An Ensemble of CNN and UMAP Automates Emission Line Selection in HETDEX DR1

Nao Sakai

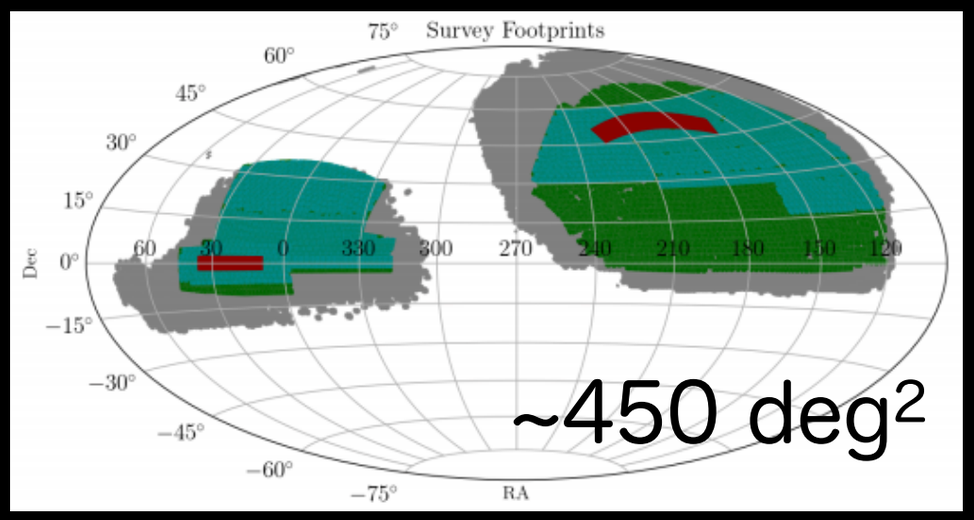
5th August, 2020

The galaxy IGM workshop

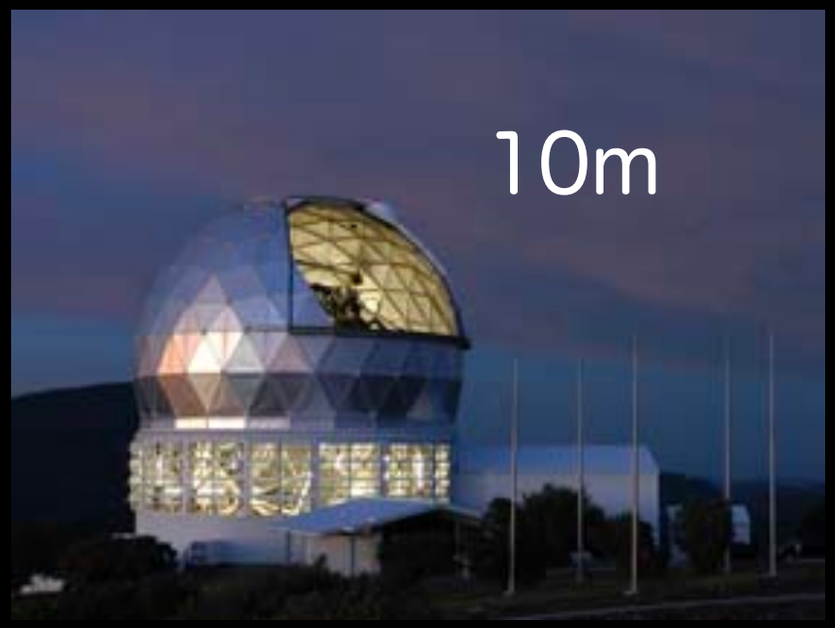
What is HETDEX?

Hobby-Eberly Telescope Dark Energy Experiment

➡ The expansion history of the universe at z=1.9~3.5

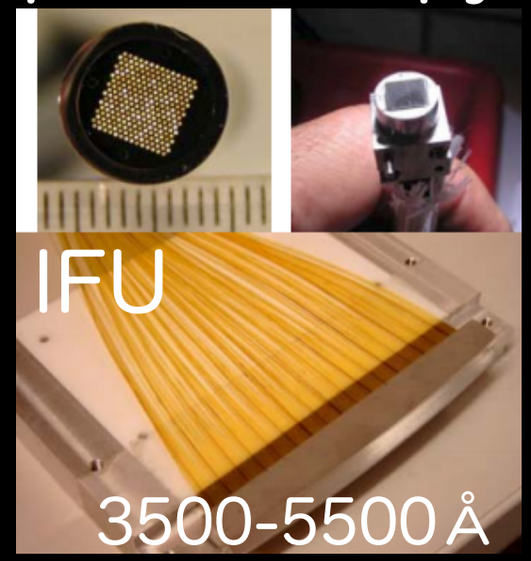


Gebhardt, HETDEX MT 2020@ Austin



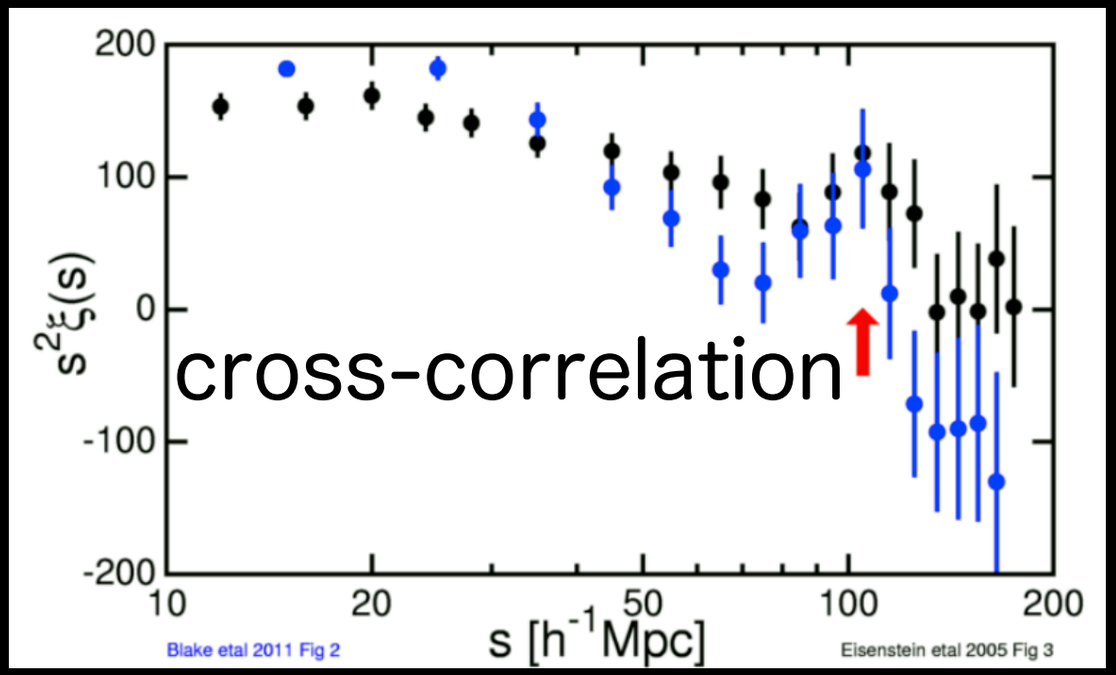
HETDEX official HP

Spectroscopy

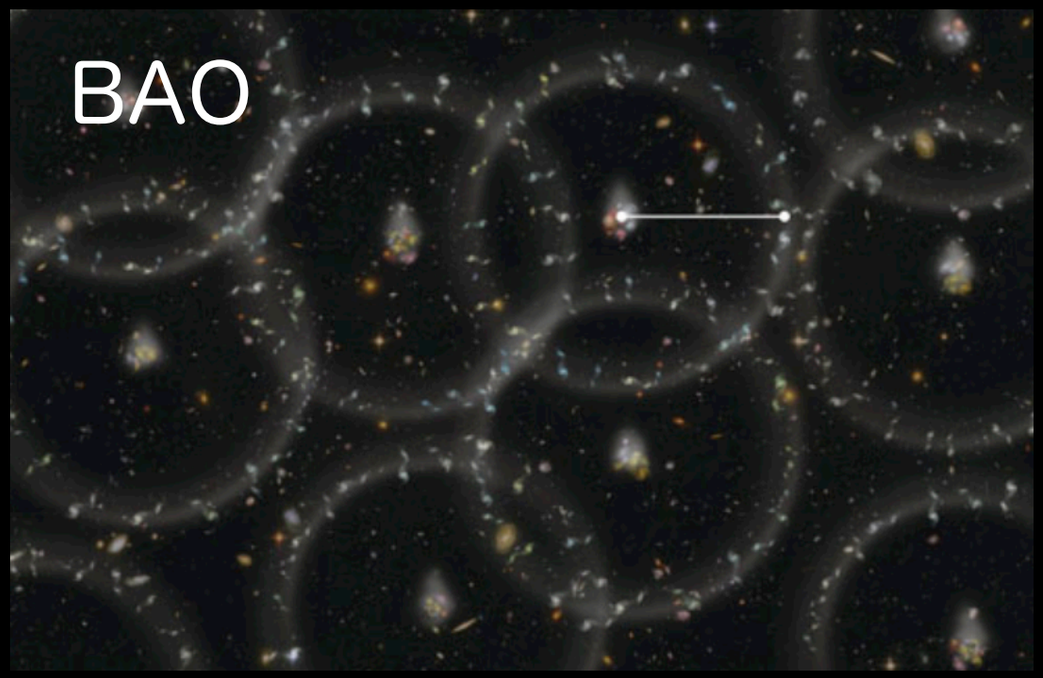


Hill, HETDEX MT 2009@ Munich

1 million LAEs



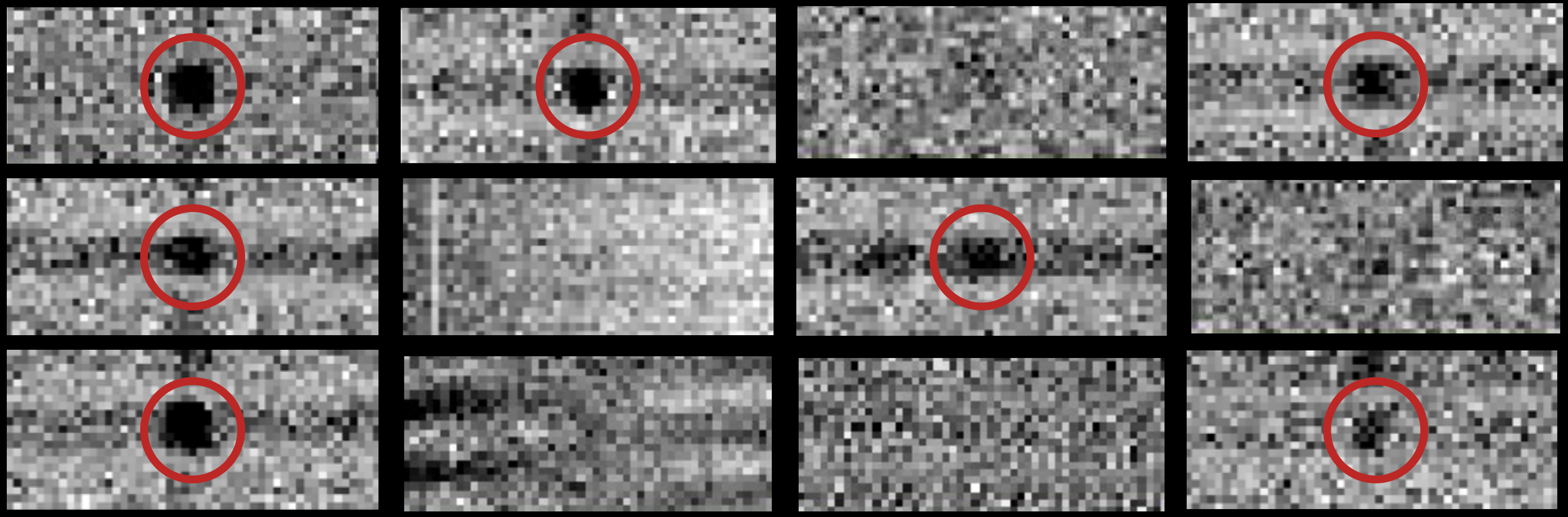
Edward L. Wright, UCLA



BOSS official HP

HETDEX pipe line is not so complete

Detected sources



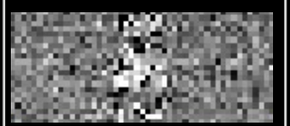
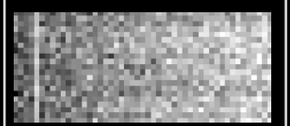
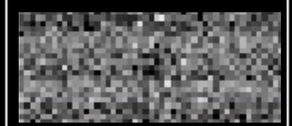
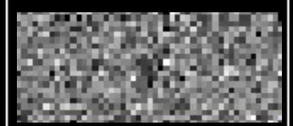
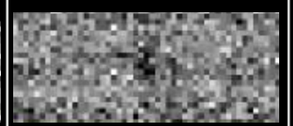
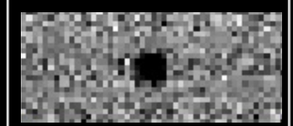
~45% are fake…

Need to do visual classification

We did visual classification

Consider as real

20 HETDEX members, 2 months

Class	-2	-1	0	1	2	3	4	5
Description	Bad Amp	Bad Pixel	Spurious	Likely Spurious	Possibly Spurious	Possibly Real	Likely Real	Confidentialy Real
2D spectrum								
Number	20640	843	9195	1343	651	755	3240	34252

Only ~3% of the whole data

Problems

Not doable ➡ Takes several years

Poor reproducibility ➡ Members have their own criteria

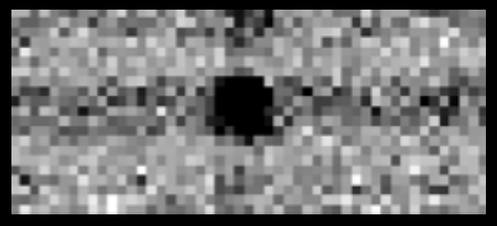
Automate this work with machine learning technique

1st round: CNN selection

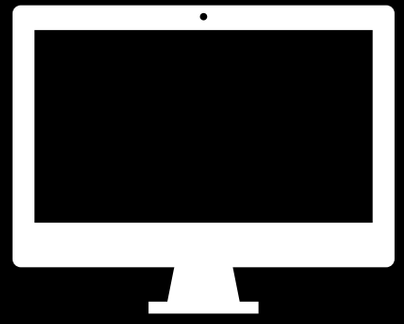
CNN: Convolutional Neural Network

- classification algorithm
- for image recognition
- Non-linear functions

input: 2D spectra (64Å)



CNN model



output: real or fake

real
0.976

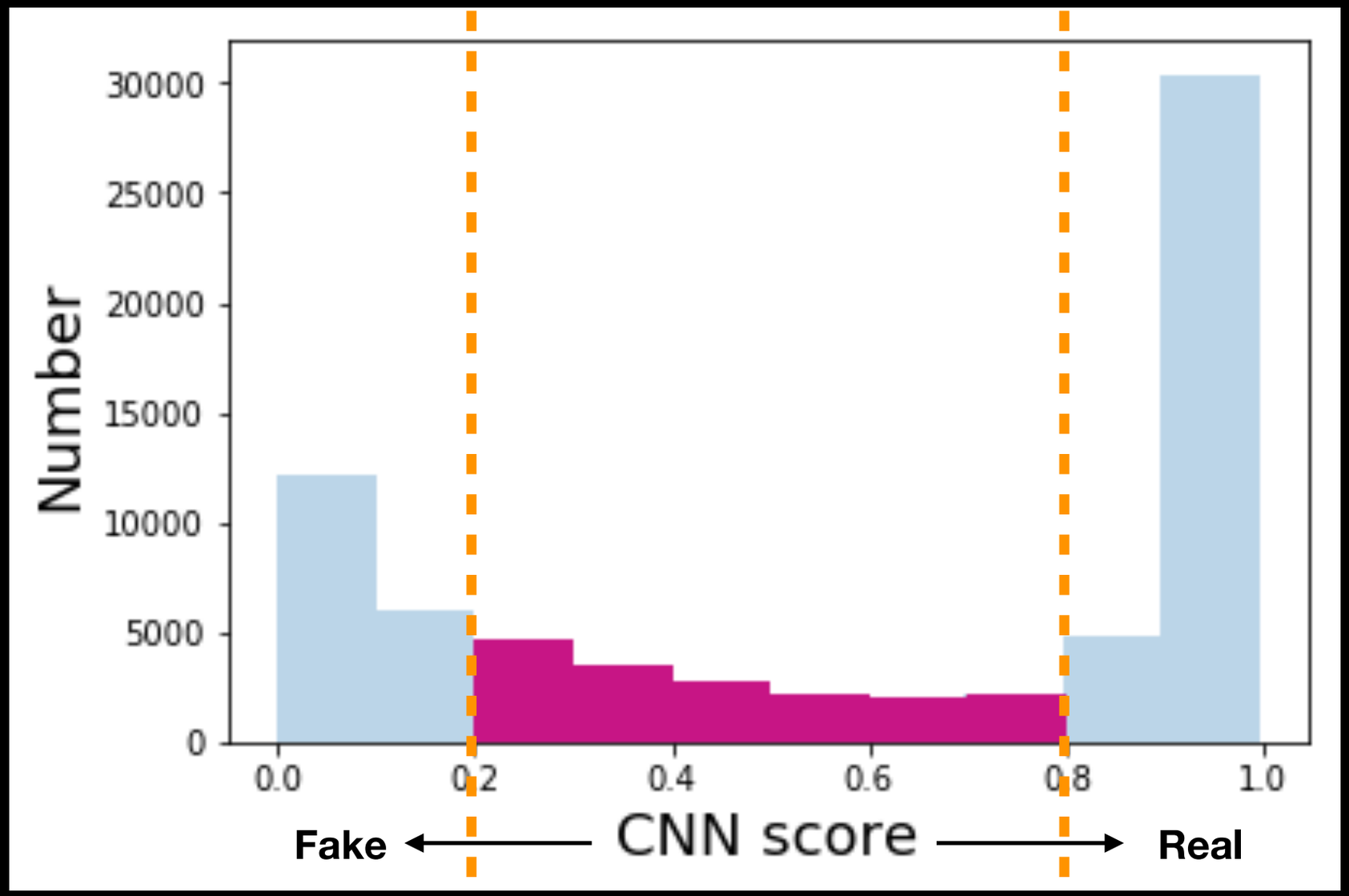
0 to 1 score
Put a threshold

Training

10k real sources (Class 4 & 5)

10k fake sources (Class 0 & 1)

sources CNN is not good at classifying



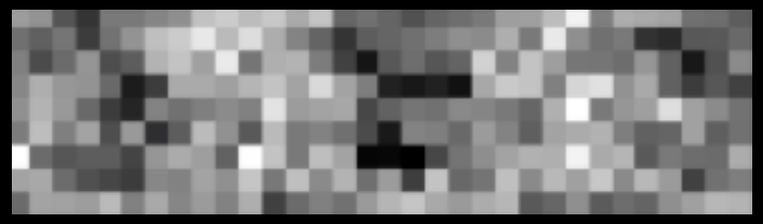
Number	18200	17448	35029
Contamination	98.84%	77.40%	4.85%
	discard them	What to do?	keep them

Many obvious fakes and reals

Consider as real

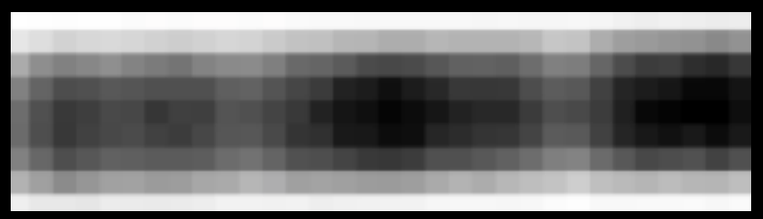
Class	-2	-1	0	1	2	3	4	5
Number	8406	408	2891	812	459	529	1504	2439

Fake
-2

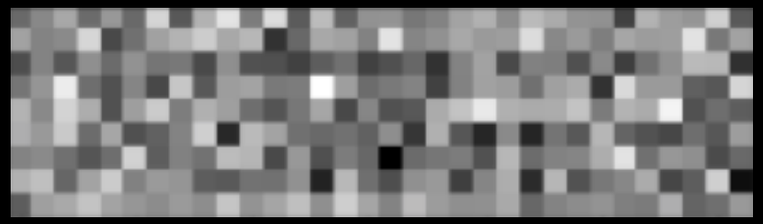


-2

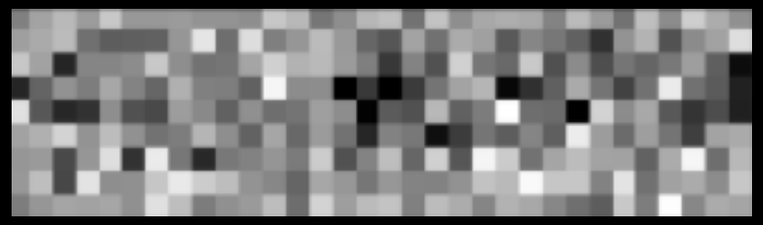
-2



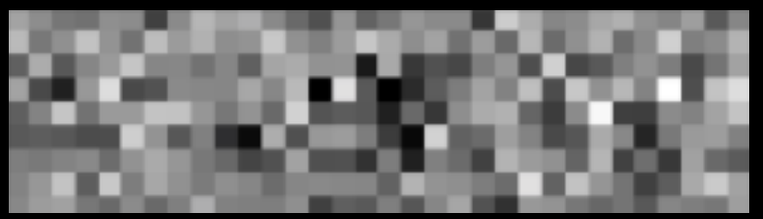
Fake
0



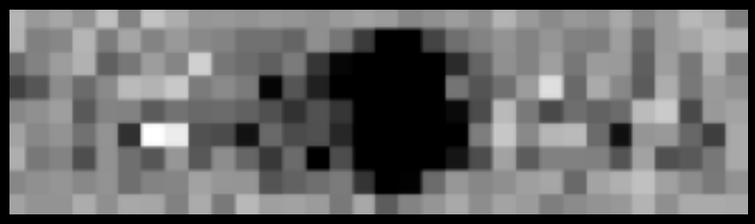
0



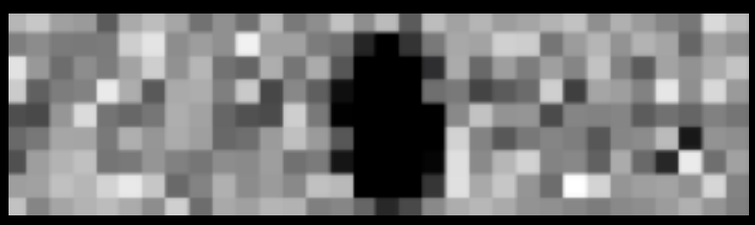
1



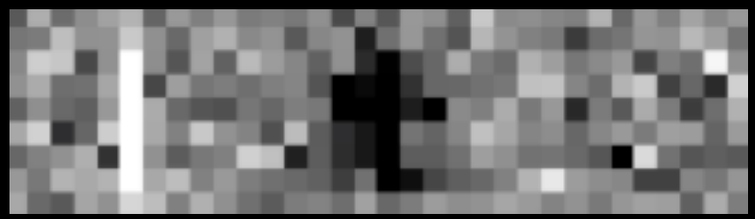
Real
5



5



4

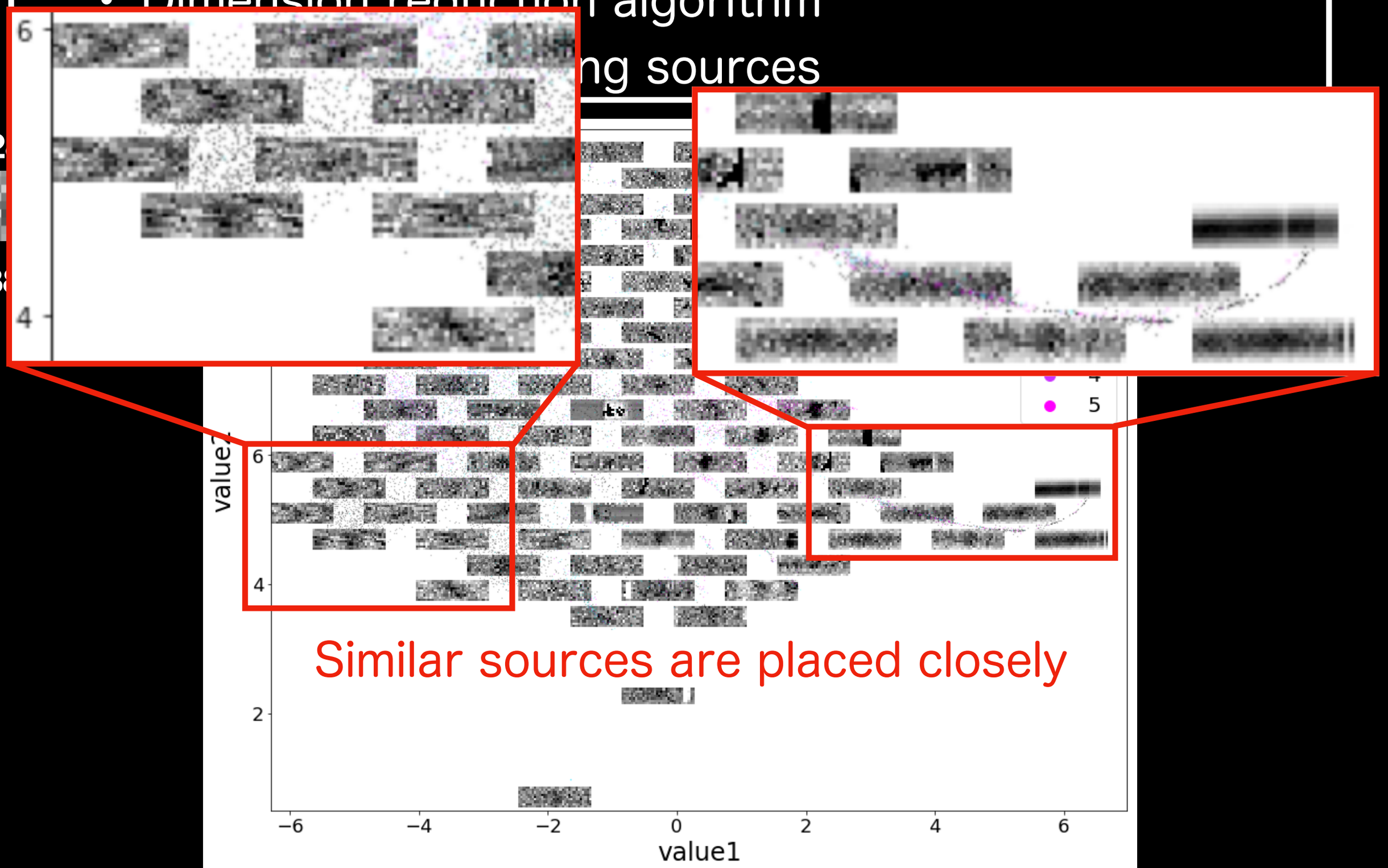


➔ Reclassification

2nd round: UMAP clustering

UMAP: Uniform Manifold Approximation and Projection

- Dimension reduction algorithm
- ng sources

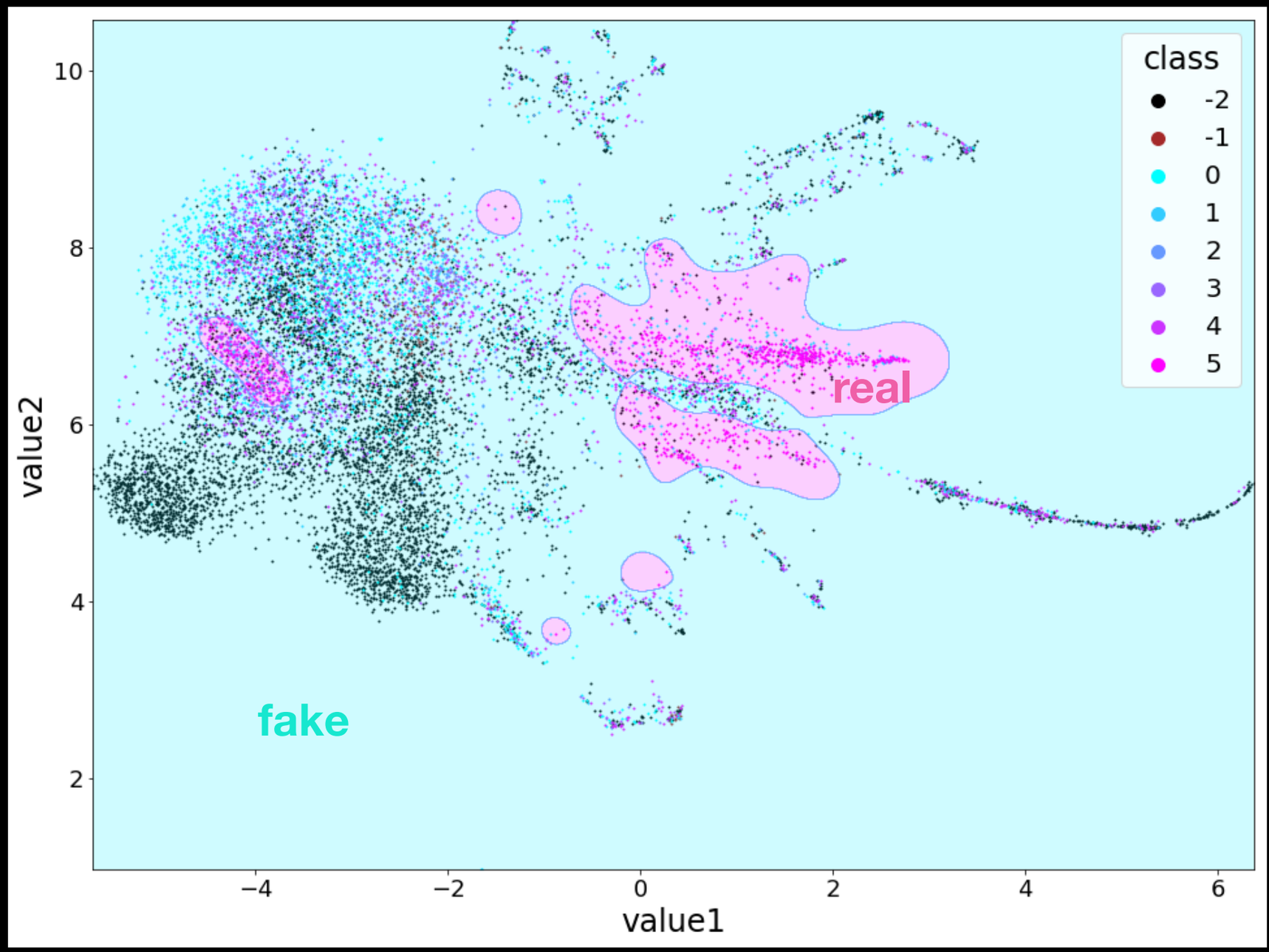


Similar sources are placed closely

SVM boundary decision

SVM: Support Vector Machine

- Boundary decision algorithm
- Margin maximization



Doable & unique method, good quality

	Completeness	Contamination
HETDEX Pipe line	100.0	46.96
CNN + UMAP	91.48	5.73

(%)

Requirement of our cosmology

$$completeness \times (1 - contamination) \geq 0.94$$

- In proportion to the S/N the power spectrum

Current value

$$0.9148 \times (1 - 0.0573) \simeq 0.8624$$

★ ~90% Figure-of-Merit ★ Doable time ★ Unique criterion

Appendix

Wave

0.251

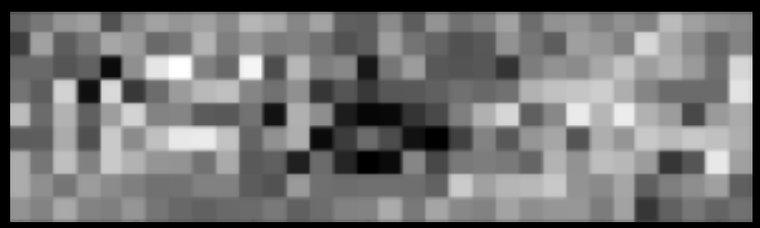
0.28

0.455

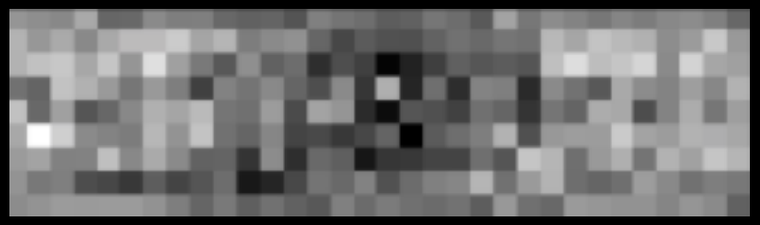
0.458

0.465

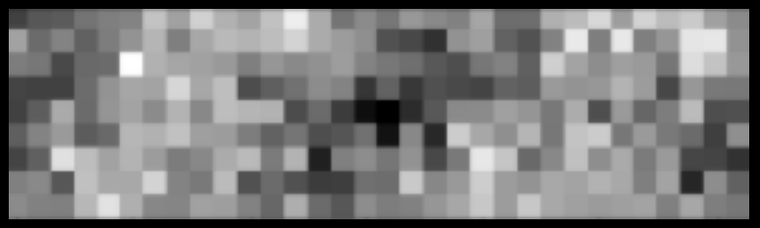
0.557



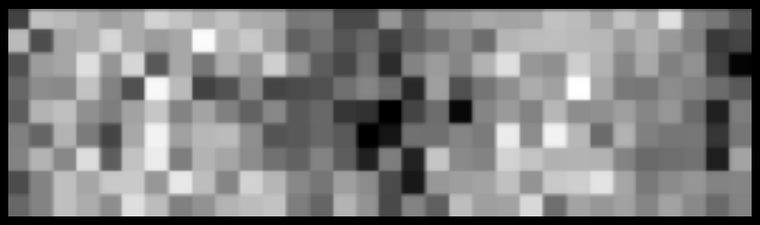
0.658



0.658

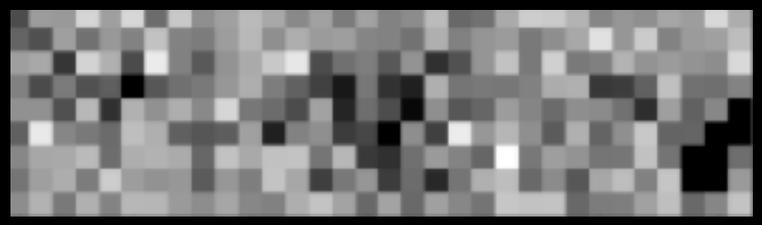


0.665

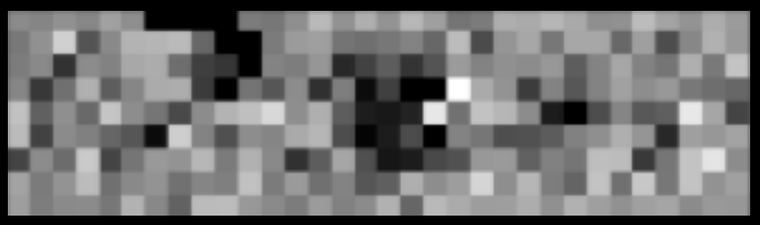


Real

0.602



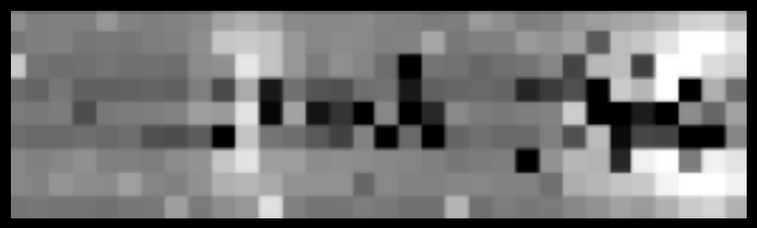
0.706



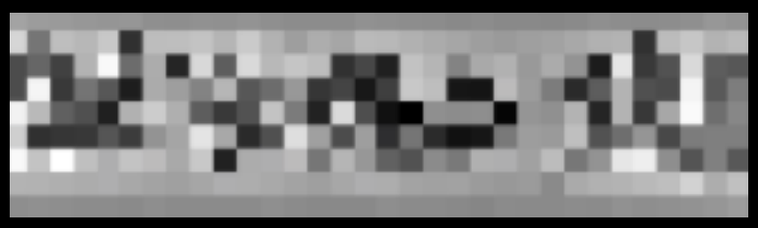
0.743

Noisy

0.237



0.255



0.346

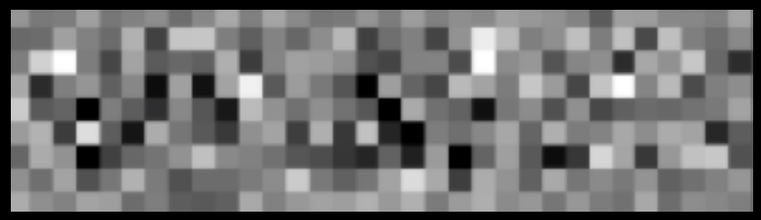
0.389

0.419

0.437

0.442

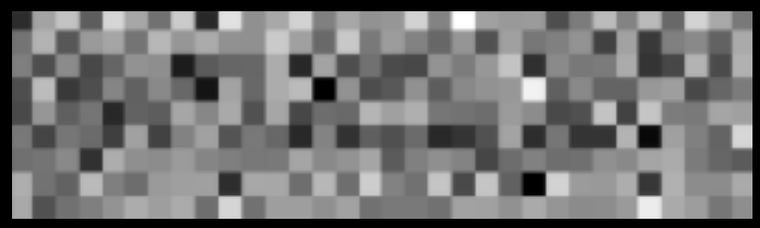
0.468



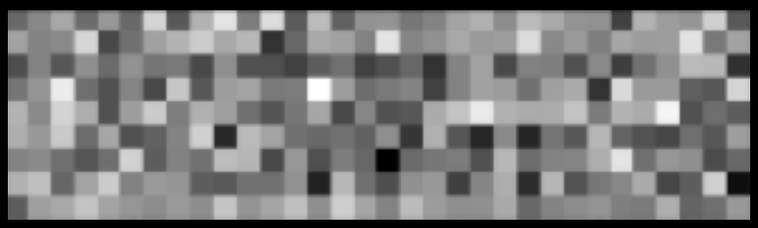
0.470

Random nosie

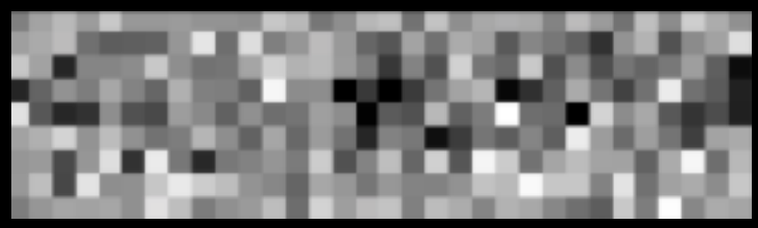
0.236



0.236



0.251



Continuum like

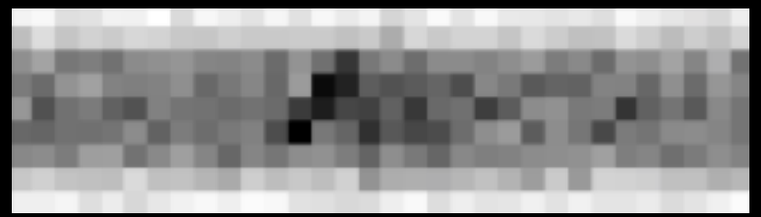
0.201

0.216

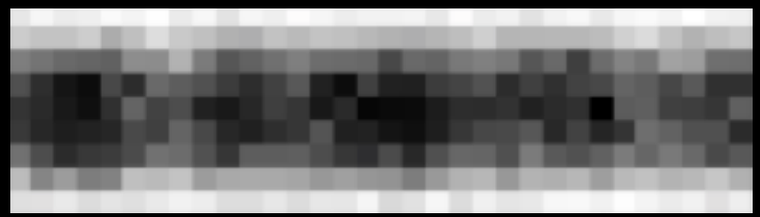
0.223

0.269

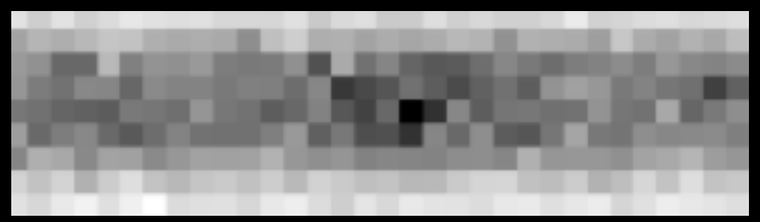
0.295



0.313

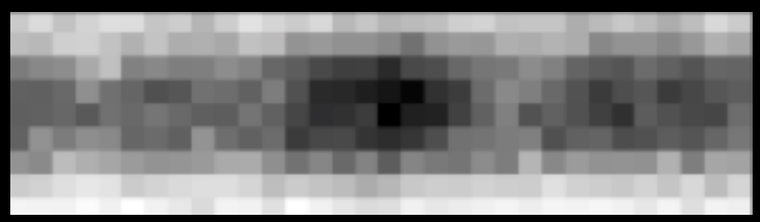


0.314



0.453

0.651

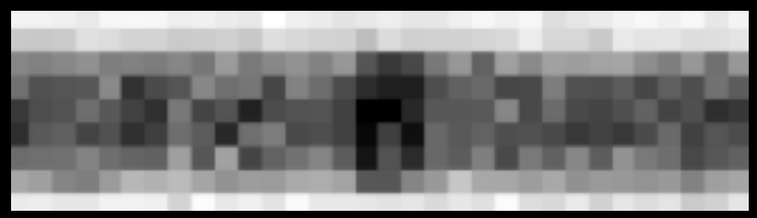


Real continuum

0.461

0.673

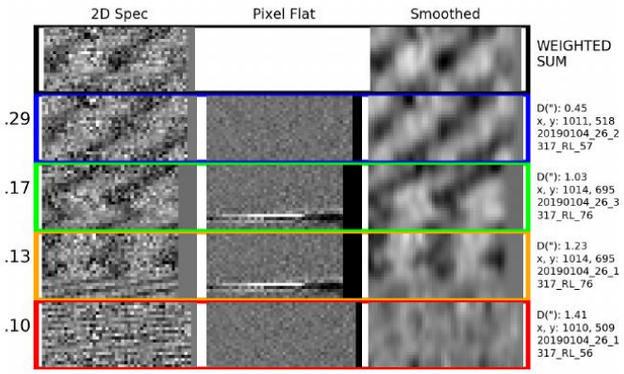
0.708

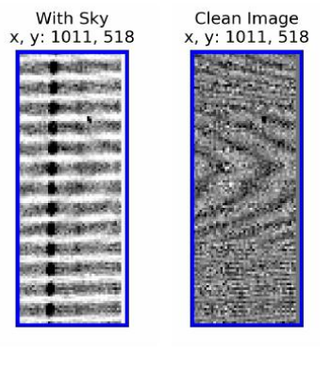


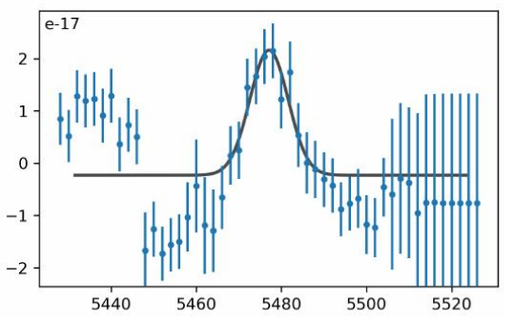
BA example

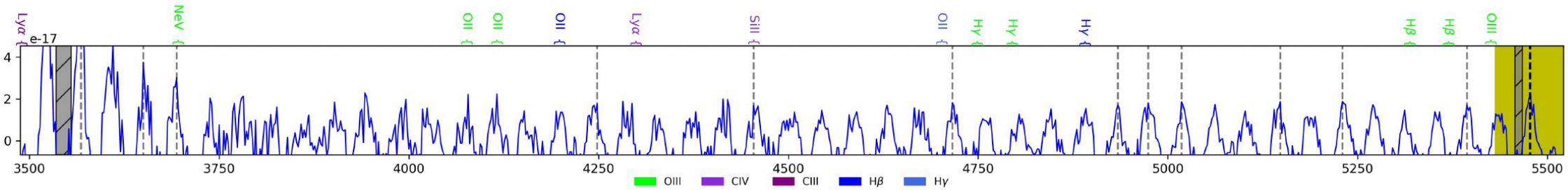
2019-03-28 23:51:22 Version 1.6.3

20190104v026_181
Obs: 20190104v026_1000558810
Entry# (1000558810), Detect ID (66)
Primary IFU SpecID (317) SlotID (024)
RA,Dec (215.628357,51.860802)
λ = 5477.09Å FWHM = 10.5985Å
EstFlux = 1.40(±0.49)e-16
EstCont = -2.30(±1.00)e-18
EW_r(LyA) = -13.00(±6.10)Å
S/N = 7.68 χ^2 = 1.99
P(LAE)/P(OII) = 0.0619
LyA z = 3.5042 OII z = 0.4696

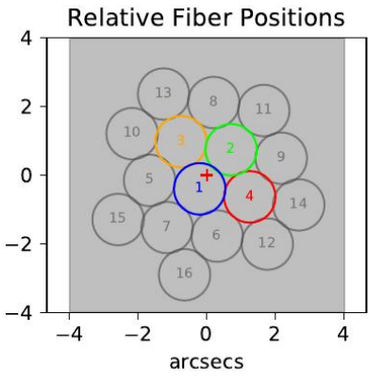






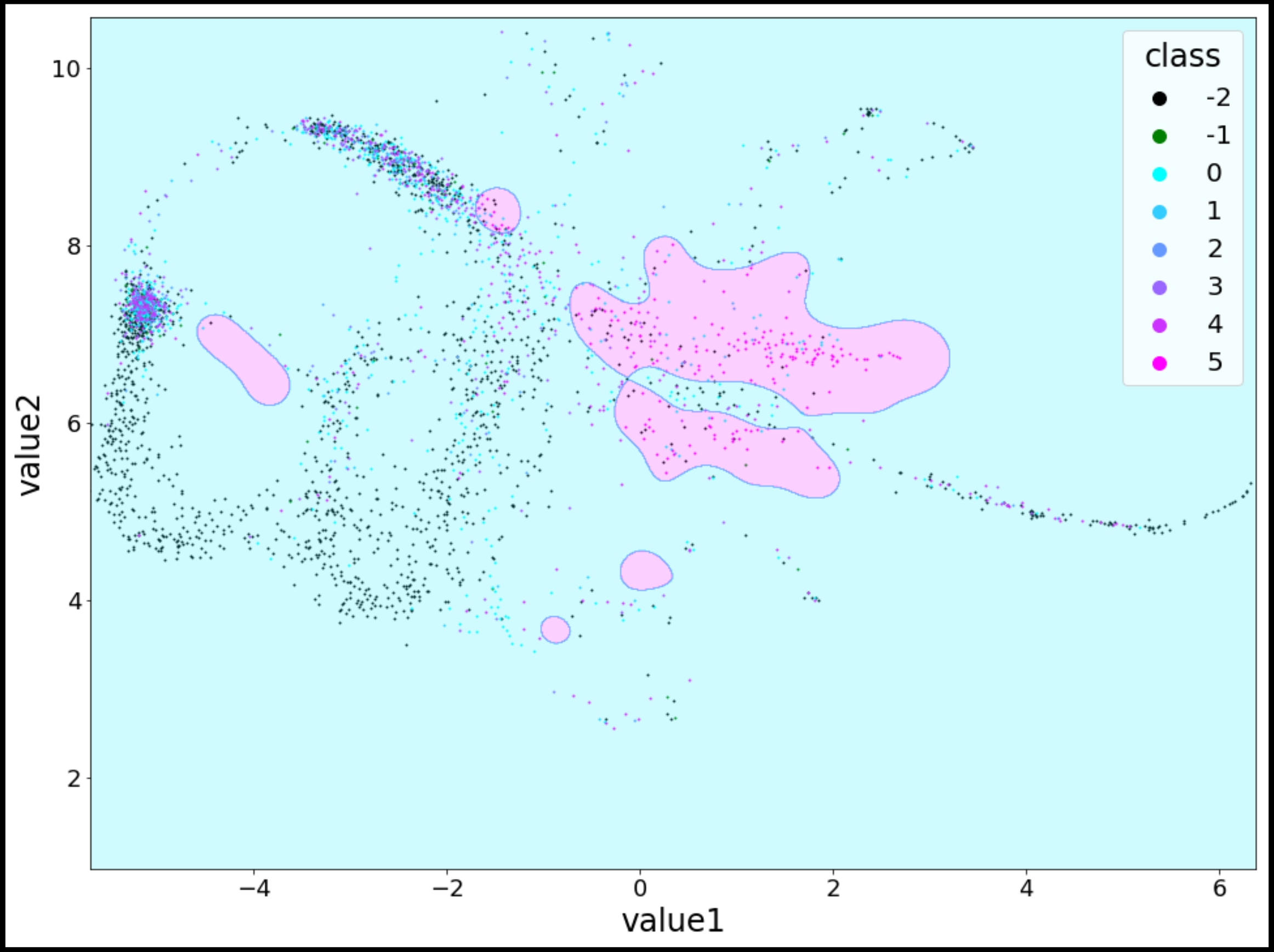


No overlapping imaging catalog.

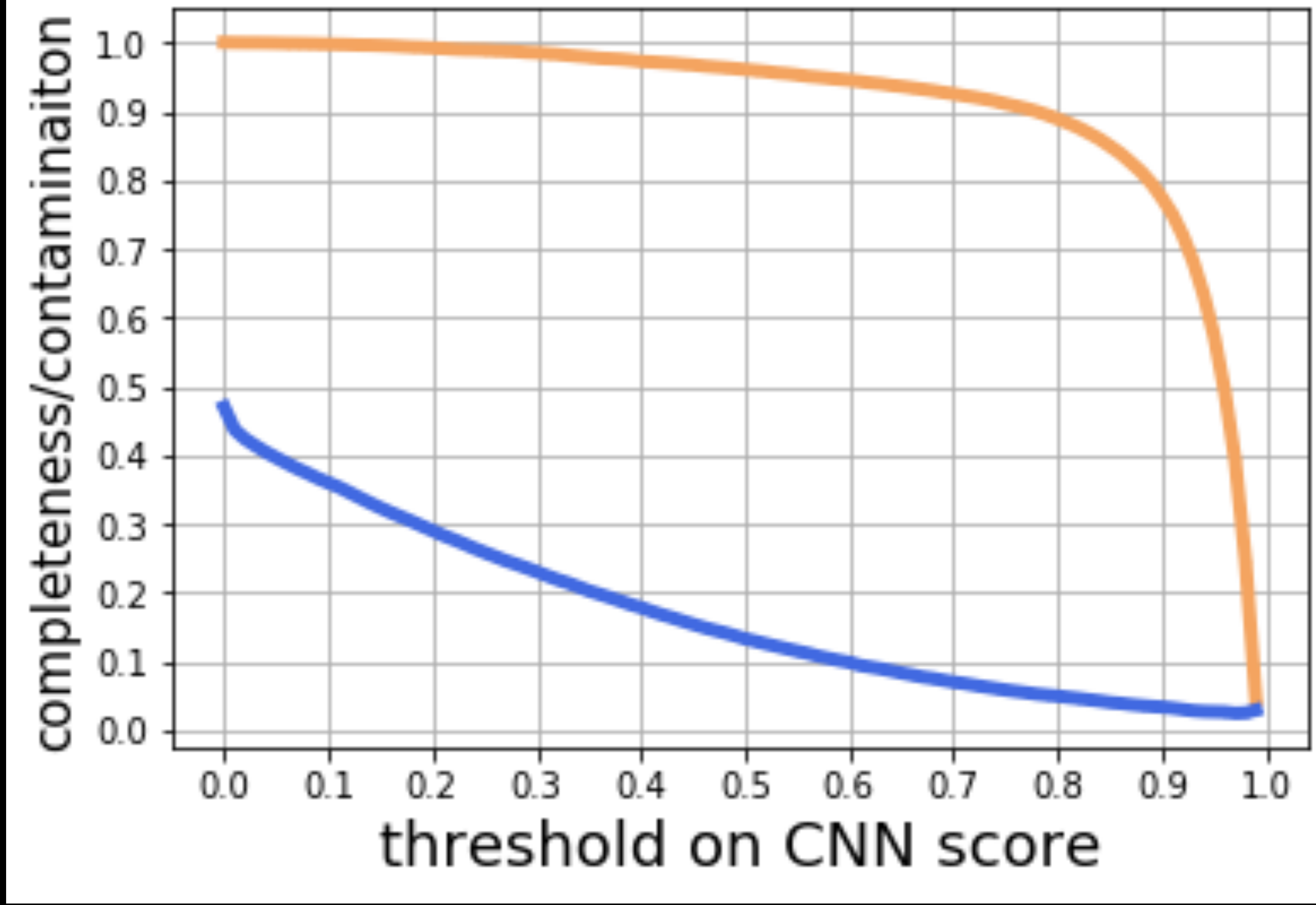


No overlapping imagaing catalog.
Row intentionally blank.

Test data



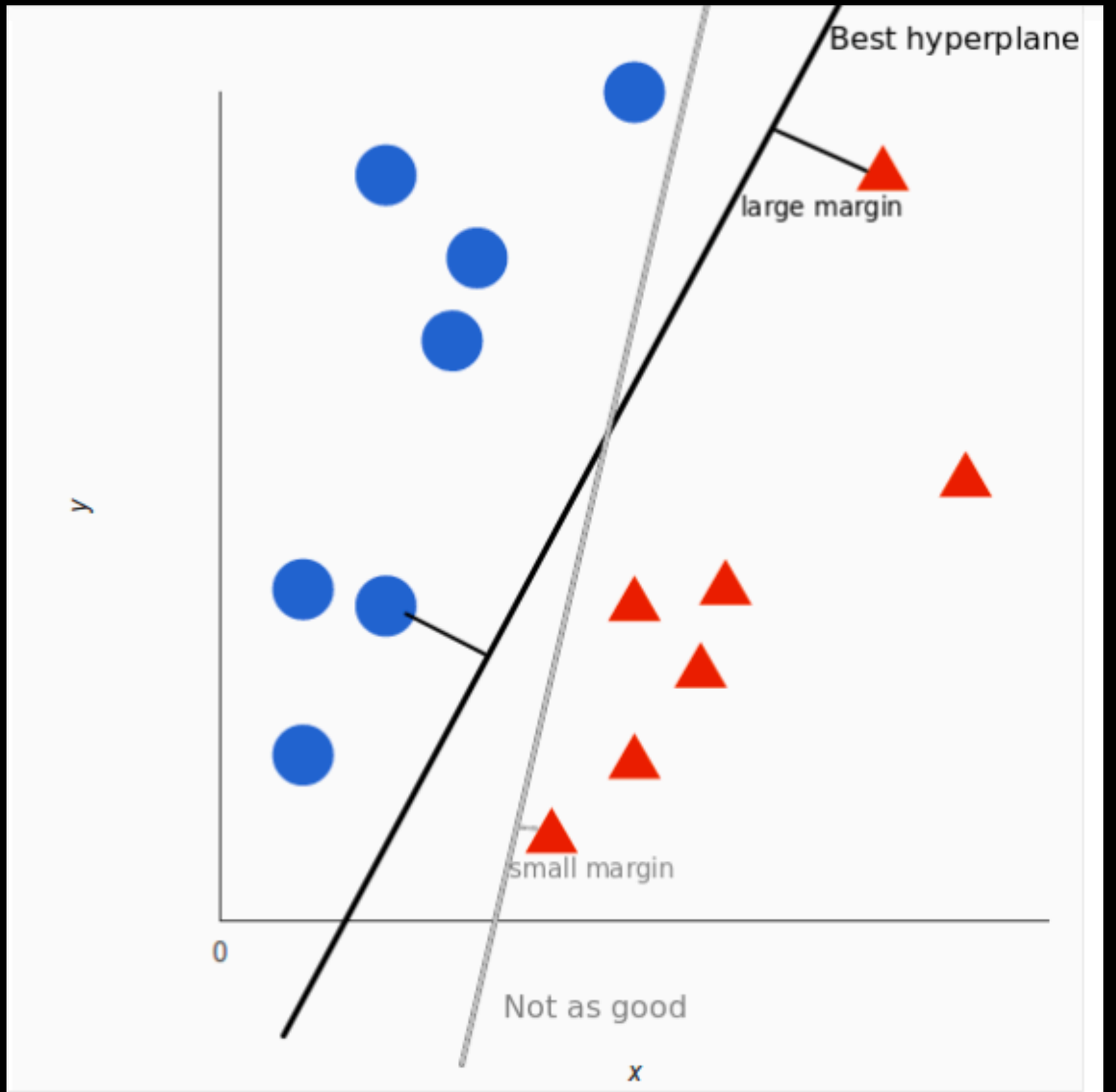
CNN score v. Comp. & contami.



Experimental Requirements

Focus on these, as the most challenging

1) Survey volume of 9 Gpc^3
2) 7 deg minimum on a side
3) 1.2 million LAEs, 2% of LAEs mis-classified as [OII], 10% false positive
4) Quantifying non-linear and radiative transfer effects down to 5 Mpc/h scales
5) LAE number counts dominated by counting statistics, imply systematics 5% over all other calibrations
6) Redshift accuracy less than 180 km/s
7) Flux limit 5e-17 above 4300AA, up to 1e-16 at blue end
8) Survey area of 440 sq deg
9) Fill factor at 1/4.6
10) Dither pattern accuracy of 0.05”
11) Average flux calibration to 5%, with average blue-to-red to 10%
12) Wavelength calibration to 30 km/s
13) Astrometric accuracy better than 0.5” for any source
14) 78 IFUs, with 34944 fibers
15) Imaging survey to 25.1 at 10-sigma for a point source
16) No more than 20% units inactive
17) Overhead no more than 5 minutes per dither set



Not all hyperplanes are created equal

in an article by Ashwani Bhardwaj

https://medium.com/@ashwanibhardwajcodevita16/from-zero-to-hero-in-depth-support-vector-machine-264931a1e135

High S/N Flags and 2D Spectrum

For science

5
Confidently real

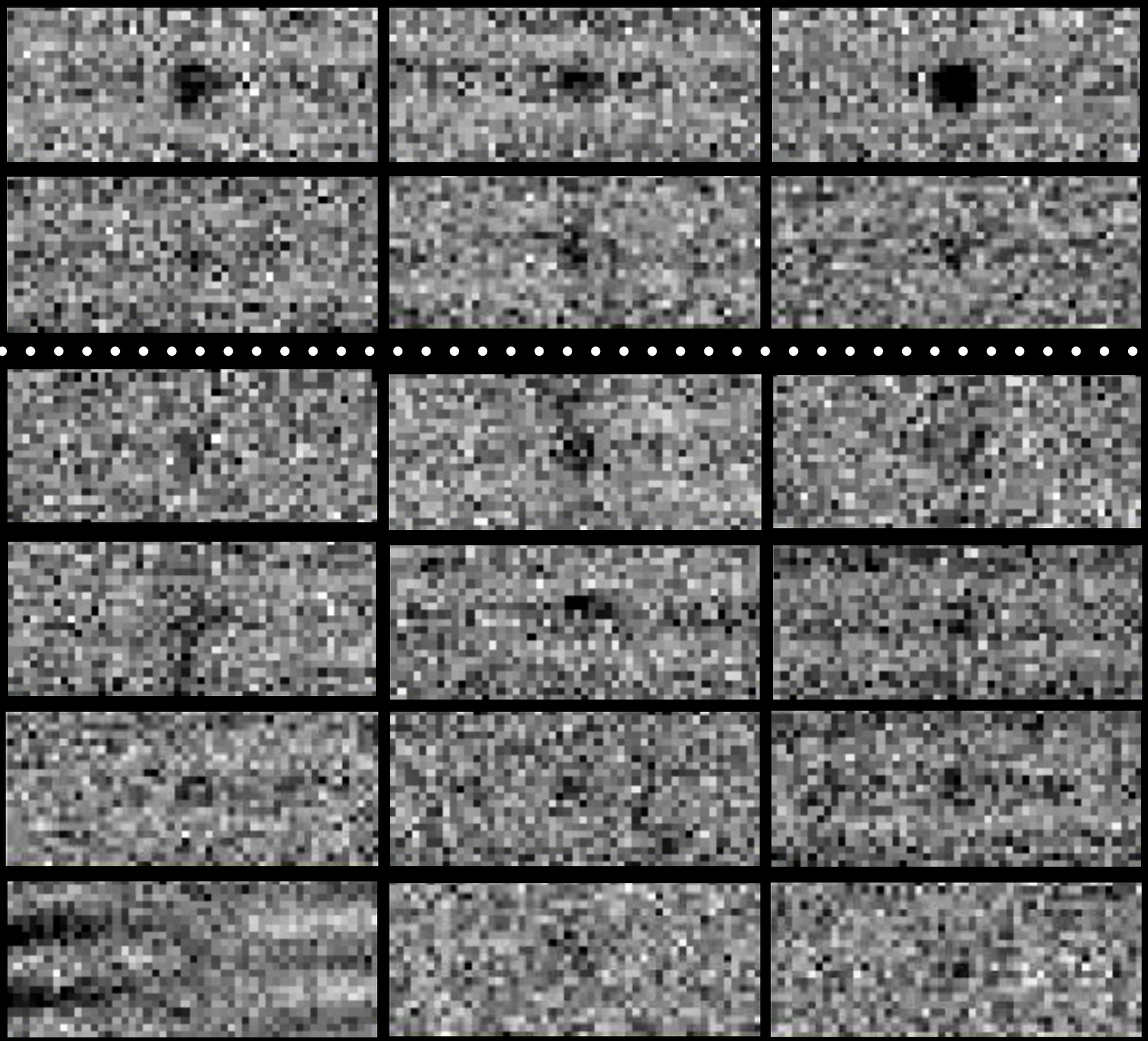
4
Likely real

3
Possibly real

2
Possibly spurious

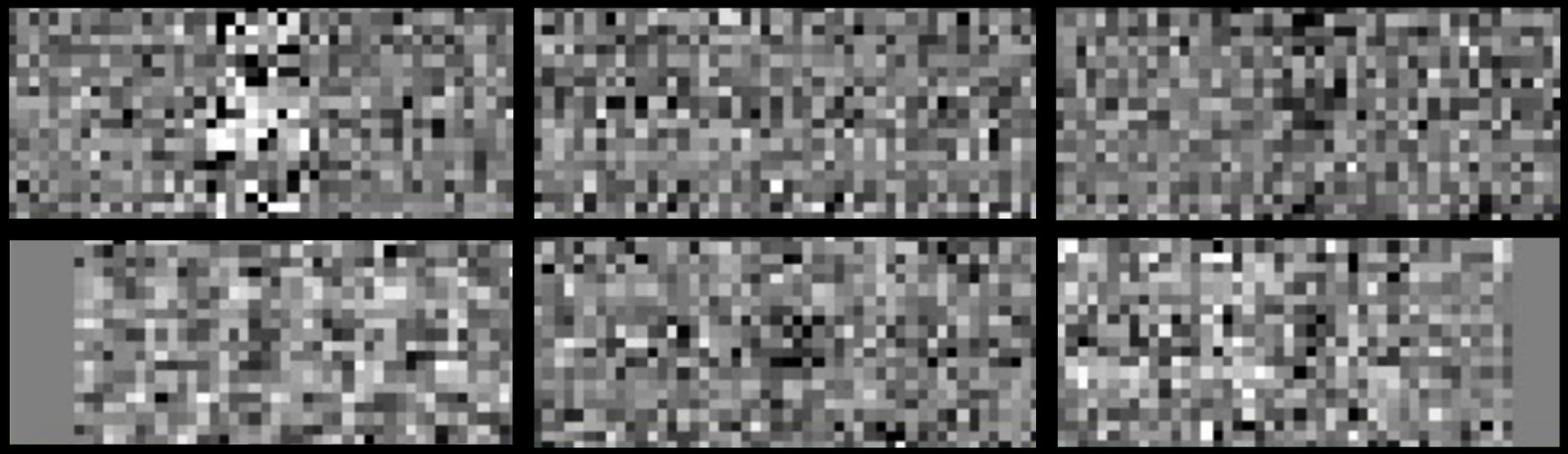
1
Likely spurious

0
Spurious



-1
Bad pixel

-2
Bad amp



LAE/notLAE

Training data balance

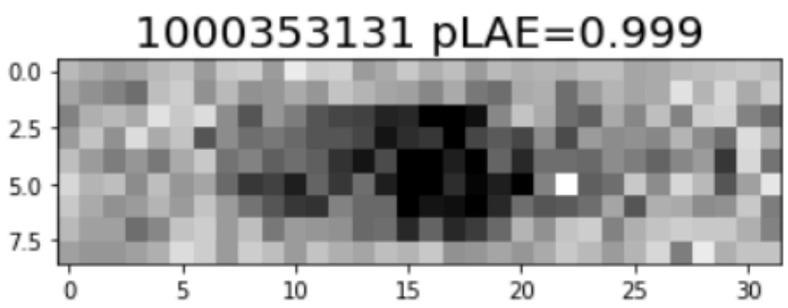
	-1	0	1	2	3	4	5
Sources in the catalogue	145	4085	1080	411	448	645	558
Unique Sources Used in Training	-	4000	1000	-	-	595	505
Augment	-	x1	x4	-	-	x7	x8

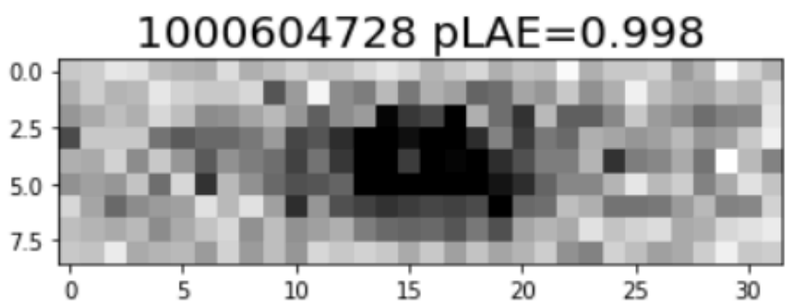
80% for training, 20% for test

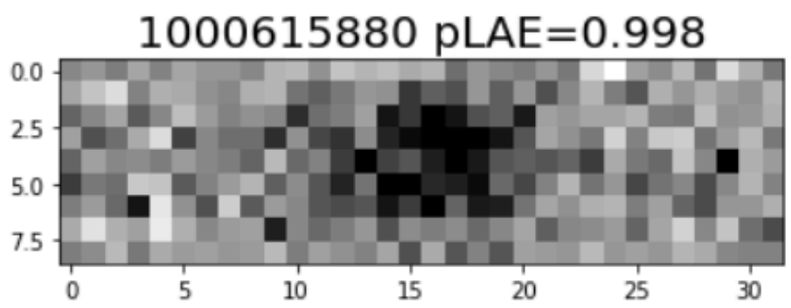
Accuracy

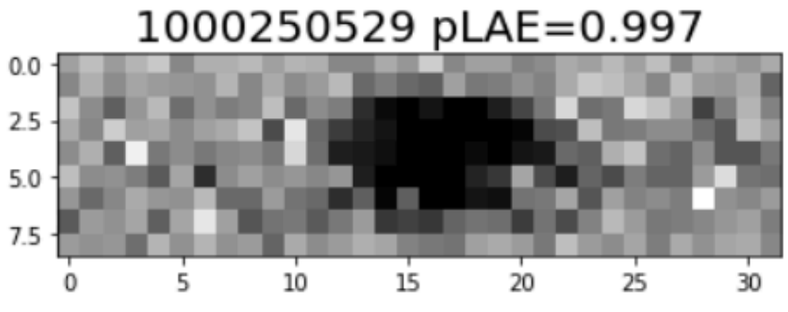
	-1	0	1	2	3	4	5
Accuracy (%)	46.2	89.4	83.5	63.5	40.6	89.9	89.1

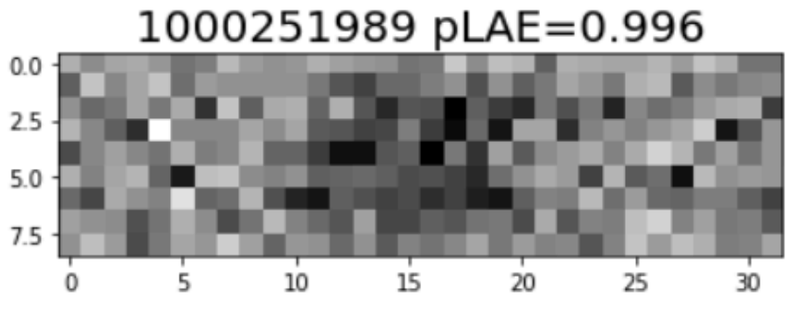
5’s LAE

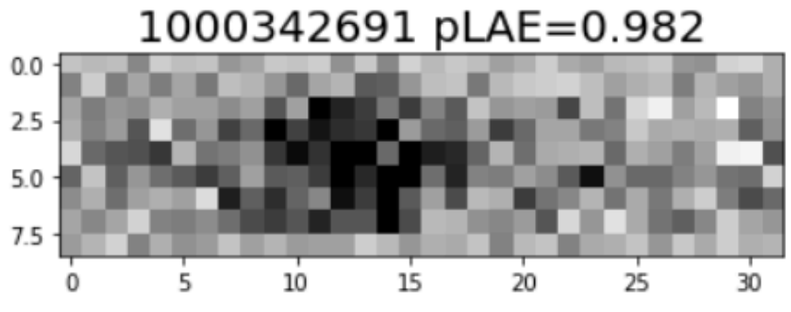

1000353131 pLAE=0.999

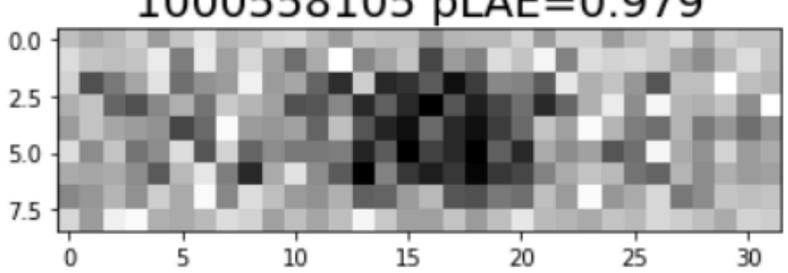

1000604728 pLAE=0.998

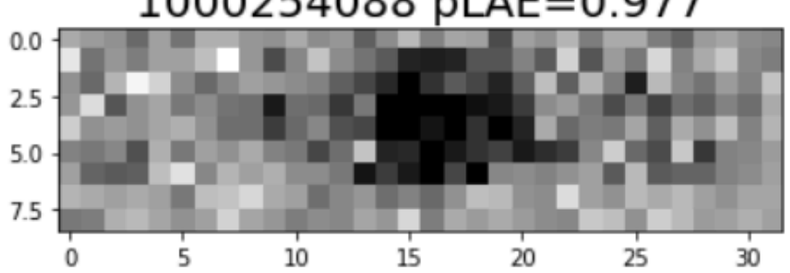

1000615880 pLAE=0.998

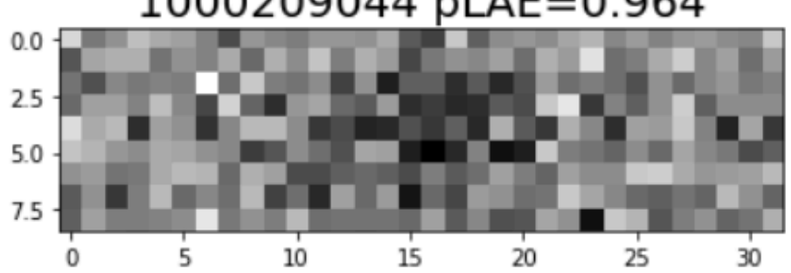

1000250529 pLAE=0.997

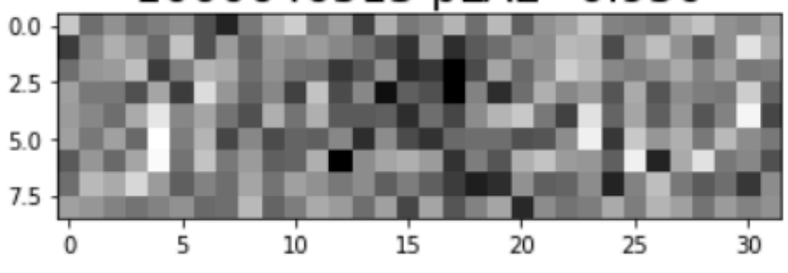

1000251989 pLAE=0.996

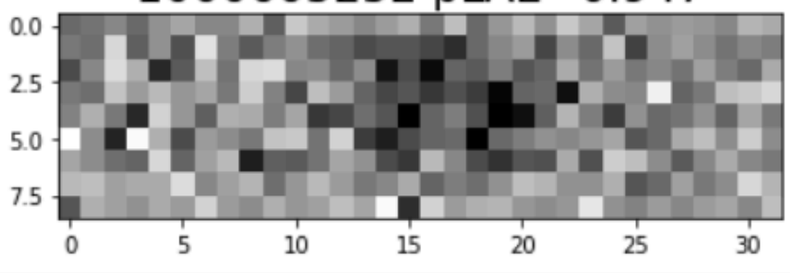

1000342691 pLAE=0.982

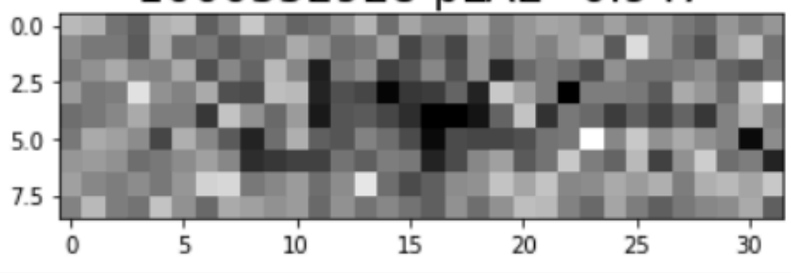

1000558105 pLAE=0.979


1000254088 pLAE=0.977

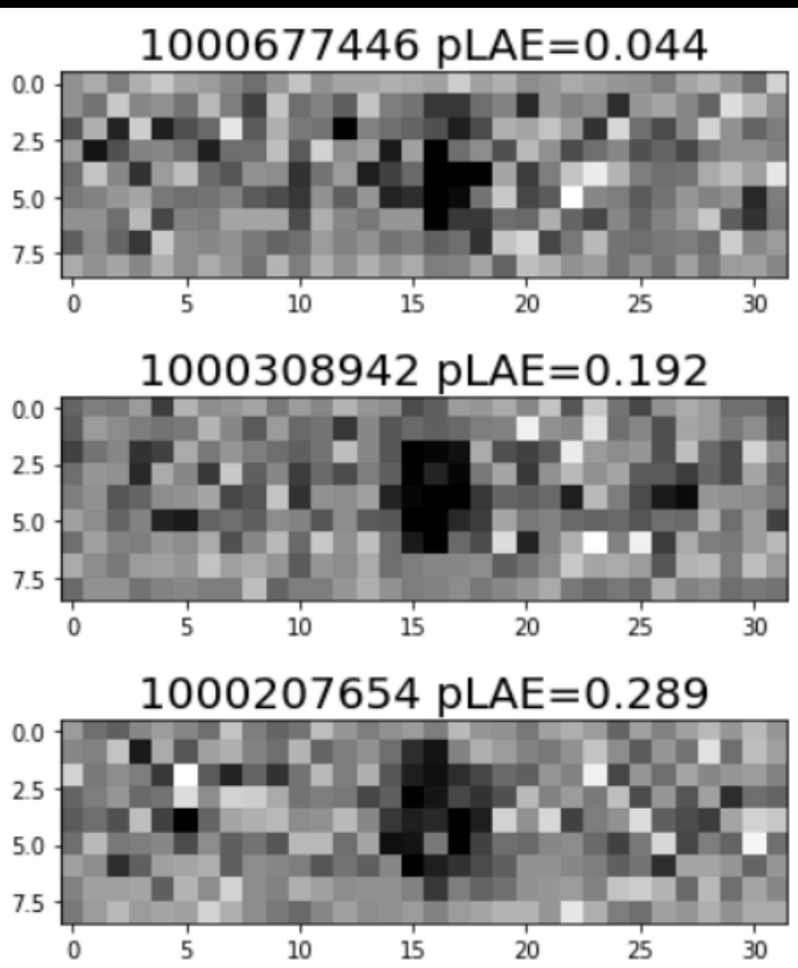

1000209044 pLAE=0.964

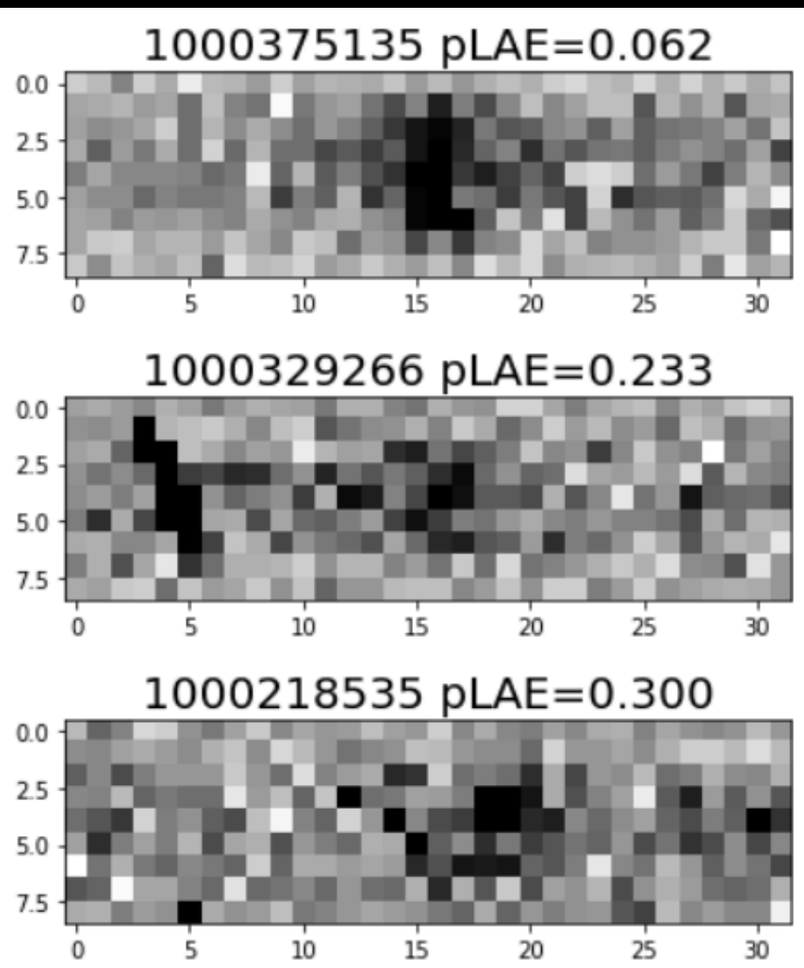

1000646513 pLAE=0.950

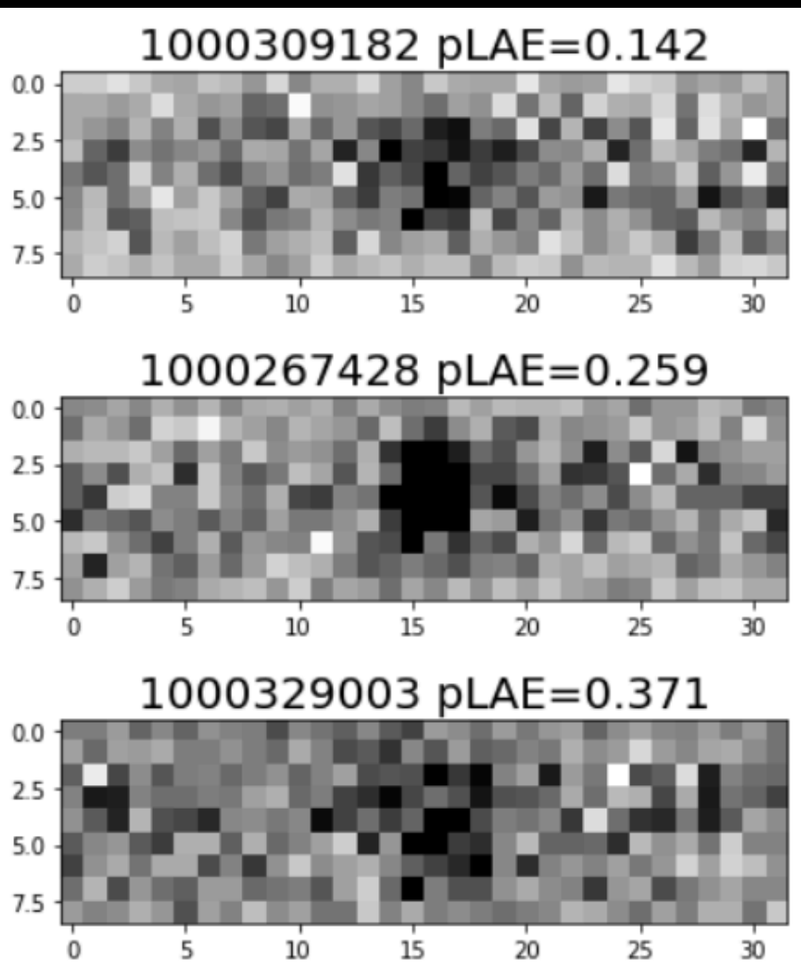

1000605252 pLAE=0.947


1000352928 pLAE=0.947

5’s notLAE


1000677446 pLAE=0.044


1000375135 pLAE=0.062


1000309182 pLAE=0.142

1000308942 pLAE=0.192

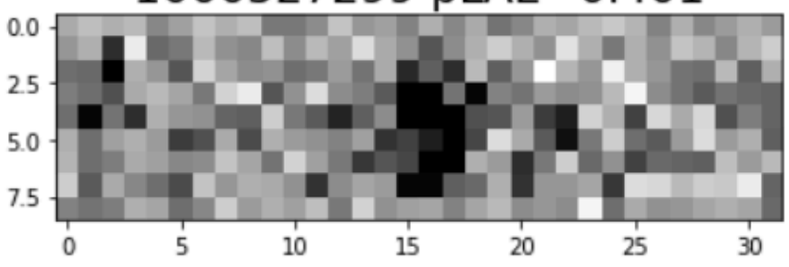
1000329266 pLAE=0.233

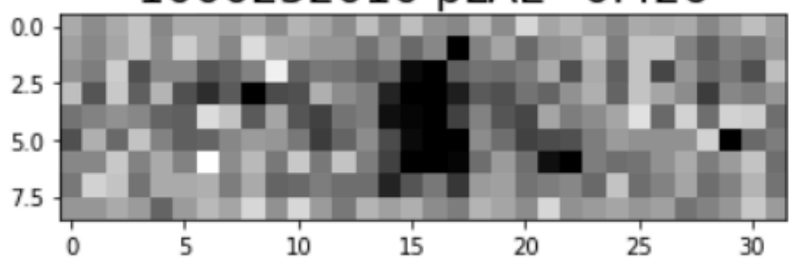
1000267428 pLAE=0.259

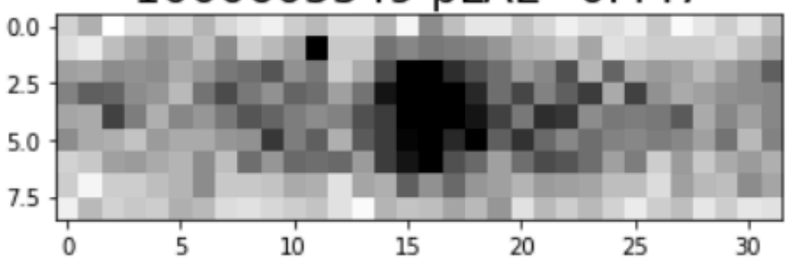
1000207654 pLAE=0.289

1000218535 pLAE=0.300

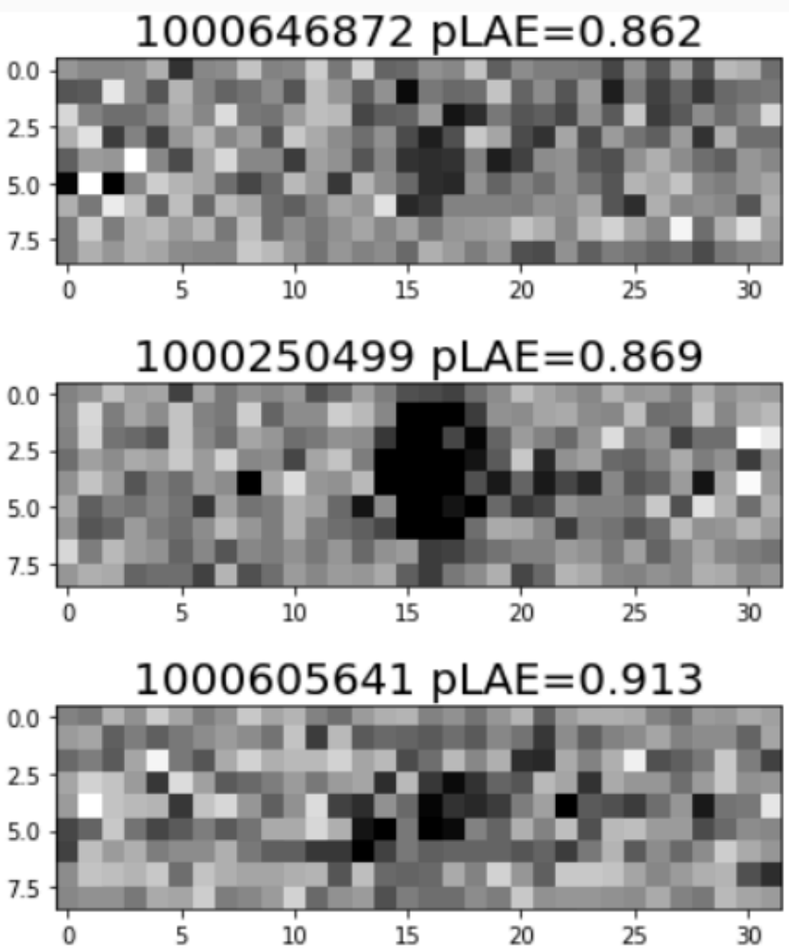
1000329003 pLAE=0.371

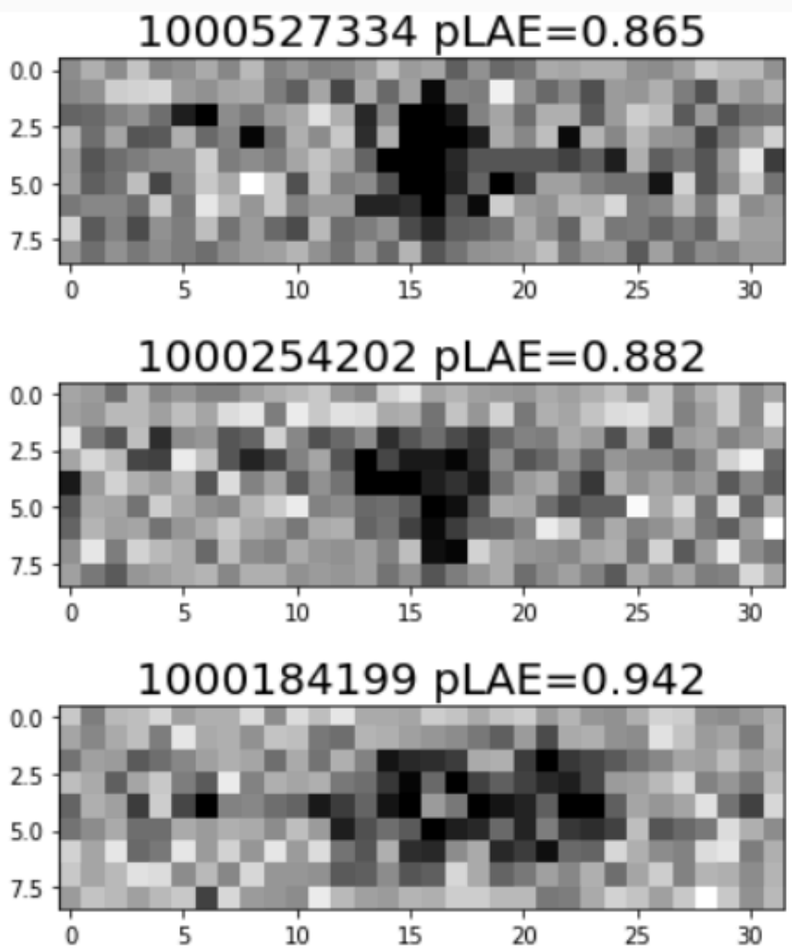

1000527299 pLAE=0.401

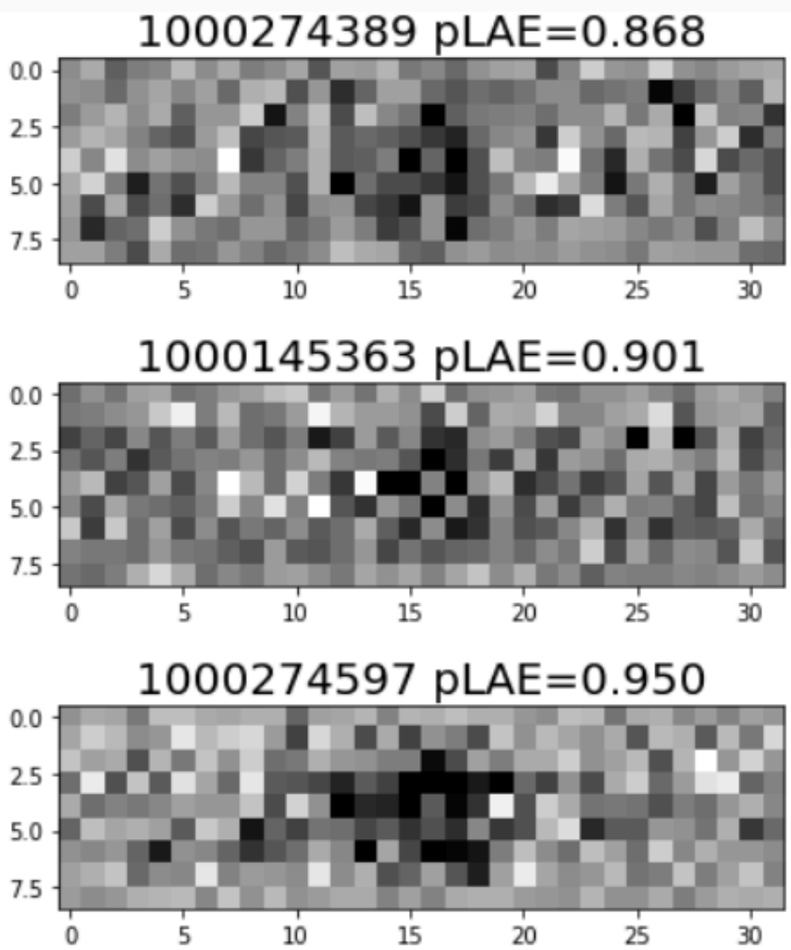

1000252010 pLAE=0.426


1000605549 pLAE=0.447

0’s LAE


1000646872 pLAE=0.862


1000527334 pLAE=0.865


1000274389 pLAE=0.868

1000250499 pLAE=0.869

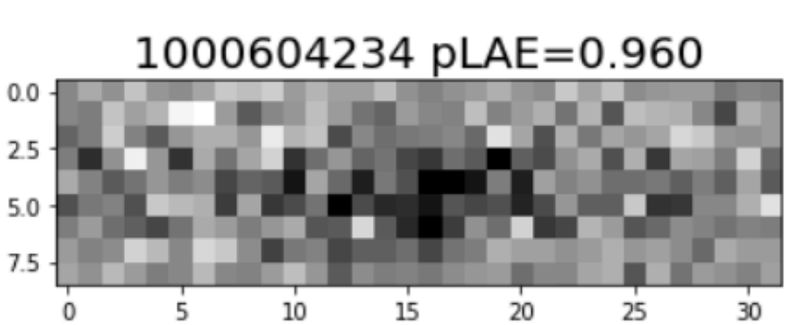
1000254202 pLAE=0.882

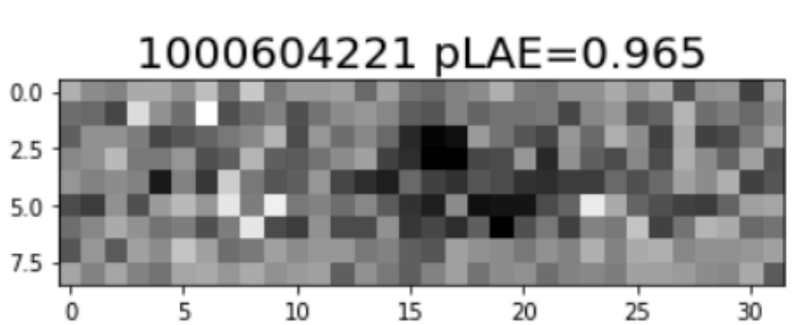
1000145363 pLAE=0.901

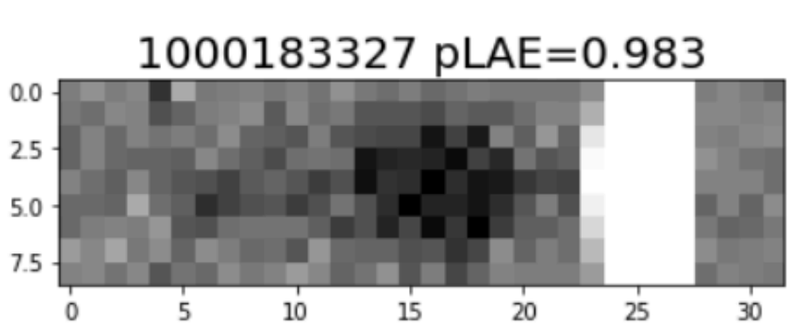
1000605641 pLAE=0.913

1000184199 pLAE=0.942

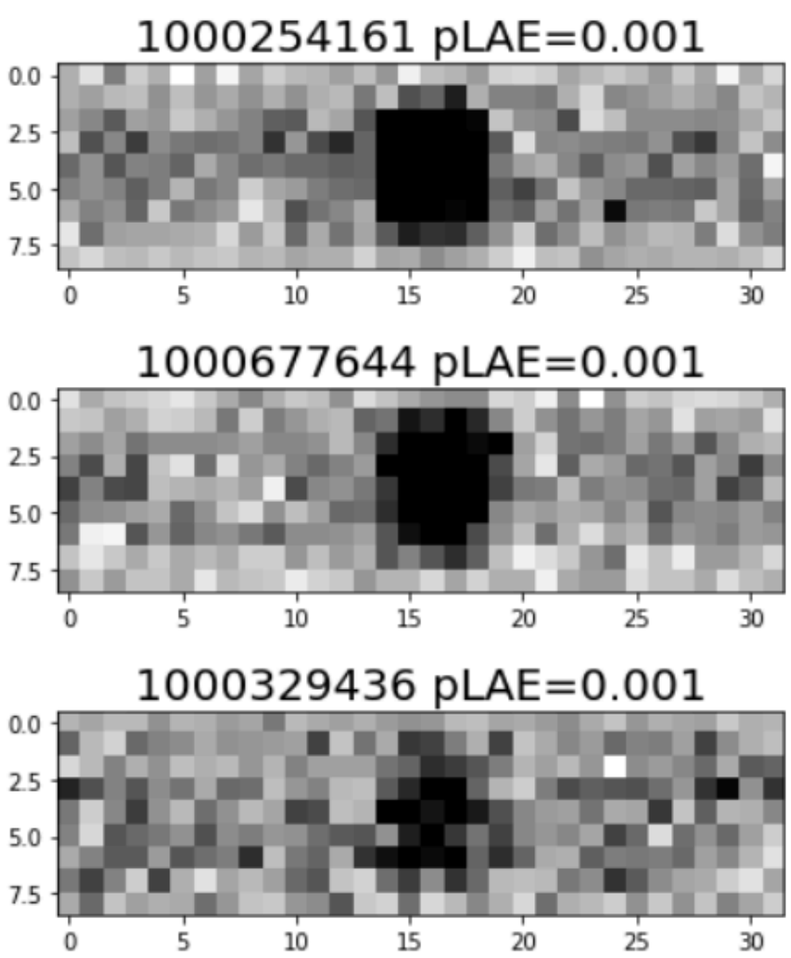
1000274597 pLAE=0.950

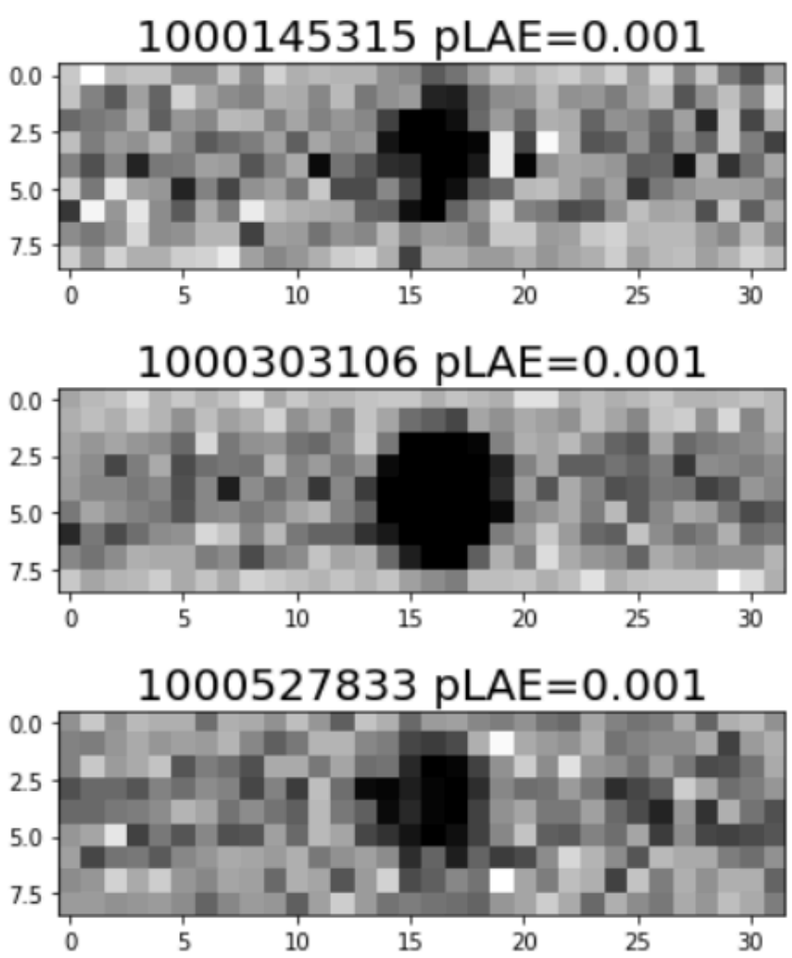

1000604234 pLAE=0.960

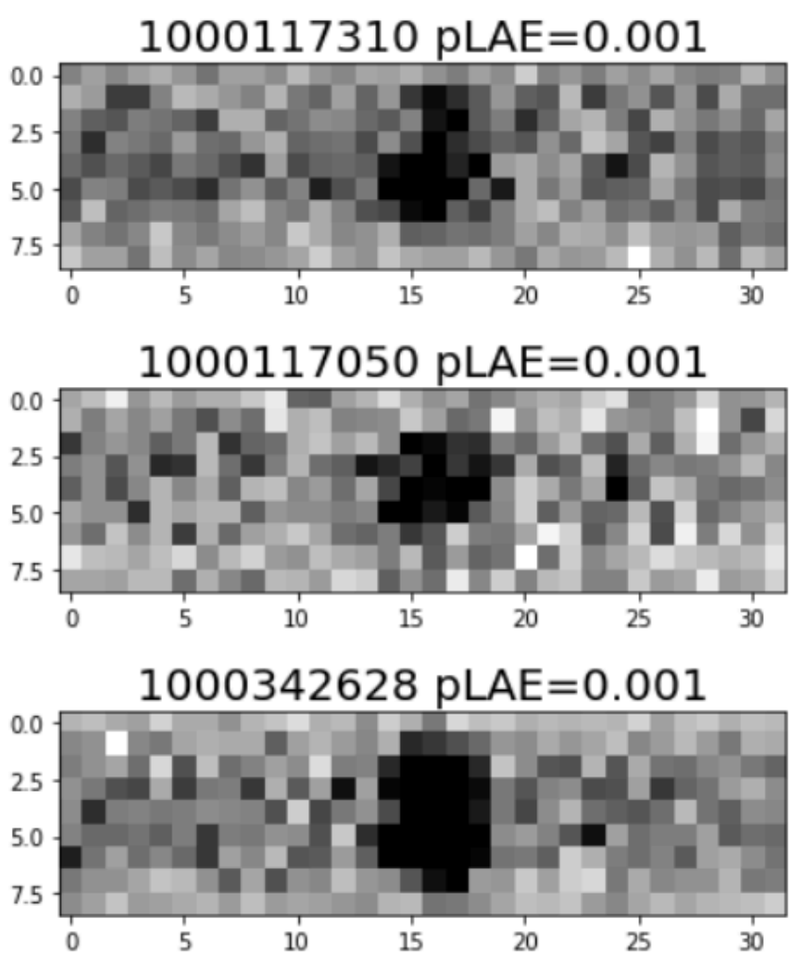

1000604221 pLAE=0.965


1000183327 pLAE=0.983

0’s notLAE


1000254161 pLAE=0.001


1000145315 pLAE=0.001


1000117310 pLAE=0.001

1000677644 pLAE=0.001

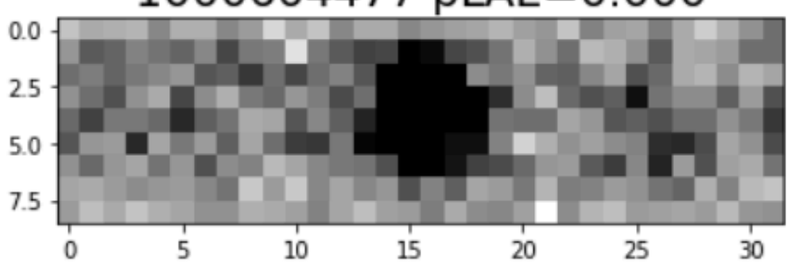
1000303106 pLAE=0.001

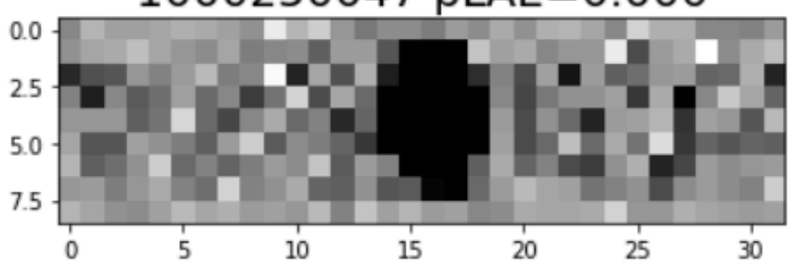
1000117050 pLAE=0.001

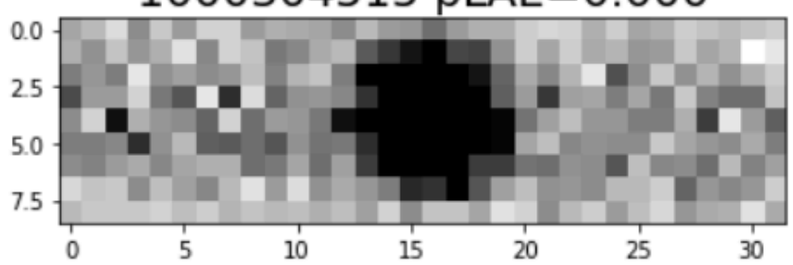
1000329436 pLAE=0.001

1000527833 pLAE=0.001

1000342628 pLAE=0.001


1000604477 pLAE=0.000


1000250047 pLAE=0.000


1000304515 pLAE=0.000