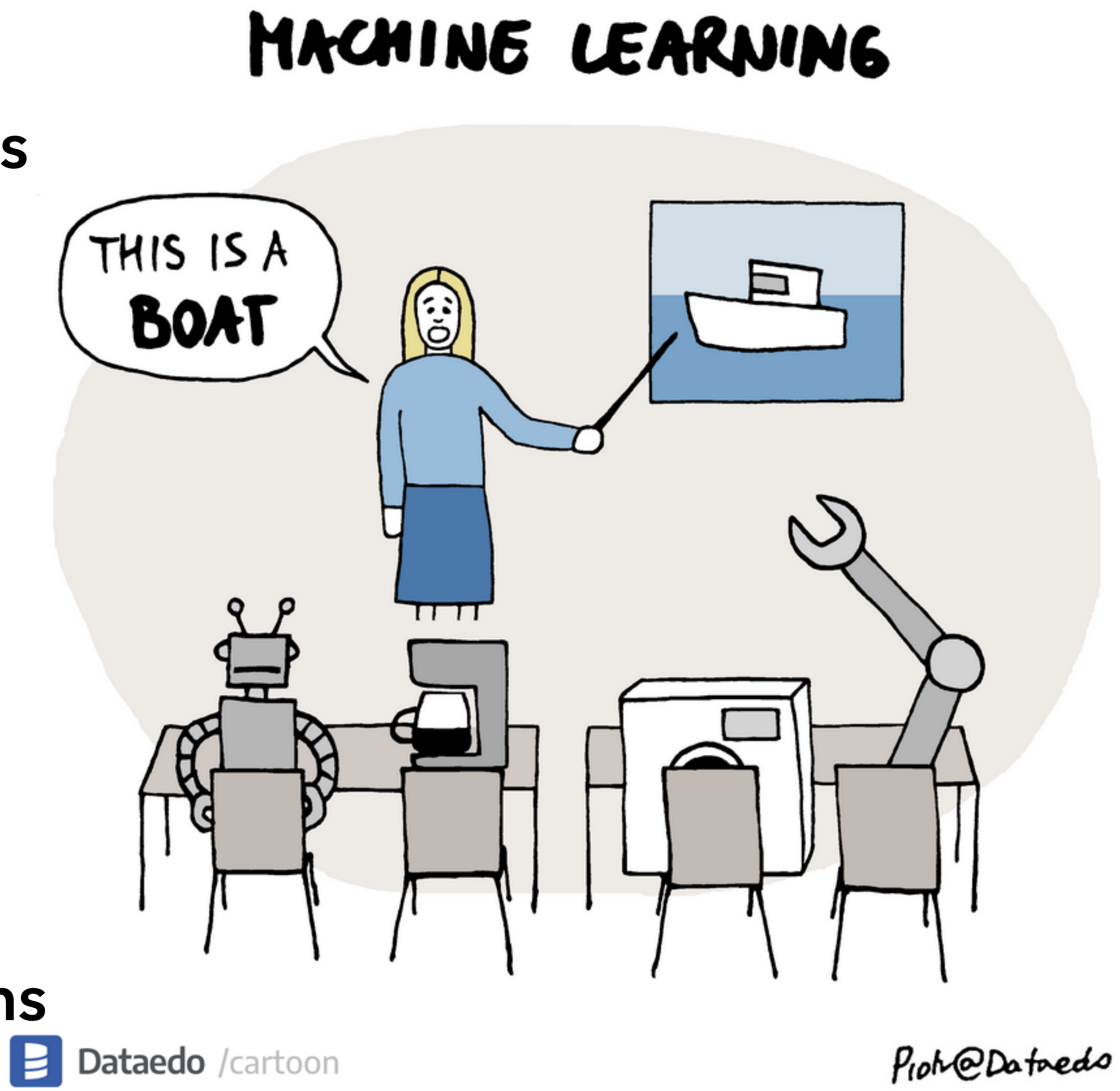
MACHINE LEARNING

It can take advantage of very high dimensional datasets

It can determine patterns and trends within data and apply them to new measurements

We can examine predictions to further understand physical processes

It can guide us towards otherwise hidden research paths



Our goal

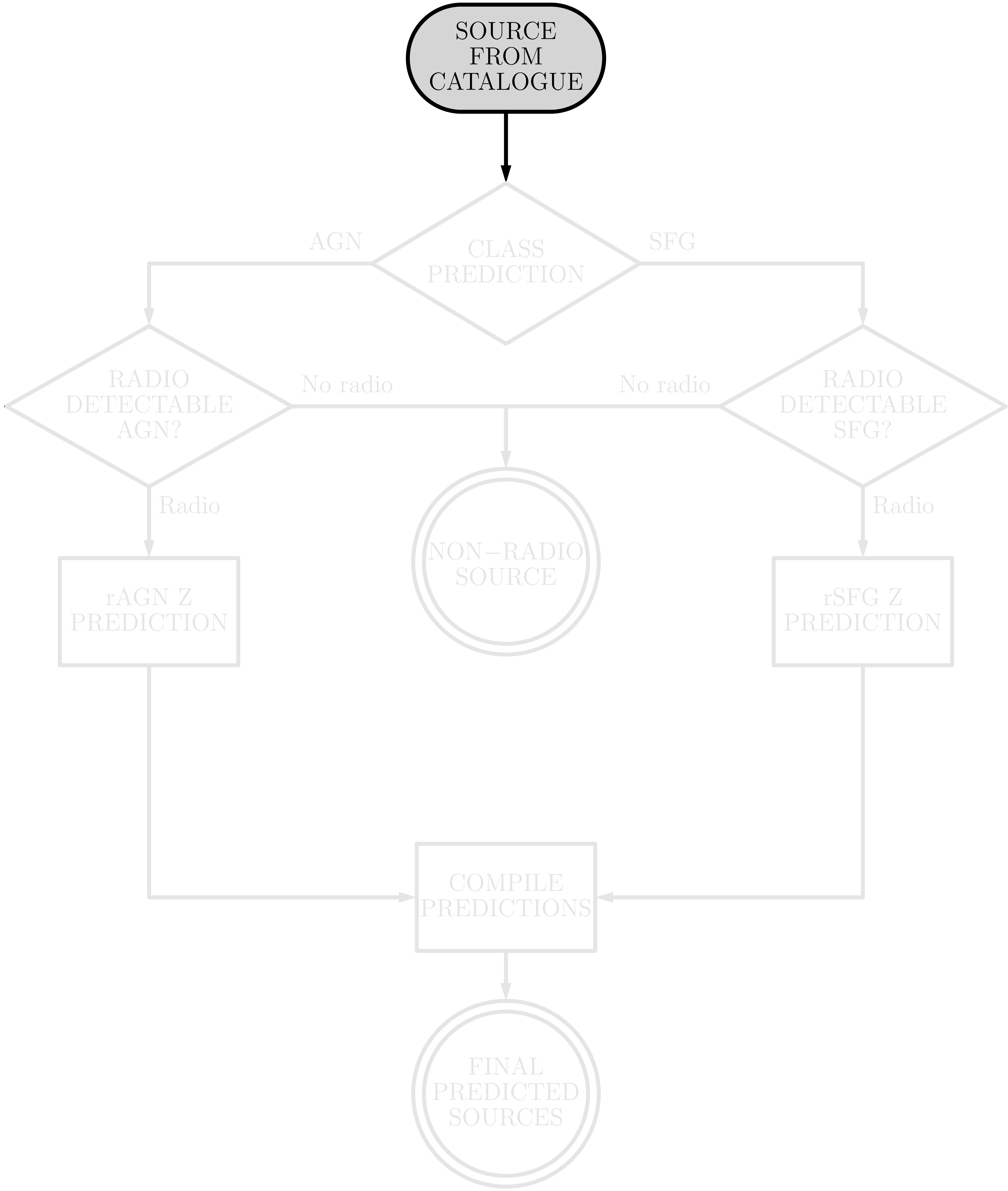
Establish a method to select and characterise radio-detected sources from large datasets

PREDICTION PIPELINE

OUR APPROACH TO SELECTING AGN AND SFG WITH ML

Three different levels

- Classify as AGN or SFG
- Select radio-detectable sources
- Estimate redshift for radio-detectable sources



PREDICTION PIPELINE

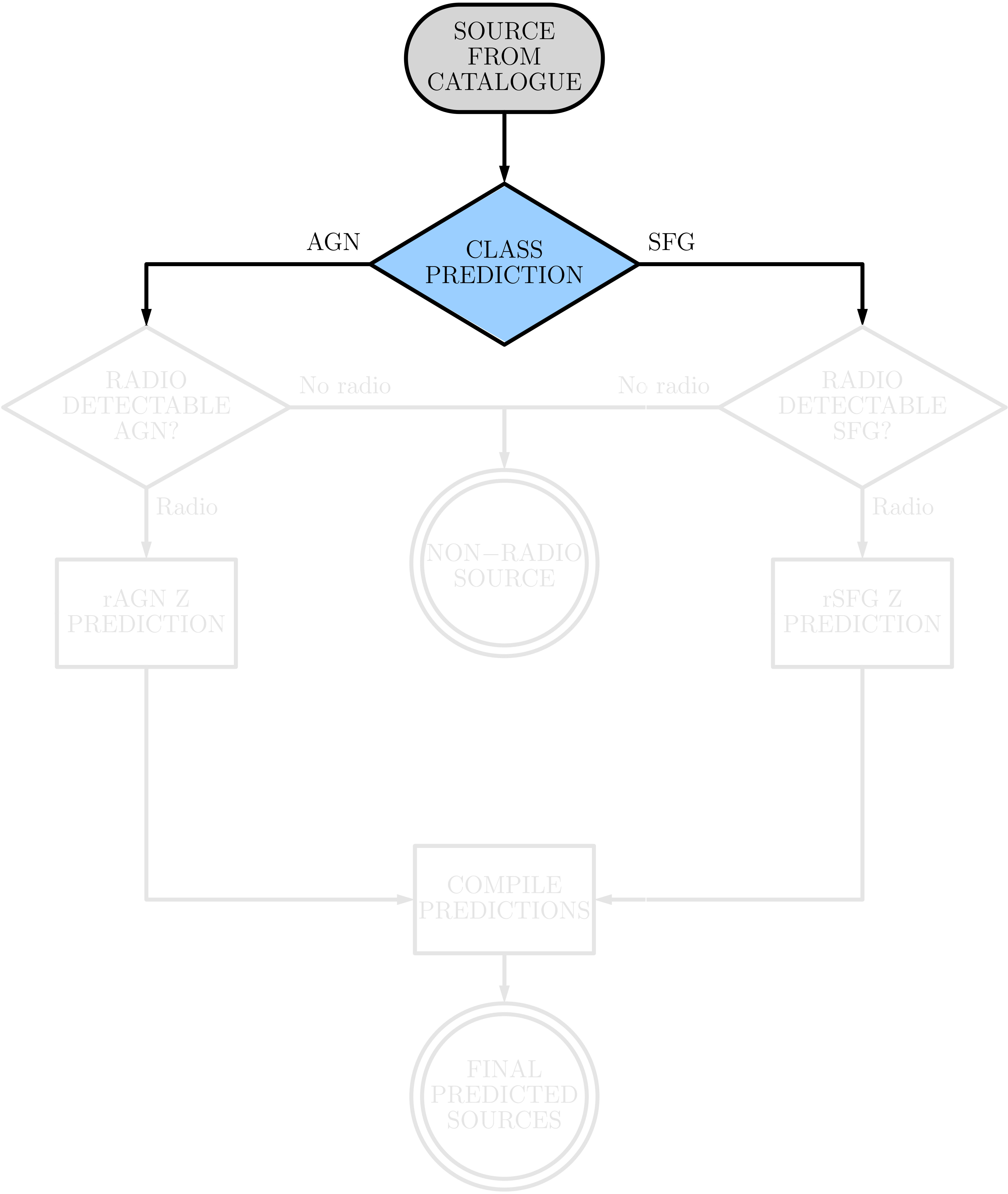
OUR APPROACH TO SELECTING AGN AND SFG WITH ML

Three different levels

Classify as AGN or SFG

Select radio-detectable sources

Estimate redshift for radio-detectable sources



PREDICTION PIPELINE

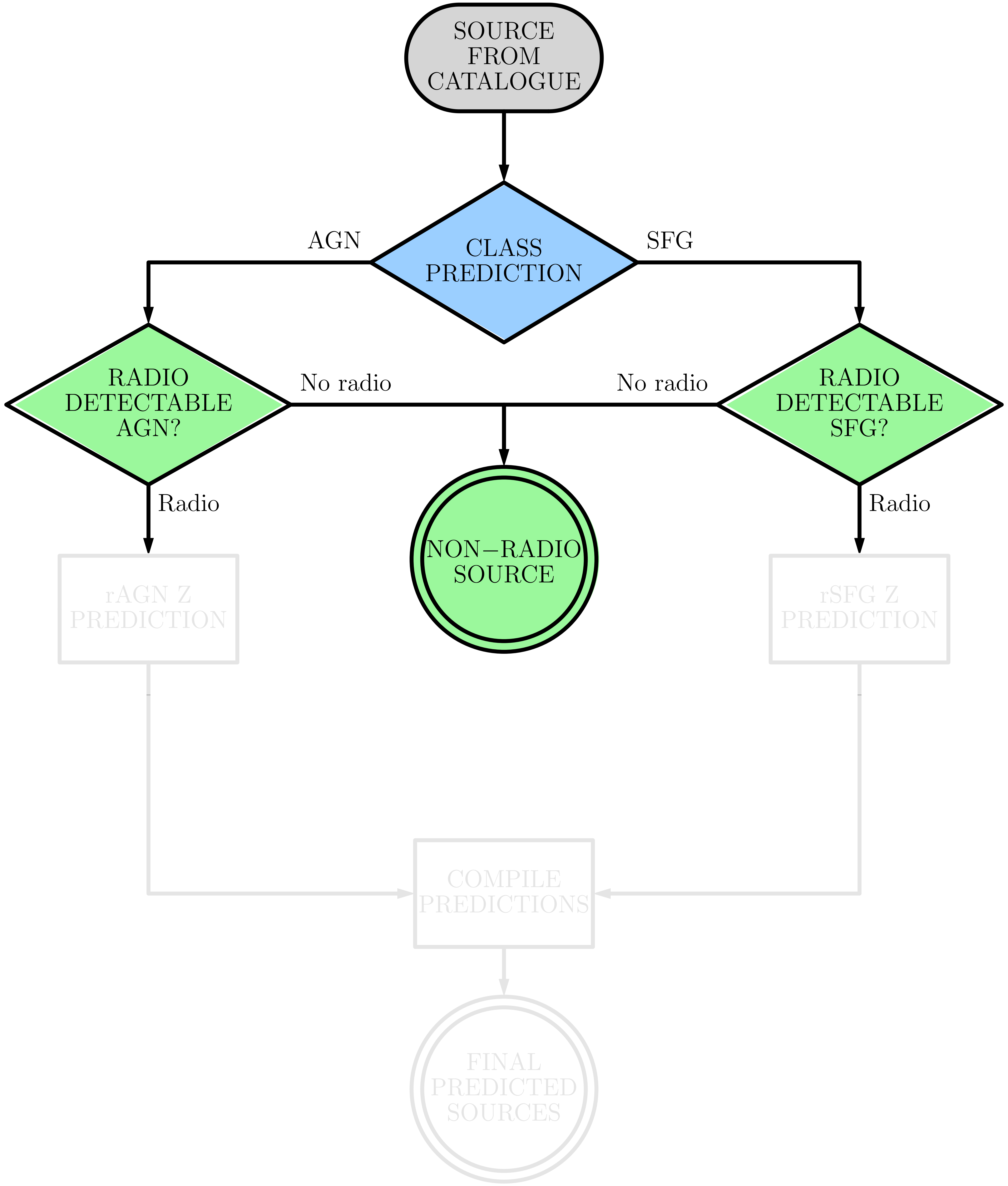
OUR APPROACH TO SELECTING AGN AND SFG WITH ML

Three different levels

Classify as AGN or SFG

Select radio-detectable sources

Estimate redshift for radio-detectable sources



Prediction Pipeline

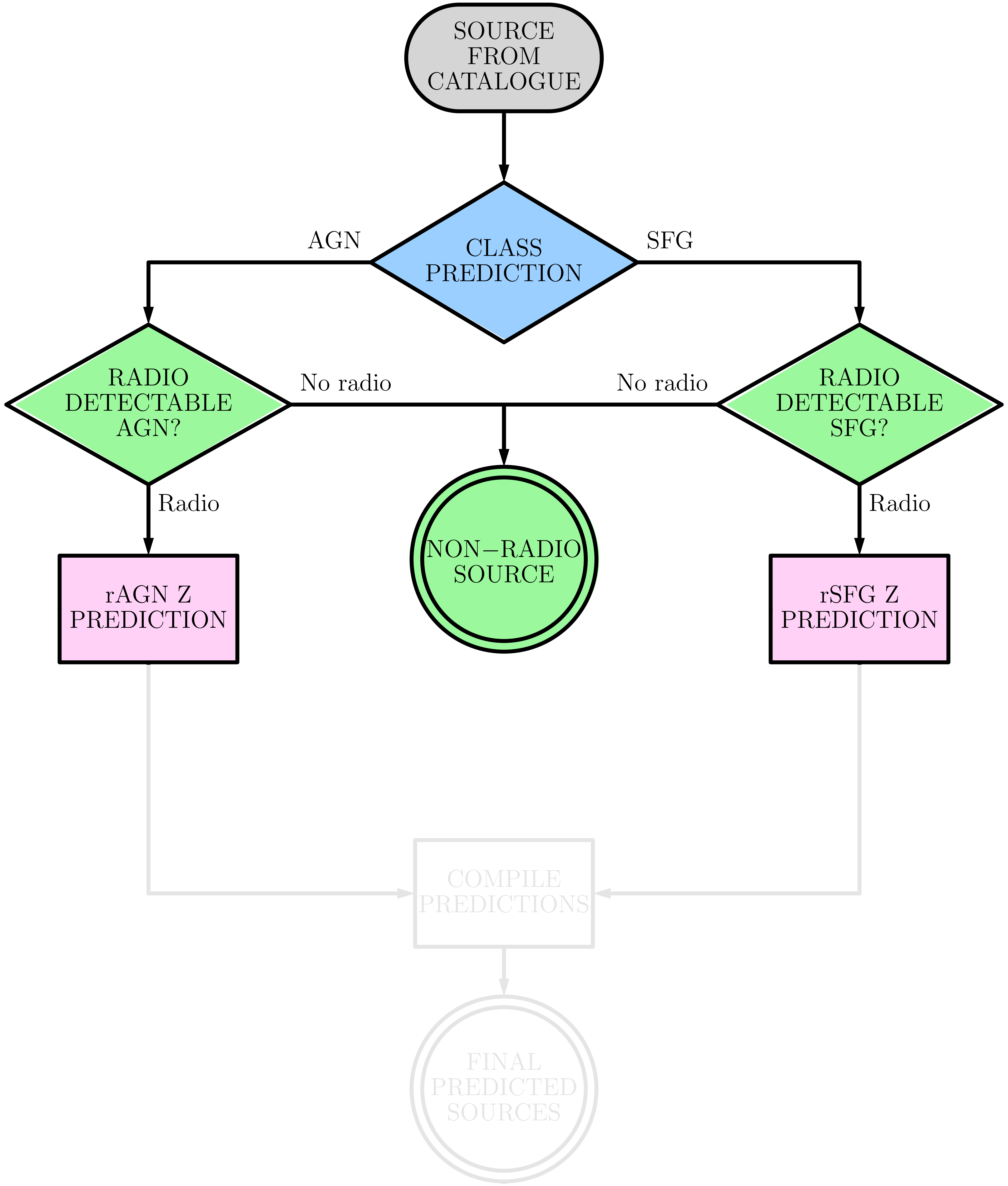
Our approach to selecting AGN and SFG with ML

Three different levels

Classify as AGN or SFG

Select radio-detectable sources

Estimate redshift for radio-detectable sources



PREDICTION PIPELINE

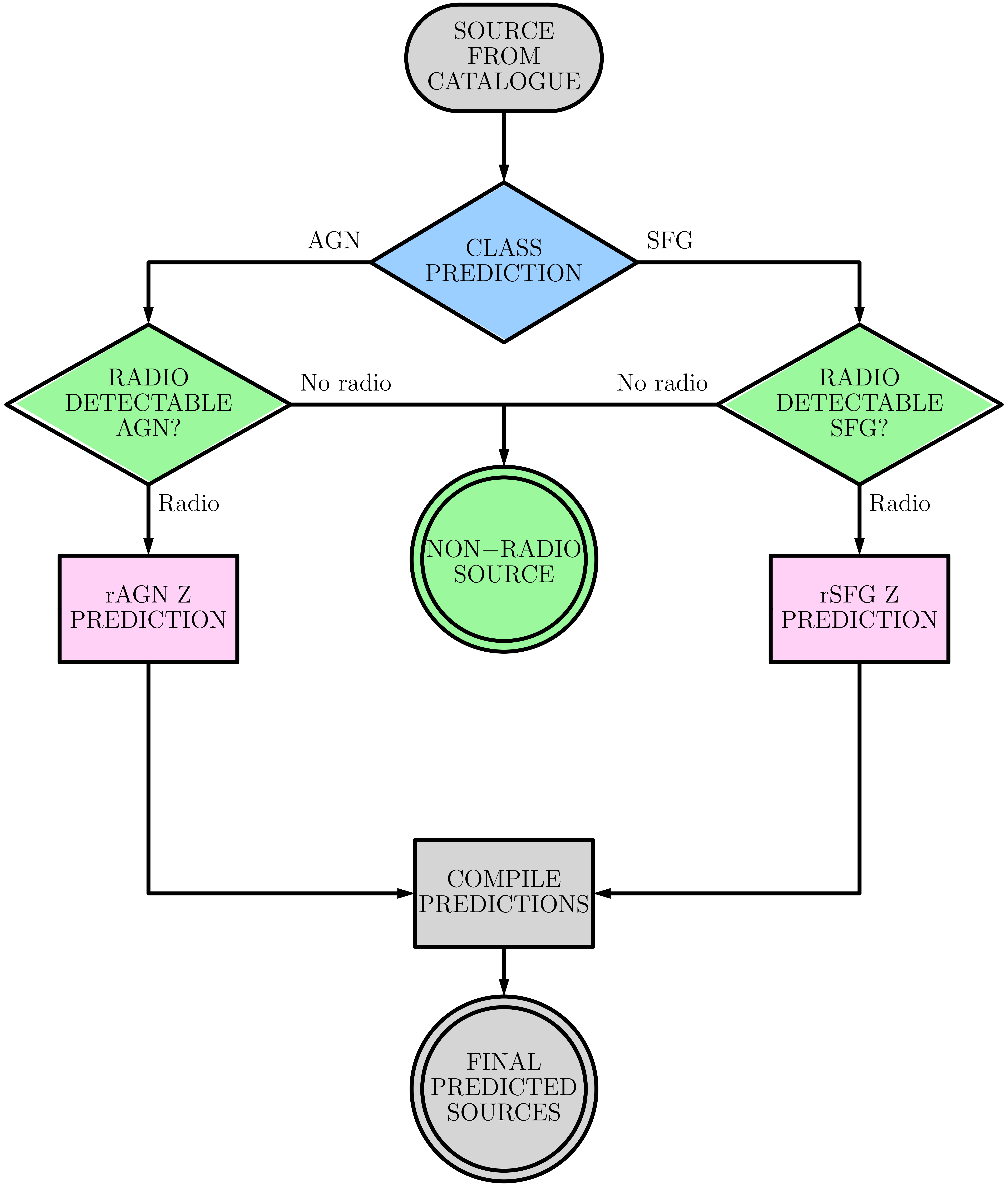
OUR APPROACH TO SELECTING AGN AND SFG WITH ML

Three different levels

Classify as AGN or SFG

Select radio-detectable sources

Estimate redshift for radio-detectable sources



PREDICTION PIPELINE

OUR APPROACH TO SELECTING AGN AND SFG WITH ML

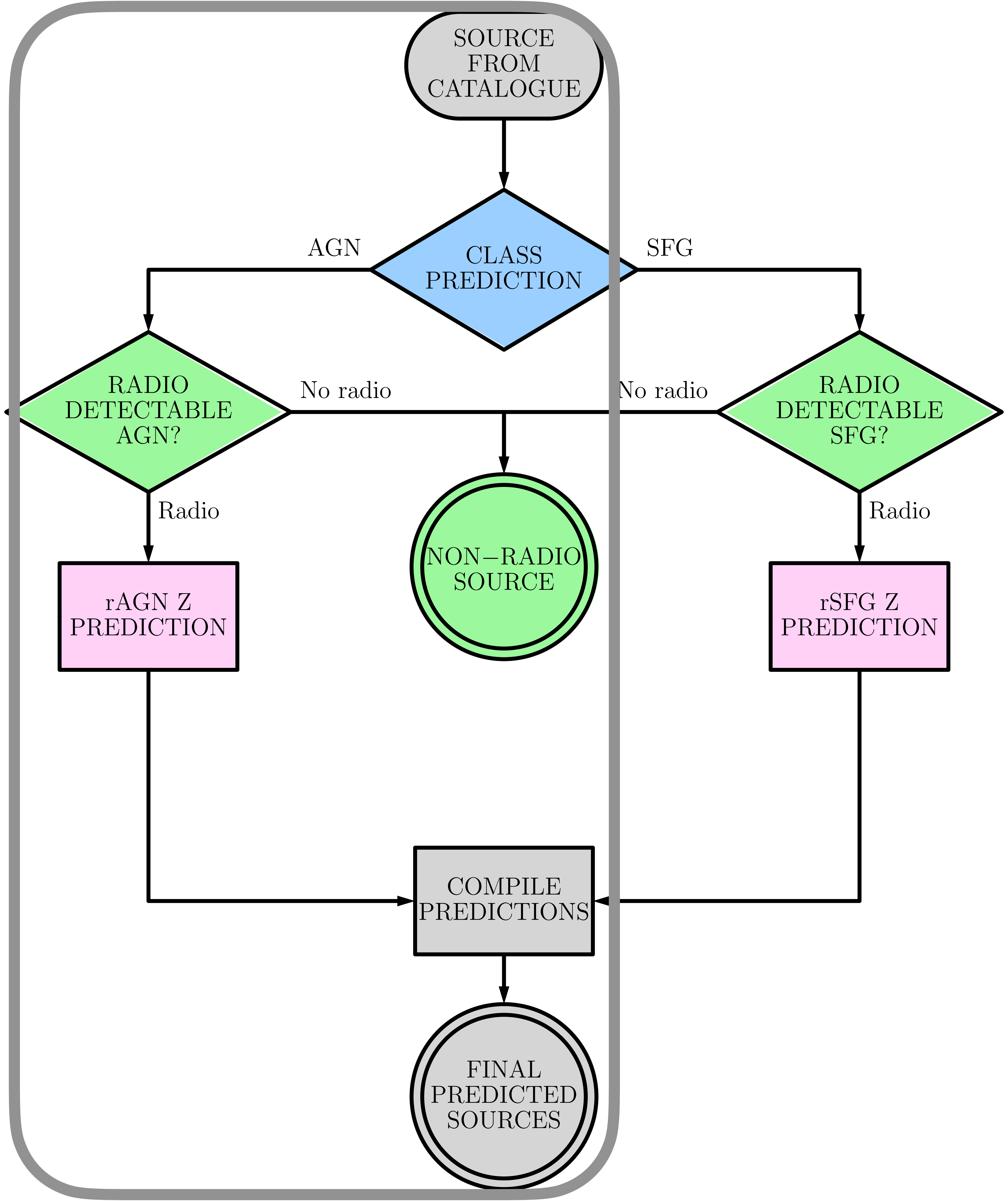
Three different levels

Classify as AGN or SFG

Select radio-detectable sources

Estimate redshift for radio-detectable sources

Carvajal et al. (2021)
+
Carvajal et al. (2023)



PREDICTION PIPELINE

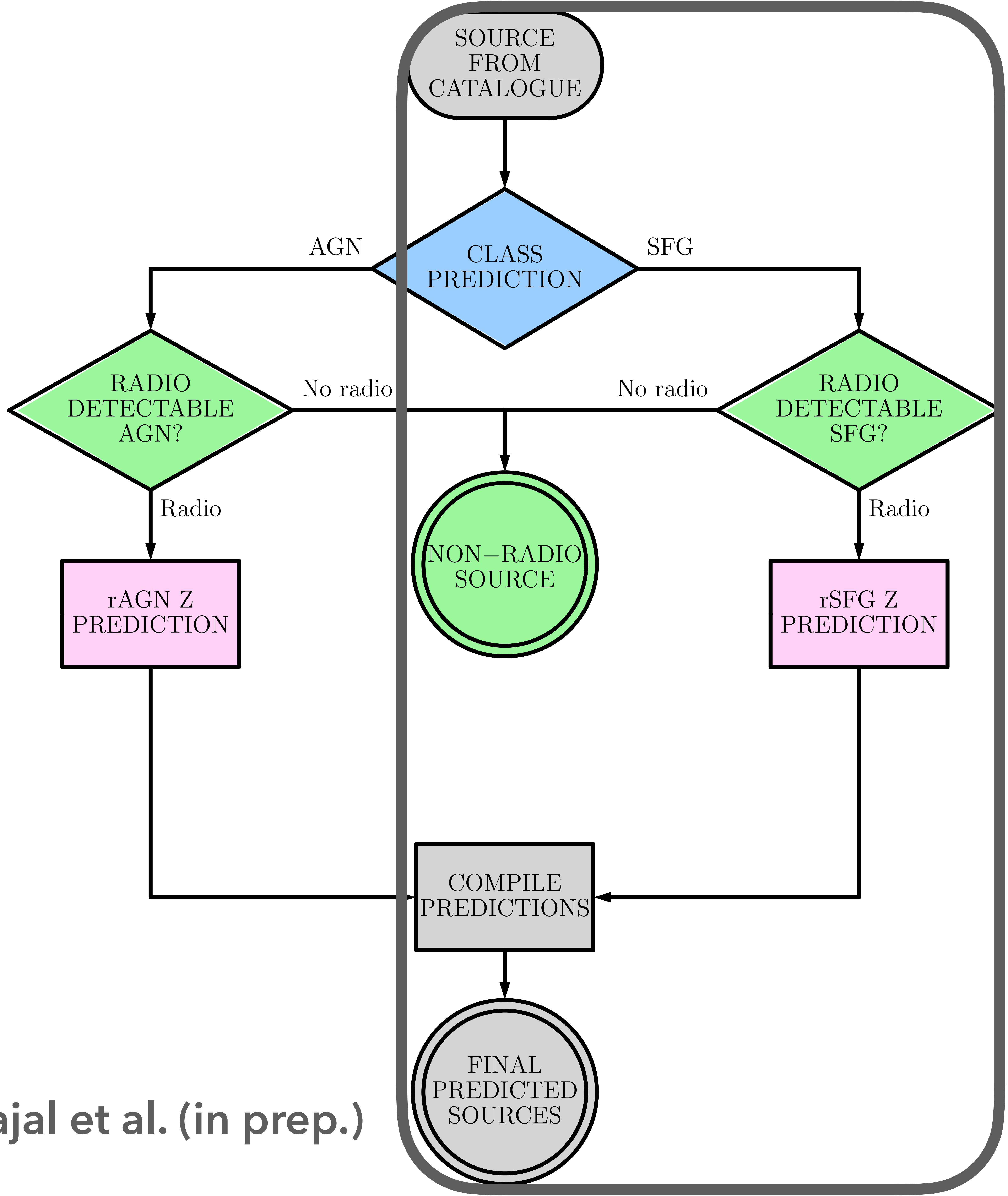
OUR APPROACH TO SELECTING AGN AND SFG WITH ML

Three different levels

Classify as AGN or SFG

Select radio-detectable sources

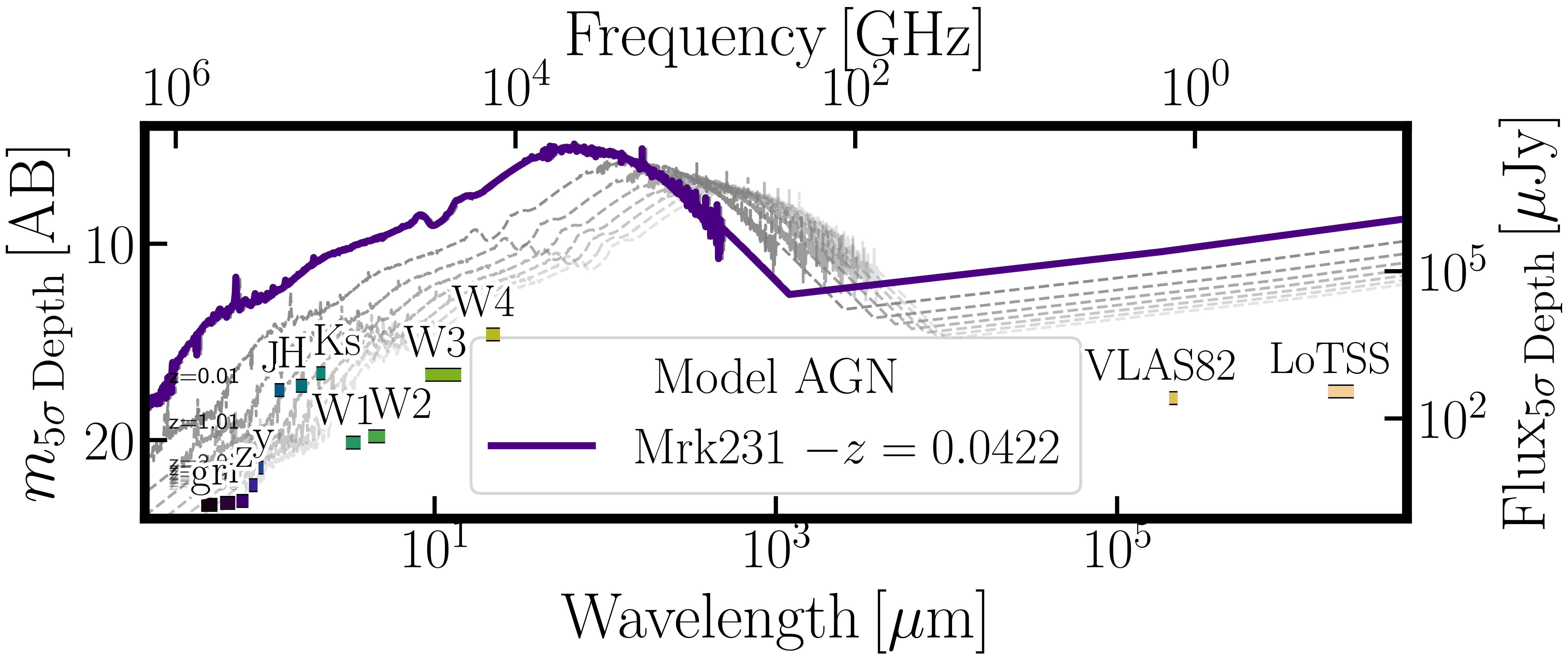
Estimate redshift for radio-detectable sources



Carvajal et al. (in prep.)

Our data

DATASET – PHOTOMETRY



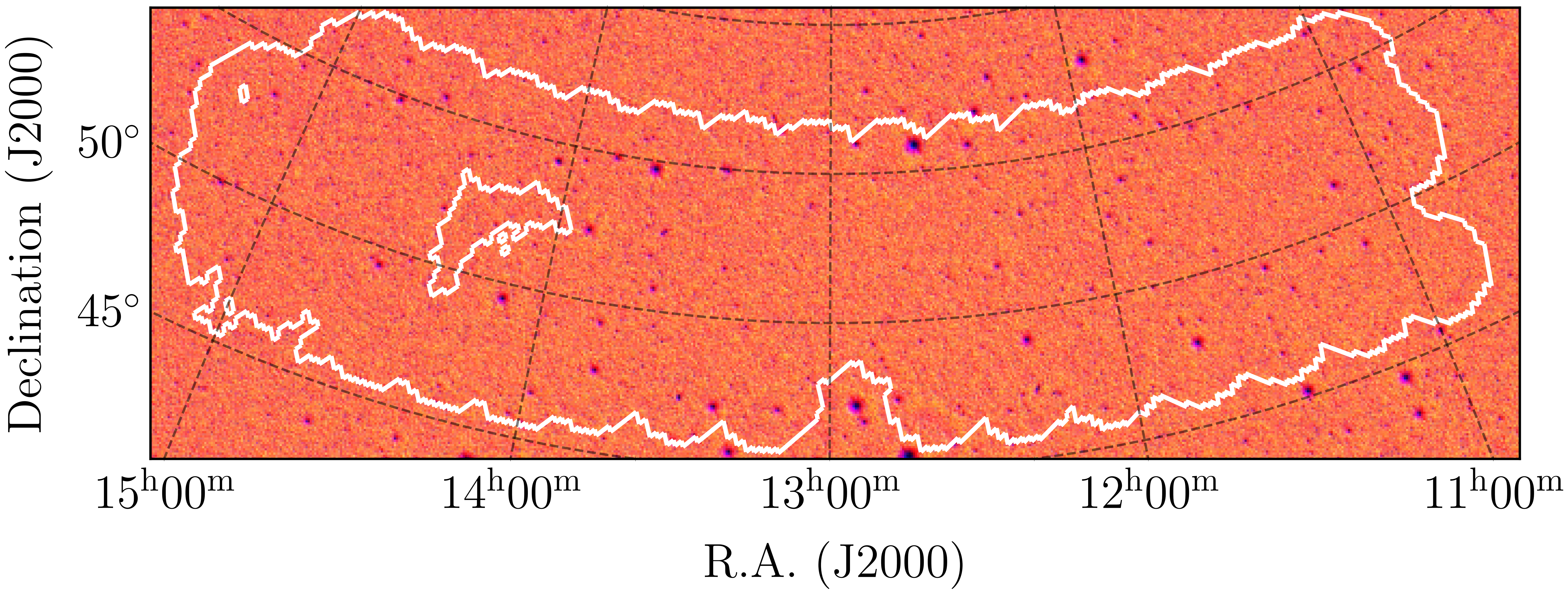
Carvajal et al. (2023)

Base catalogue: CatWISE2020 (Marocco et al. 2021, W1, W2)

Counterparts from: Pan-STARRS DR1, 2MASS, AllWISE

Colours from all bands (selected subset per model)

Target labels: class (AGN or SFG), radio detection, redshift



HETDEX Spring Field

424 deg² covered by LoTSS-DR1 @ 144 MHz, 71μJy, 6″ resolution

~15 million CatWISE2020 detections (~190k with LoTSS counterpart, 1%)

~50k spec. confirmed AGN (~6.4k radio, 13%) + ~70k spec. confirmed SFGs (~6.6k radio, 9%)

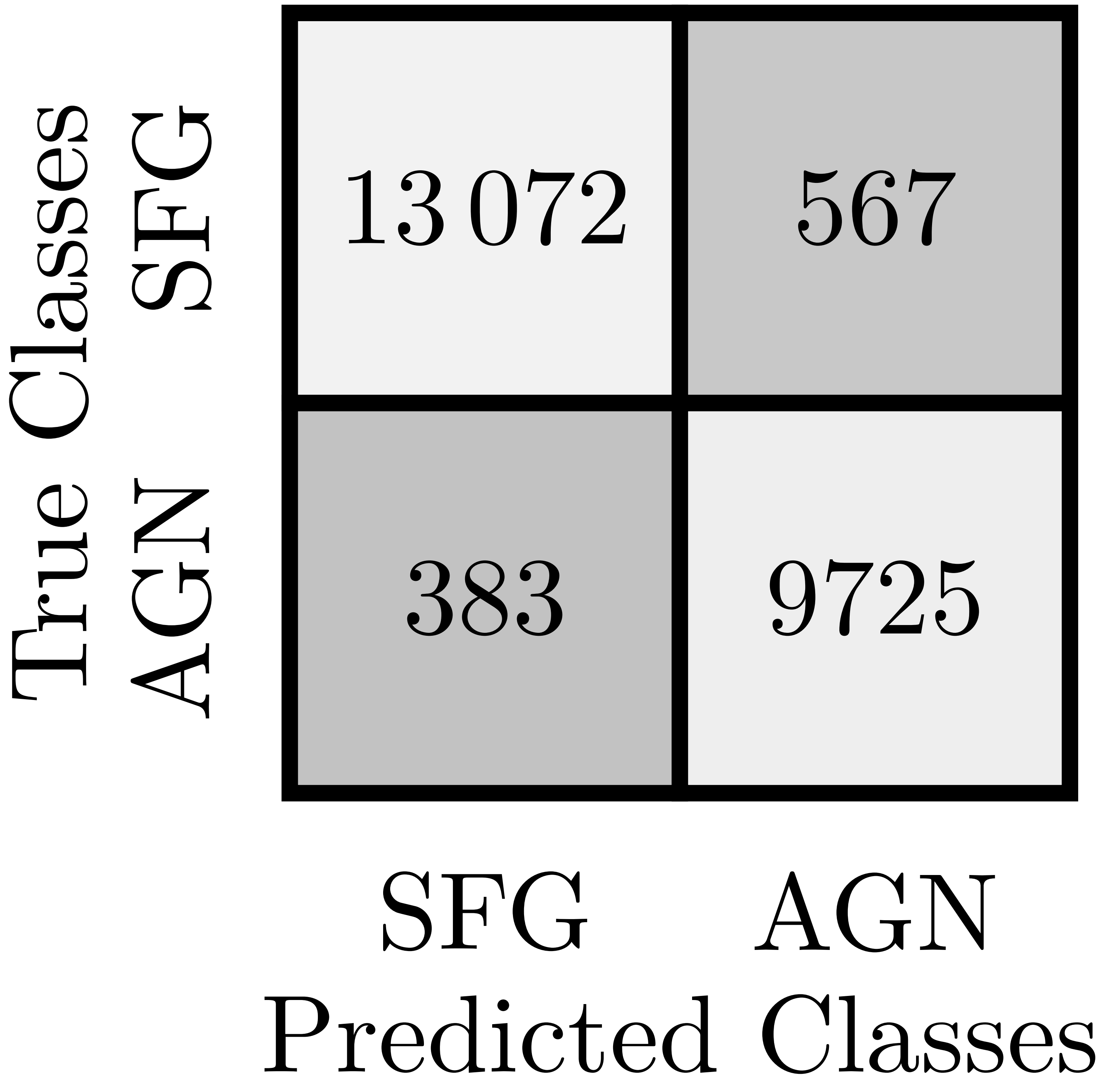
Models' Results

AGN Selection Completeness: 96 %

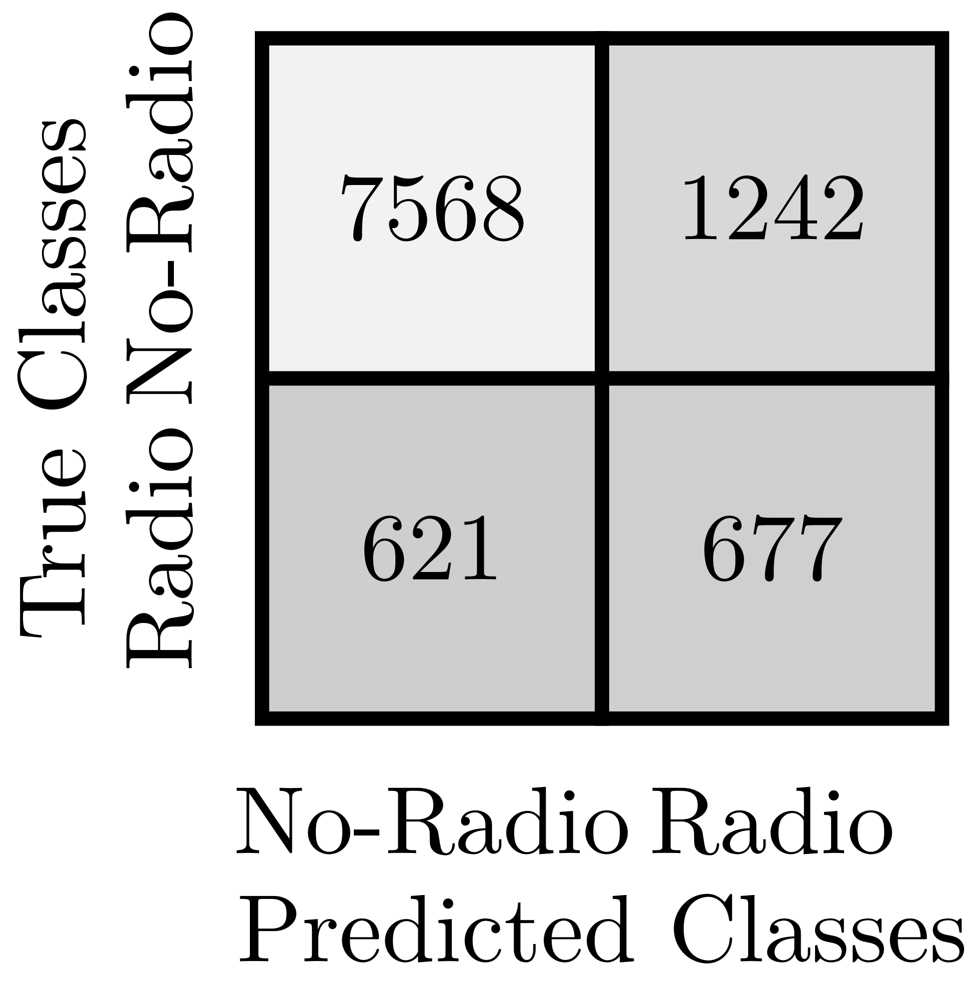
Increase from baseline AGN fraction of 43%

SFG Selection Completeness: 96 %

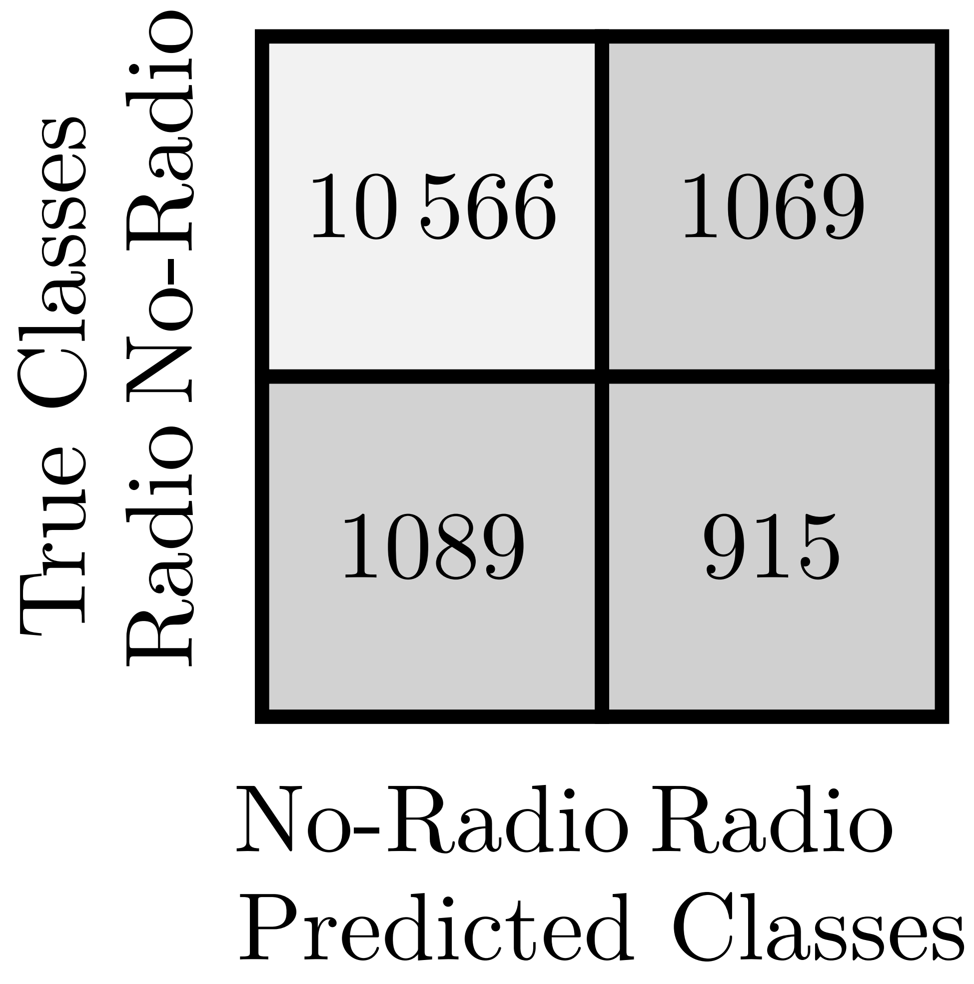
Increase from baseline SFG fraction of 57%



Radio Selection in AGN
Completeness: 52 %
Increase from baseline radio fraction in AGN of 13%

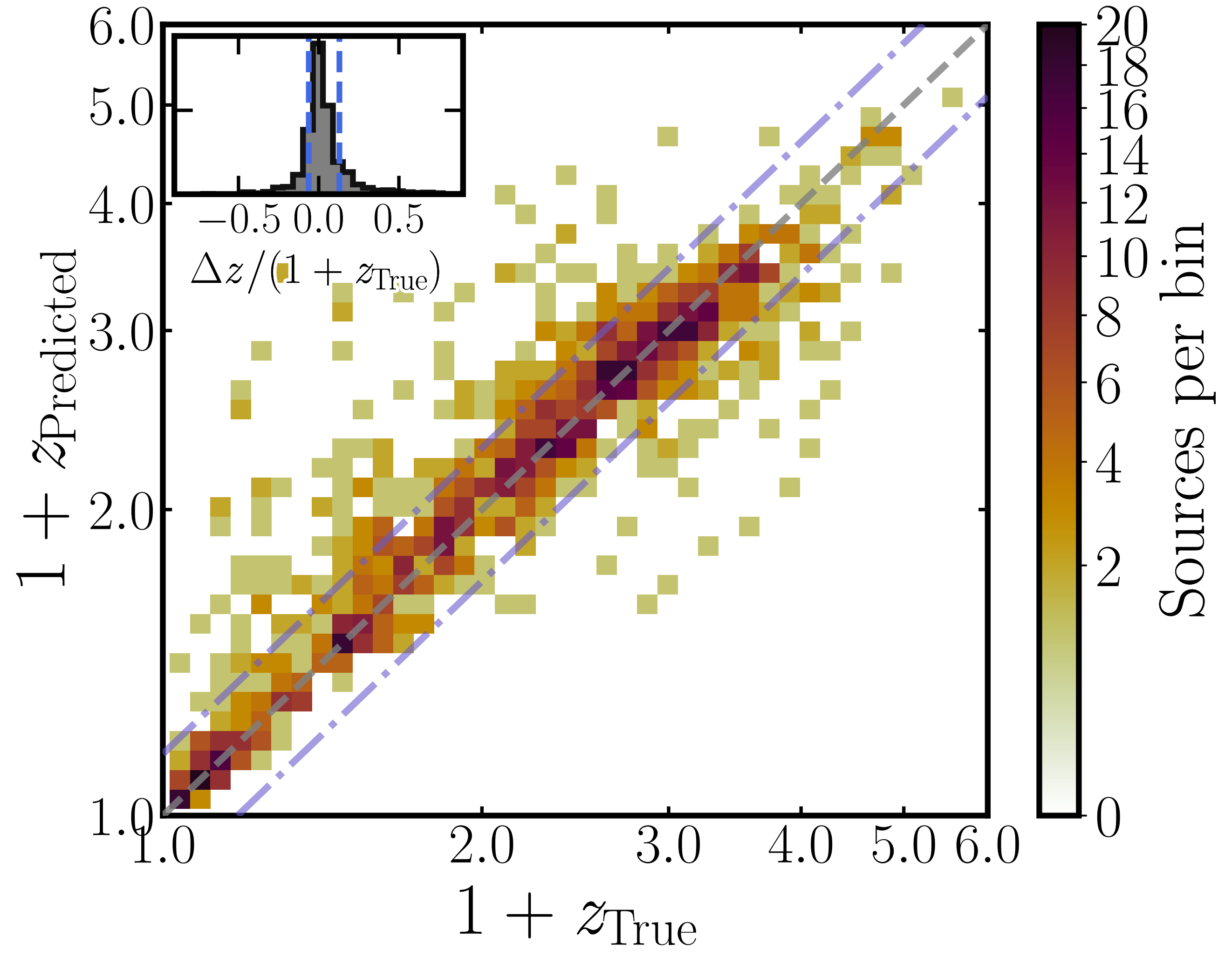


Radio Selection in SFG
Completeness: 46 %
Increase from baseline radio fraction in SFGs of 9%



REDSHIFT IN rAGN: 81 % ACCURATE

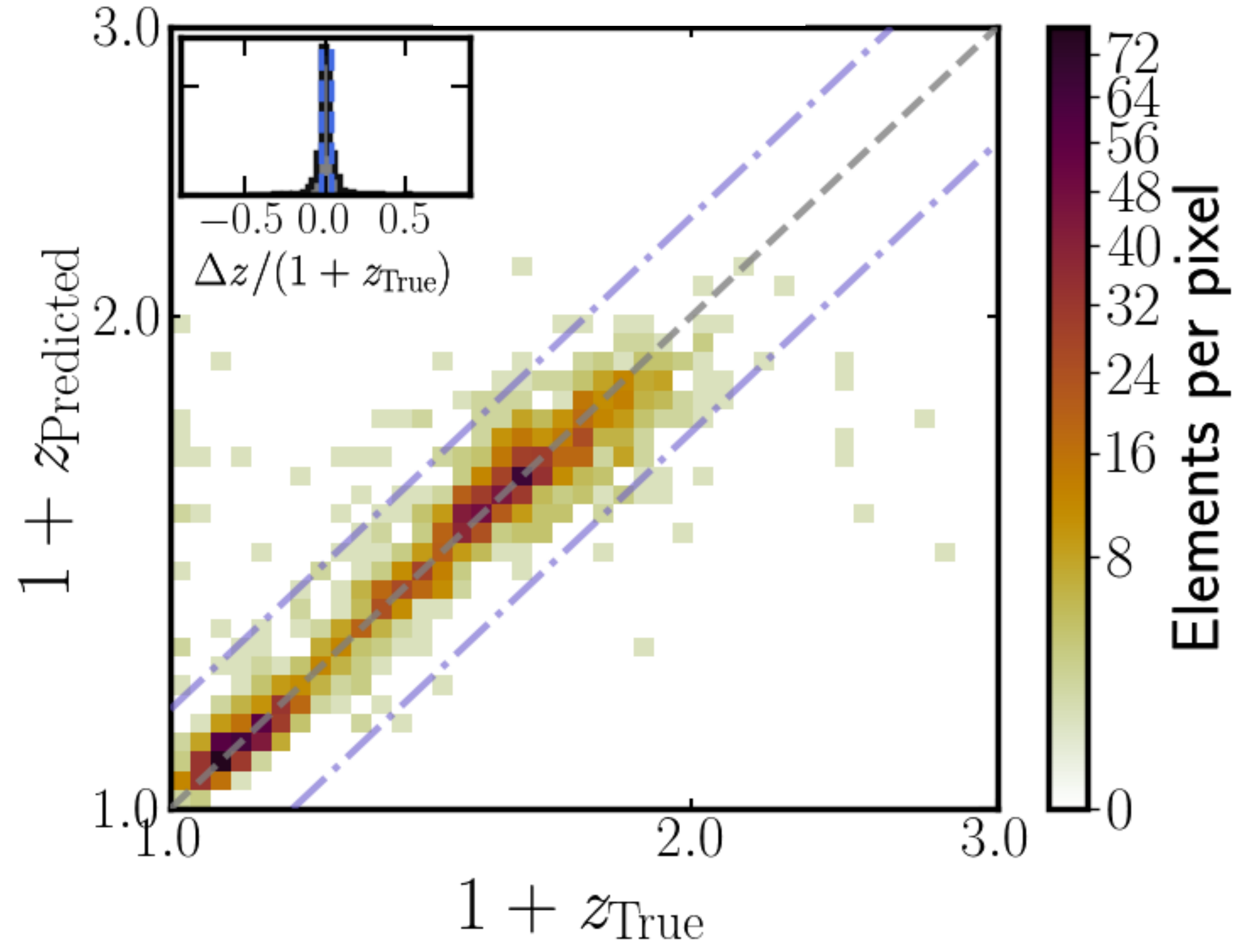
OUTLIER FRACTION OF 19%



Carvajal et al. (2023)

REDSHIFT IN rSFG: 97 % ACCURATE

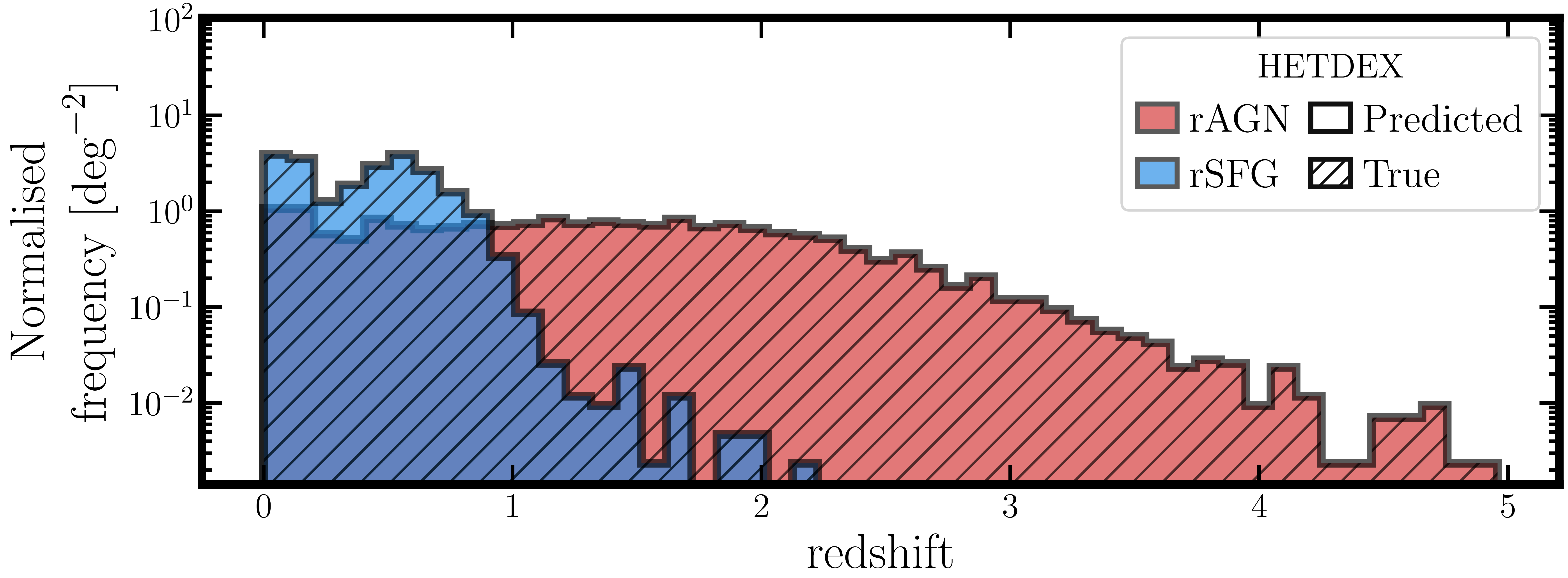
OUTLIER FRACTION OF 3%



Applying pipeline to full HETDEX dataset

New sample of 68k rAGN candidates 115k rSFG candidates

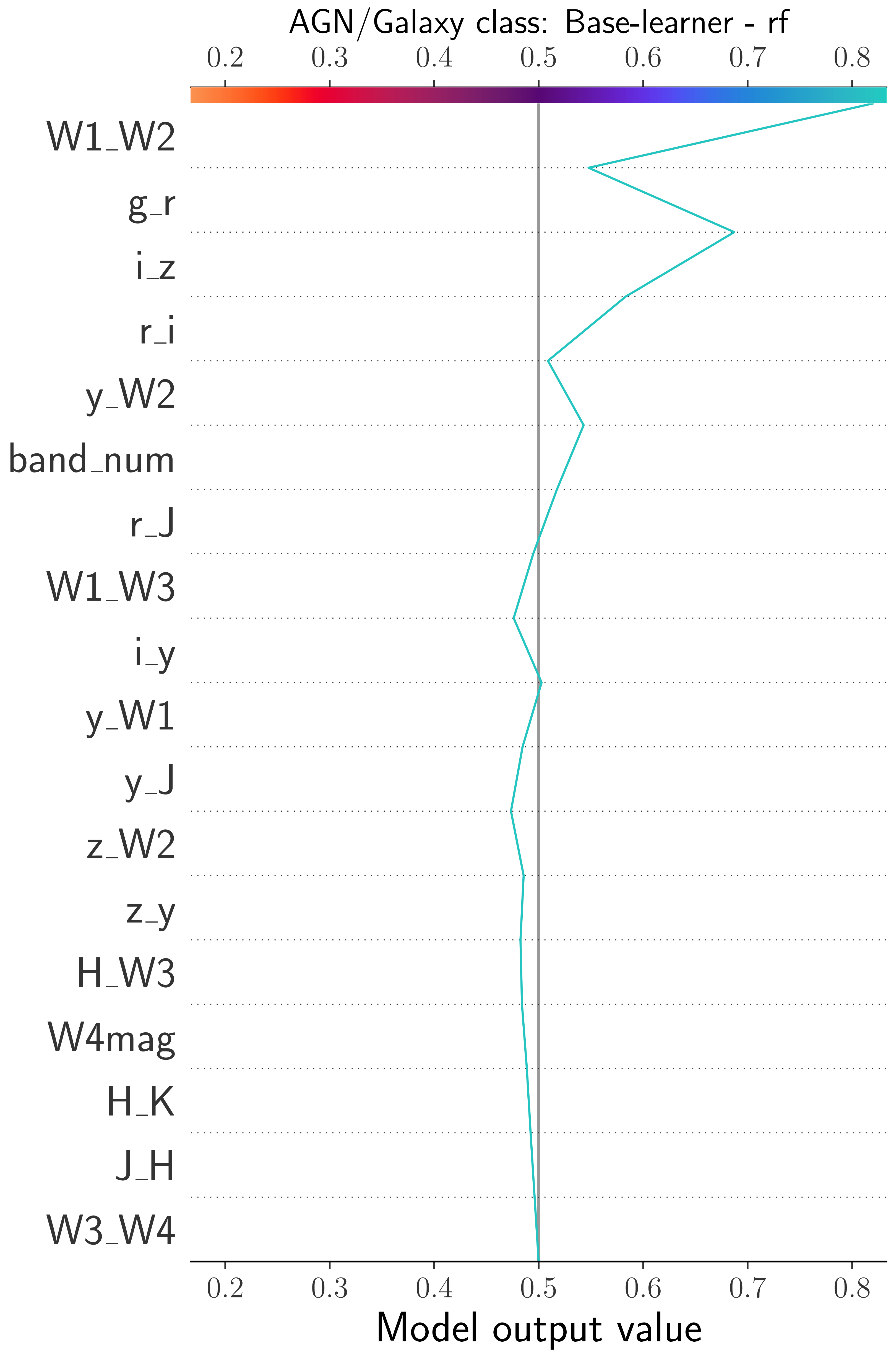
More than 10x original sample (6.4k rAGN and 6.6k rSFG)



UNDERSTANDING OUR MODELS AND PREDICTIONS

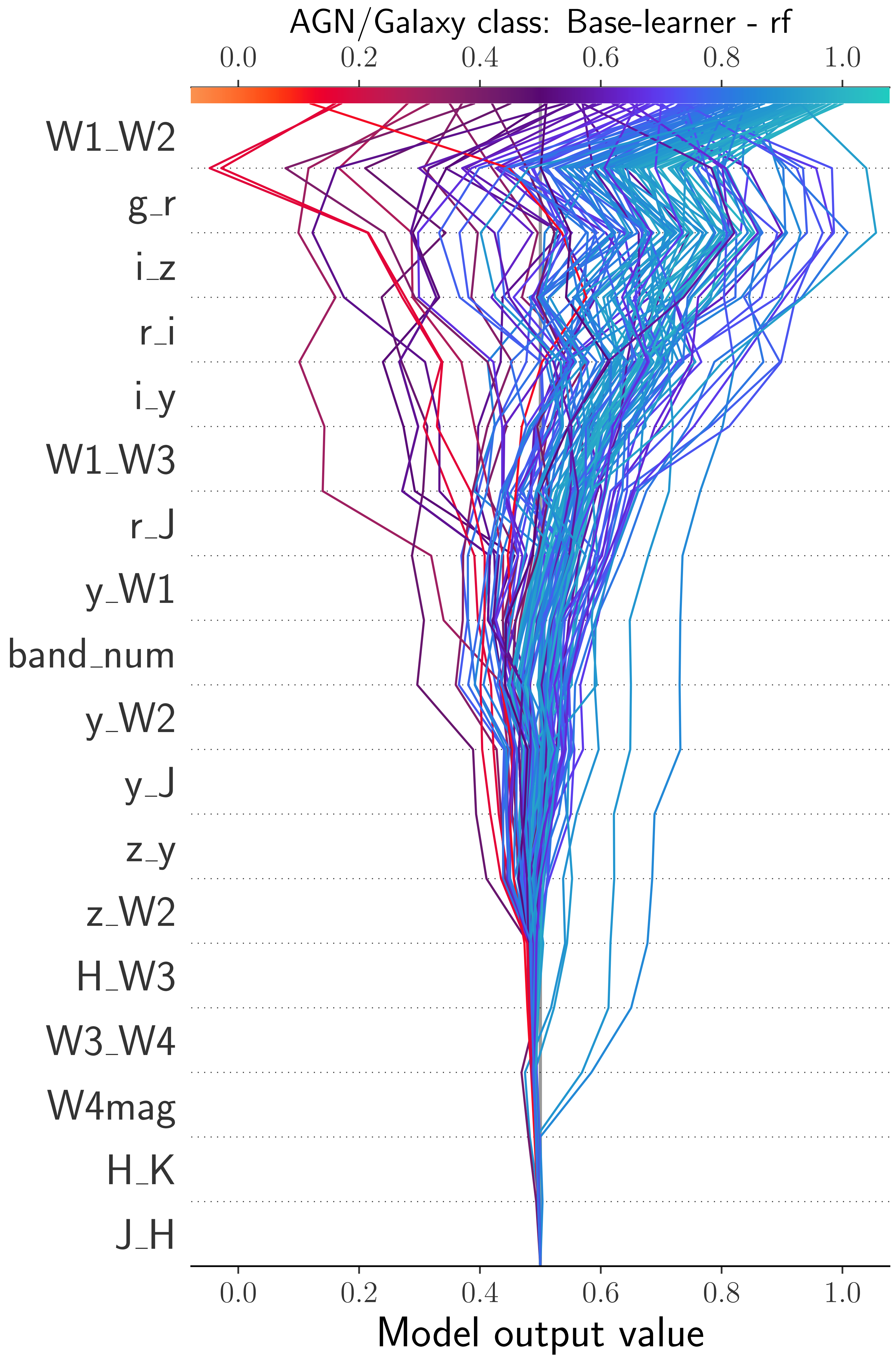
Feature Importance

Understand which features drive predictions more strongly



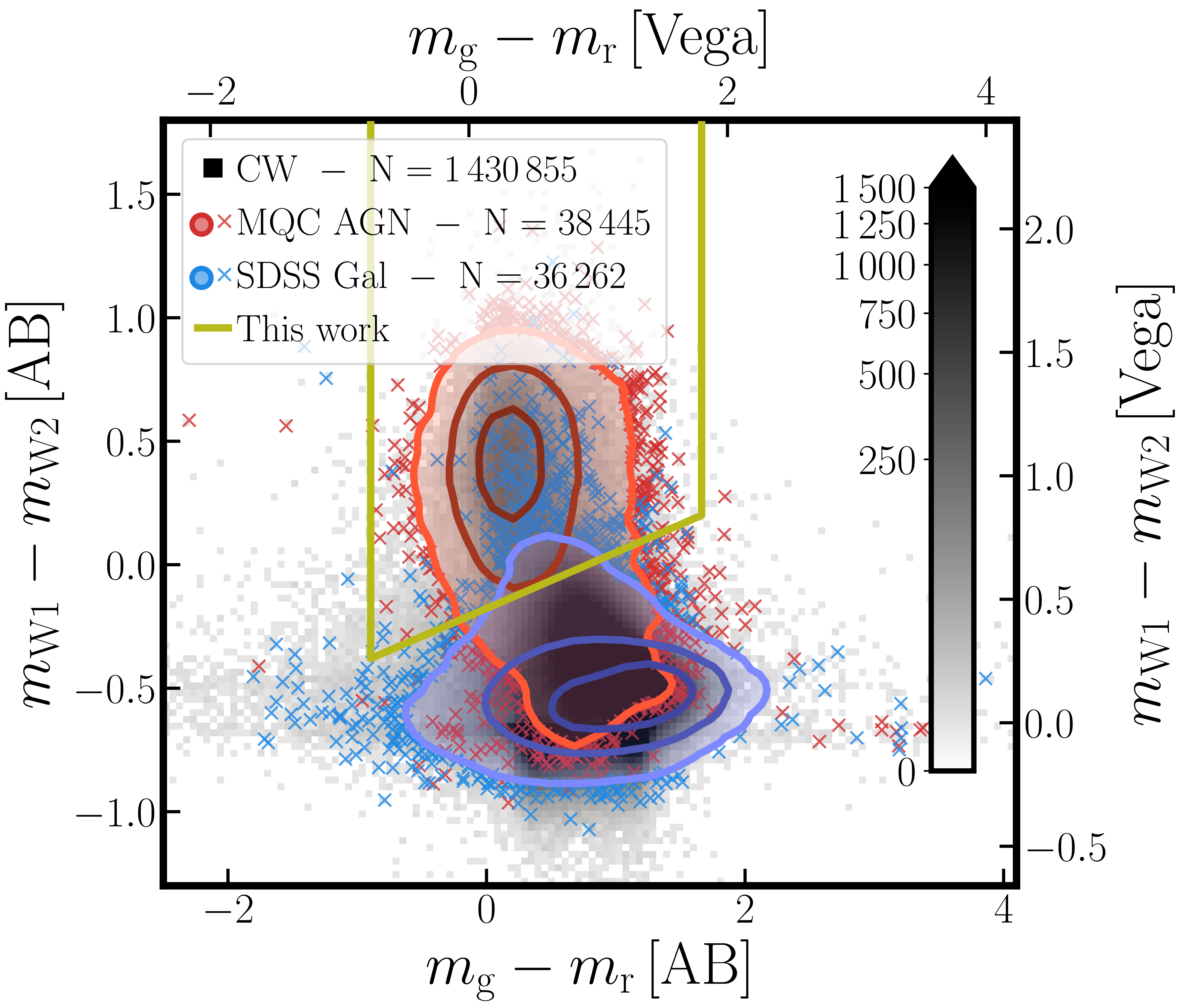
Feature Importance

Understand which features drive predictions more strongly



ML-BASED COLOUR-COLOUR DIAGRAM

g-r – W1-W2 DIAGRAM



Carvajal et al. (2023)

Applying pipeline elsewhere?

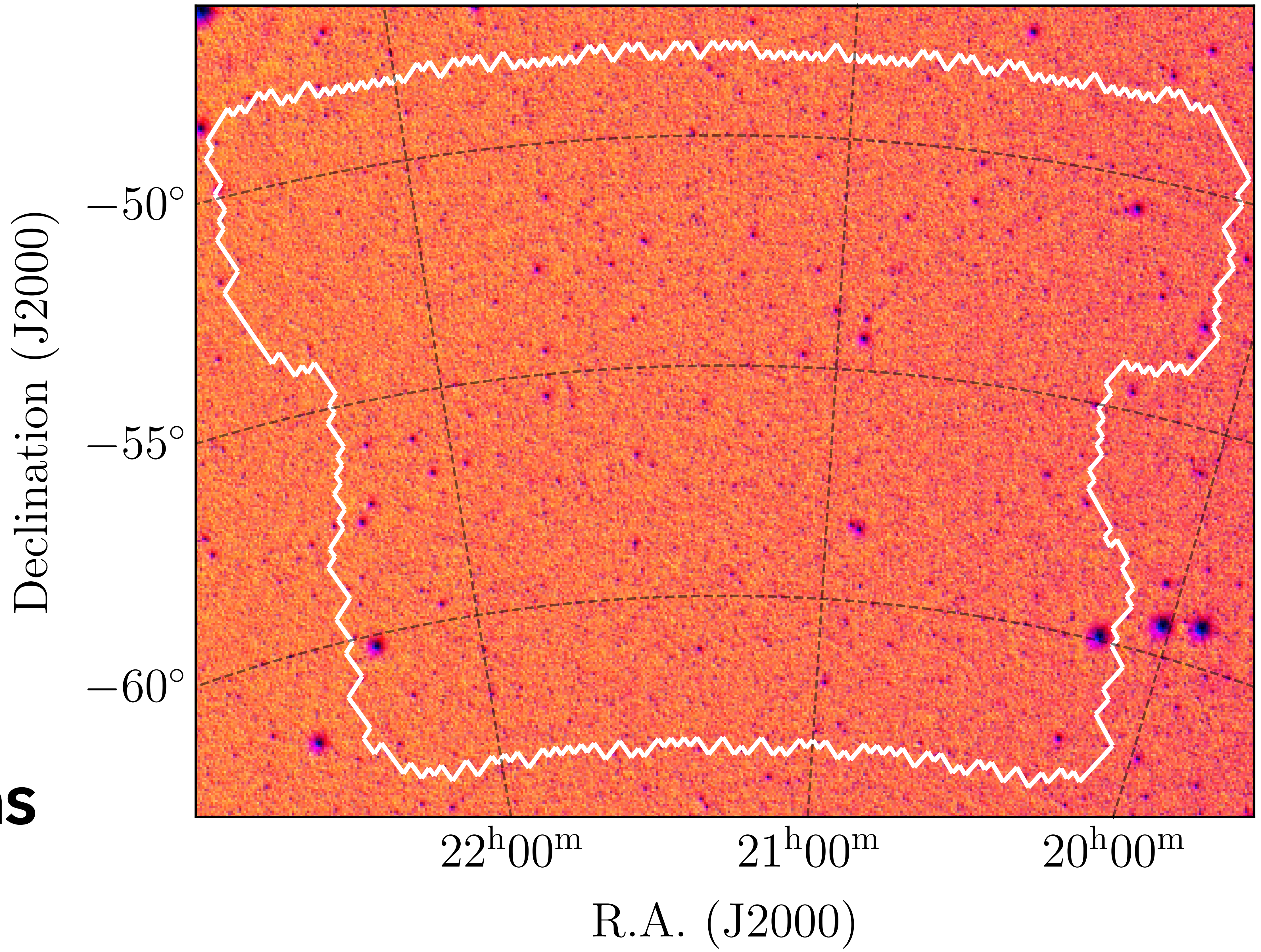
APPLY PIPELINE IN THE SOUTHERN SKY

EMU Pilot Survey (EMU-PS) – 270 deg² –

18 arcsec resolution @ 944 MHz – 25-30 μJy rms

SKA-like conditions

~10M CatWISES2020 sources



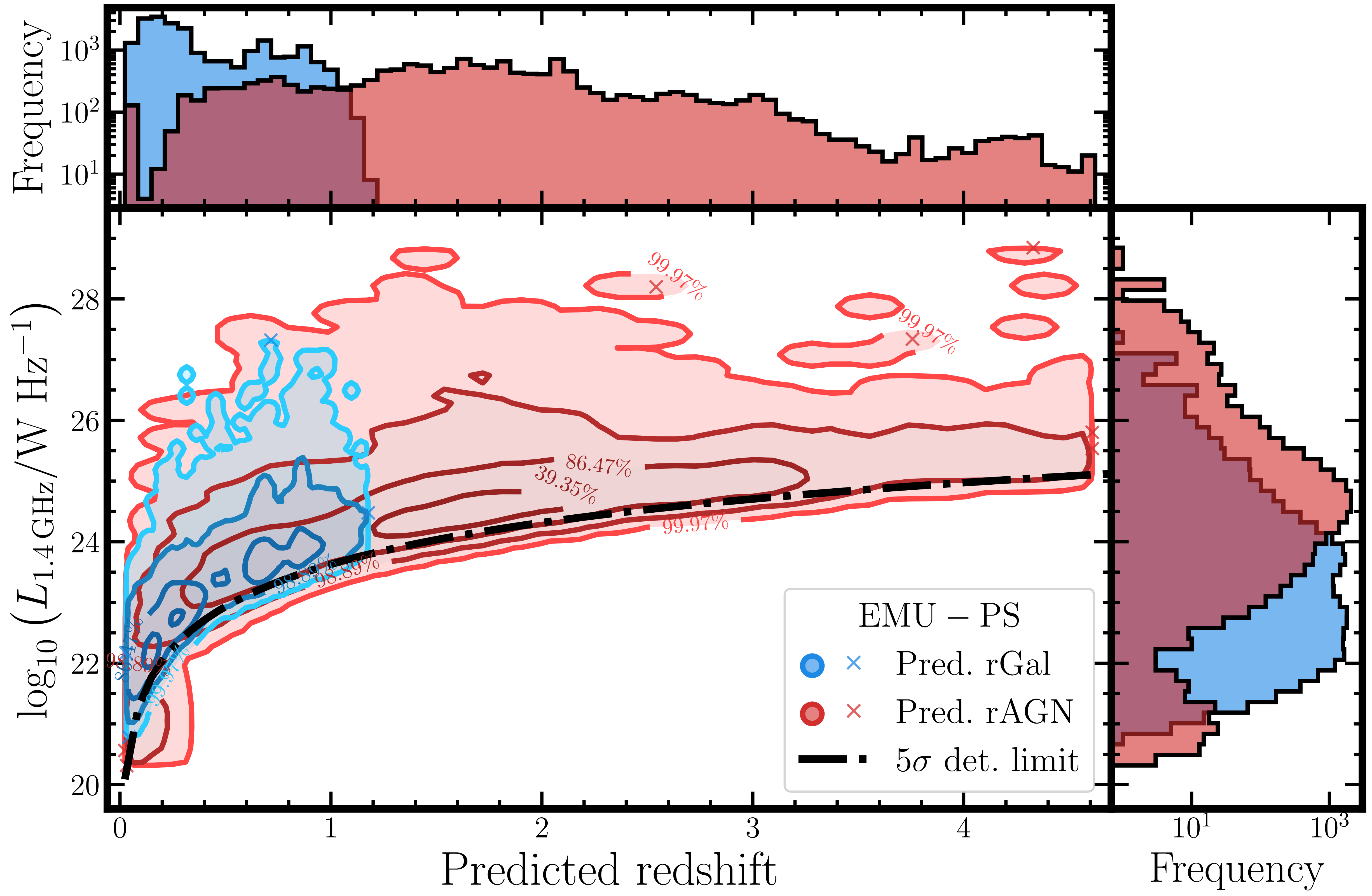
EMU-PS

More than 10x new rAGN and rSFG candidates

rAGN: 92 113 (originally 2367)

rSFG: 128 249 (originally 870)

Such numbers allow for population studies



Radio Luminosity Function

Radio Luminosity Function

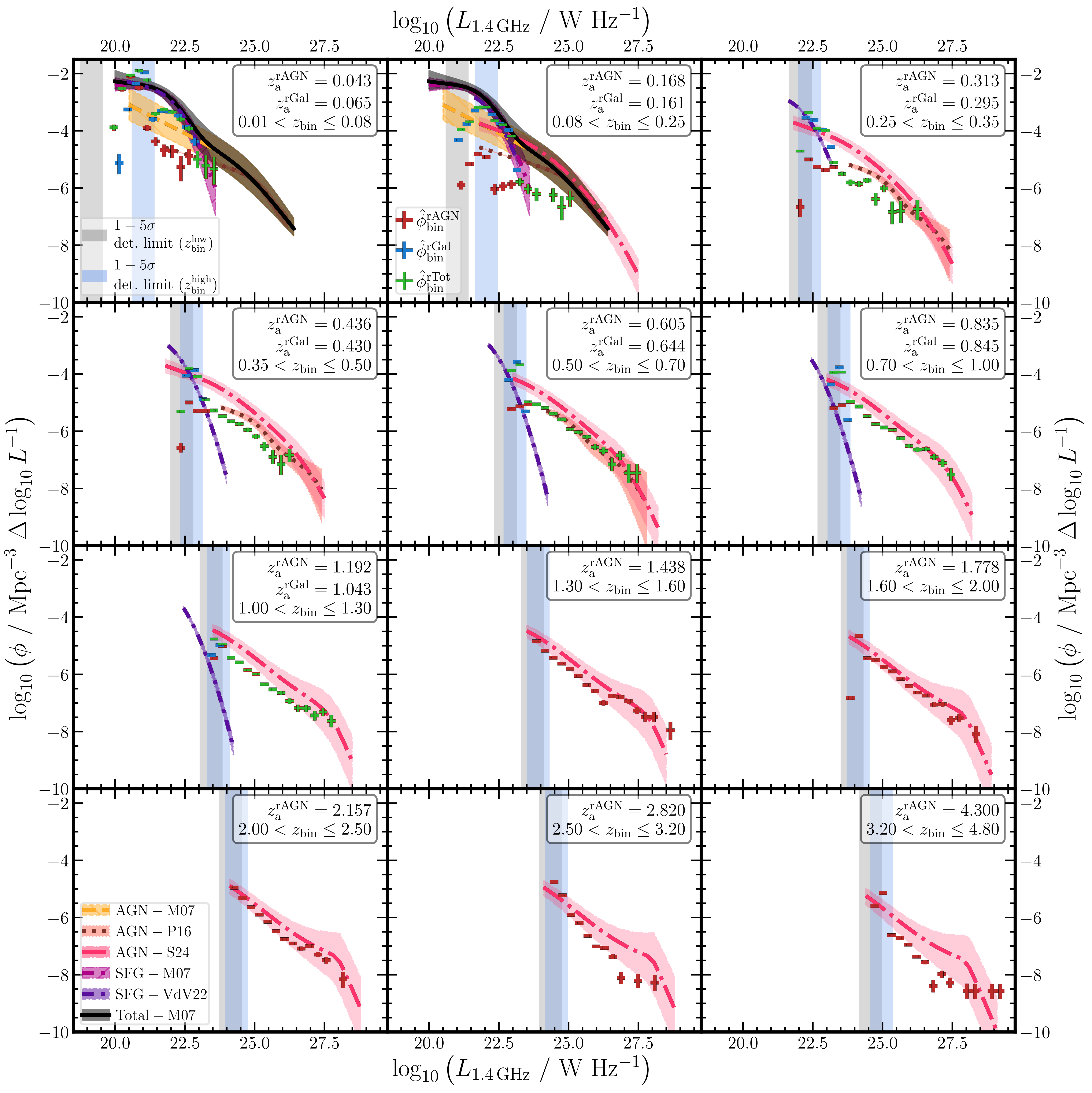
Number density candidate SFGs and AGN from pipeline in EMU-PS

Binned luminosity function ($1/V_{max}$ approach, Page & Carrera, 2000)

Calculated over twelve redshift bins [0.1, 4.8]

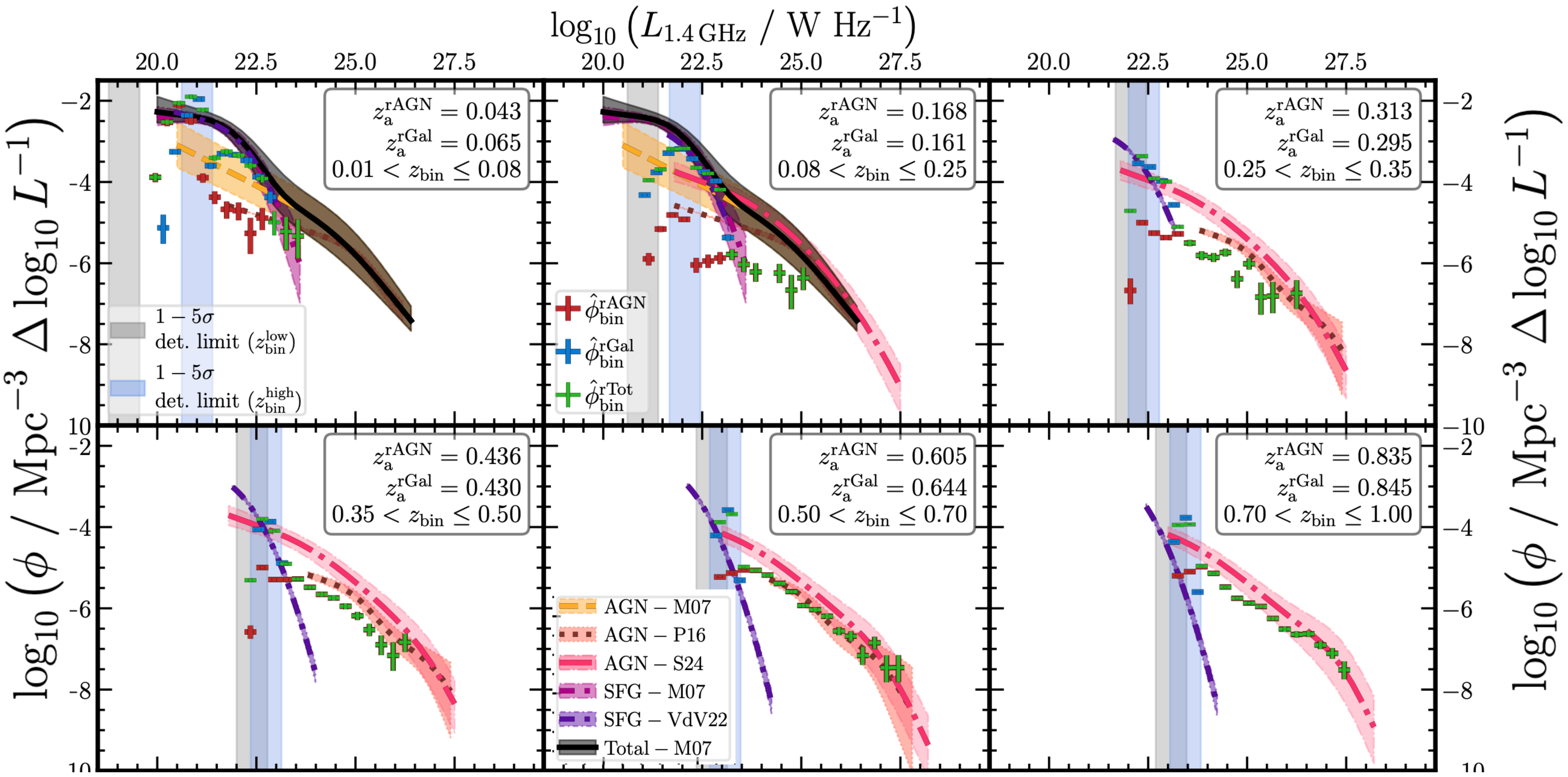
Correct individual counts by completeness and purity from pipeline

Compared with previous parametric functions

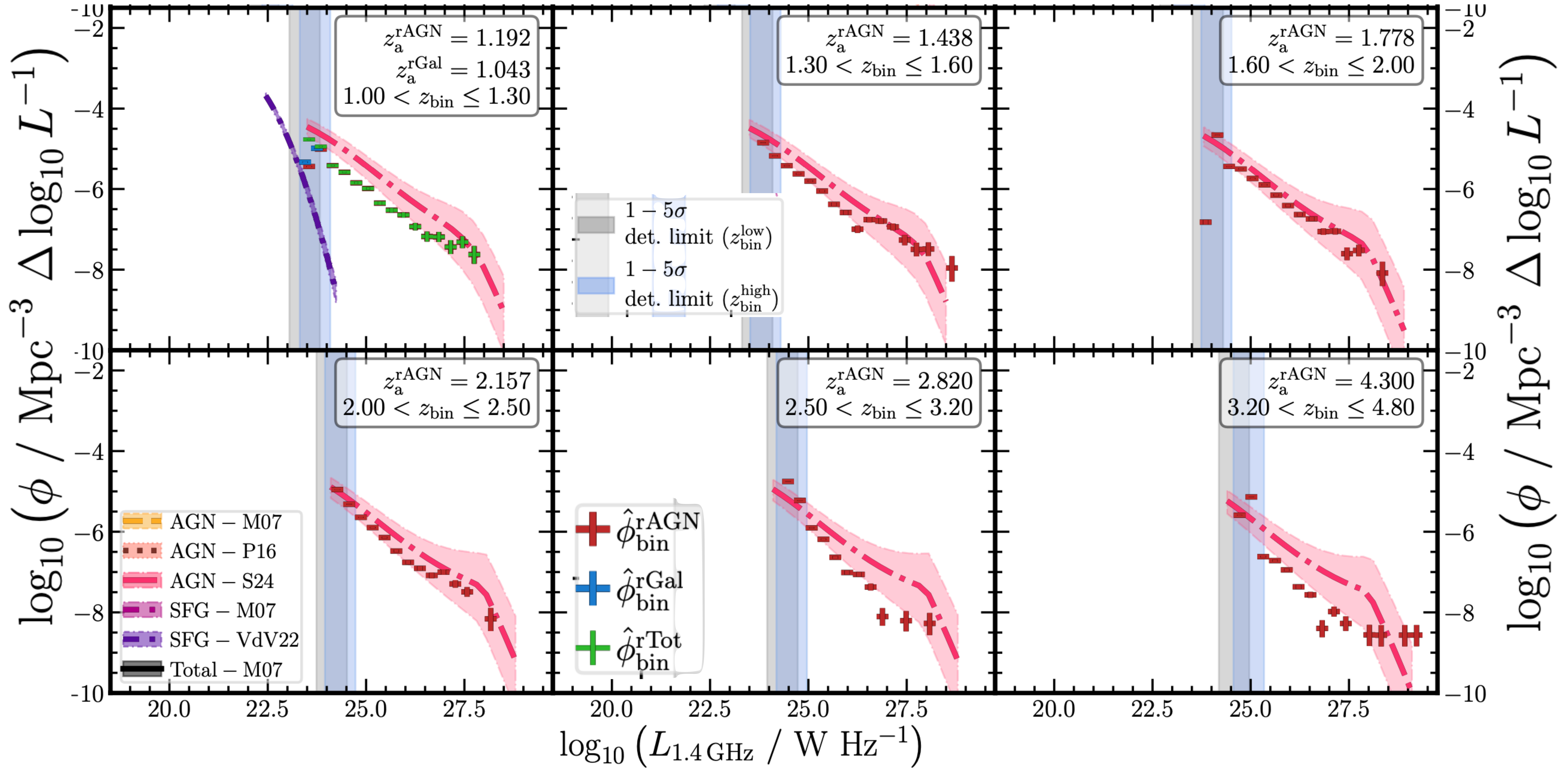


Close to previous work only from ML corrections

Small error bars from large sample (at all z)



$\log_{10}\left(L_{1.4\,\mathrm{GHz}}\ /\ \mathrm{W\ Hz^{-1}}\right)$
20.0
22.5
25.0
27.5
$\log_{10}\left(\phi\ /\ \mathrm{Mpc^{-3}}\ \Delta\log_{10}L^{-1}\right)$
$z_a^{\rm rAGN} = 0.043$
$z_a^{\rm rGal} = 0.065$
$0.01 < z_{\rm bin} \leq 0.08$
$z_a^{\rm rAGN} = 0.168$
$z_a^{\rm rGal} = 0.161$
$0.08 < z_{\rm bin} \leq 0.25$
$z_a^{\rm rAGN} = 0.313$
$z_a^{\rm rGal} = 0.295$
$0.25 < z_{\rm bin} \leq 0.35$
$z_a^{\rm rAGN} = 0.436$
$z_a^{\rm rGal} = 0.430$
$0.35 < z_{\rm bin} \leq 0.50$
$z_a^{\rm rAGN} = 0.605$
$z_a^{\rm rGal} = 0.644$
$0.50 < z_{\rm bin} \leq 0.70$
$z_a^{\rm rAGN} = 0.835$
$z_a^{\rm rGal} = 0.845$
$0.70 < z_{\rm bin} \leq 1.00$
$1-5\sigma$ det. limit ($z_{\rm bin}^{\rm low}$)
$1-5\sigma$ det. limit ($z_{\rm bin}^{\rm high}$)
$\hat{\phi}_{\rm bin}^{\rm rAGN}$
$\hat{\phi}_{\rm bin}^{\rm rGal}$
$\hat{\phi}_{\rm bin}^{\rm rTot}$
AGN – M07
AGN – P16
AGN – S24
SFG – M07
SFG – VdV22
Total – M07



$z_a^{rAGN} = 1.192$
$z_a^{rGal} = 1.043$
$1.00 < z_{bin} \leq 1.30$
$z_a^{rAGN} = 1.438$
$1.30 < z_{bin} \leq 1.60$
$z_a^{rAGN} = 1.778$
$1.60 < z_{bin} \leq 2.00$
$z_a^{rAGN} = 2.157$
$2.00 < z_{bin} \leq 2.50$
$z_a^{rAGN} = 2.820$
$2.50 < z_{bin} \leq 3.20$
$z_a^{rAGN} = 4.300$
$3.20 < z_{bin} \leq 4.80$
$1 - 5\sigma$ det. limit (z_{bin}^{low})
$1 - 5\sigma$ det. limit (z_{bin}^{high})
AGN – M07
AGN – P16
AGN – S24
SFG – M07
SFG – VdV22
Total – M07
$\hat{\phi}_{bin}^{rAGN}$
$\hat{\phi}_{bin}^{rGal}$
$\hat{\phi}_{bin}^{rTot}$
$\log_{10}\left(\phi \;/\; \mathrm{Mpc}^{-3}\;\Delta\log_{10} L^{-1}\right)$
$\log_{10}\left(L_{1.4\,\mathrm{GHz}} \;/\; \mathrm{W\ Hz}^{-1}\right)$

In conclusion

CONCLUDING REMARKS

Want to determine contribution of AGN into galaxy evolution

Clear characterisation of sources is needed: class + redshift

Need to exploit large radio surveys for AGN and SFG studies

Machine-assisted pipeline to select rAGN + rSFGs candidates and redshift

Understand inner works of algorithms to extract physical insight

Increase sample sizes improving statistics

WHERE TO GO FROM HERE?

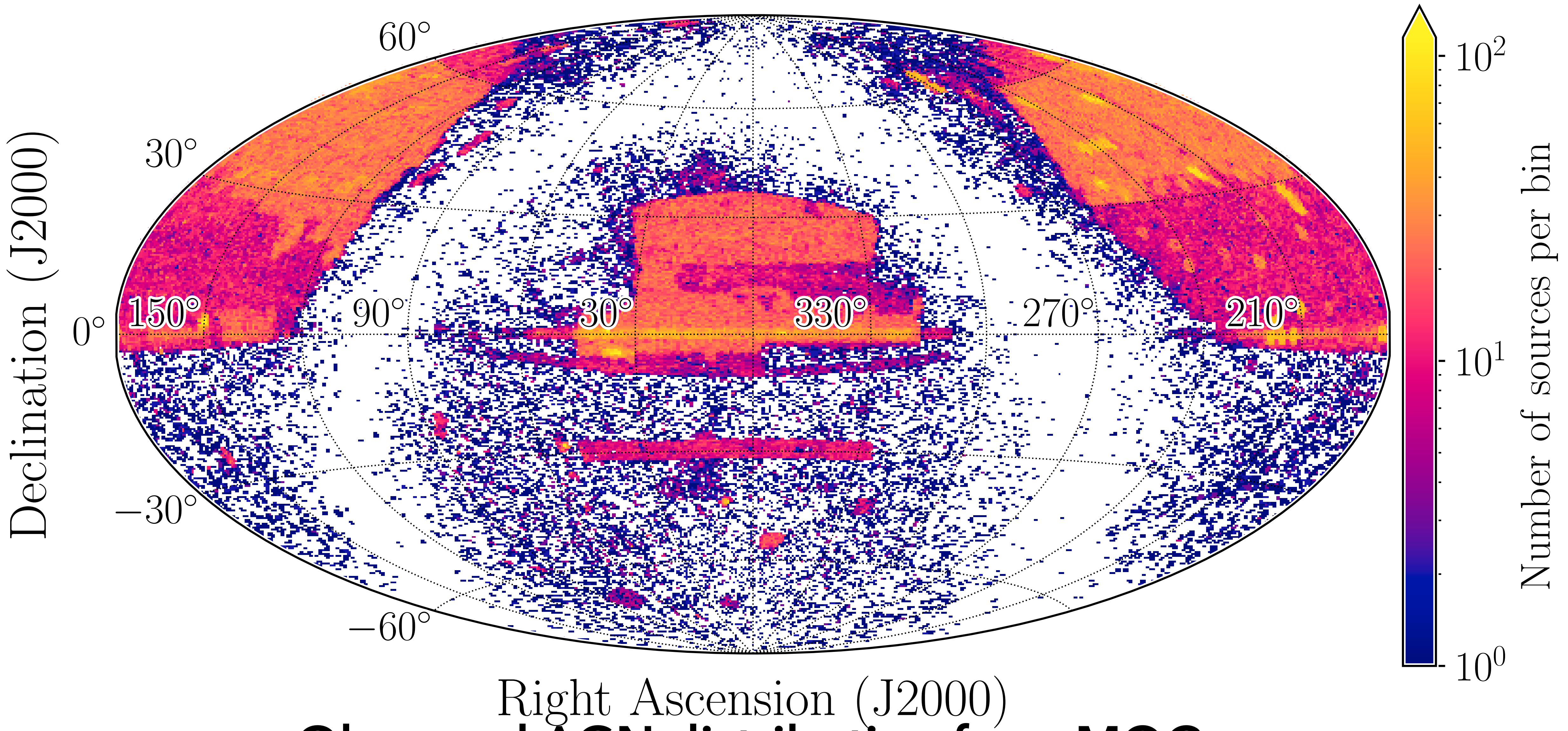
Need to go to the southern hemisphere!

AGN distribution strongly skewed to the north

SKA (and precursors: EMU, MIGTHEE) observations

Deeper surveys: VHS, KIDS-S, DES

Code/pipeline improvements



Observed AGN distribution from MQC

Ciências ULisboa

instituto de astrofísica e ciências do espaço

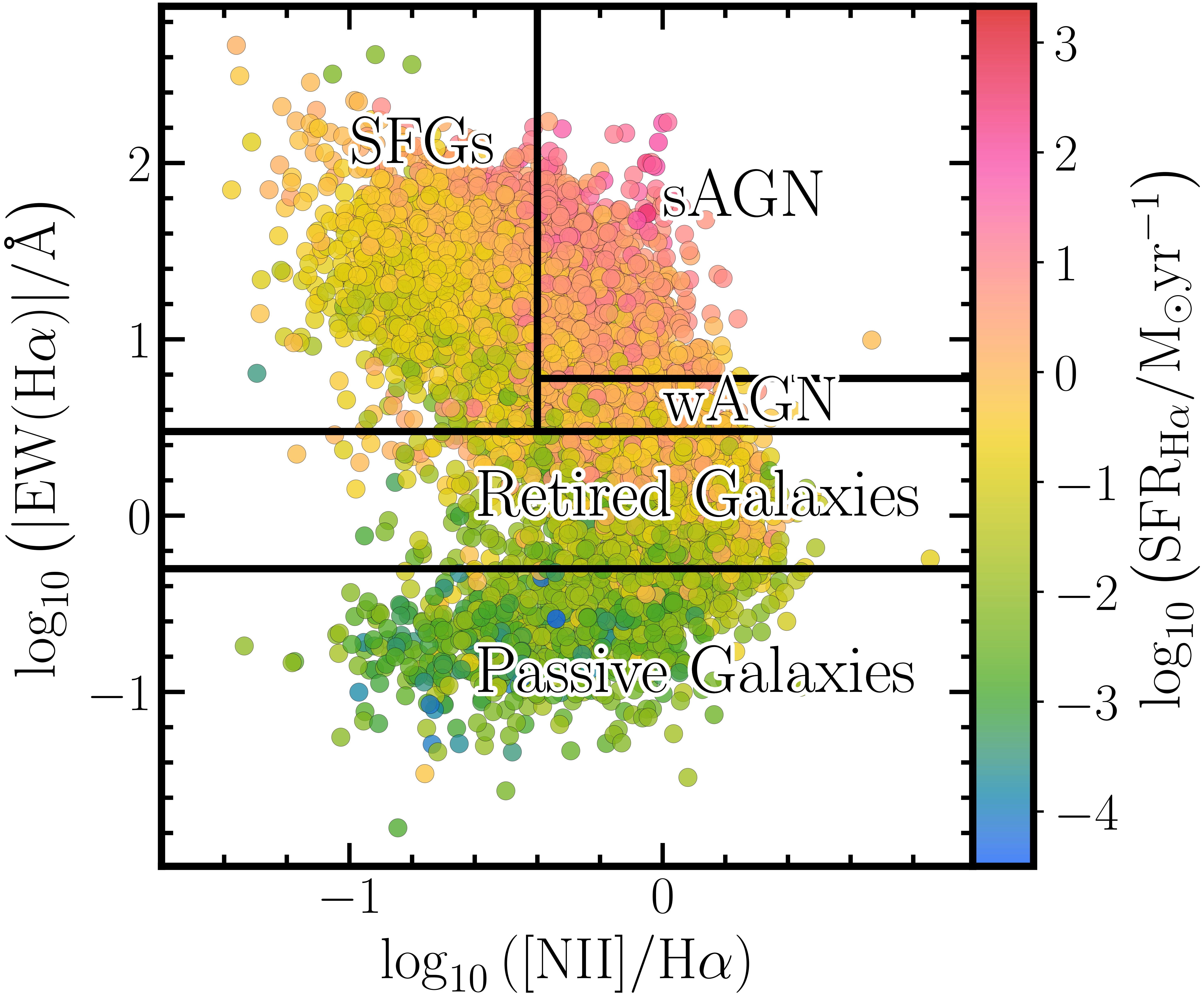
Towards Better Selection and Characterisation Criteria for High-Redshift Radio Galaxies Using Machine-Assisted Pattern Recognition

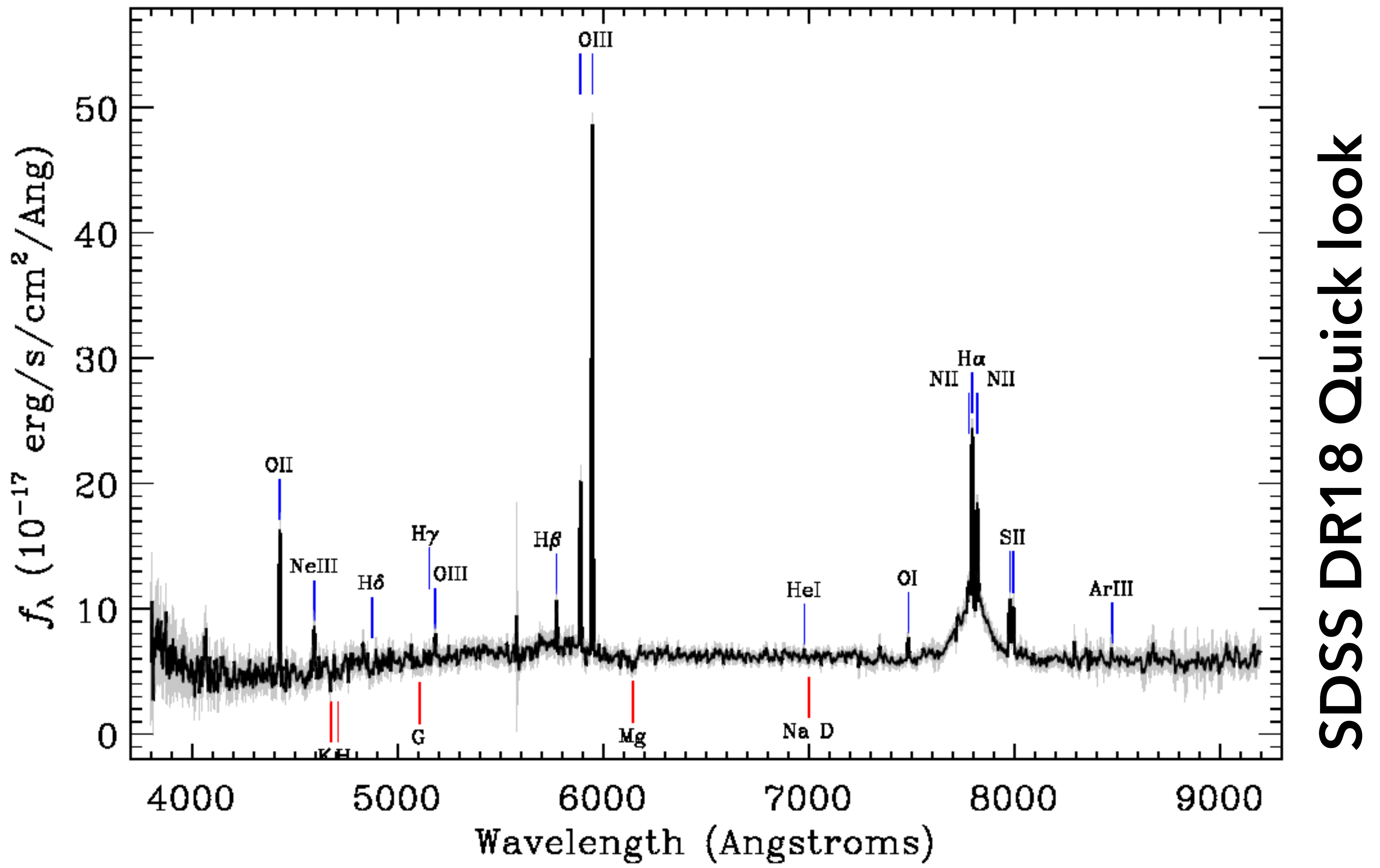
RODRIGO CARVAJAL

SUPERVISED BY
DR J. AFONSO
DR I. MATUTE
DR H. MESSIAS

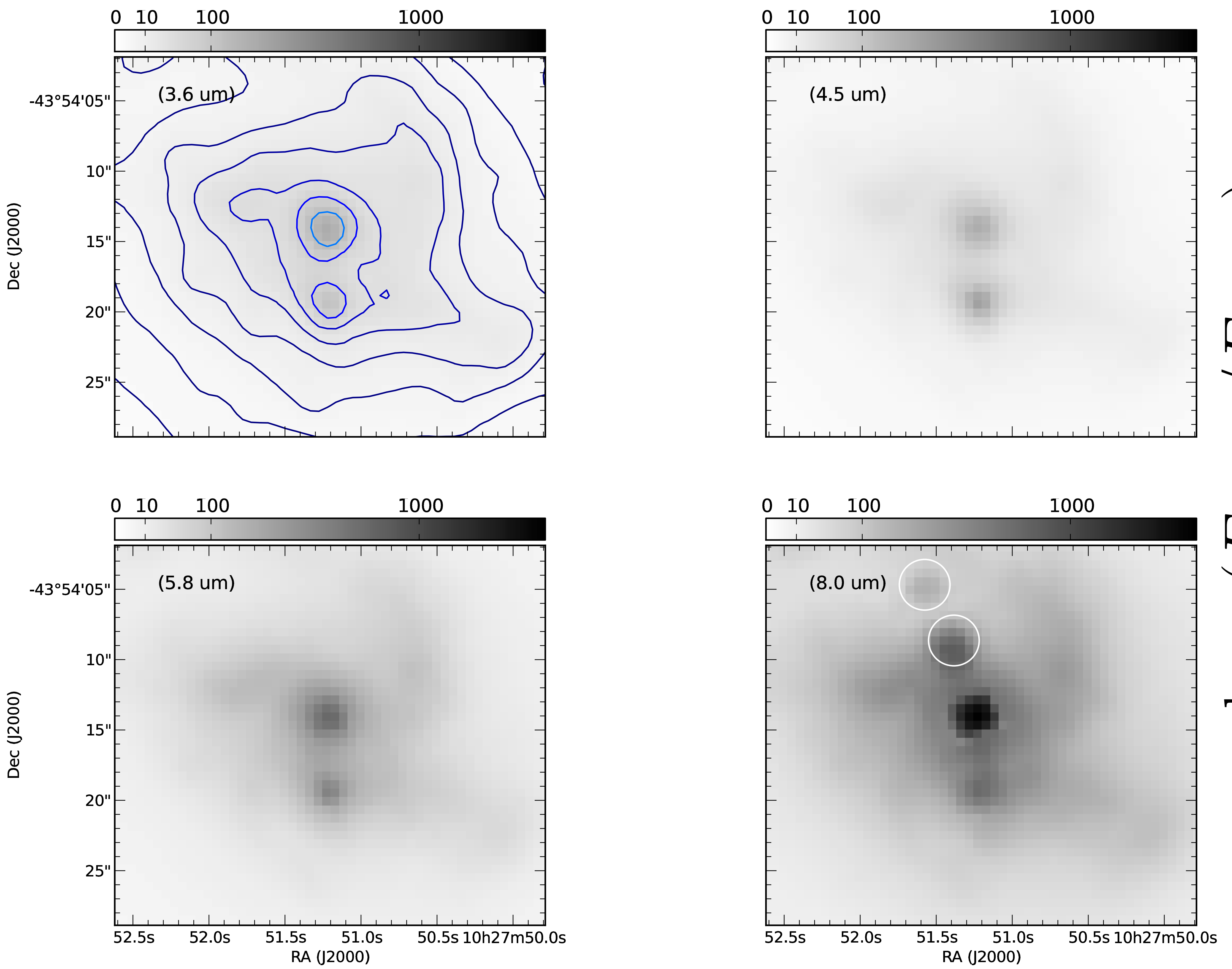
PD F FCT PhD PROGRAMMES
COFINANCIAMENTO / COFINANCING
fct Fundação para a Ciência e a Tecnologia
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

• • •

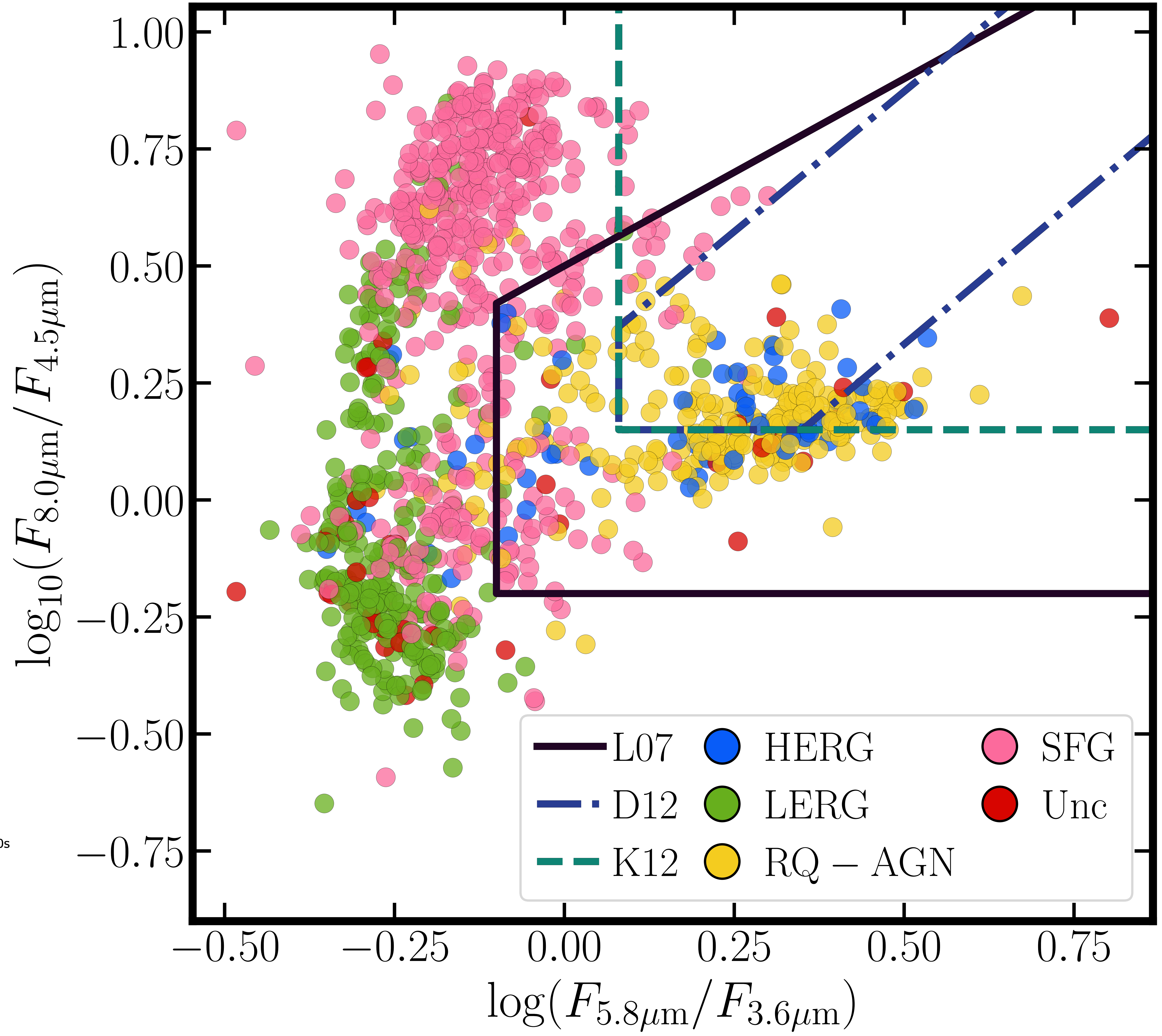




WHAN DIAGRAM

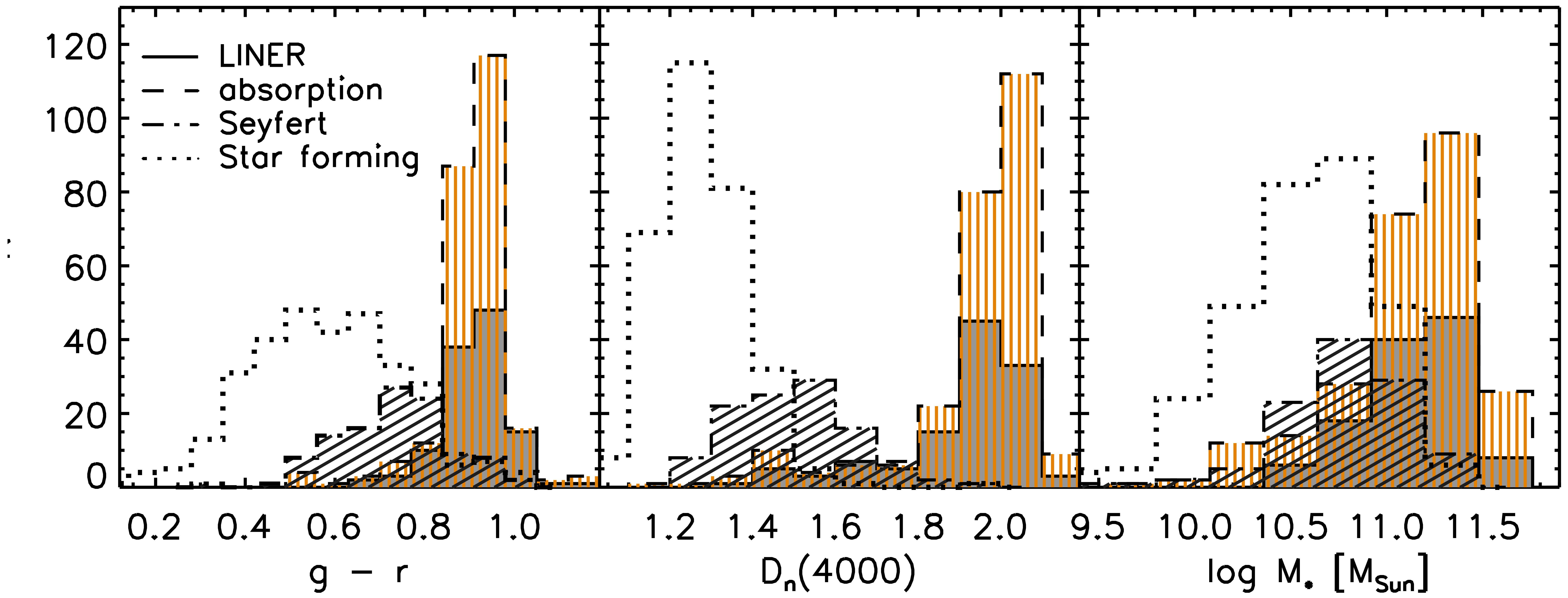


Ohyama et al. (2015)



Spitzer Colours

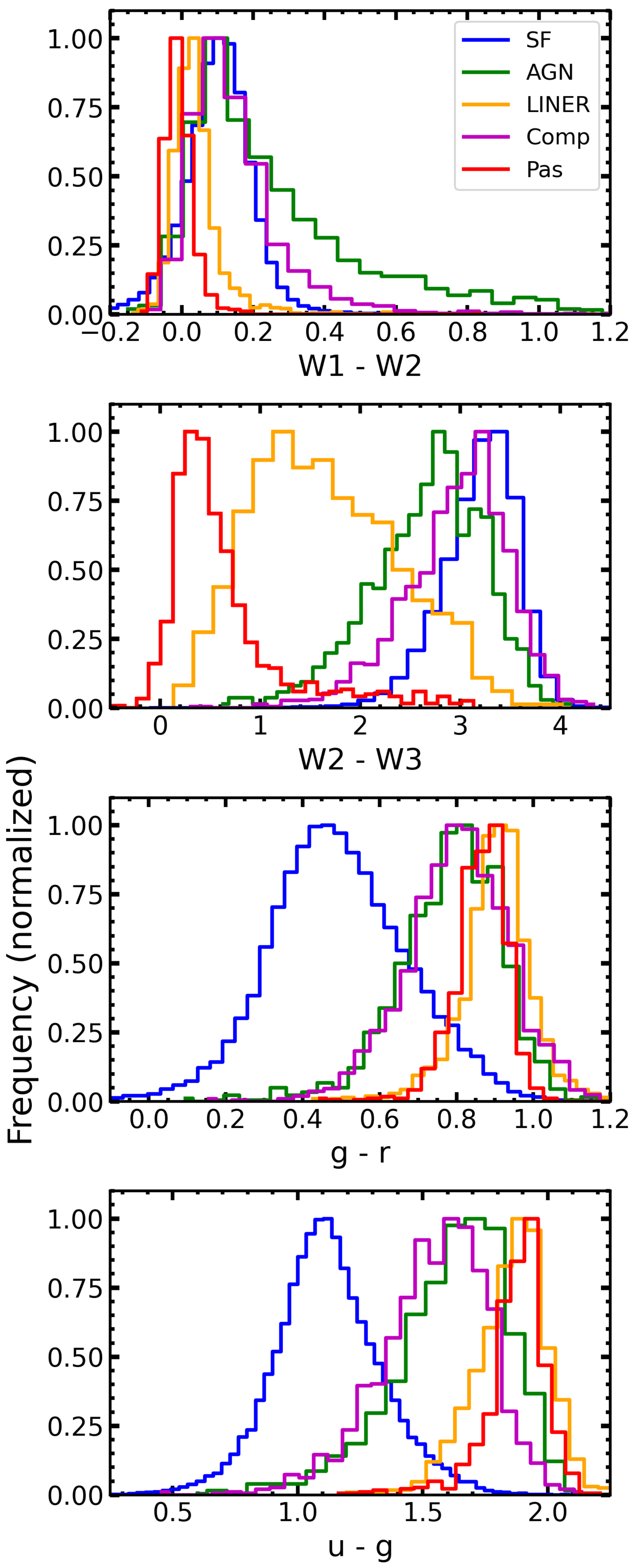
PHOTOMETRY



Smolčić et al. (2009)

Optical Colours

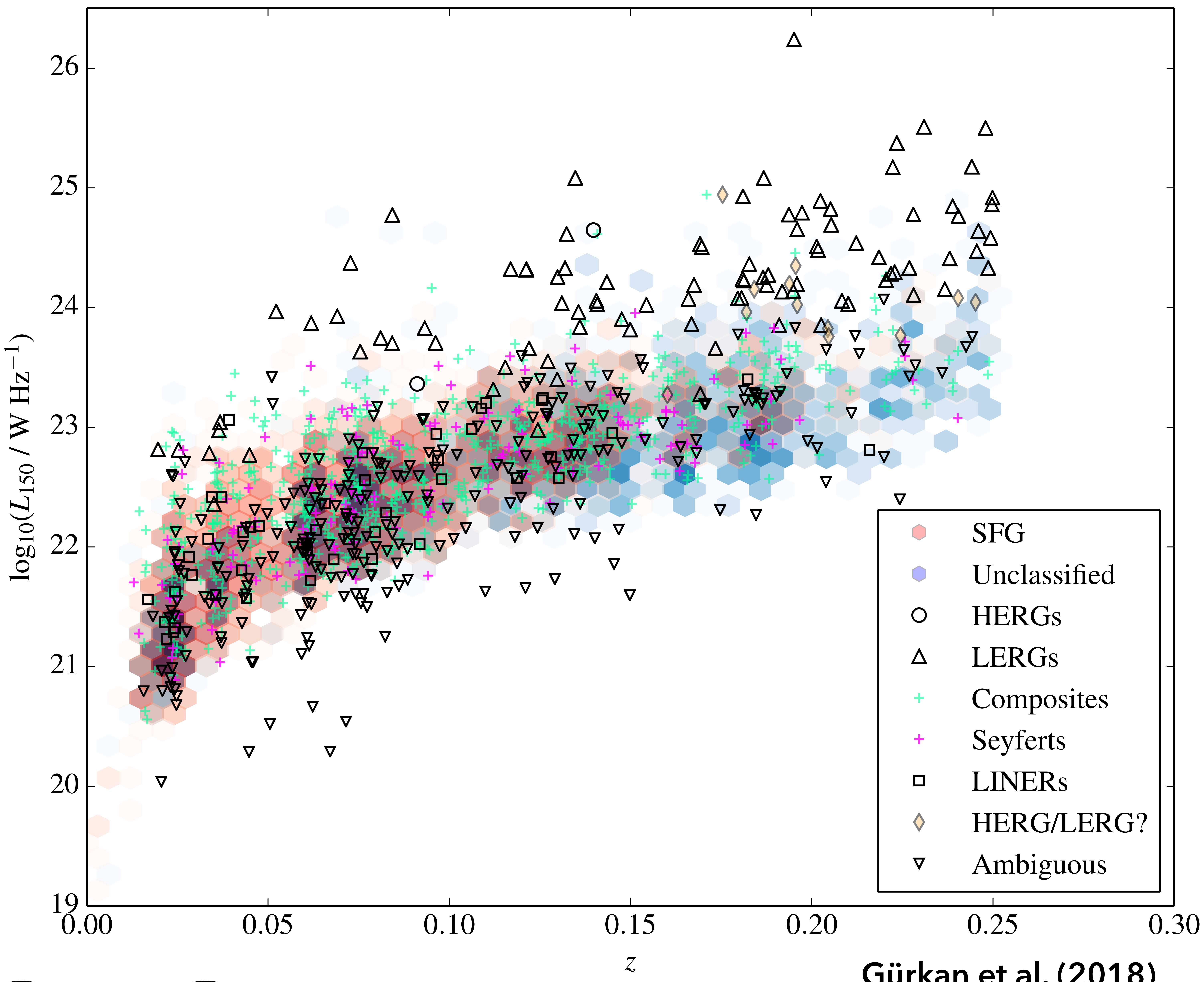
PHOTOMETRY



Daoutis et al. (2023)

Optical and MIR Colours

PHOTOMETRY

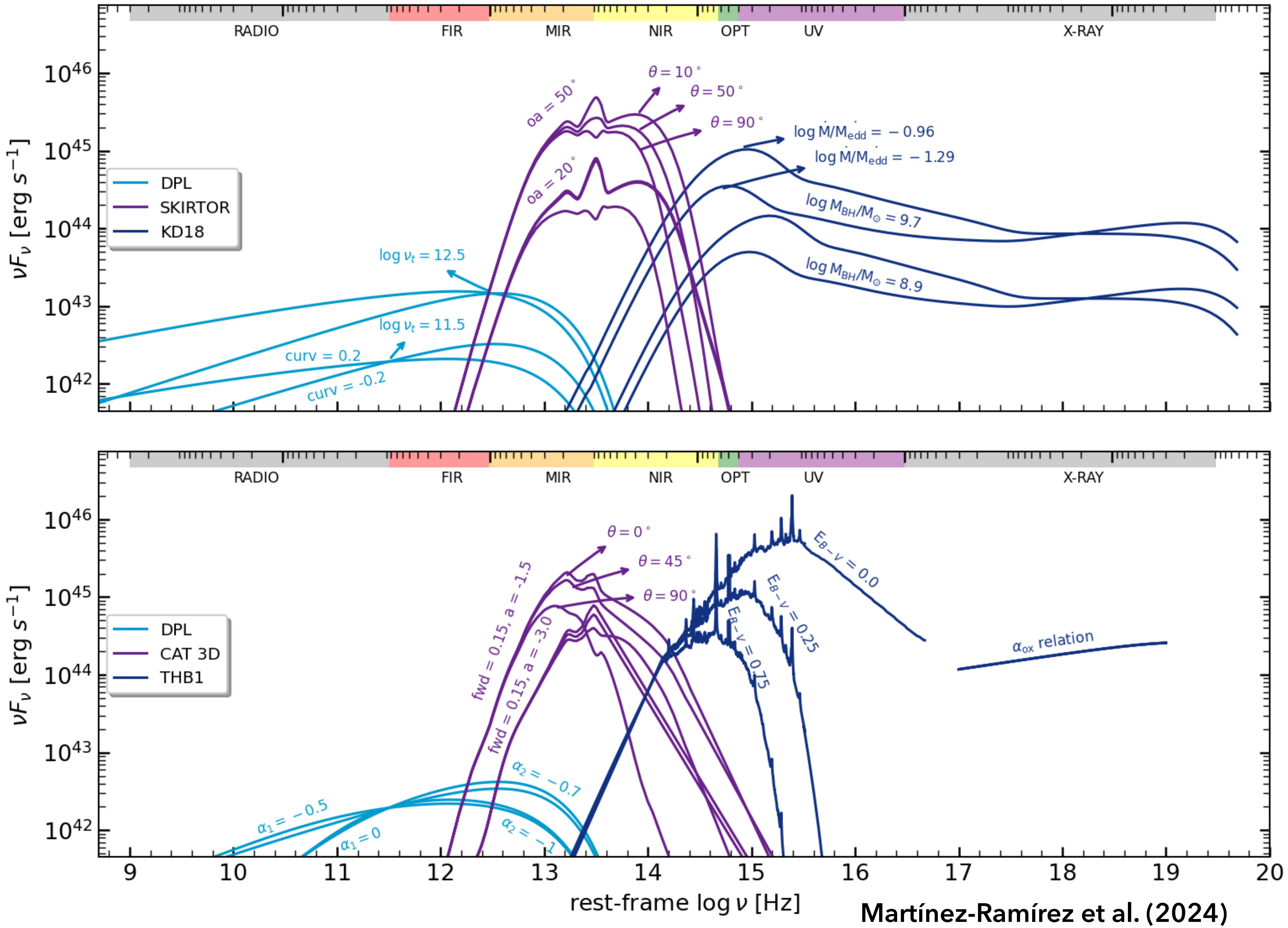


Gürkan et al. (2018)

RADIO POWER

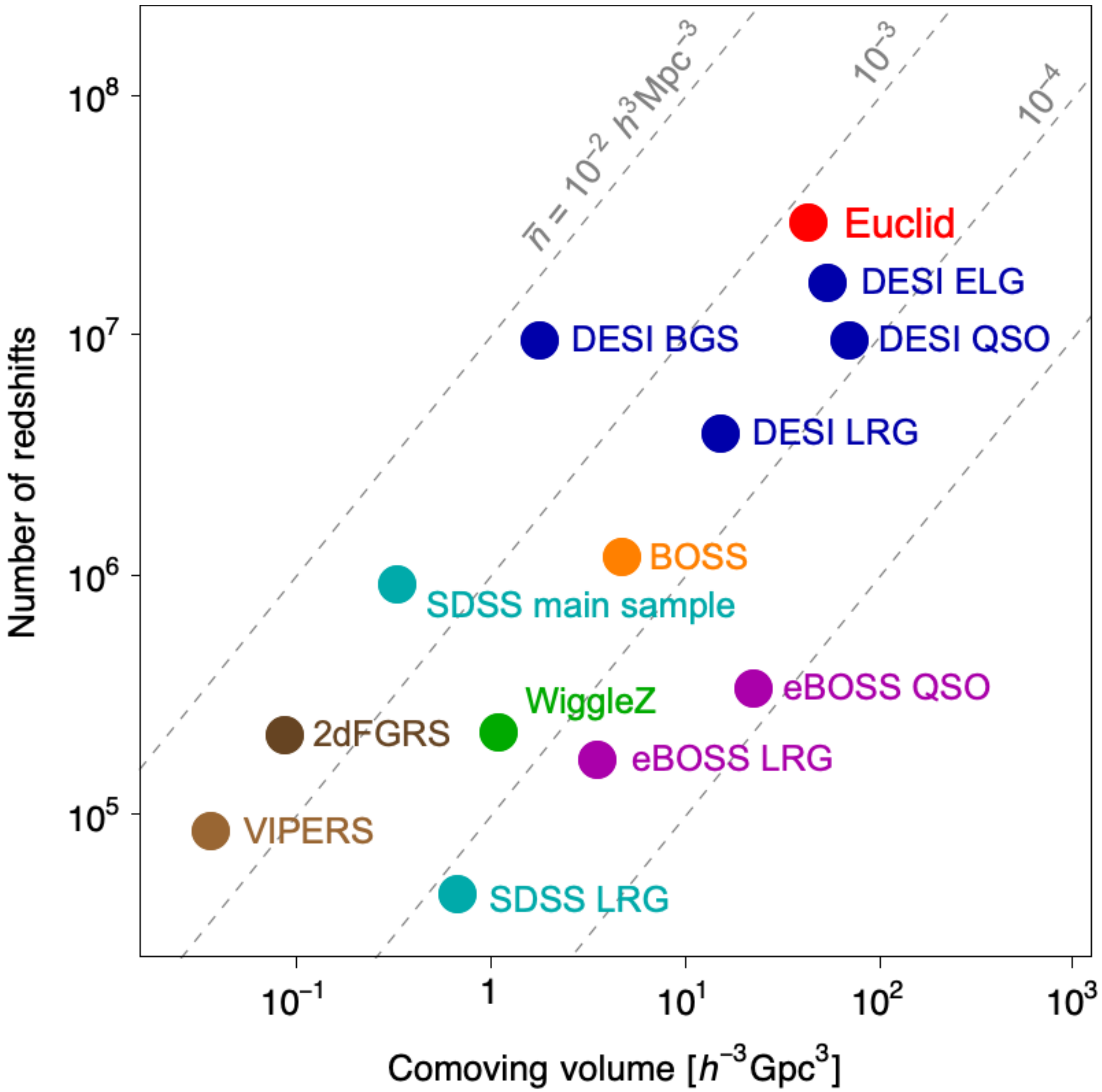
SED FITTING

ONE WAY TO DO MULTI-WAVELENGTH ANALYSIS



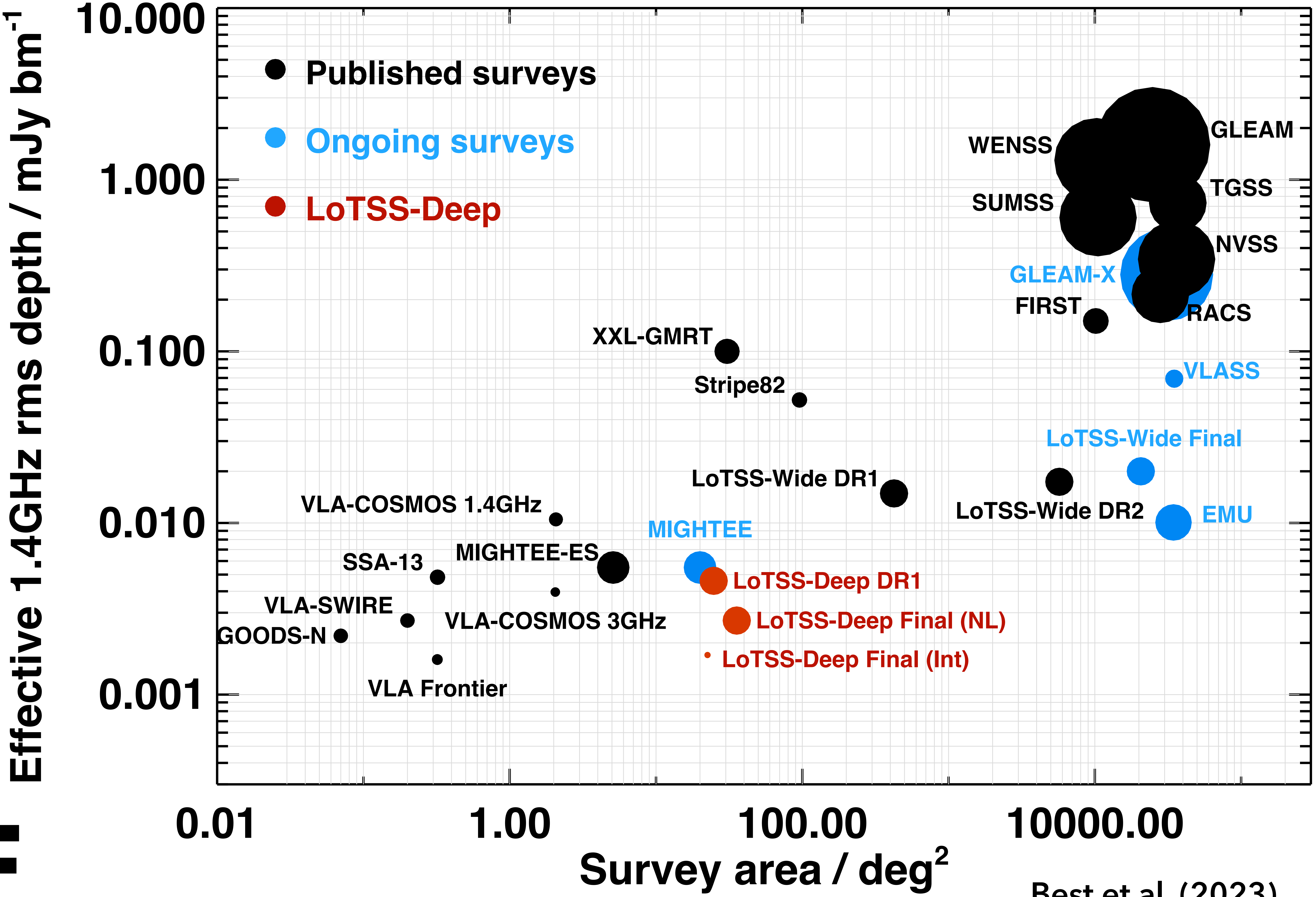
Martínez-Ramírez et al. (2024)

PLENTY OF DATA



Mellier et al. (2024)

PLENTY OF DATA



Best et al. (2023)

Some issues arise

Too much data for traditional methods

Maireles-González et al. (2023)

Data For ML

Training data – Initial fit of model parameters – Bulk of data

Validation data – Fit hyperparameters of model

Calibration data – Calibrate probabilities

Testing data – Final assessment of predictions

Prediction data – Measurements without labels

DATA FLOW

Start with CatWISE2020 sources

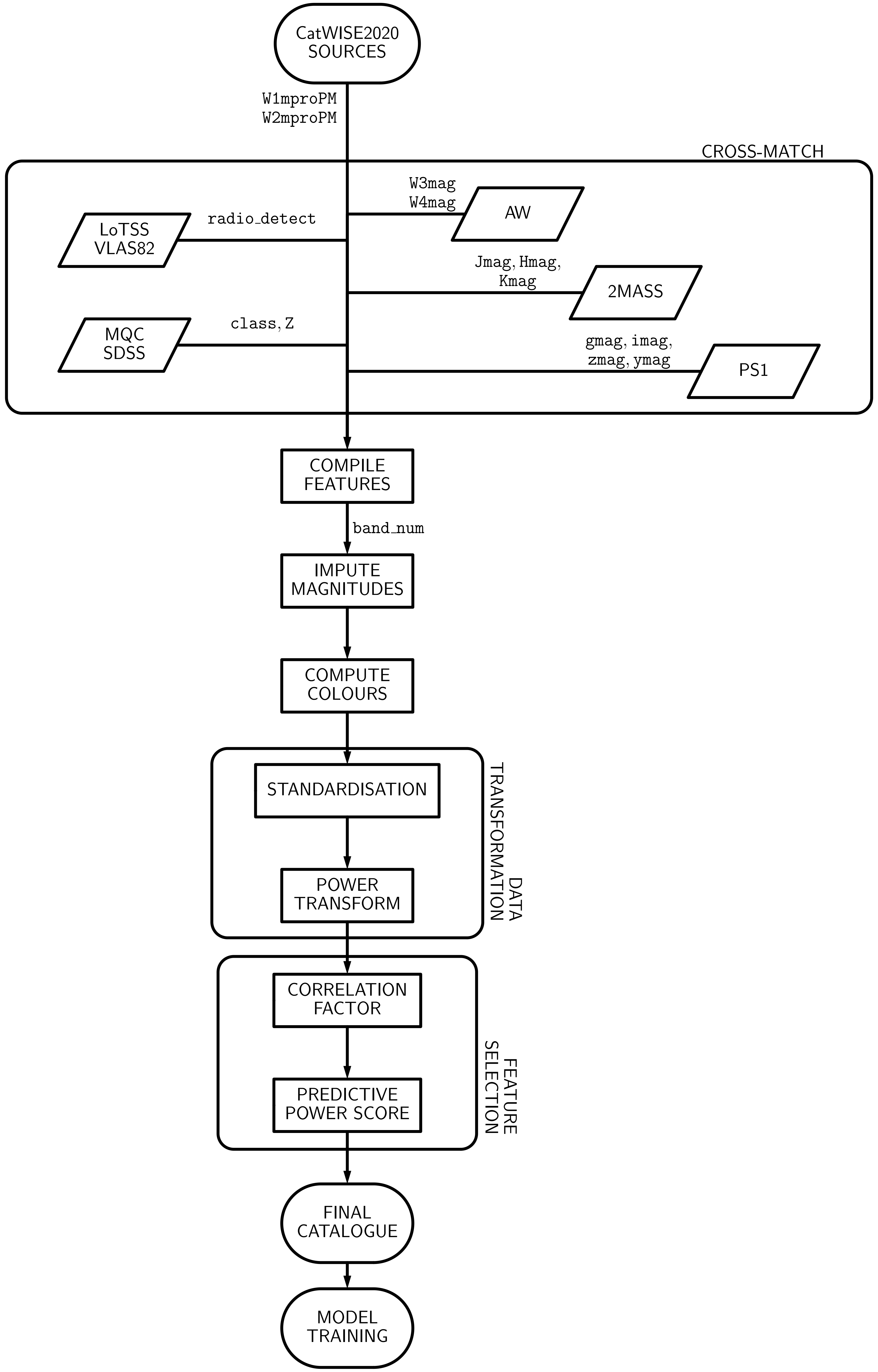
Add Optical, NIR, MIR counterparts

Impute missing values

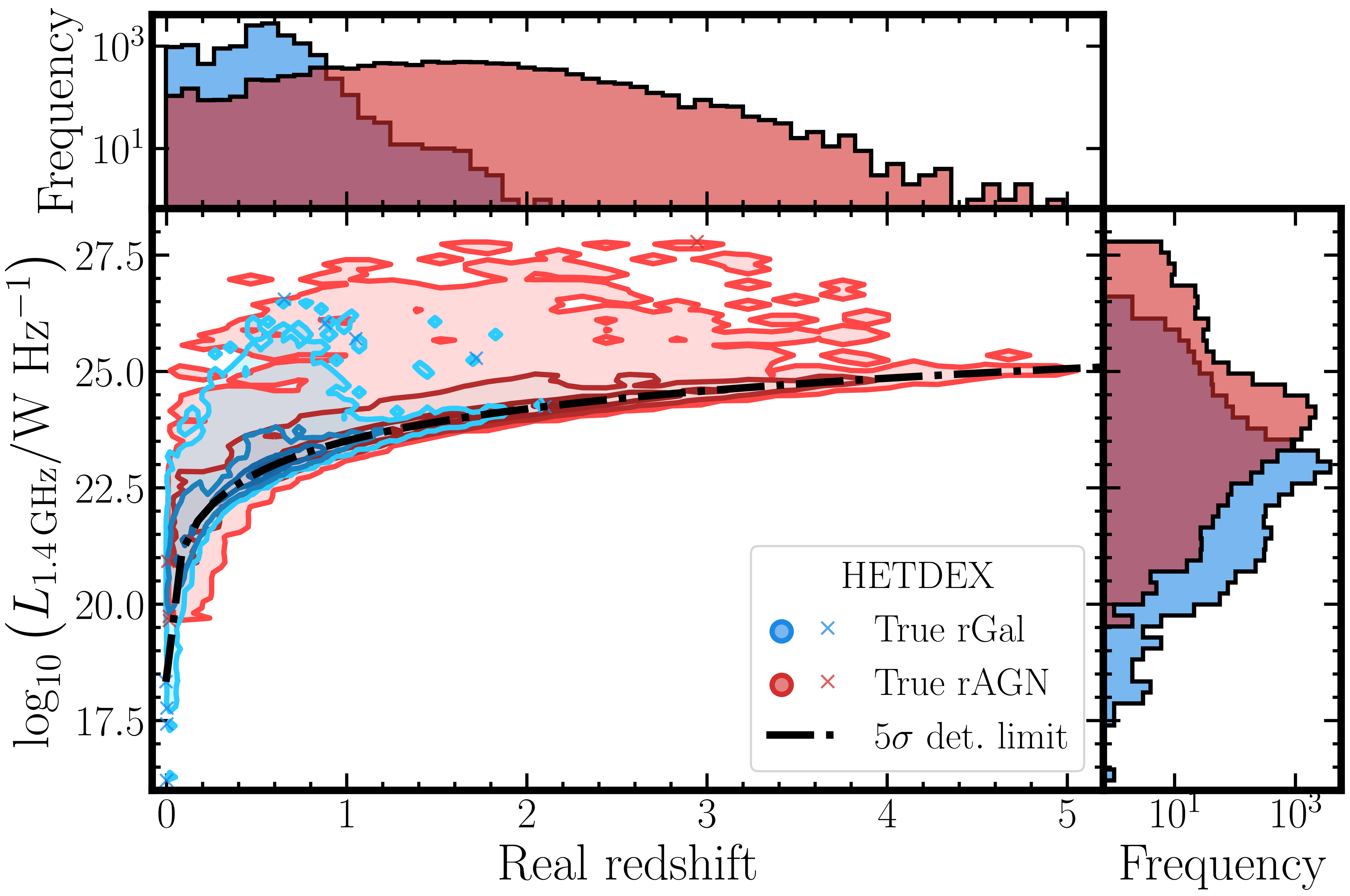
Include colours

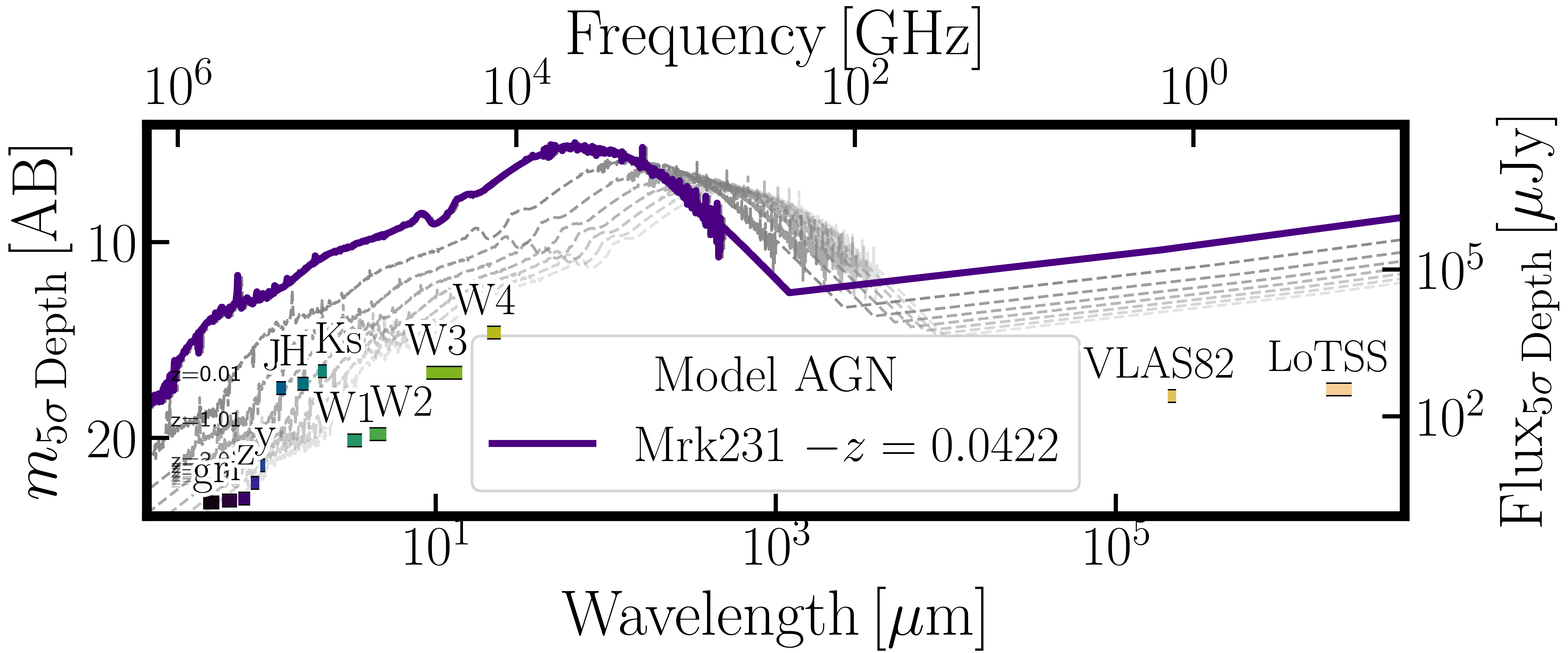
Add labels for training: Class, radio-detection, redshift

For each step, determine most informative features



Source distribution HETDEX





HETDEX Photometry

Stripe 82 (S82) Field

Only for final testing

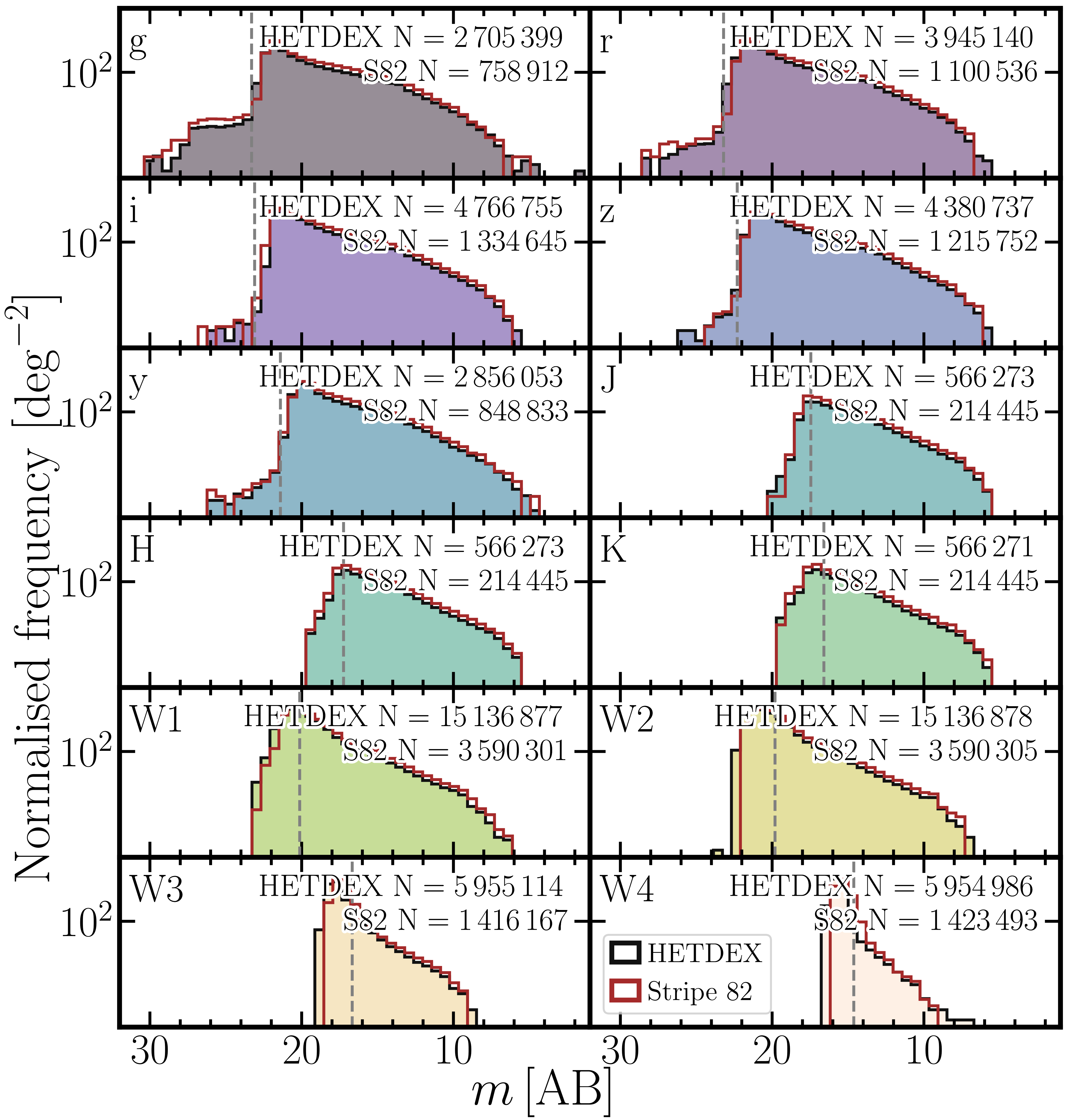
92 deg² covered by VLA @ 1.4 GHz, 52μJy, 1.8″ resolution

~3.5 million CatWISE2020 detections

~18k spectroscopically-confirmed AGN + ~4k spectroscopically-confirmed SFGs

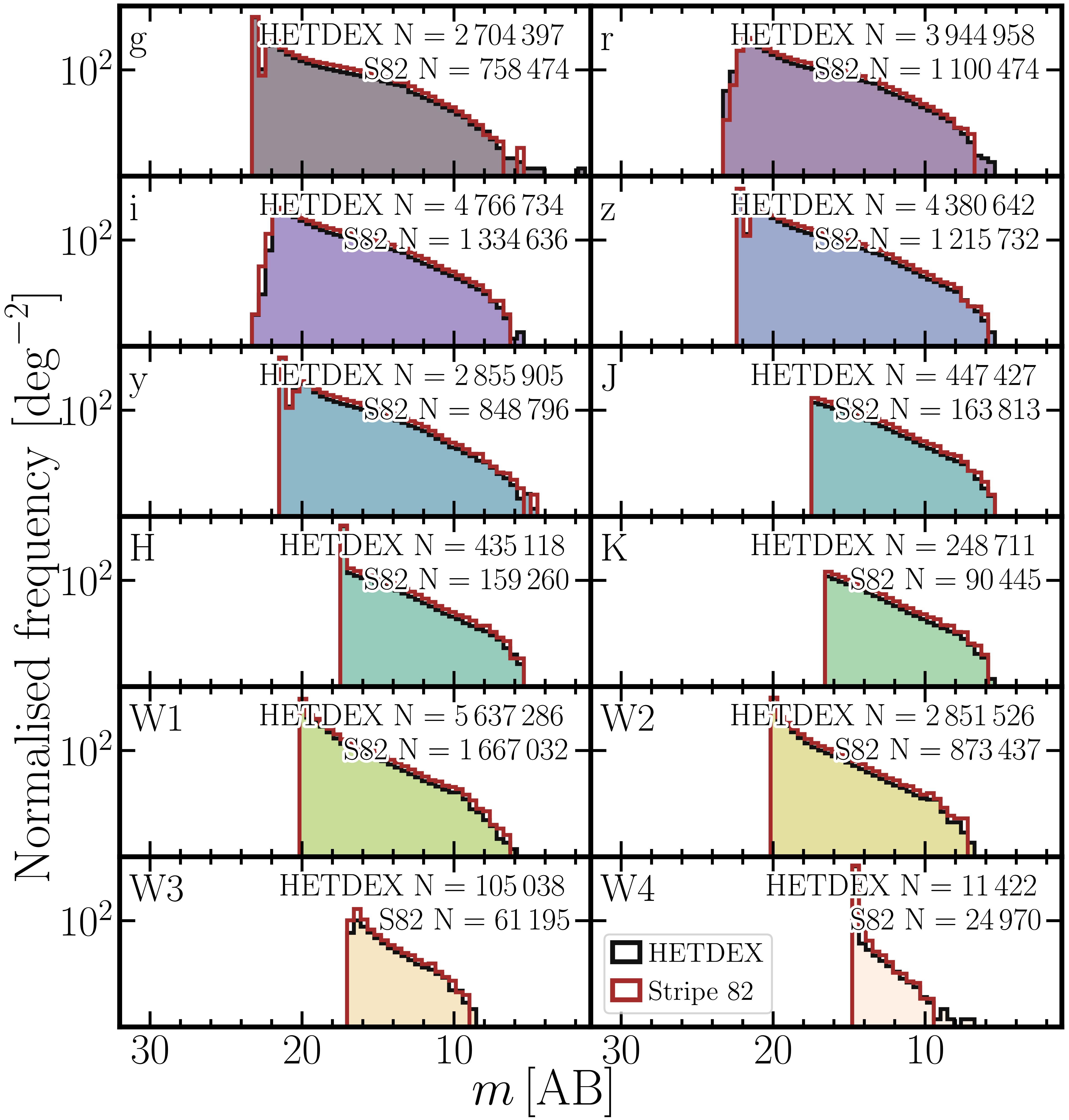
Full photometry

Similar across fields



Imputed Photometry

Similar across fields



BASELINE SELECTION

Based only on the fraction of sources in the sample:

Selection of AGN: Probability 43 % - Selection of SFG: Probability 57 %

Selection of radio in AGN: Probability 13 % - Selection of radio in SFG: Probability 13 %

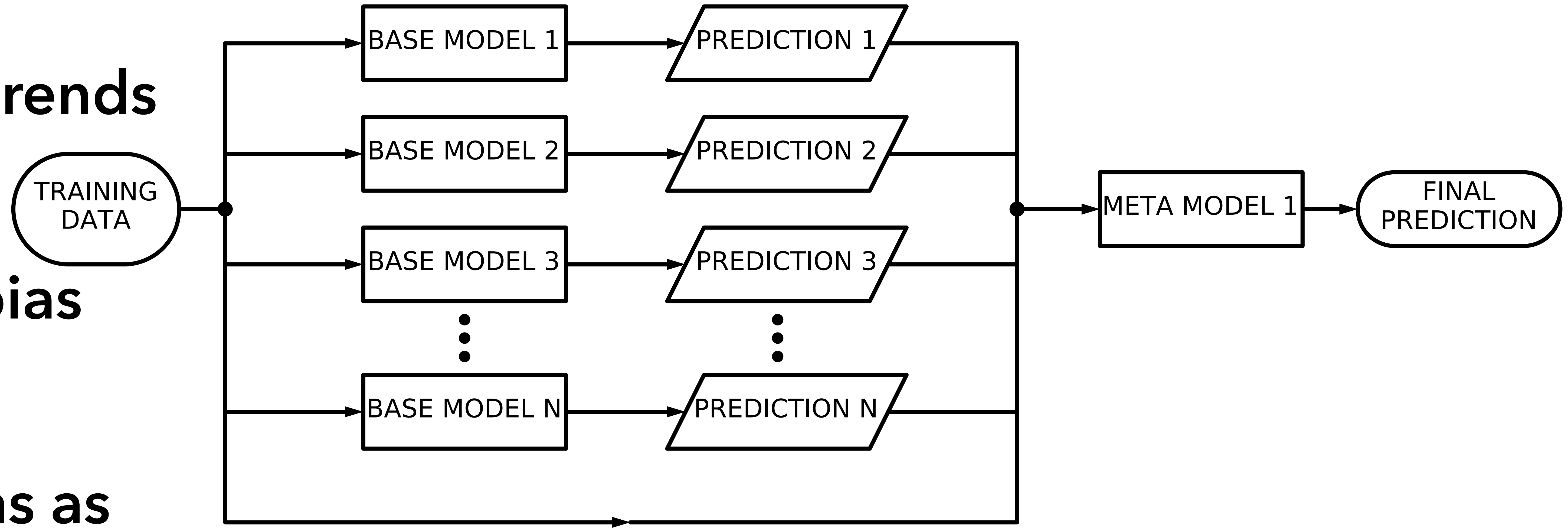
MODEL STACKING

For each step of pipeline, combine several algorithms

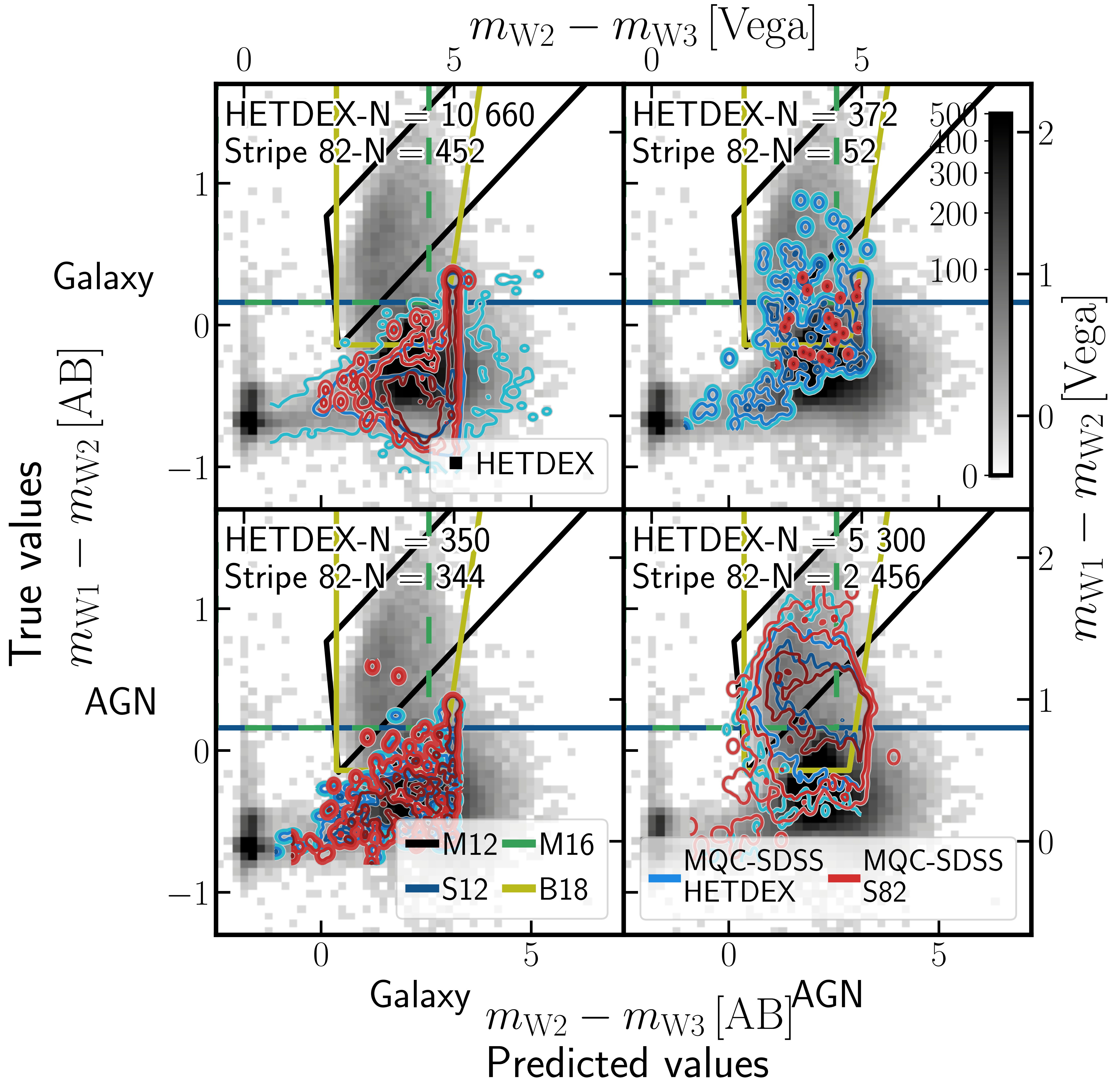
Each model learns slightly different trends

Aims at improving results reducing bias

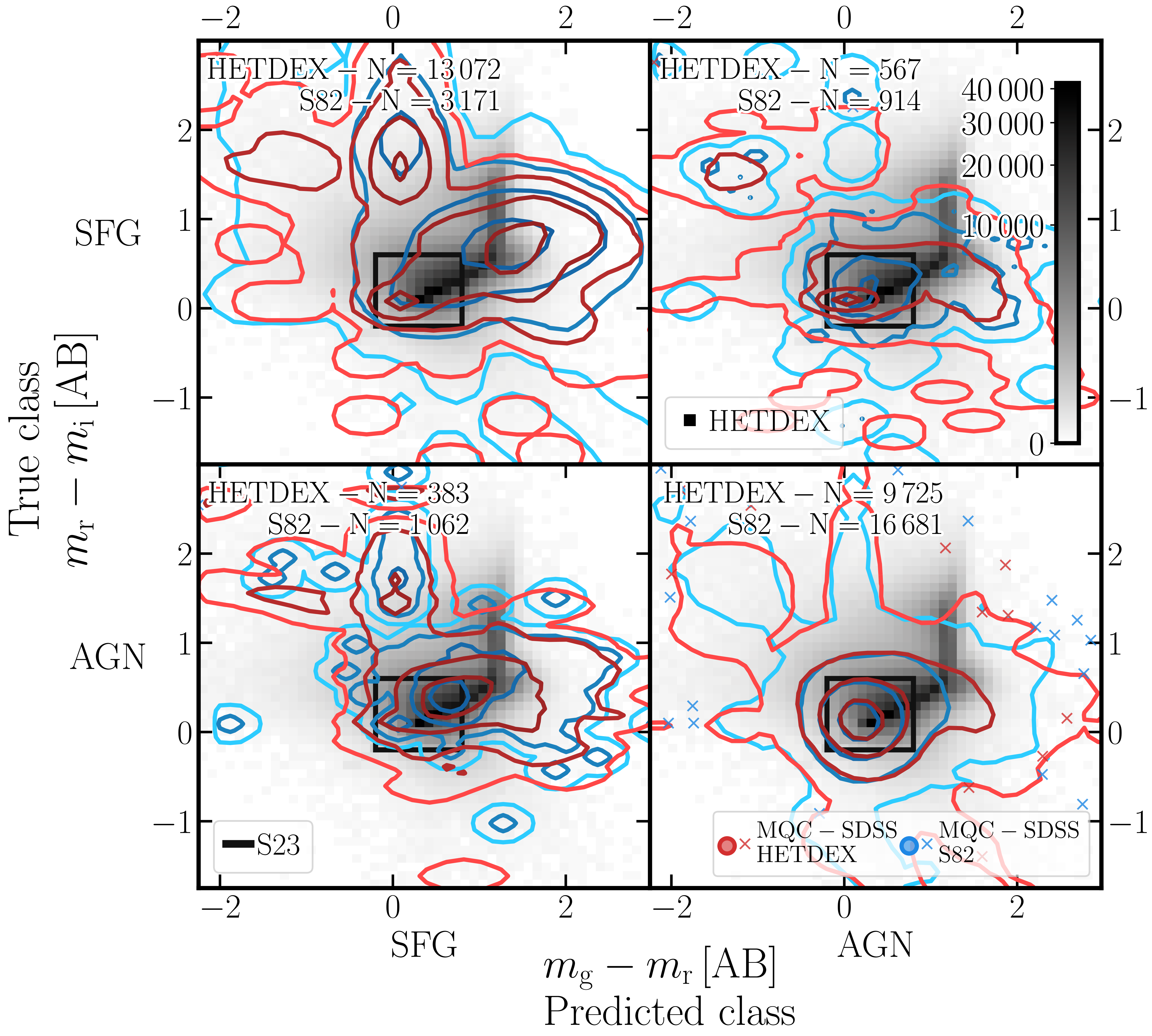
Final model uses previous predictions as input



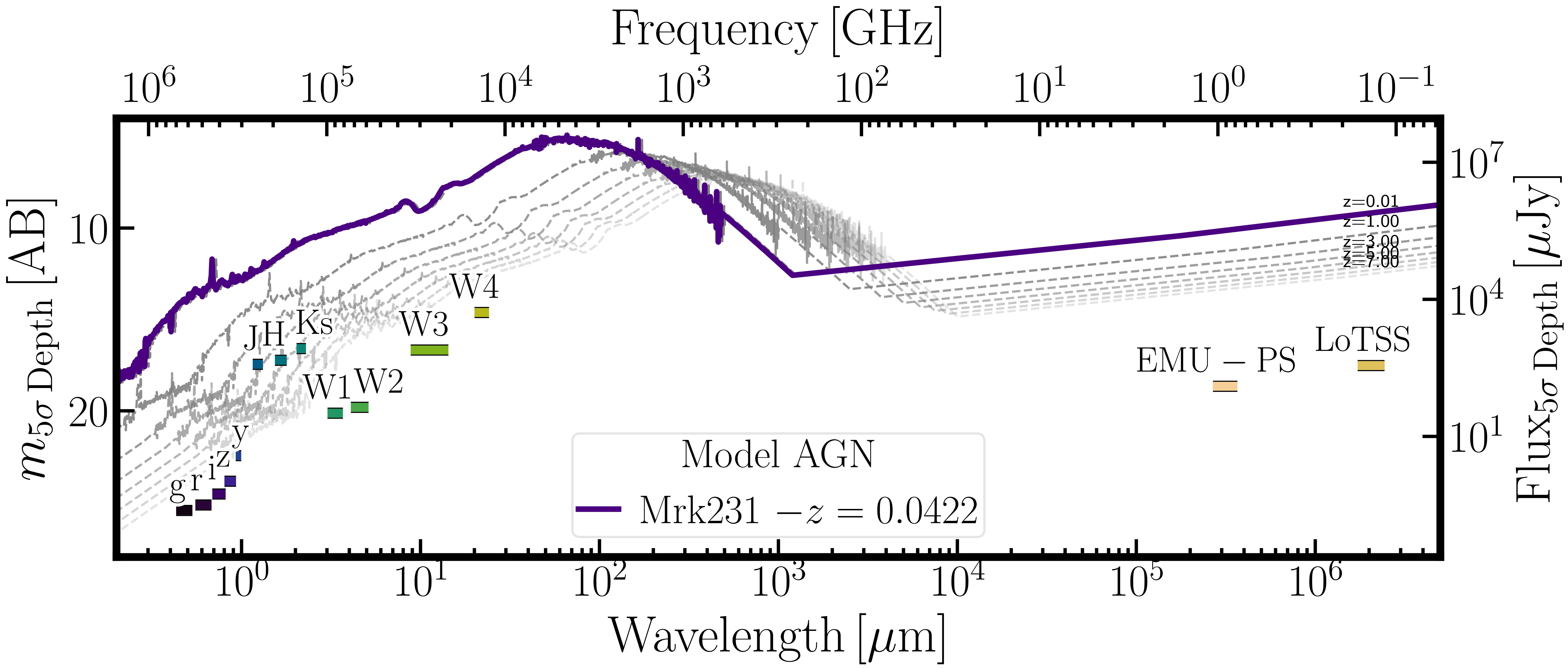
WISE Colours and Predictions



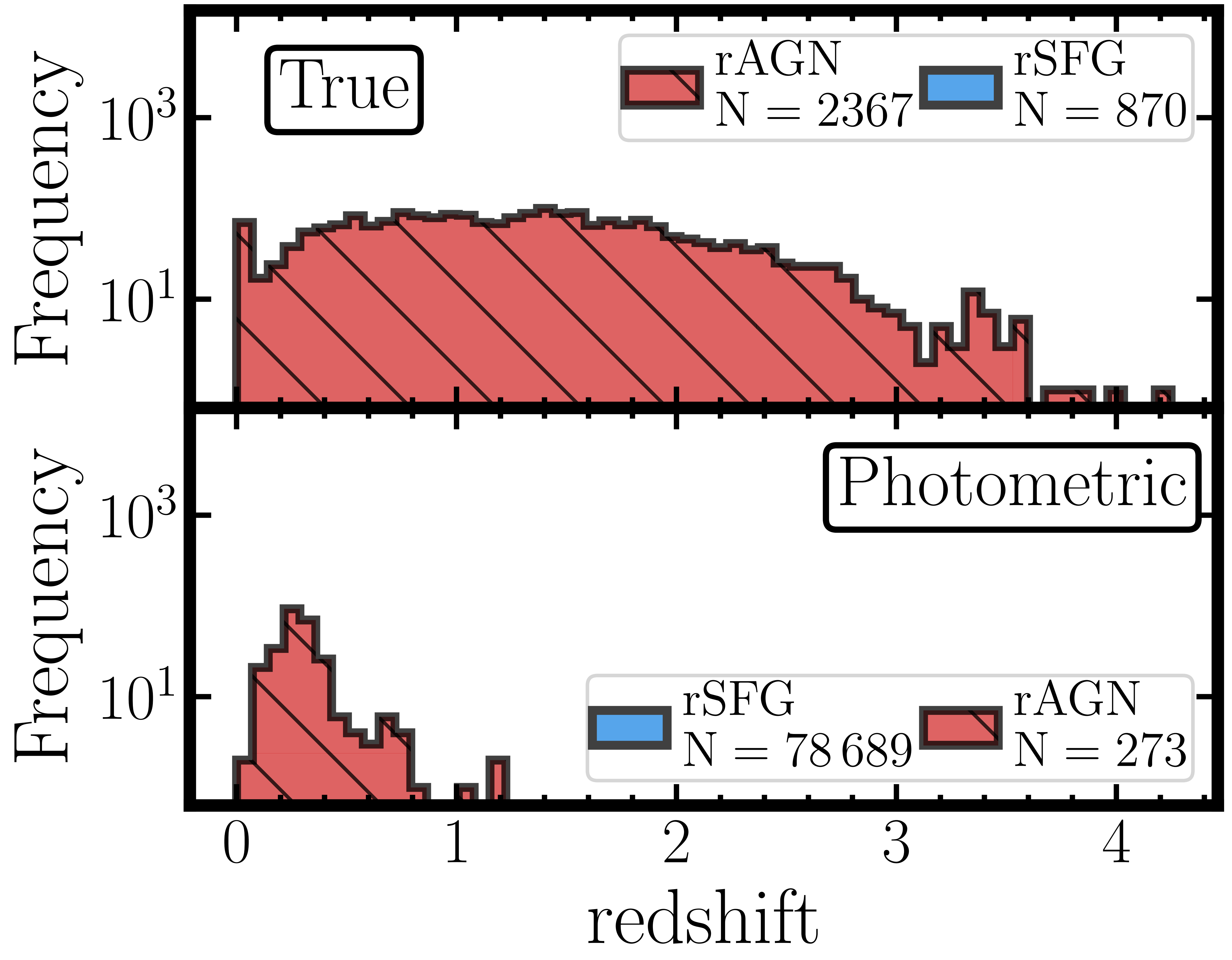
WISE + PS1 COLOURS AND PREDICTIONS



EMU PHOTOMETRY BANDS



Prediction in EMU-PS



Radio Luminosity Function corrections