Active Deep Learning In Big Spectra Archives

Petr Škoda
Astronomical Institute of the Czech Academy of Sciences Ondřejov
Faculty of Information Technology of the Czech Technical University in Prague

Ondřej Podsztavek
Faculty of Information Technology of the Czech Technical University in Prague

Supported by OP VVV, Research Center for Informatics,
CZ.02.1.01/0.0/0.0/16_019/0000765

COST TD1403 BIGSKYEARTH Conference
AstroGeoInformatics
Knowledge Discovery in Big Data from Astronomy and Earth Observation
IAC Tenerife, Spain, 17h December 2018

Thanks to

- Astronomical institute CAS- 2m
- LAMOST people (Dongwei Fan, Yue Wu)
- China-VO (Chenzhou Cui)

Be Stars - Introduction

Gamma Cas (Padre Angello Sechhi 1866)

Vatican obs predecessor – visual spectrograph

Some have or have had emission in Balmer lines

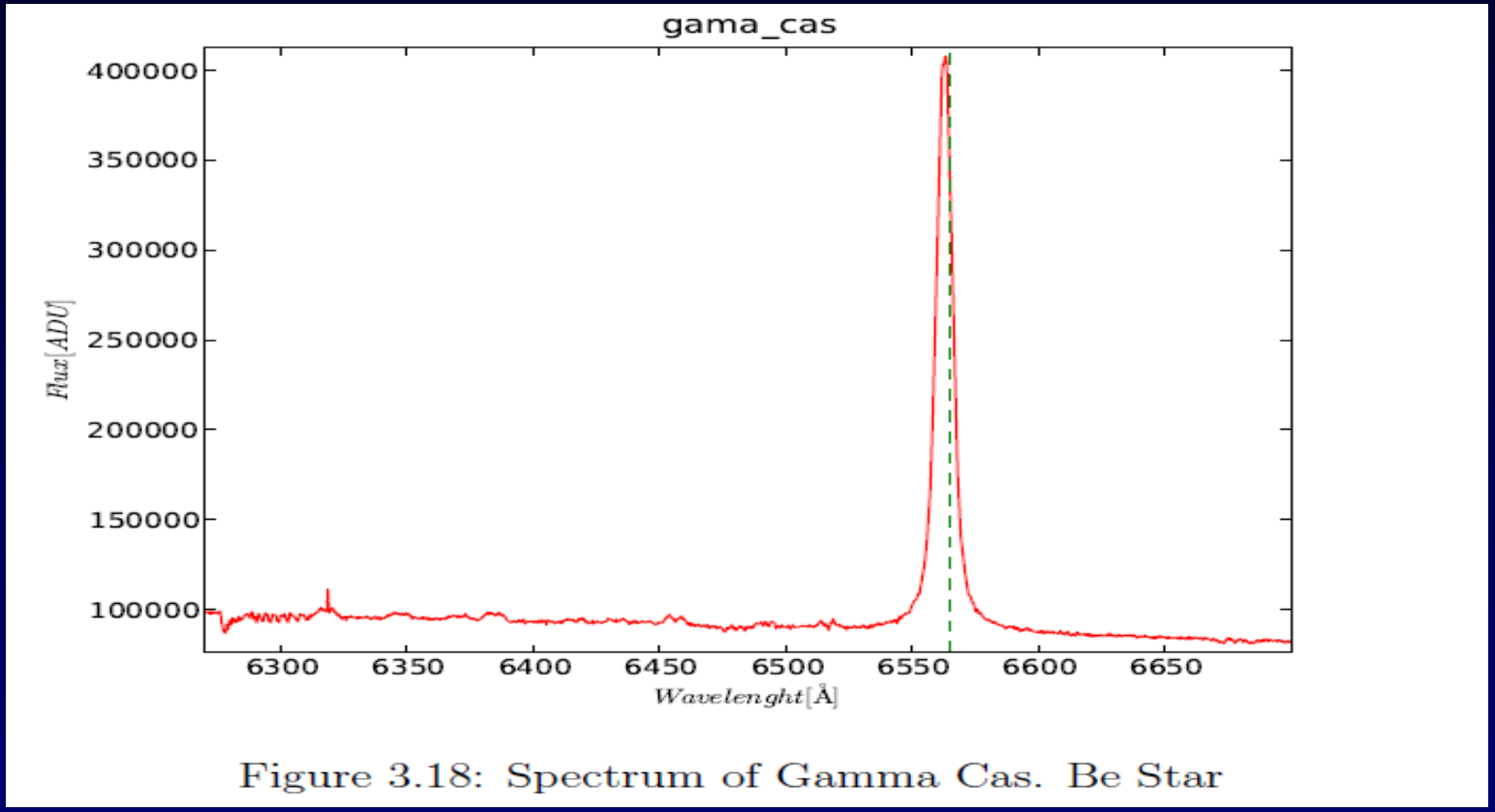
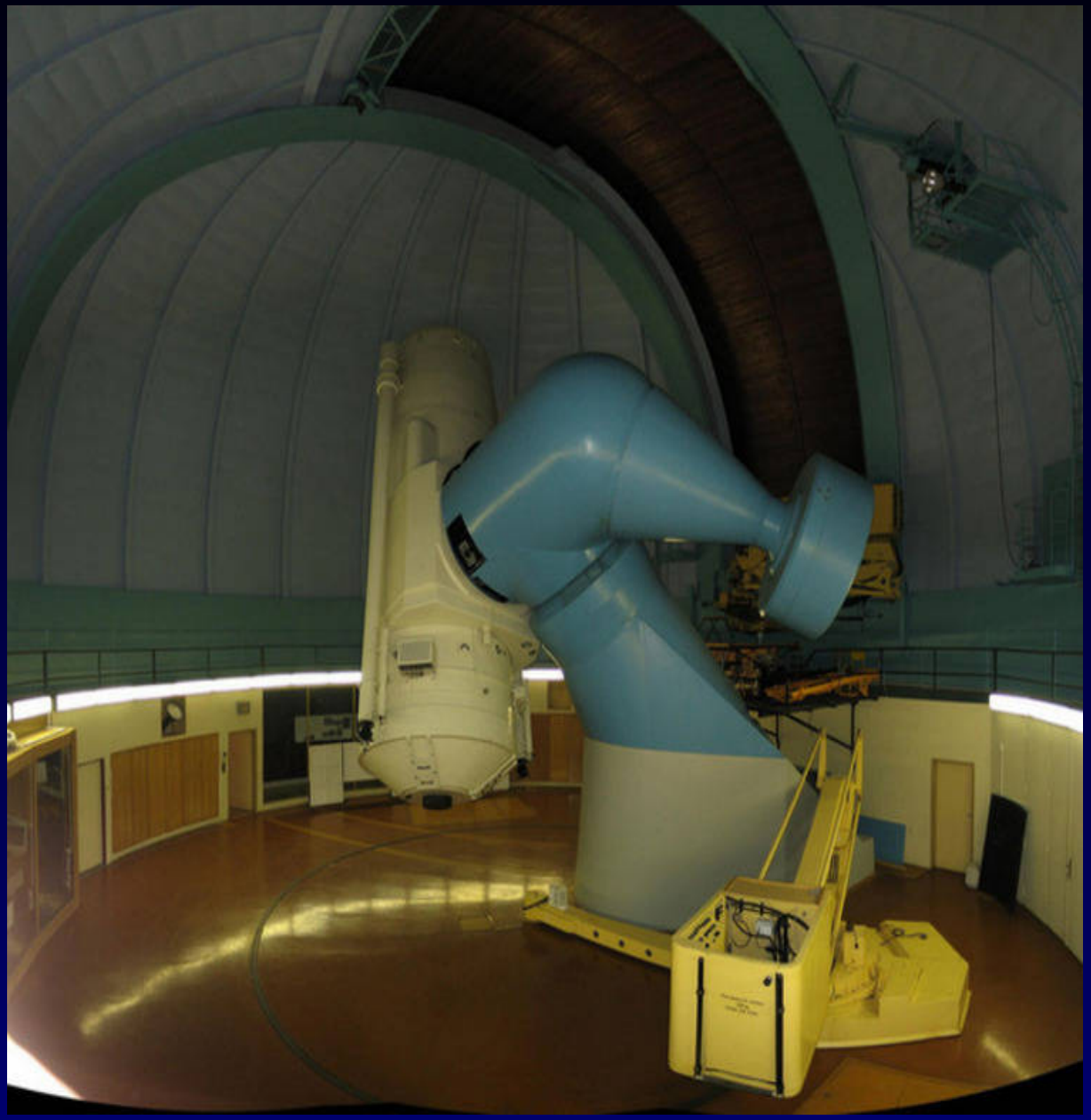


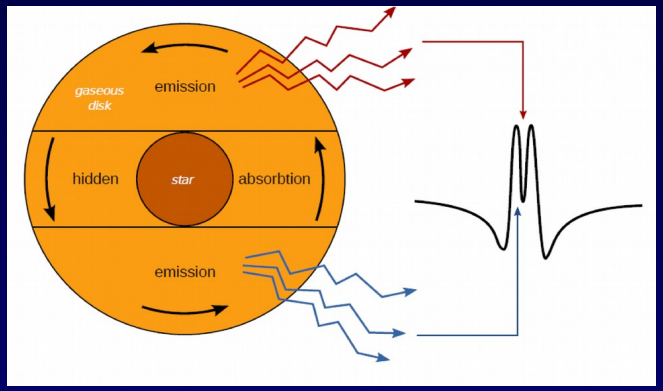
Figure 3.18: Spectrum of Gamma Cas. Be Star

Ondřejov 2m Perek Telescope (1967)

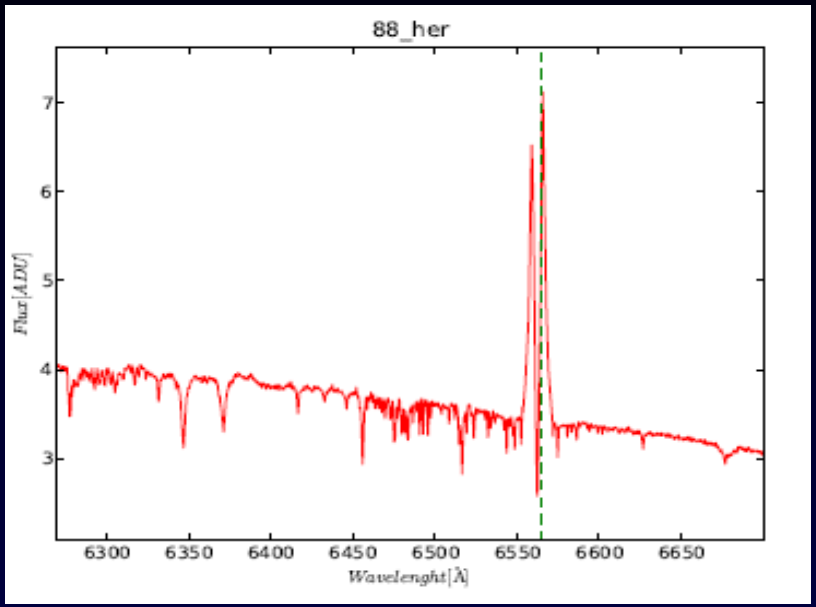


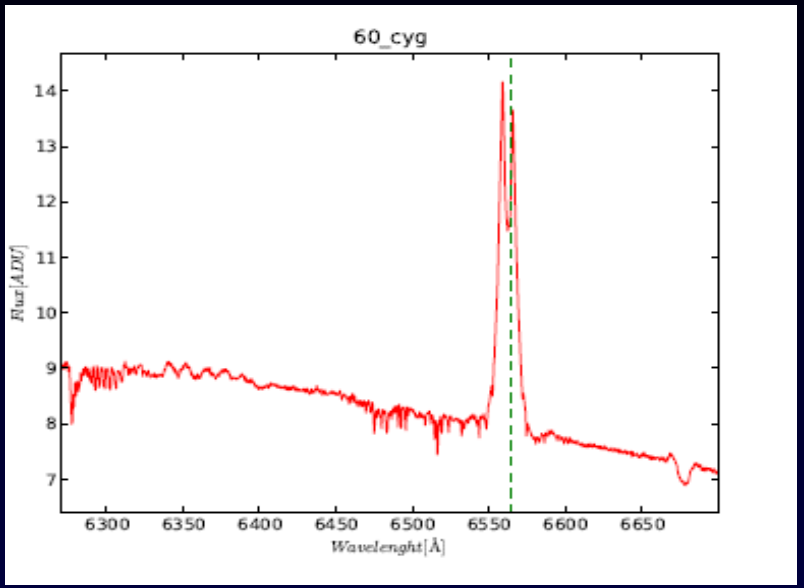
50 years tradition in Be stars
Archive over 20 000 spectra
13000 in Halpha region

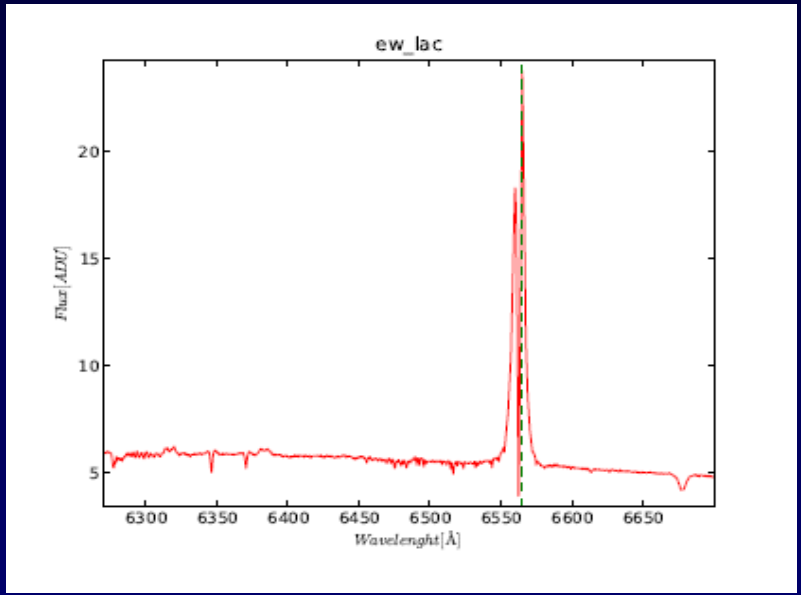
Be mysterious phenomenon
changes line profile
epizodes of emision
fast rotate
Hot (early B types)
disk or envelope

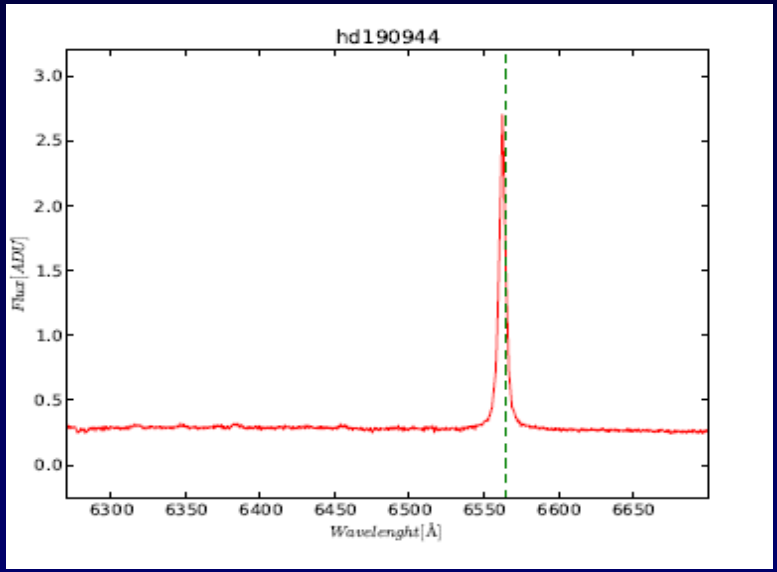


Be Stars : Shell lines vs. no absorption

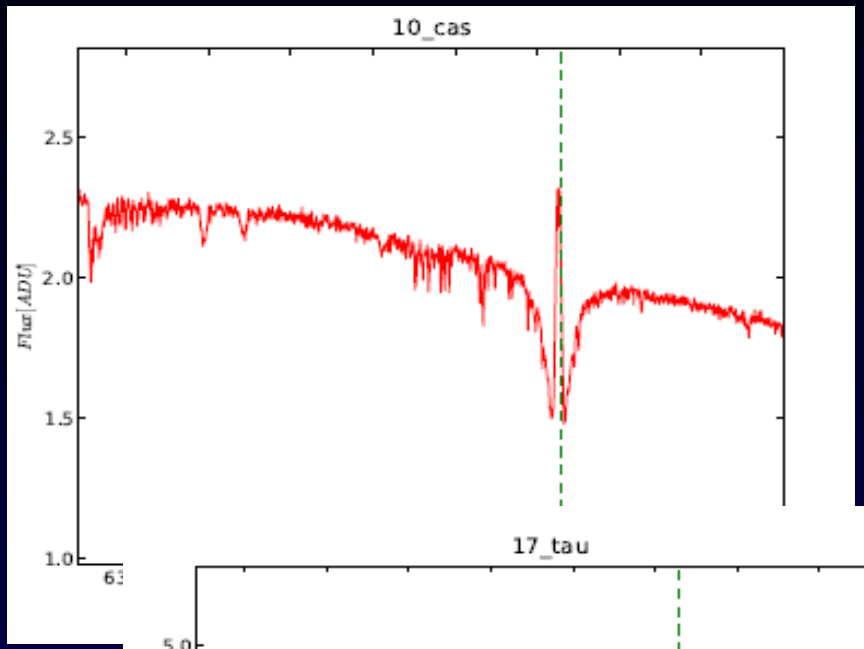


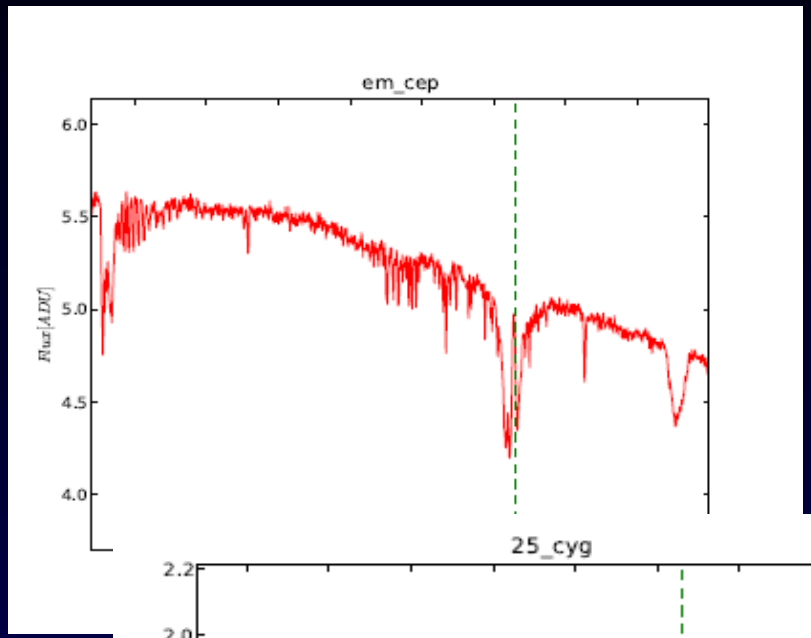


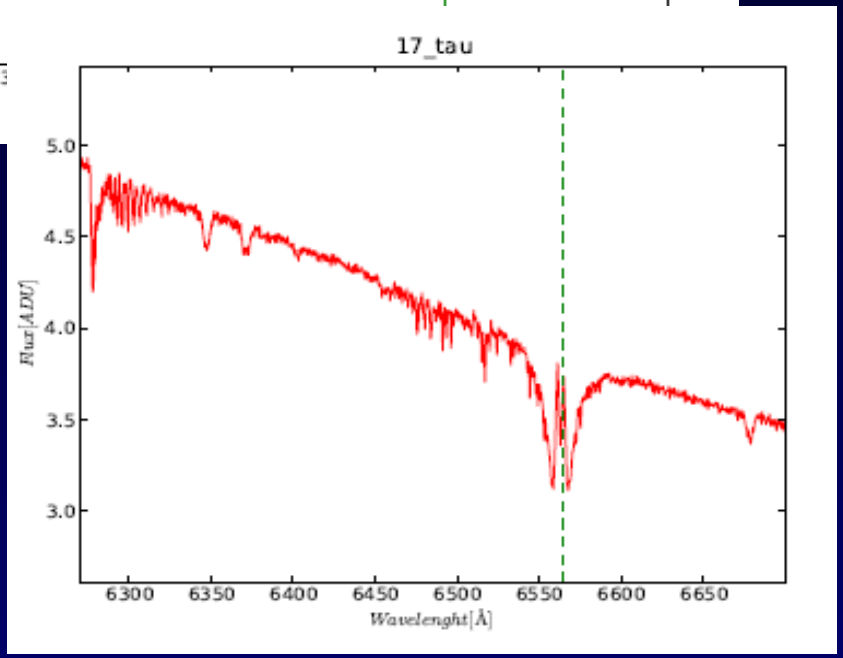


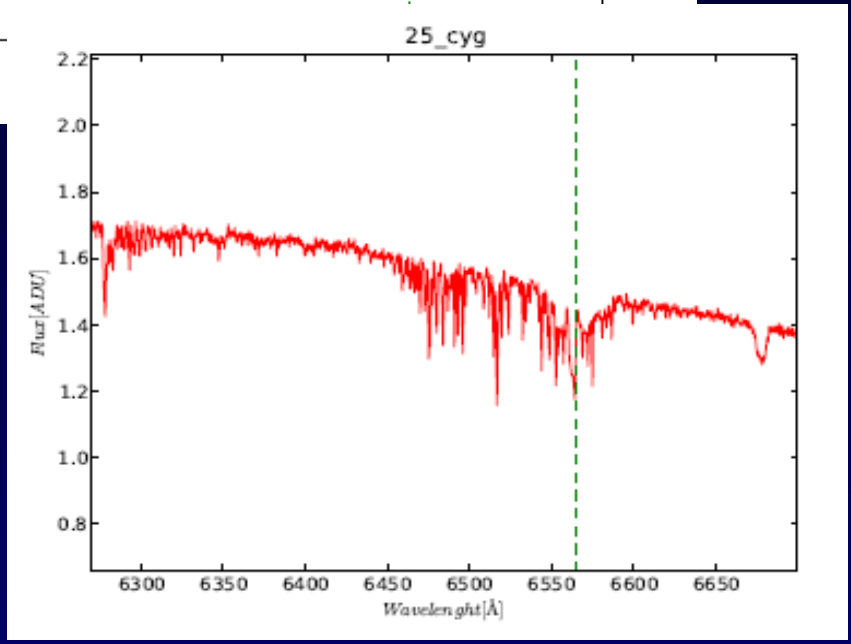


Be Stars : Emission in absorption



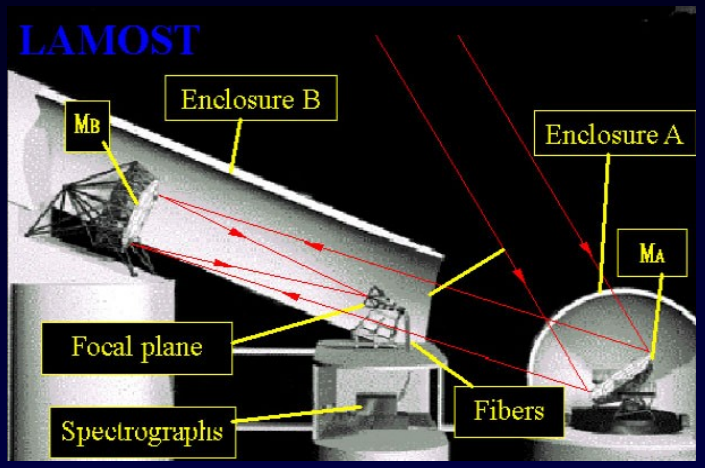


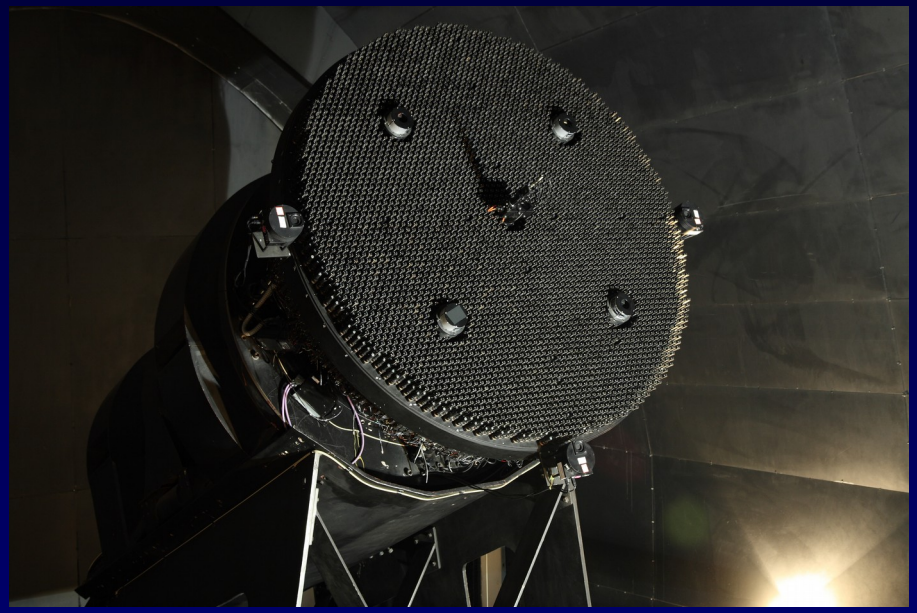




LAMOST (Guoshoujing)

Xinglong- China
4m mirror (30 deg meridian)
4000 fibers





LAMOST Spectral Surveys

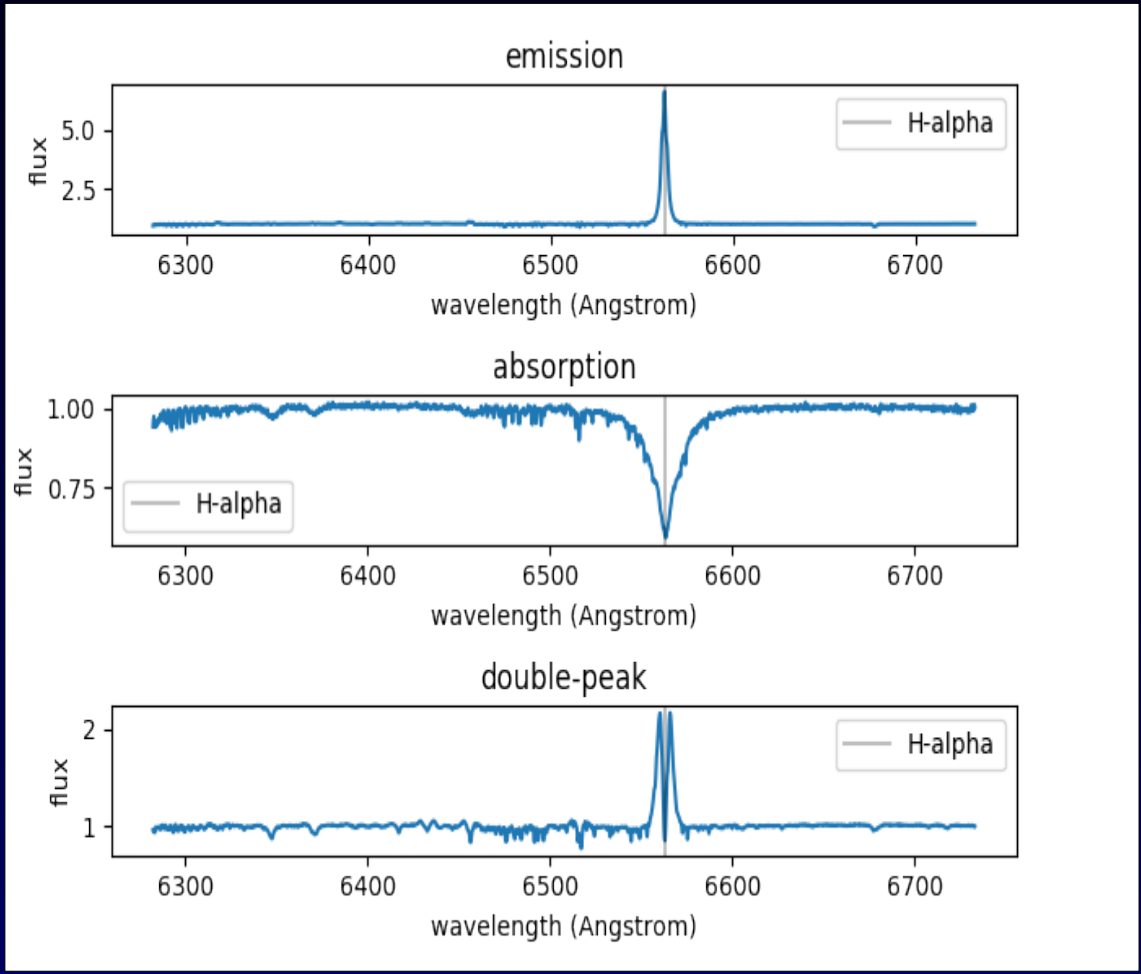
DR1 (end 2013) **2 204 860** spectra
1 085 404 stars classified by pipeline

DR5 (half 2017) **9 017 844** spectra
DR6 (half 2018) +739 006

Each Fiber – 2 motors
double arm 33mm circle

Fibre collects light from
3.3 arcsec circle on sky

Ondřejov Data Classification



13335 spectra from CCD700

Our TARGET class only Be stars =emission or double peak (2+1)

Attempts for more classes failed (subjective – Hanuschik 1996)

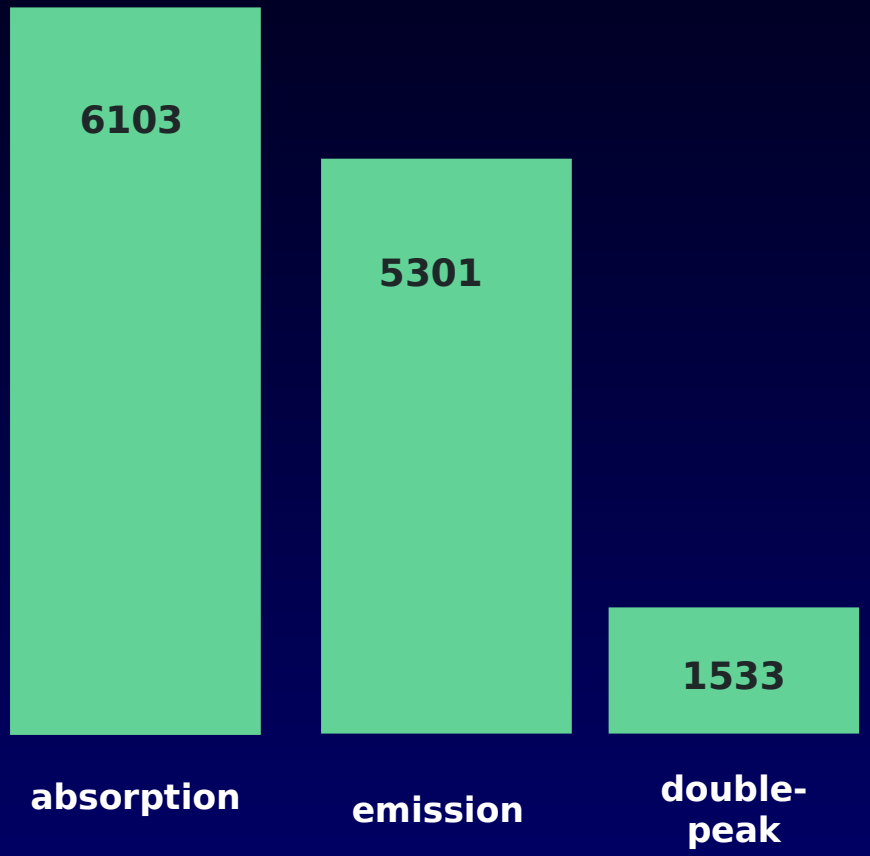
Still not enough labels for DL !

Balancing Classes in Training Set

Synthetic Minority Oversampling Technique

SMOTE

Bowyer et al. 2011



Common Spectra in Both Surveys

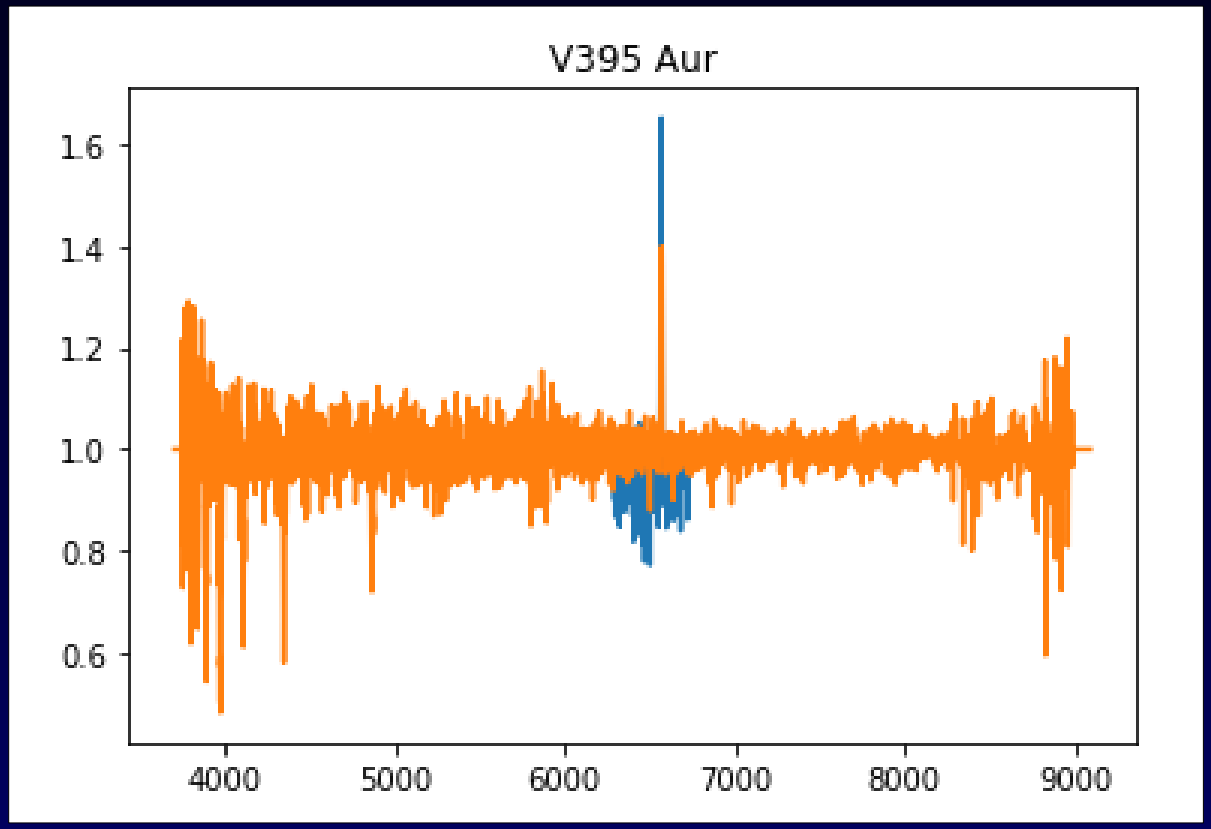
Needed cross-matching in VO Topcat, SPLAT-VO)

LAMOST (2 mil) and Ondrejov (12000) on SSAP server (DaCHS) + DATALINK

ONLY 22 common (but different time of observation = changes in profile)

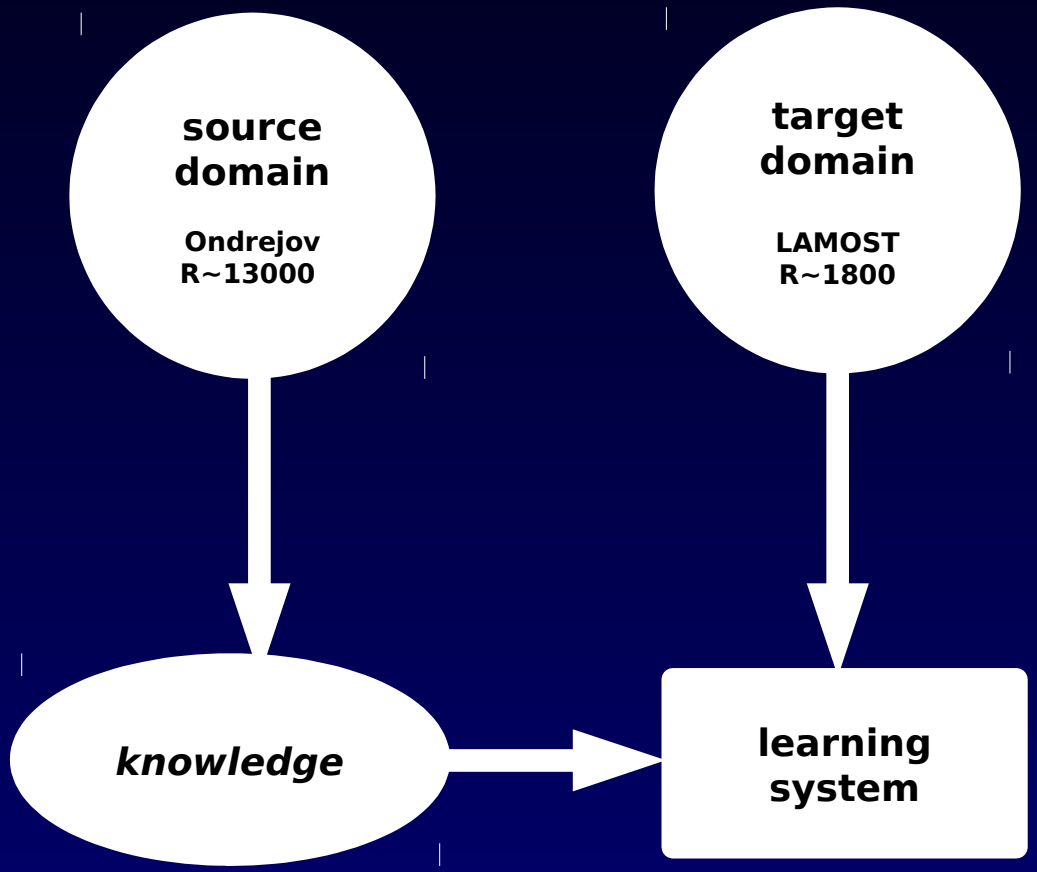
ONLY 4 emission line

No way to proper training set!

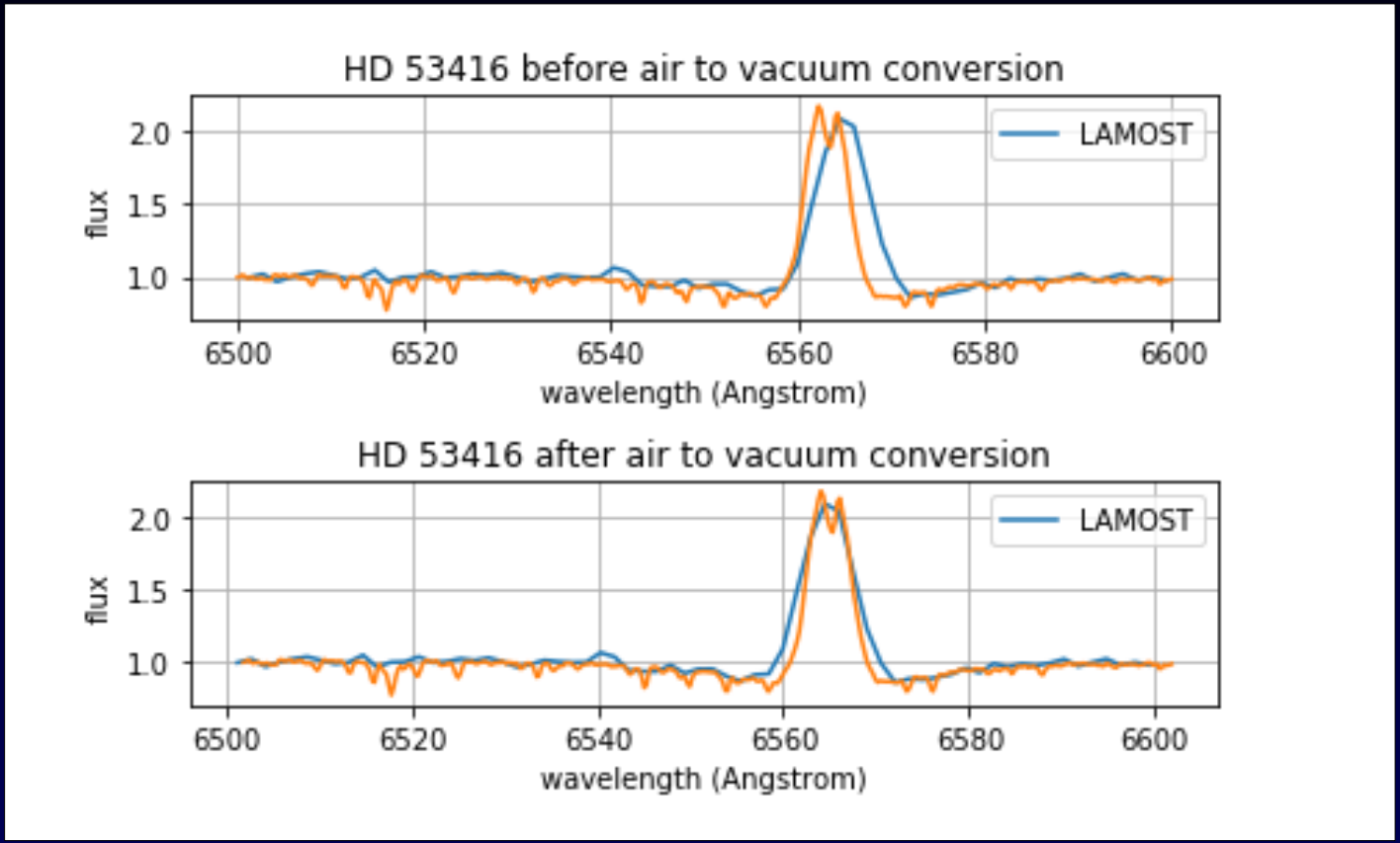


Domain Adaptation

Transfer Learning

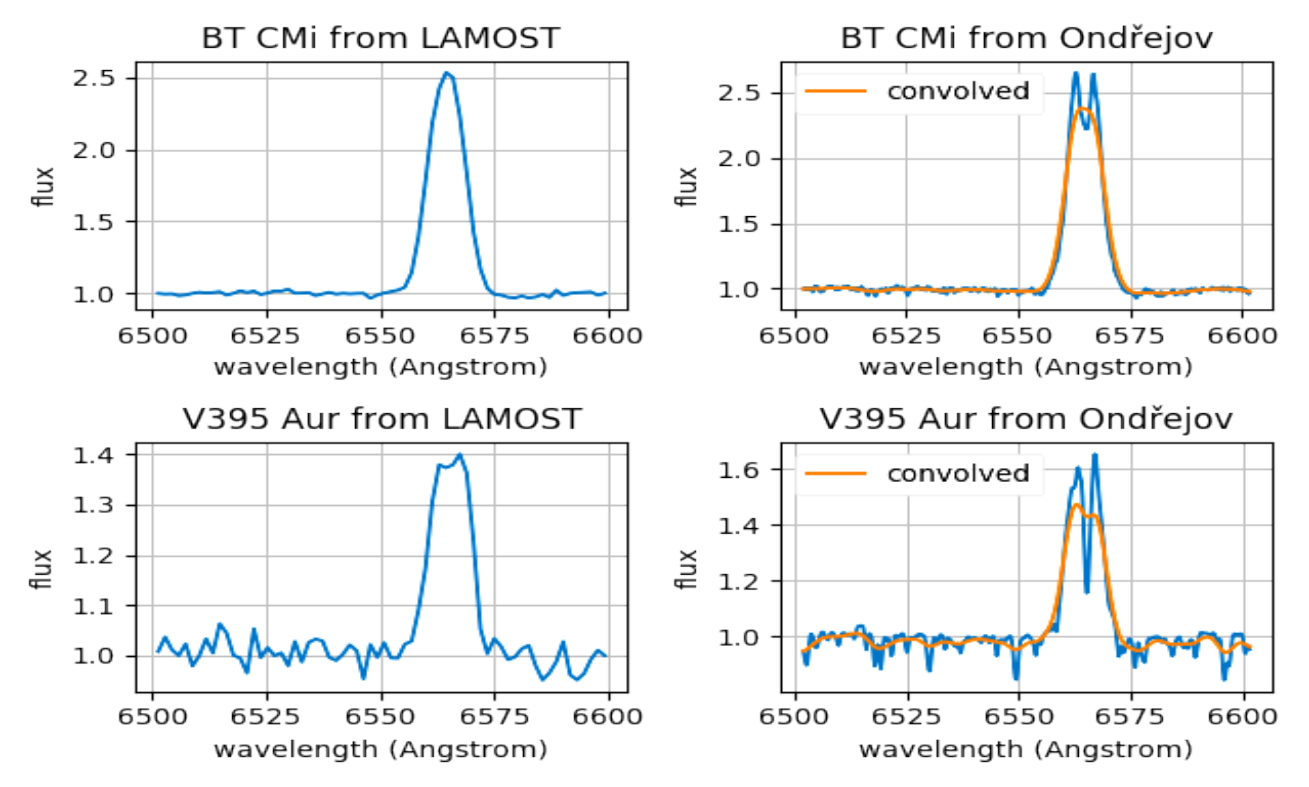


Domain Adaptation



Wavelength transform from air to vacuum

Domain Adaptation



Change of spectral resolution by Gaussian blur (13000 → 1800)
Rebinning to grid of 140 pixels in 6519 to 6732 A → FEATURE VECTOR

Normalisation od Data

Continuum normalisation – rectification (continuum on 1.0, limit at 0)

Ondřejov – automatic pipeline
LAMOST DR1 special extension – but BAD (LAMOST private com)

experiments show that (at least for our Halpha region) the continuum normalisation is not needed (training on unrectified gives same validation accuracy 0.96)

Rescaling to zero mean, unit variance - common in ML

Different emission height vs. rescaled shape … ????
The height of emission should be part of classification ???

Deep Convolutional Net

deep network to have representation power
and **dropout** to reduce overfitting.

Inspired by **VGGNet**
adapted to 1D spectrum.

No feature extraction!

Training:
TENSORFLOW+KERAS on GTX980 GPU

Validation set 20% + test set 10%

input (140 pixel spectrum)
conv3-64
conv3-64
maxpool2
conv3-128
conv3-128
maxpool2
conv3-256
conv3-256
maxpool2
fc-512
fc-512
softmax

Active Learning

Human (Oracle) involved

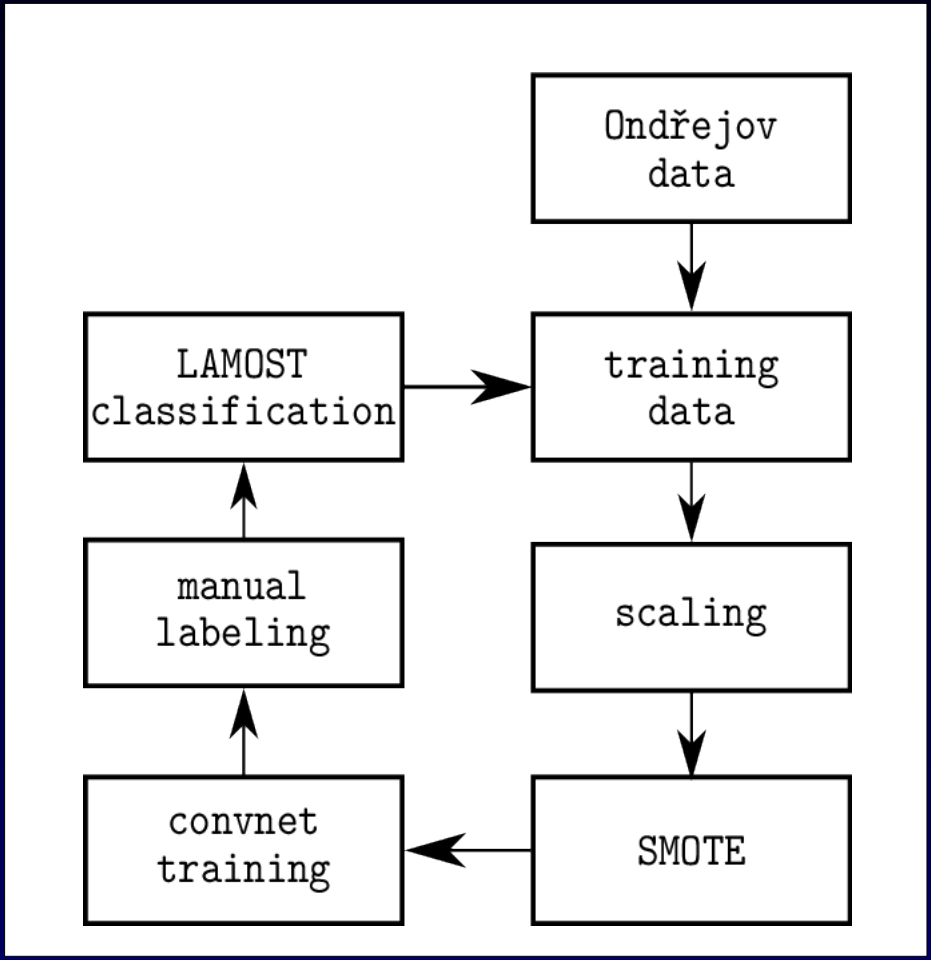
Random Sampling:

From predicted TARGET class (single or double peak) selected randomly 100

Visual check : re-classification (confirm, change, put in NOISE)

These data added to training set

Repeat
until few misclassifications
(6 times)



Results

12944 candidates (from 2 mil)

572 double peak (but complex shapes)

12372 single peak

Visual check (500 boring)

Most are correct - but physical origin (YSO, late types M)
Unreliable classification from LAMOST 1D pipeline (F, A/B)

A lot of new Be stars (xmatching – 243 unknown <12 mag r)
Many SUPRISES !!!

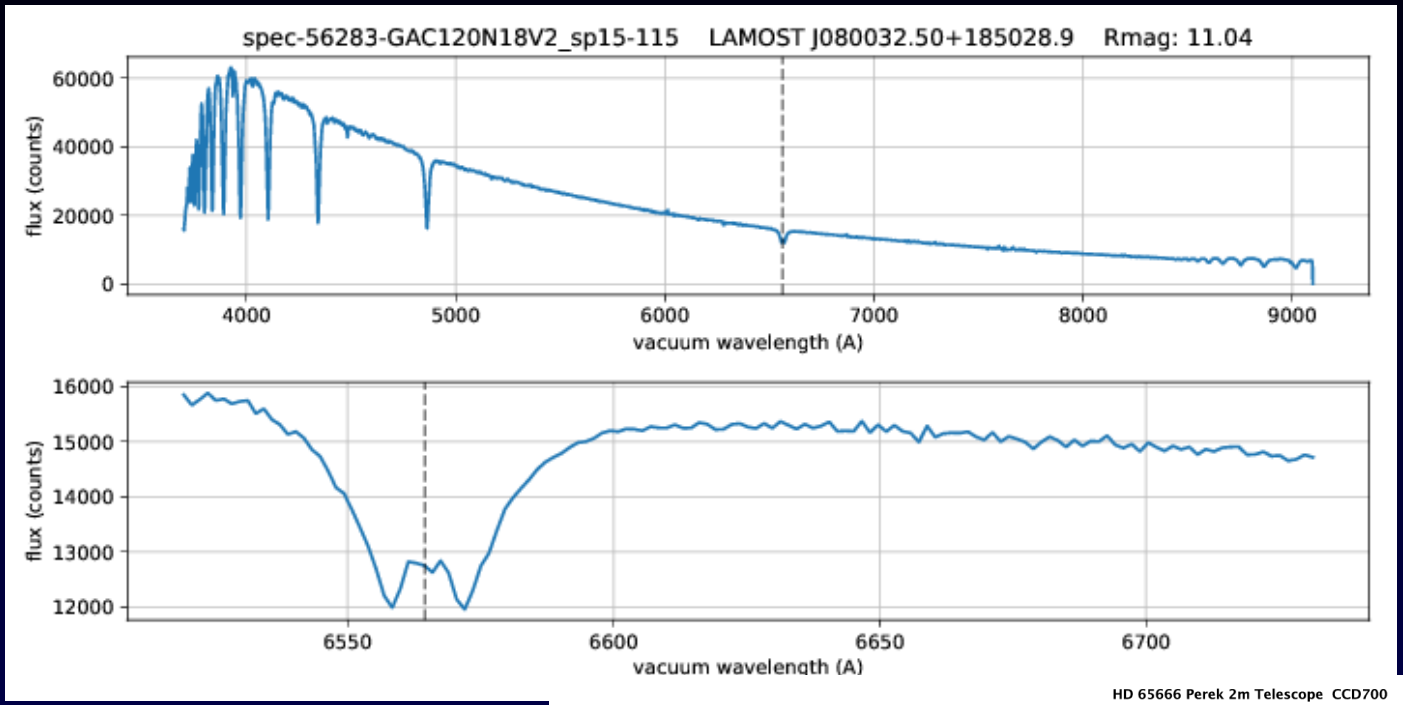
Comparison with Other Methods

Integral pixel statistics on different intervals

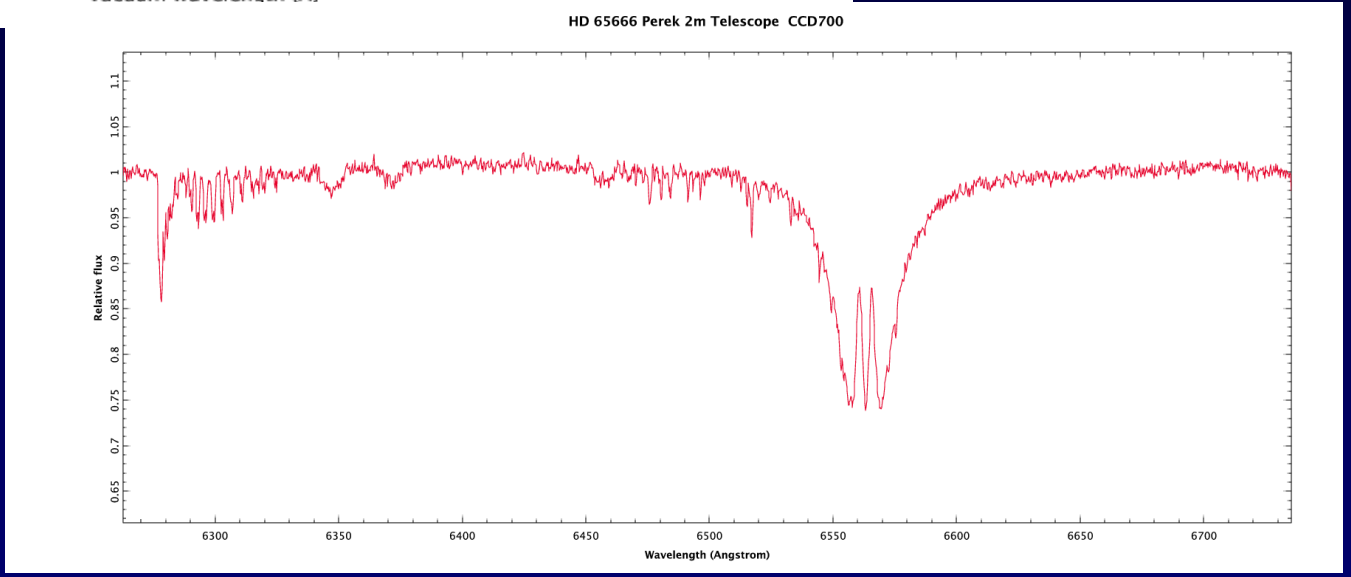
Lin (2015) DR1 (**23** classical Be + 180 new Be)
We have **20** (3 not seen in normalized) **179** OK (1 faint)

Hou (2016) DR2 (more data – over 4 mil) - 11205
we have **2078** of them

Results – Unknown Bright Be star

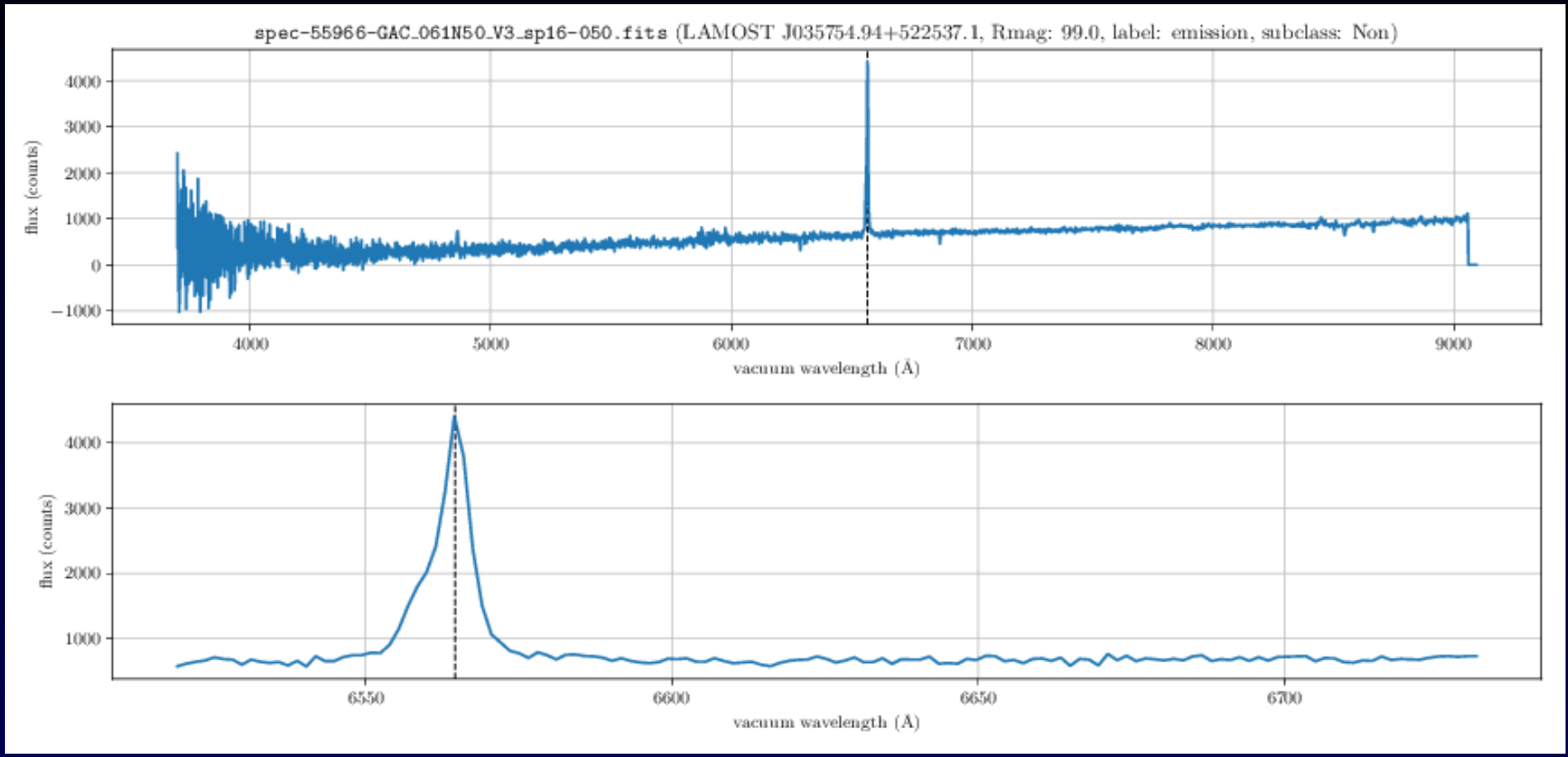


LAMOST

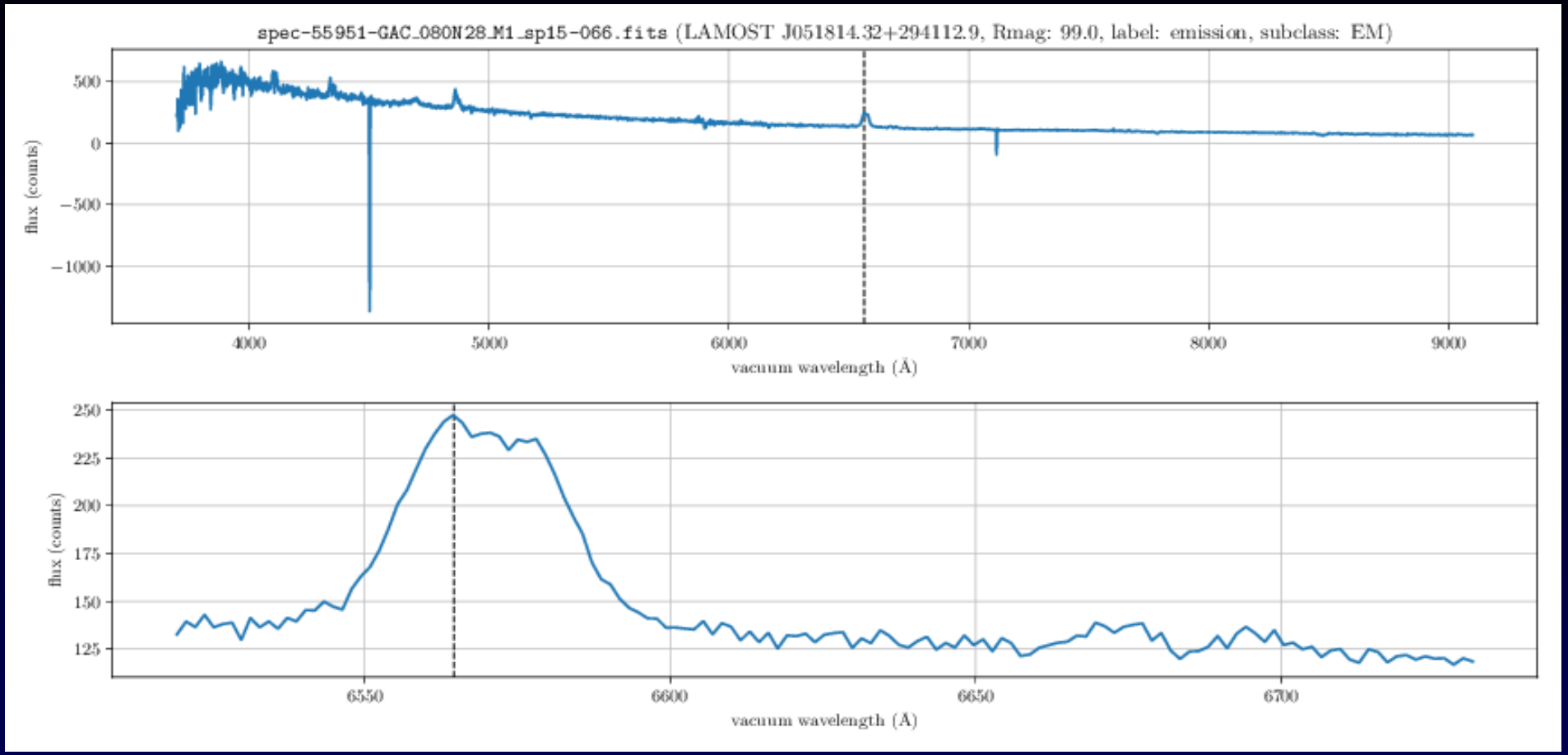


CCD700 OND

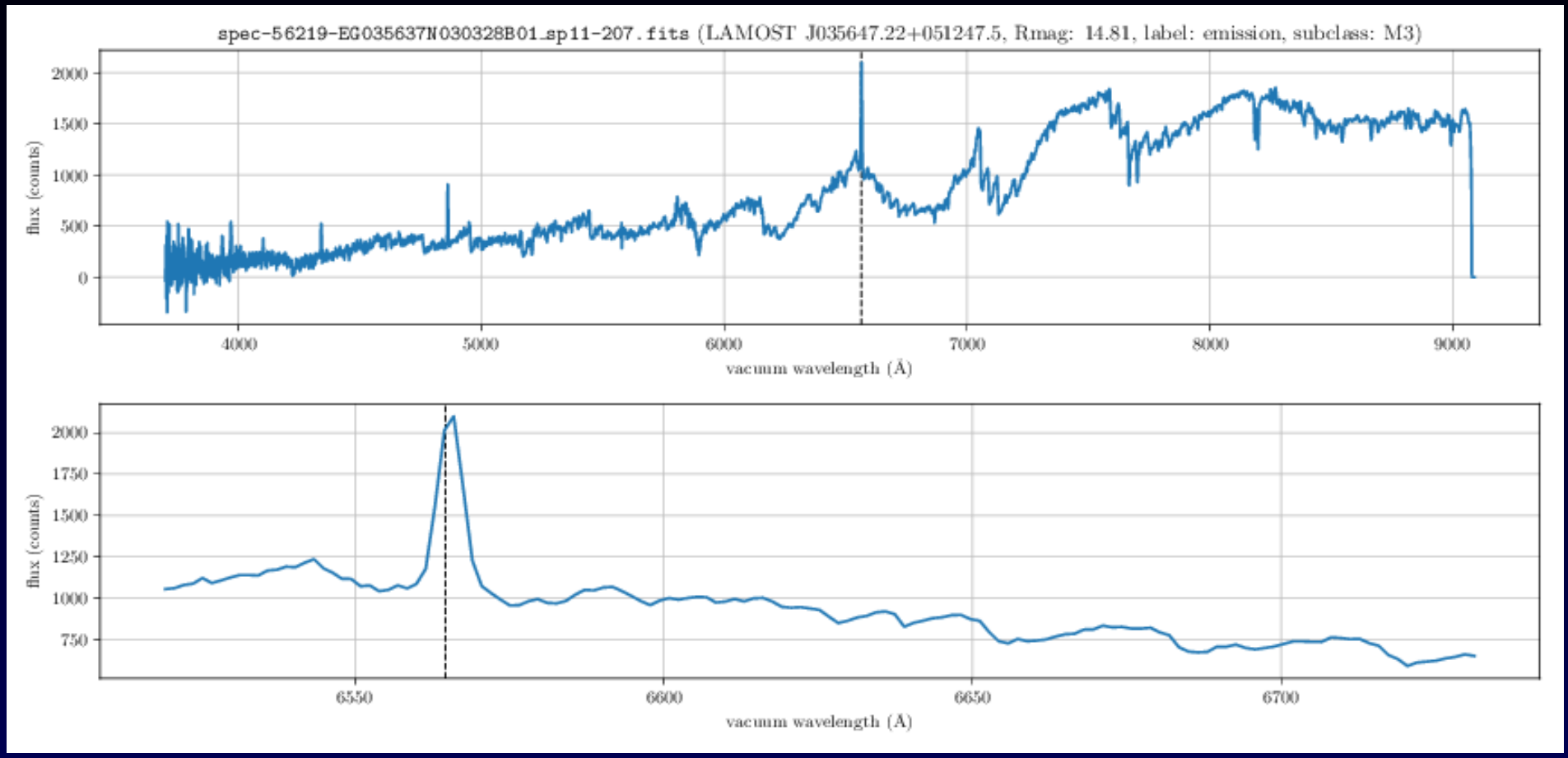
Results – Single Peak Emission



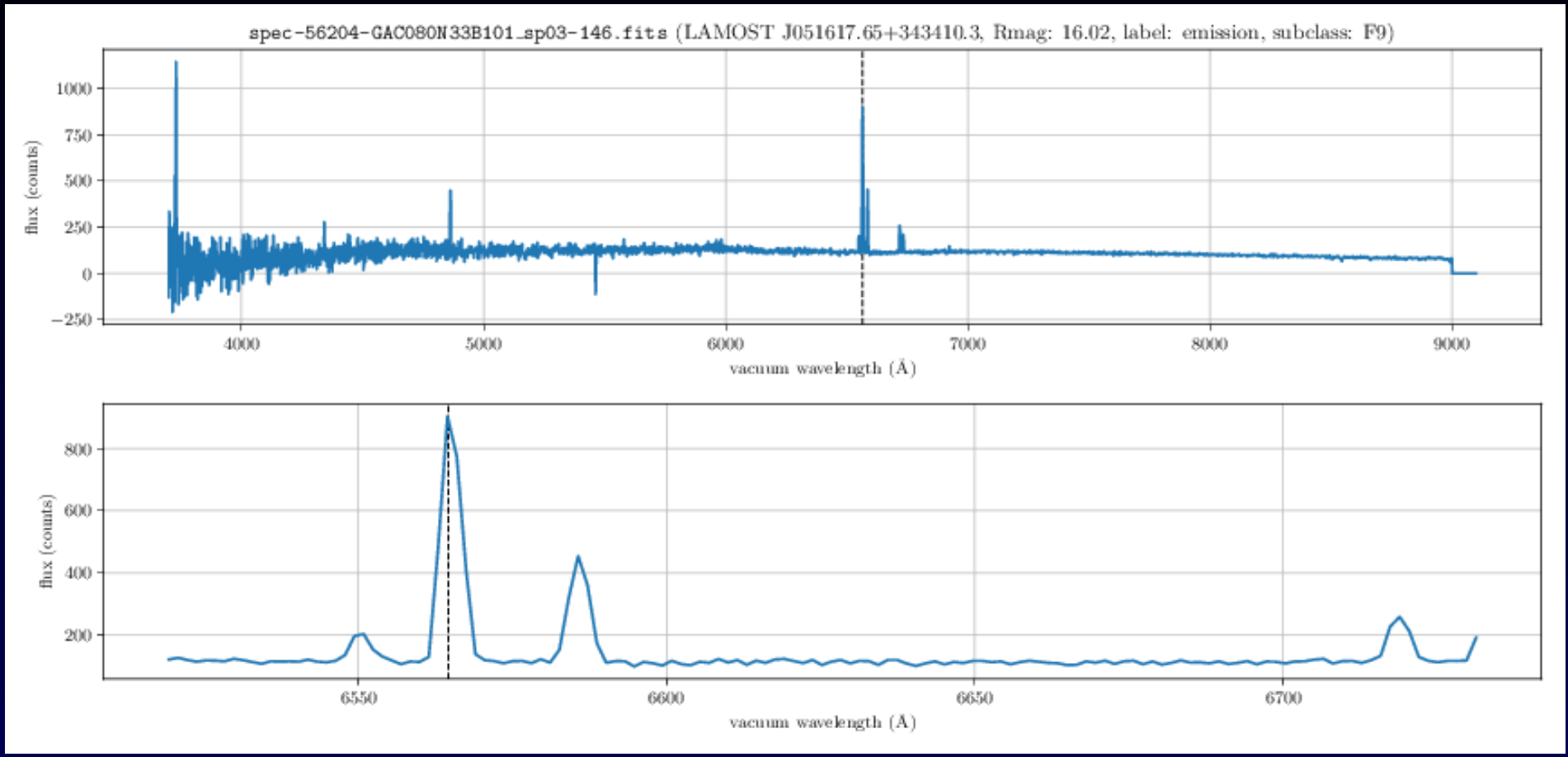
Results – Single Peak Emission



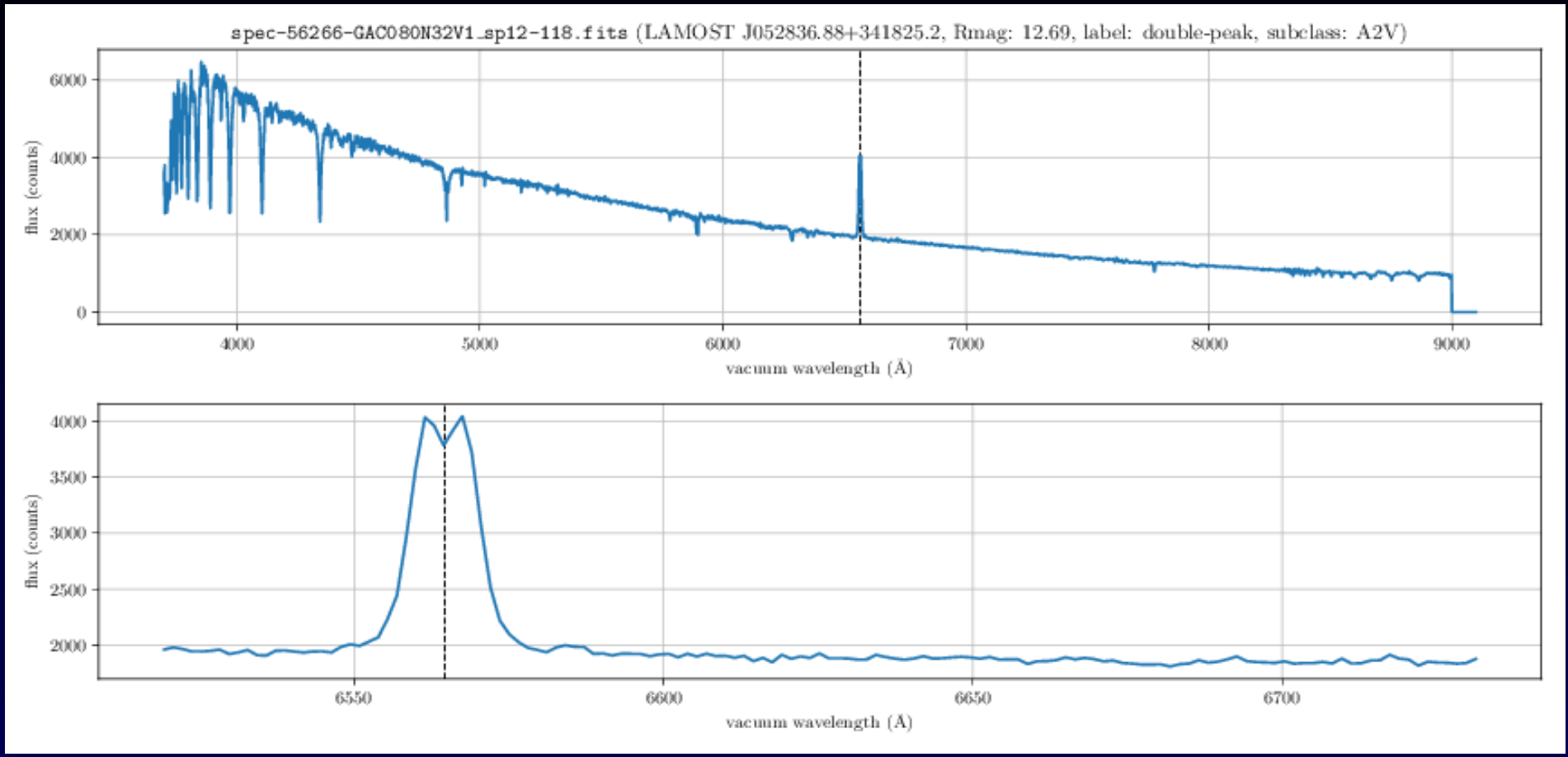
Results – Single Peak Emission



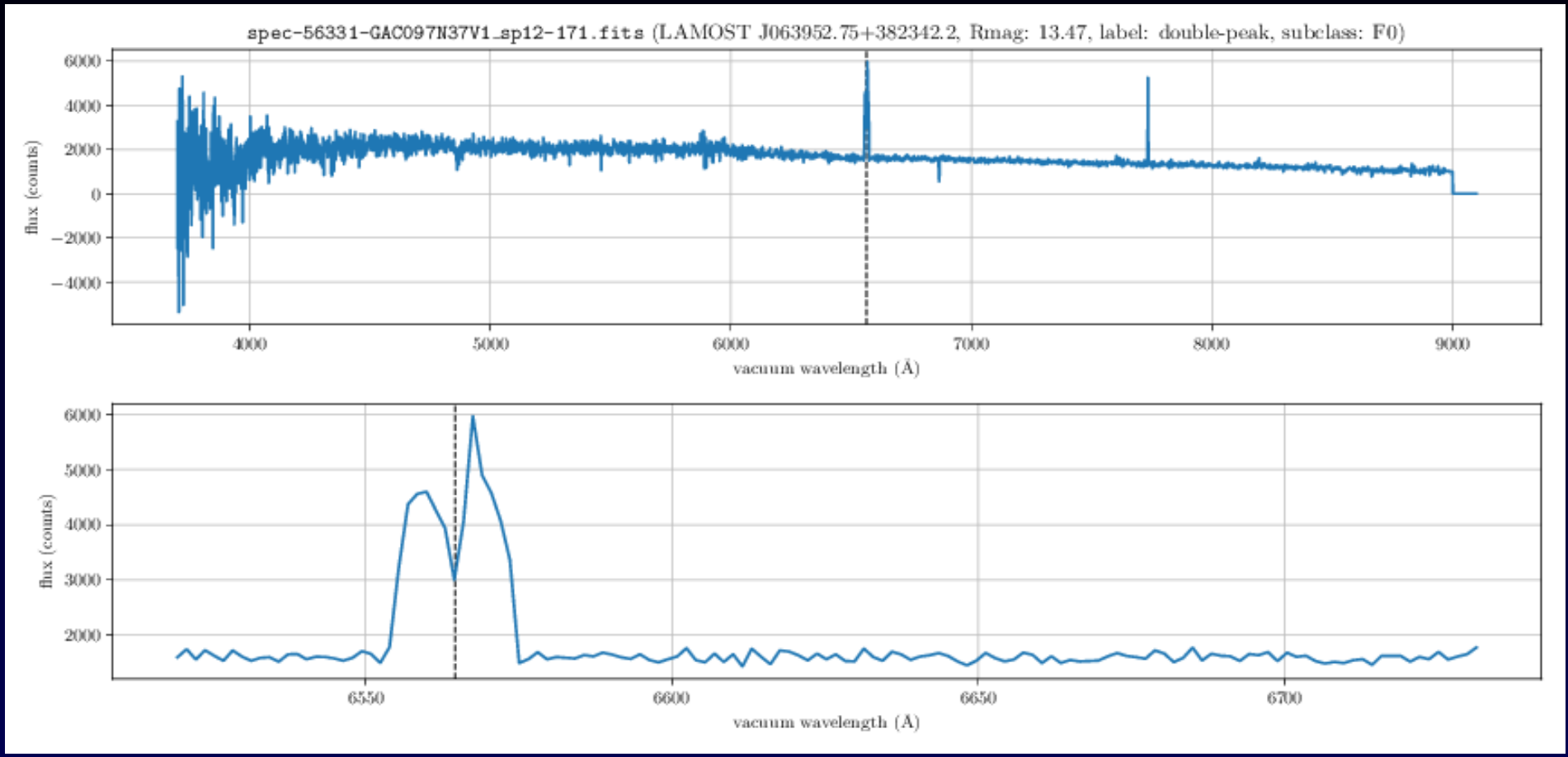
Results – Single Peak Emission



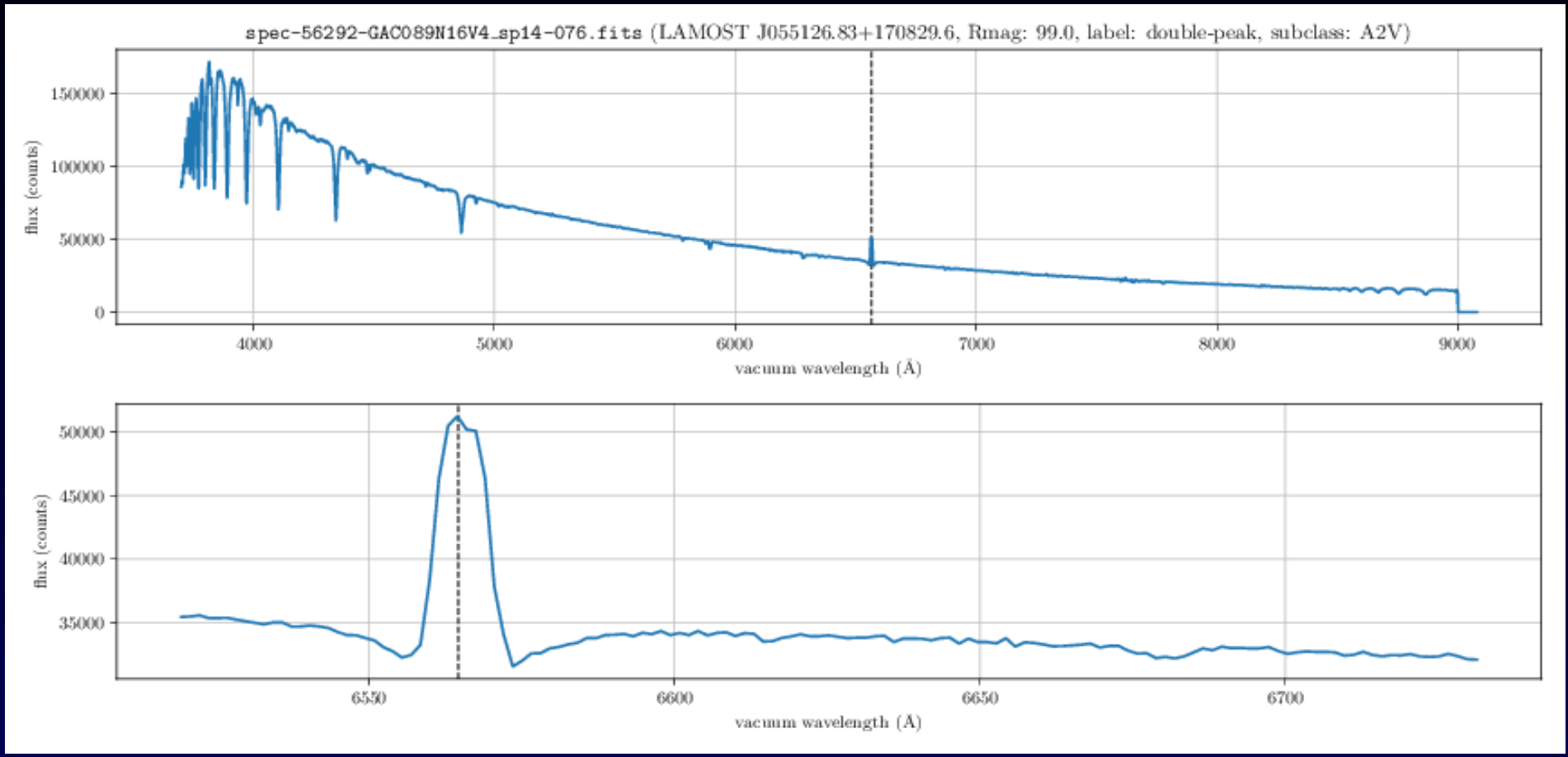
Resuts – Double Peak Emission



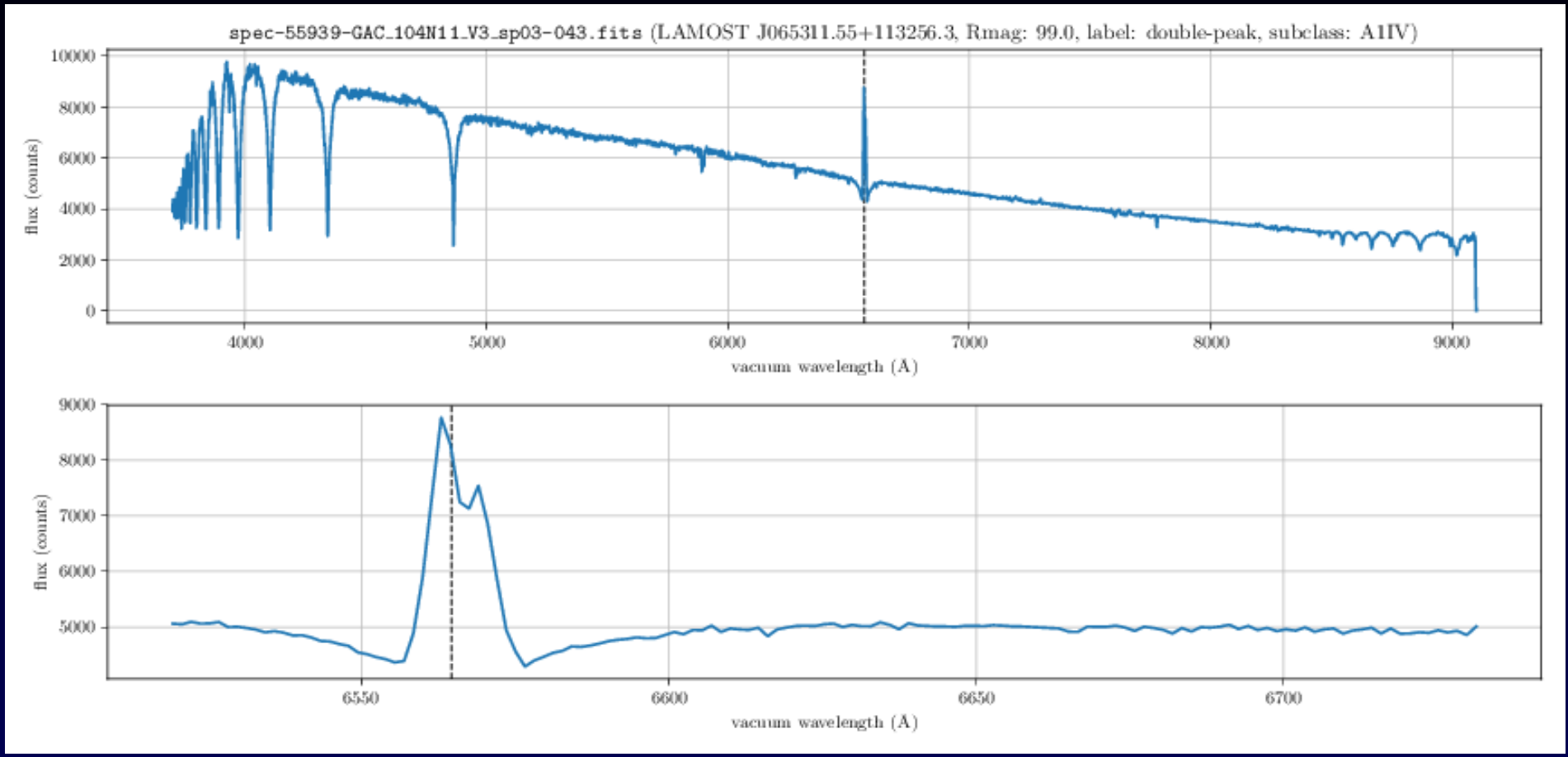
Resuts – Double Peak Emission



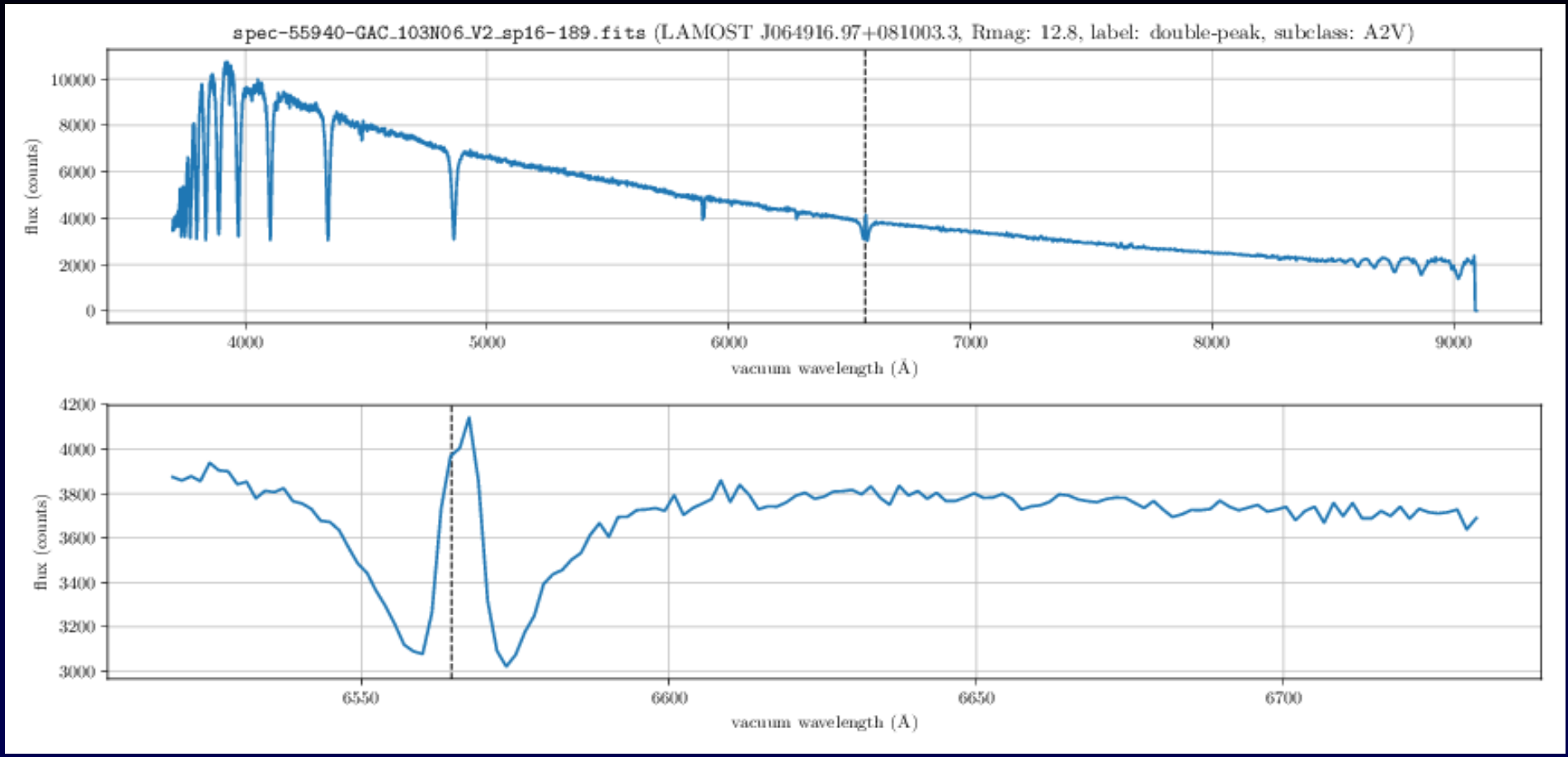
Resuts – Double Peak Emission



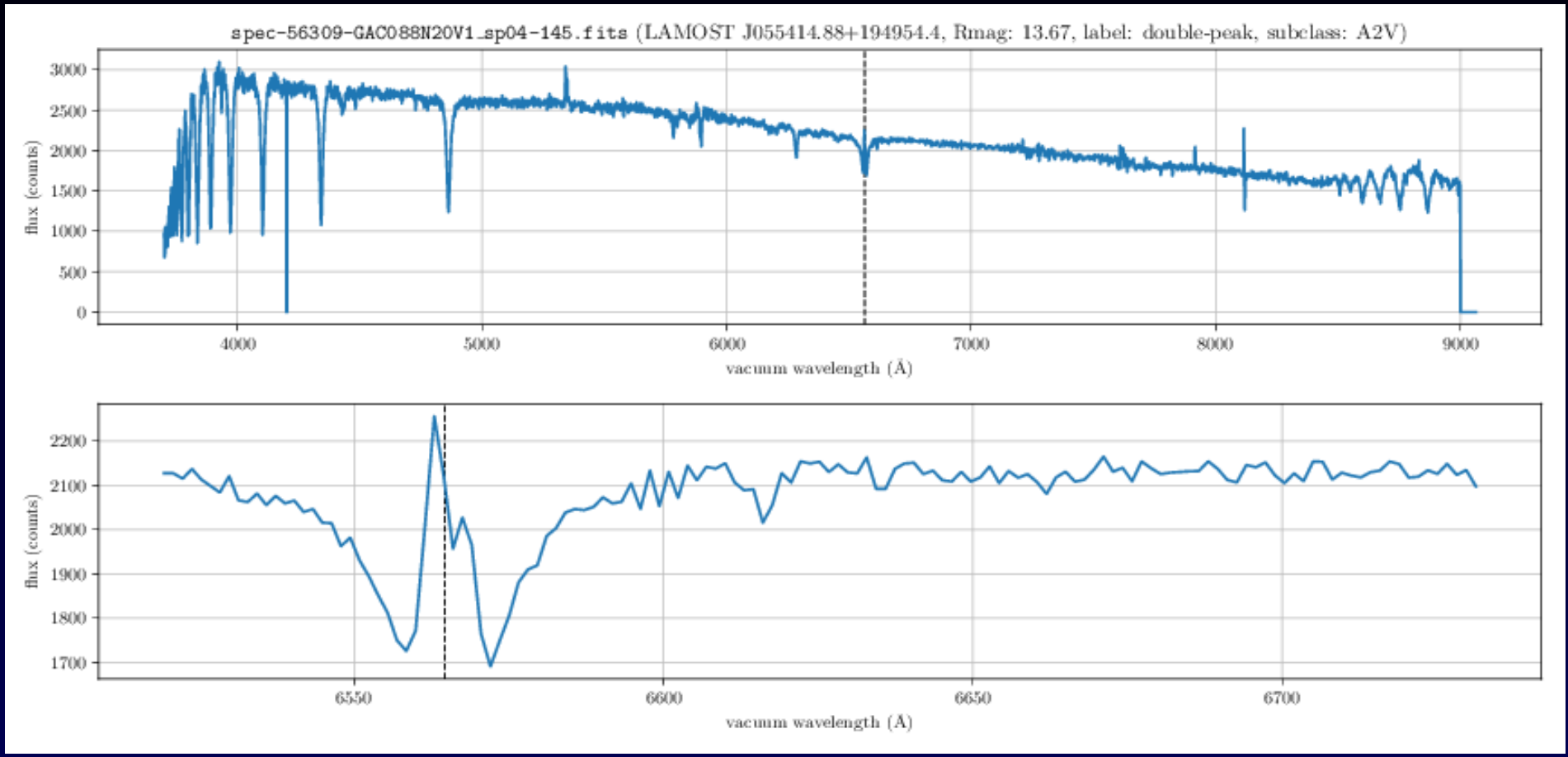
Resuts – Double Peak Emission



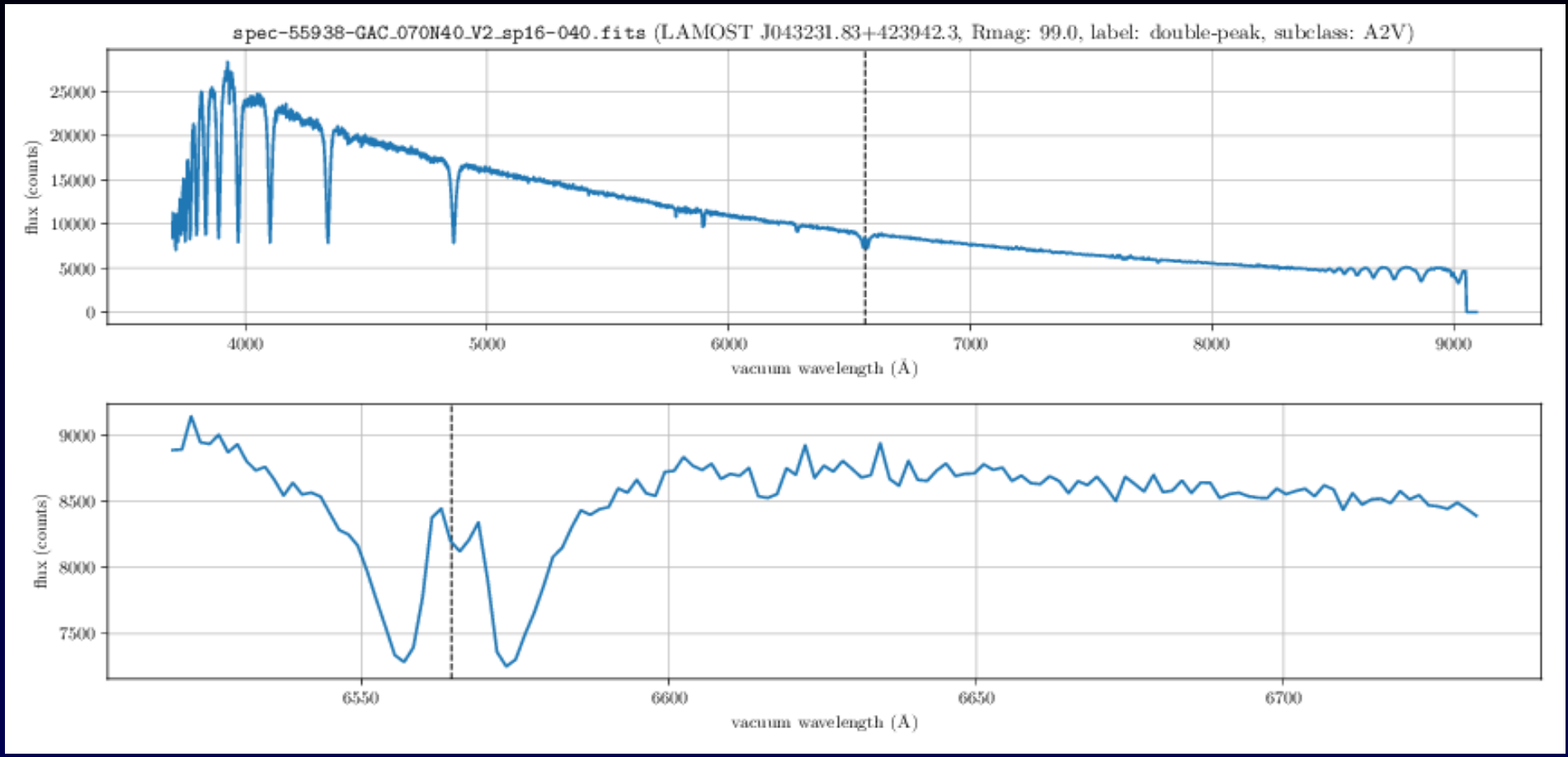
Resuts – Double Peak Emission



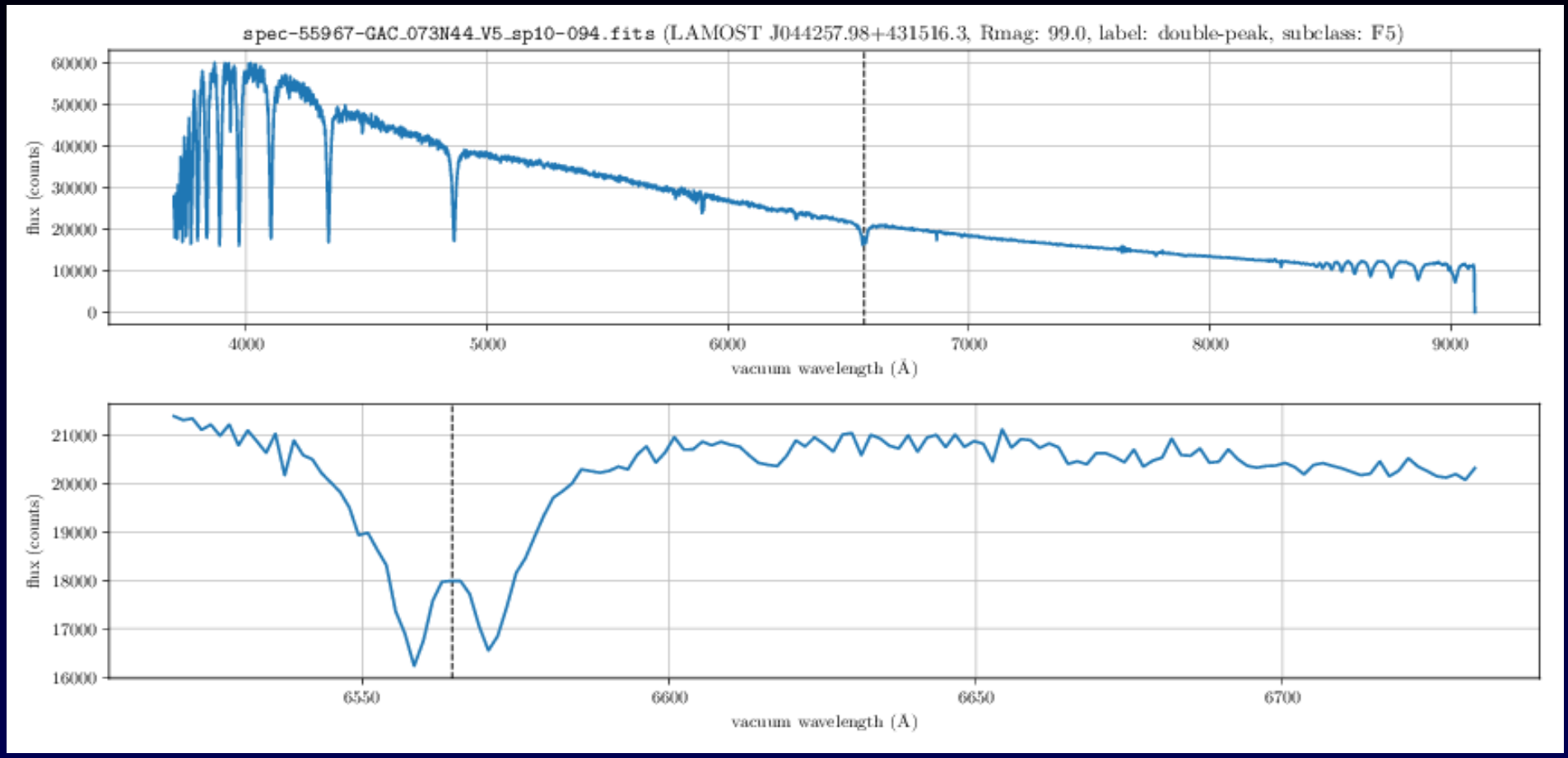
Resuts – Double Peak Emission



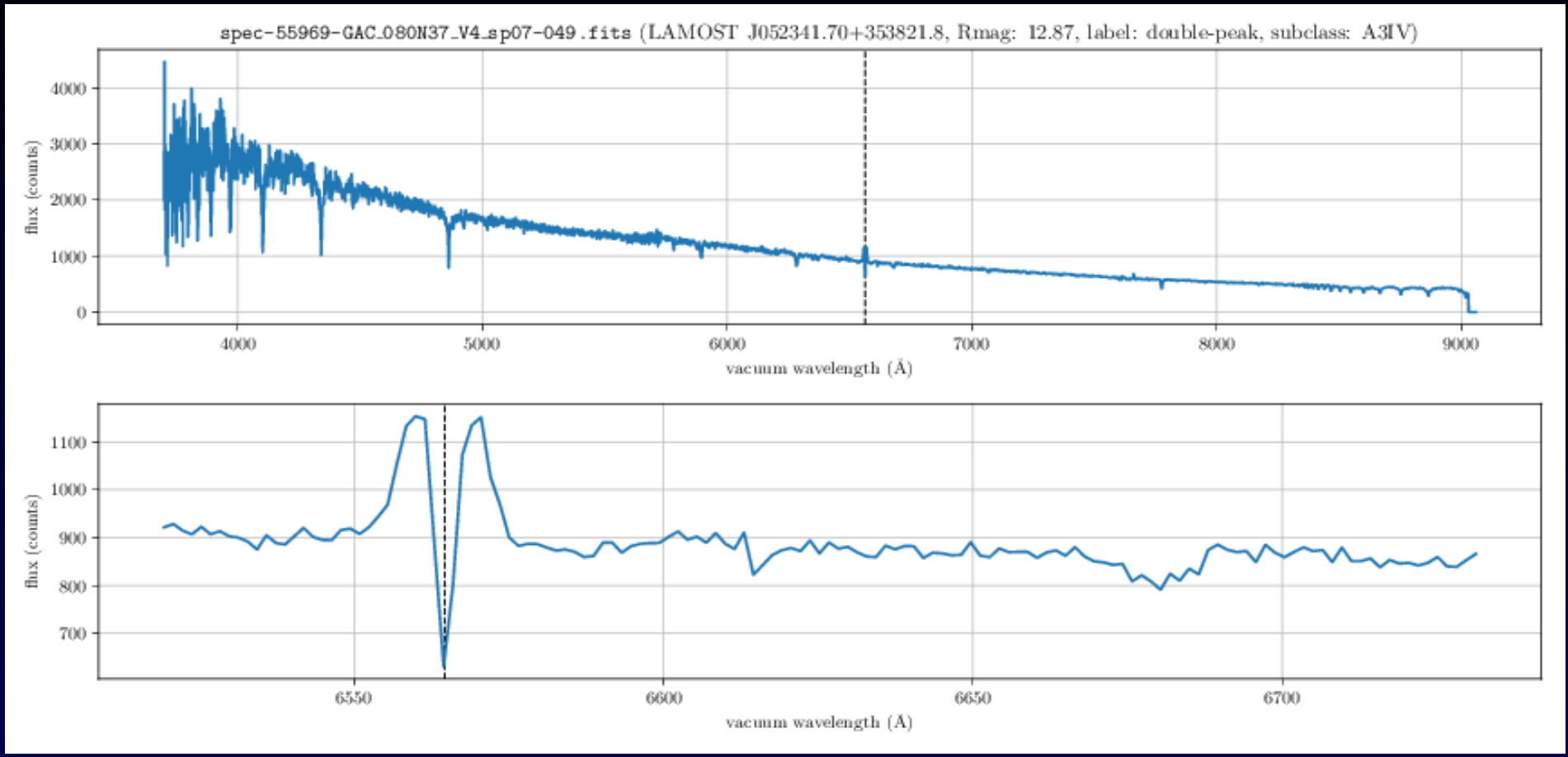
Resuts – Double Peak Emission



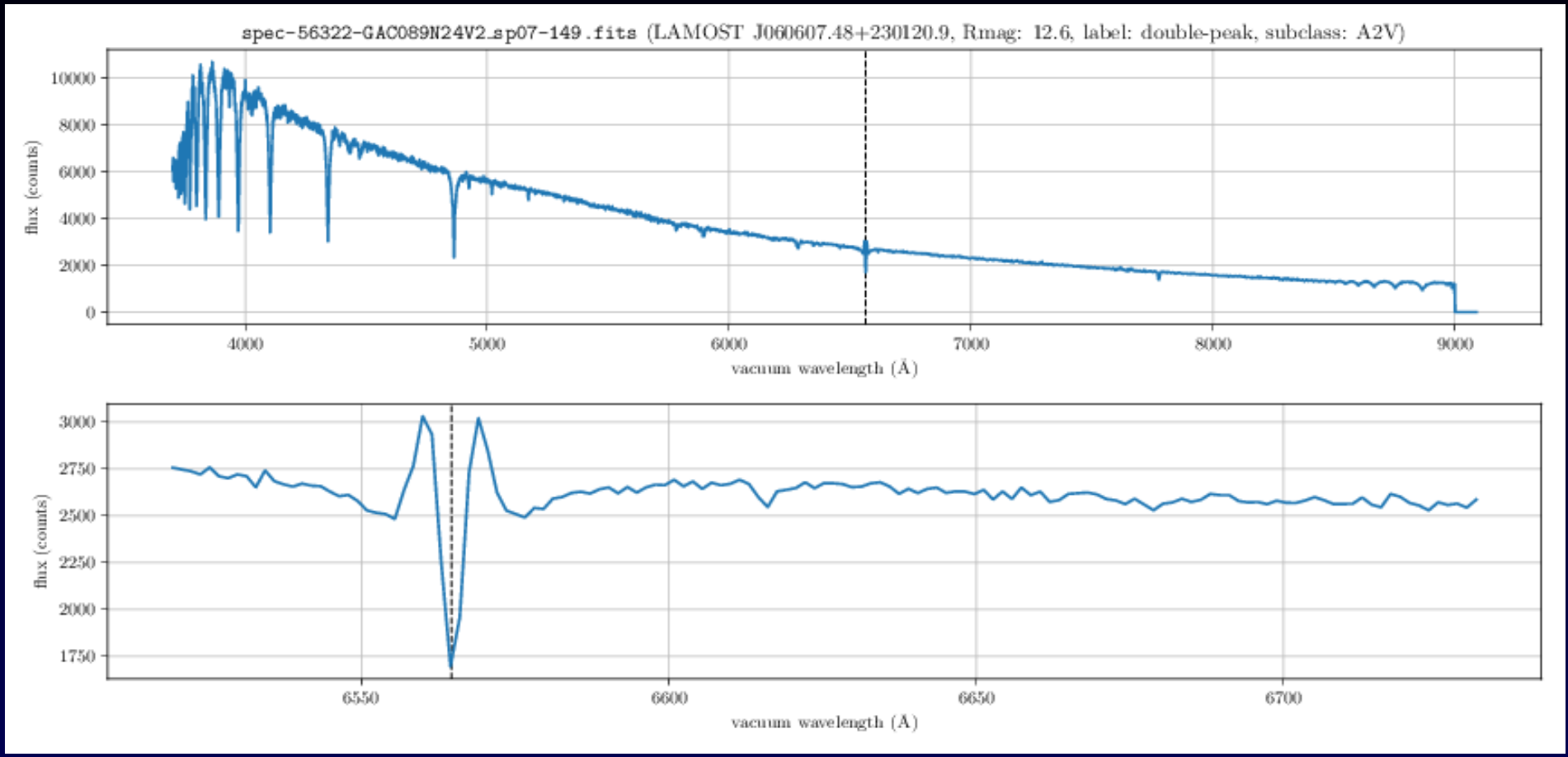
Resuts – Double Peak Emission



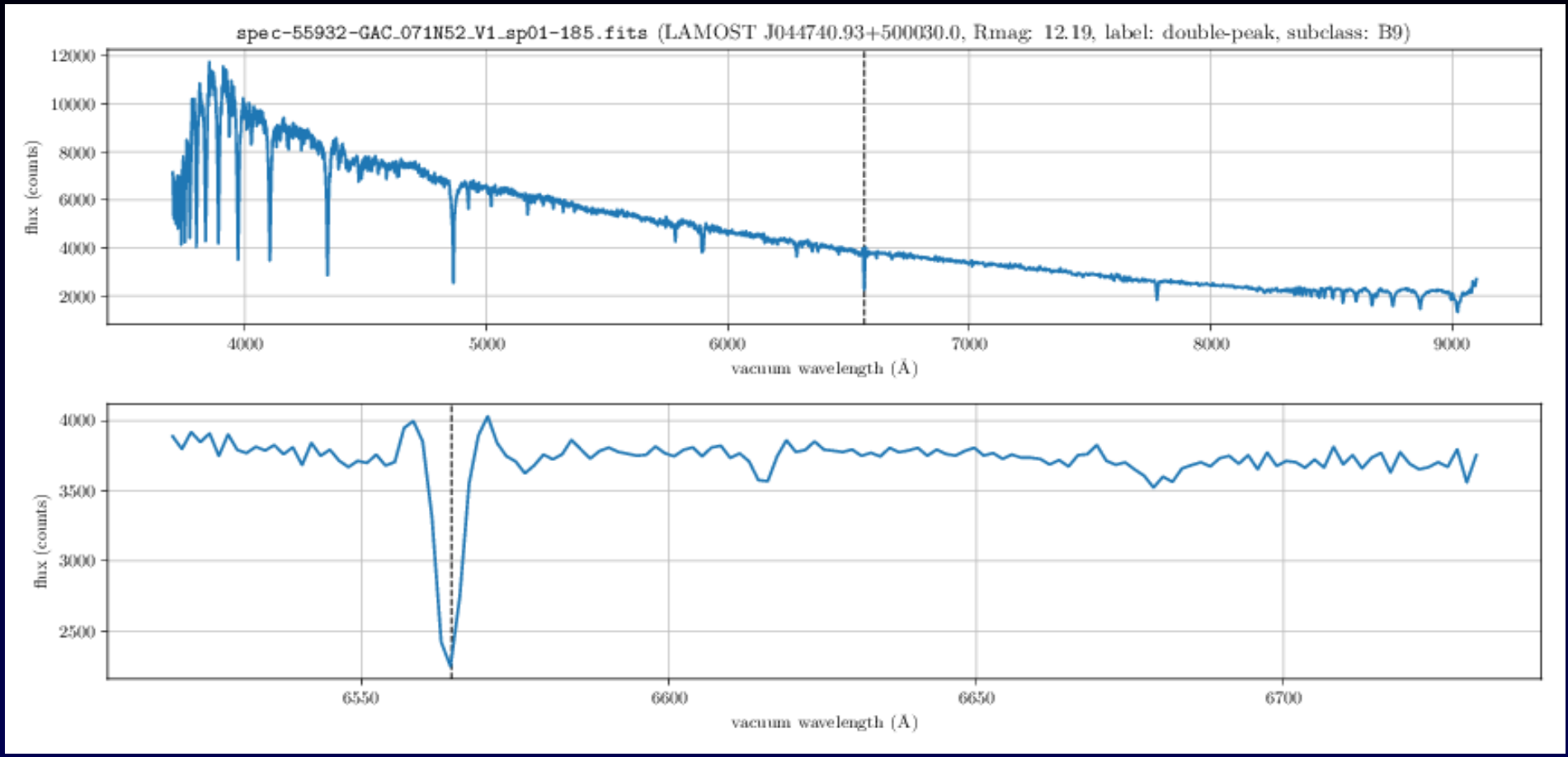
Resuts – Double Peak Emission



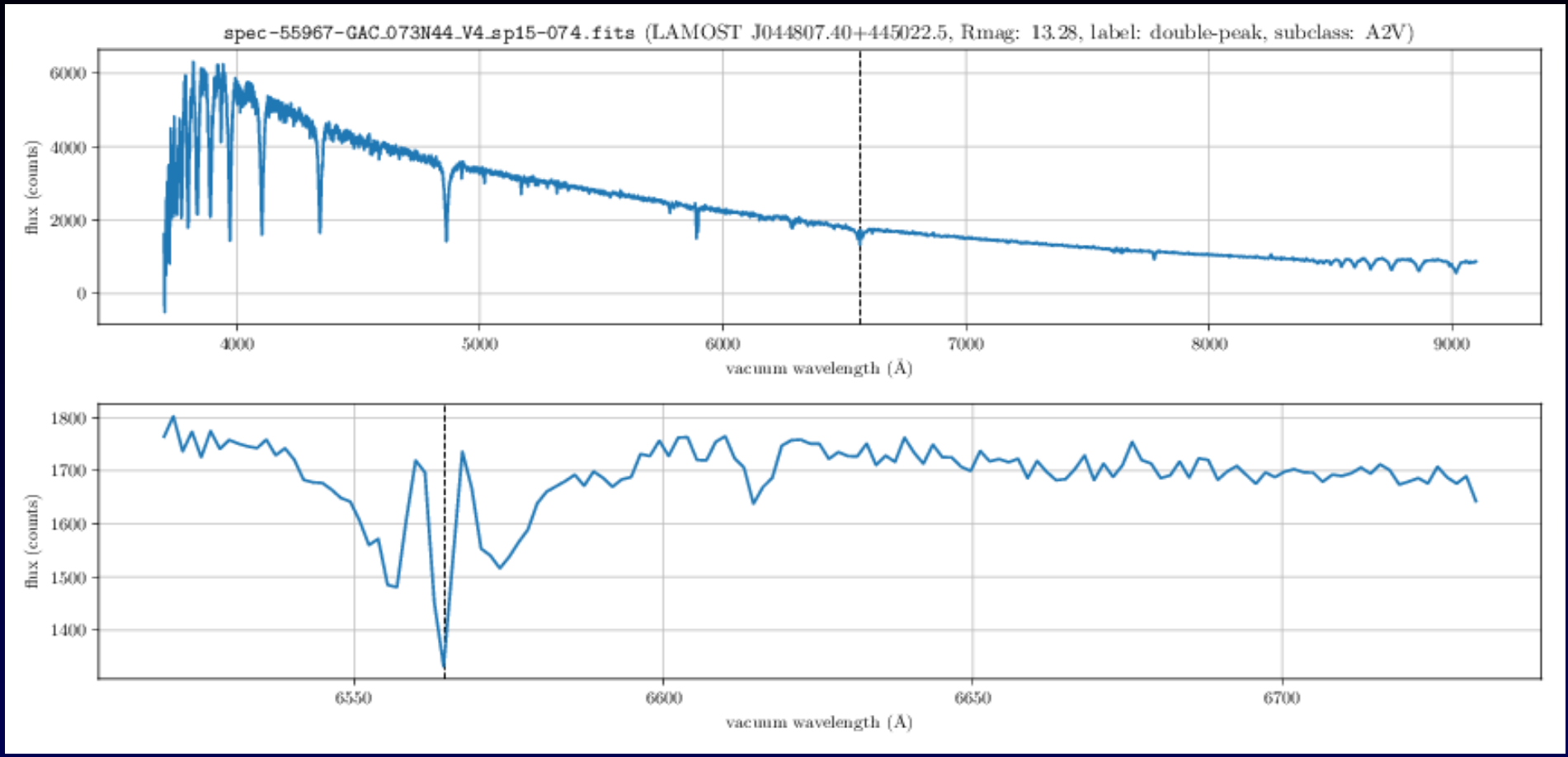
Resuts – Double Peak Emission



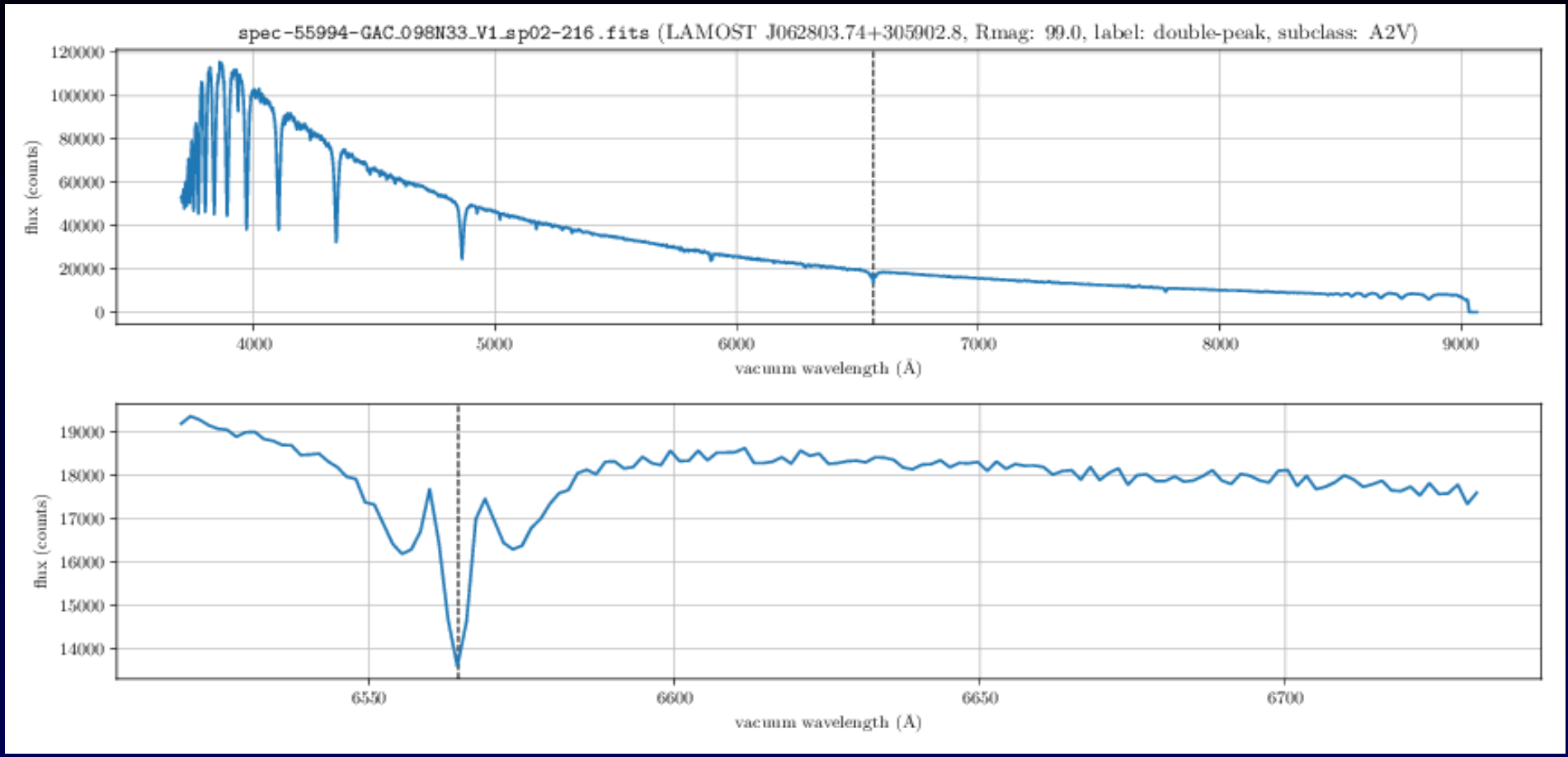
Resuts – Double Peak Emission



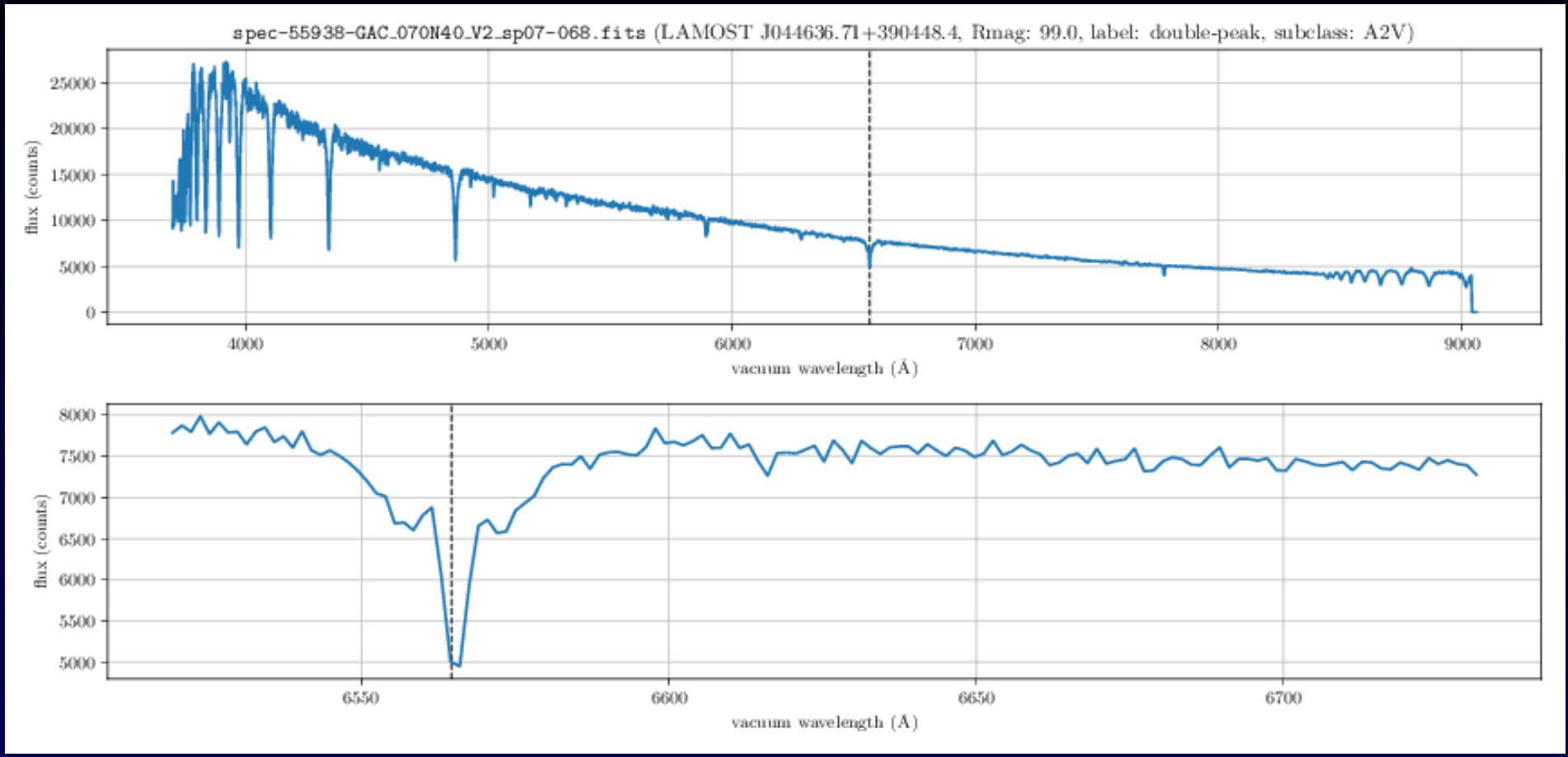
Resuts – Double Peak Emission



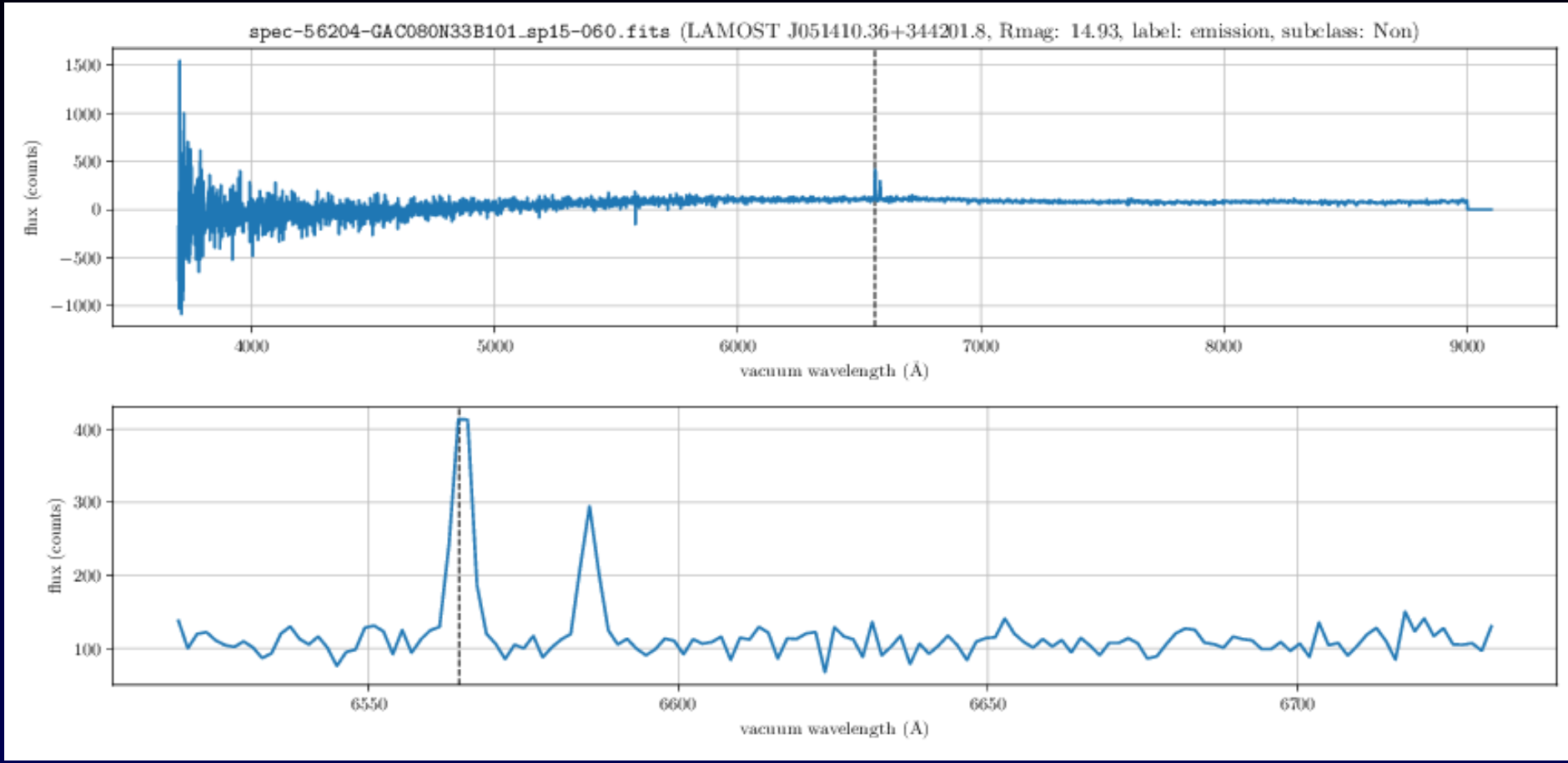
Resuts – Double Peak Emission



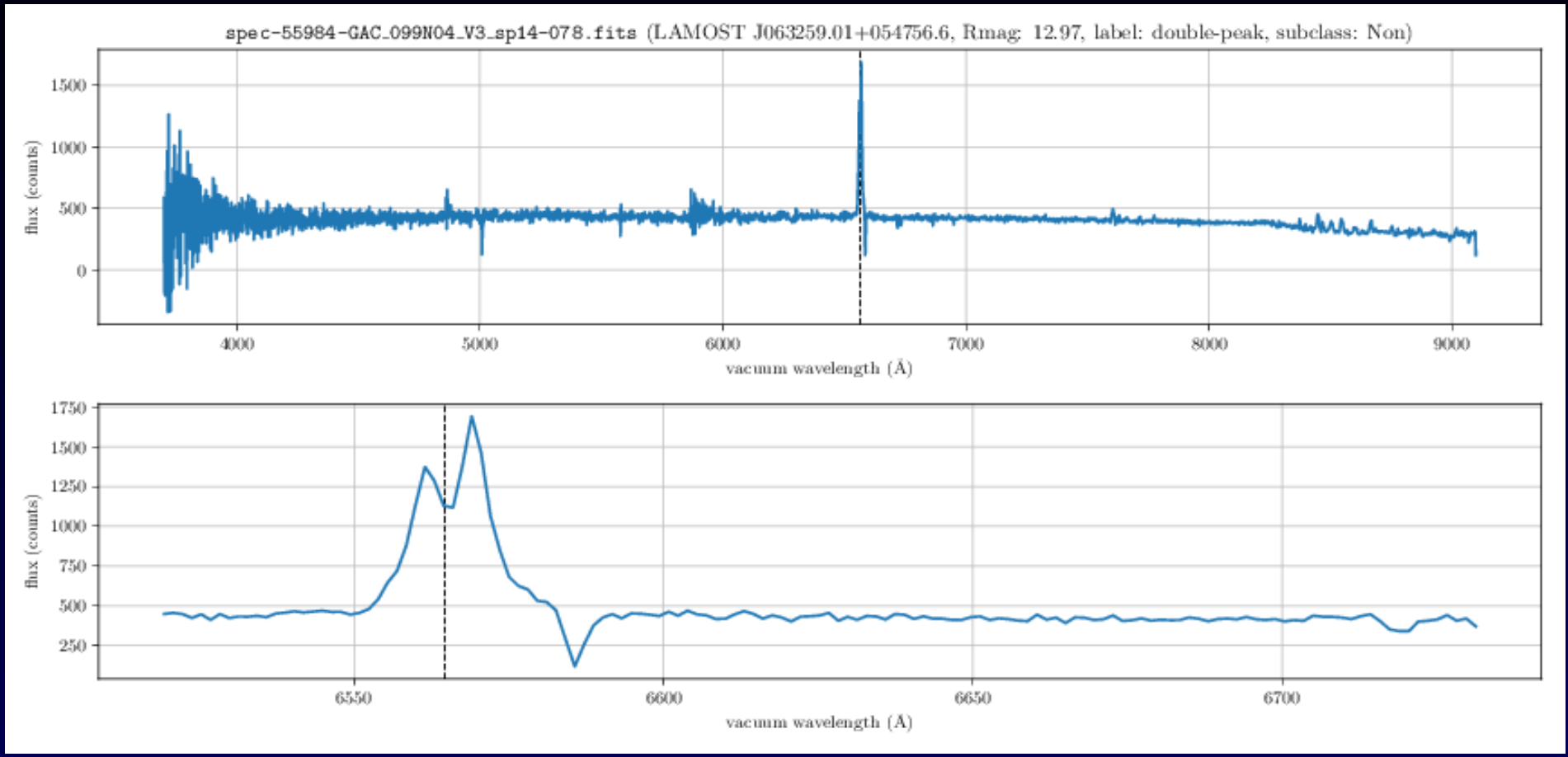
Resuts – Double Peak Emission



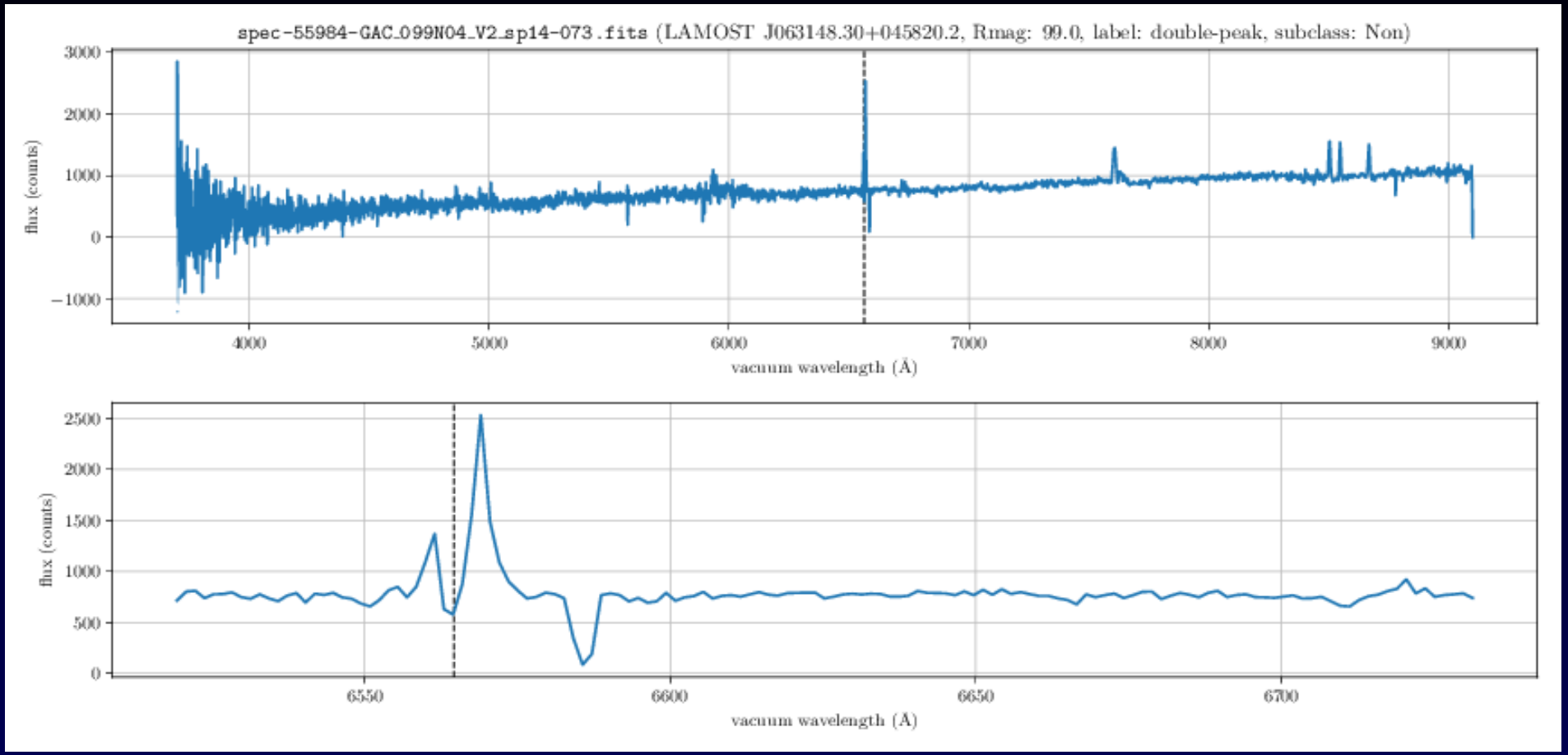
Results – Interesting



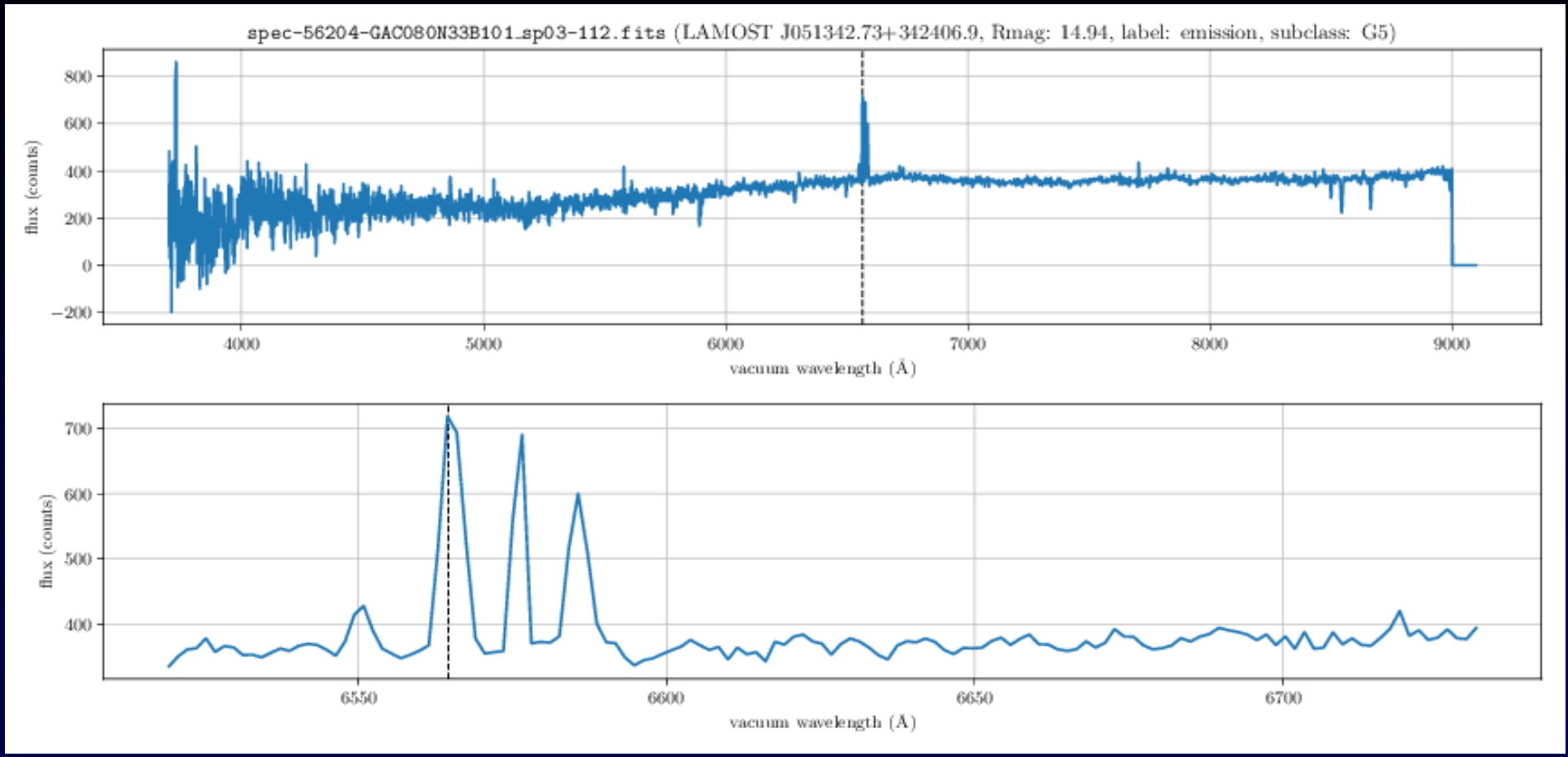
Results – Interesting



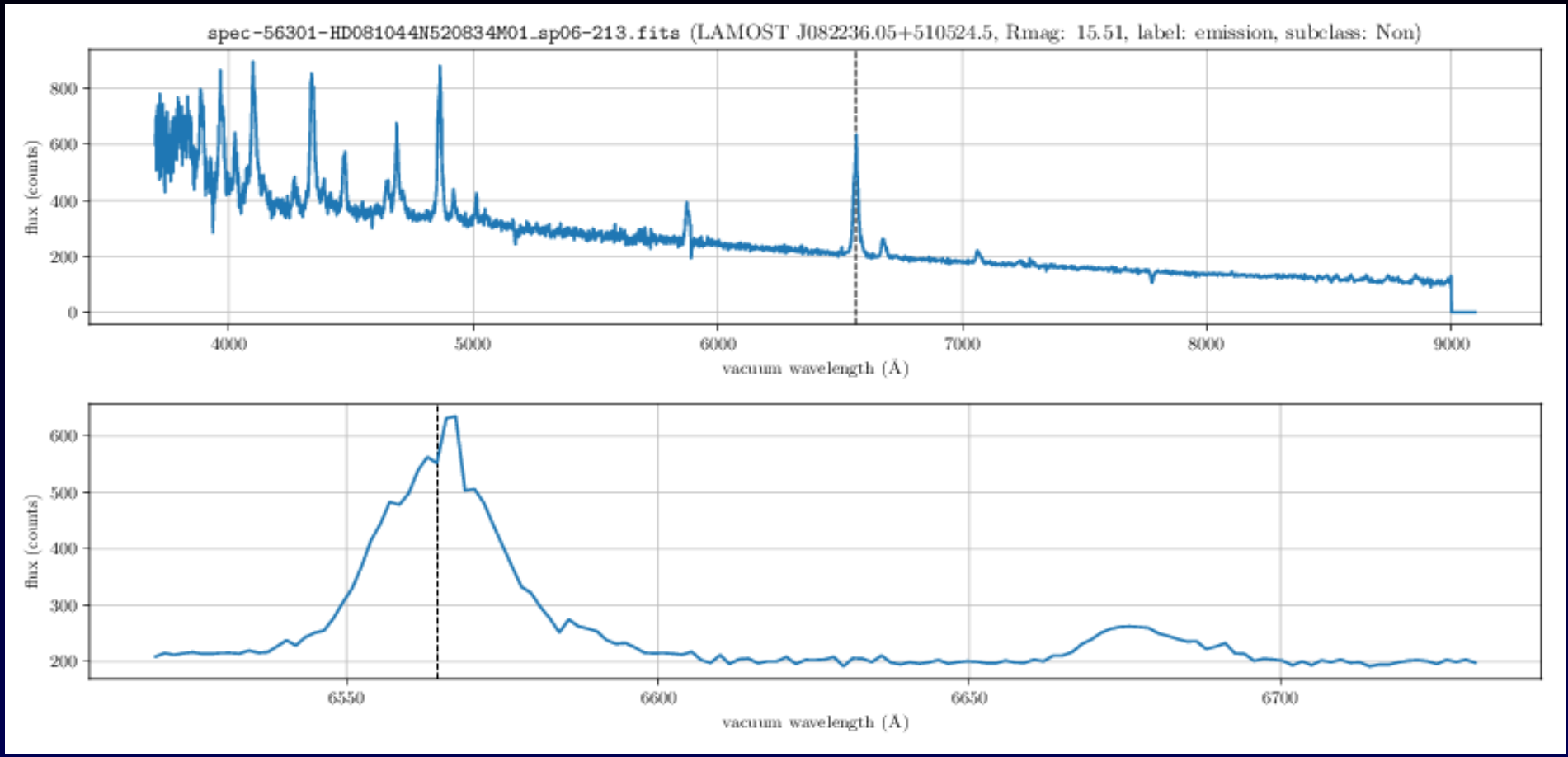
Results – Interesting



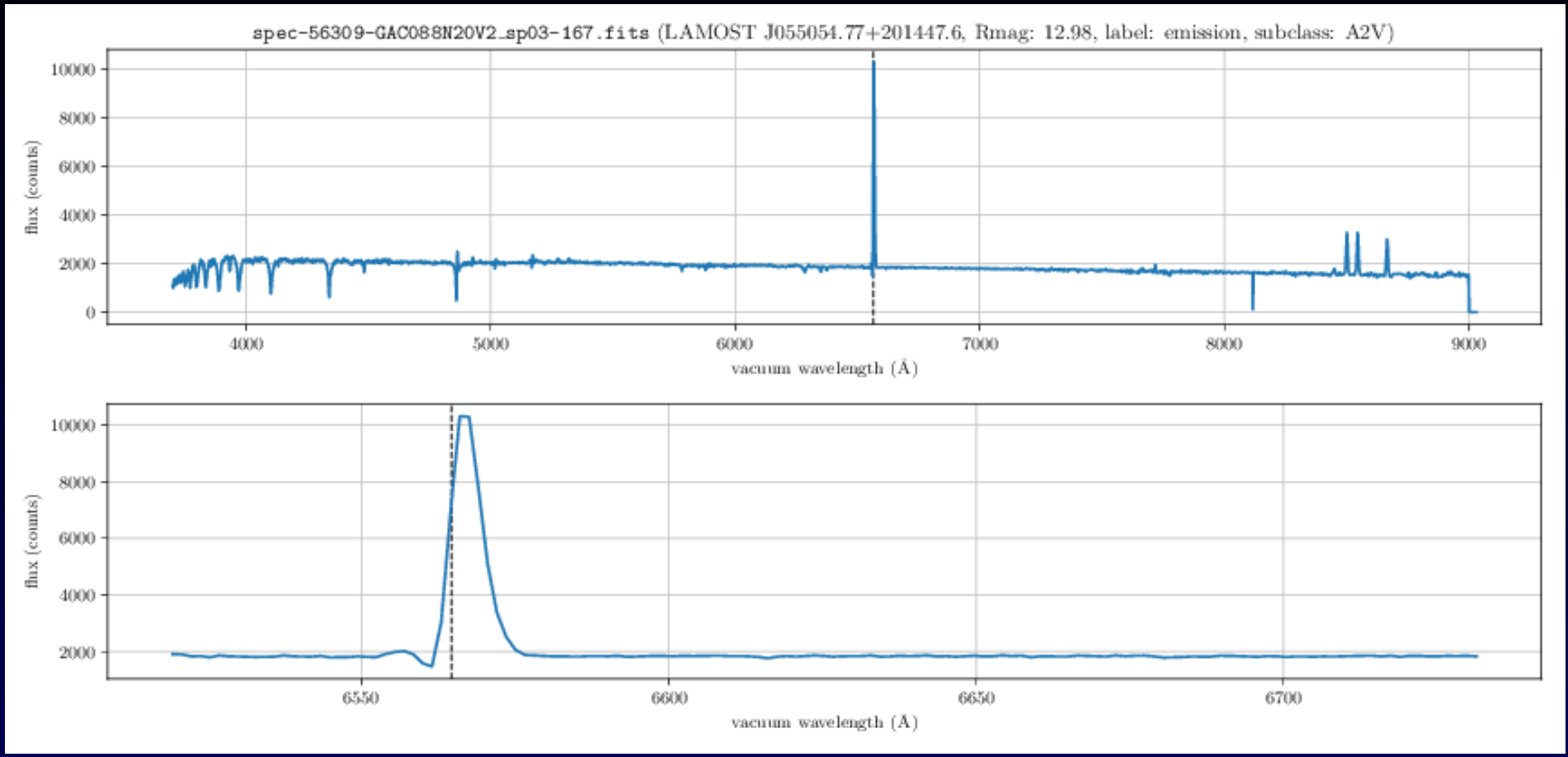
Results – Interesting



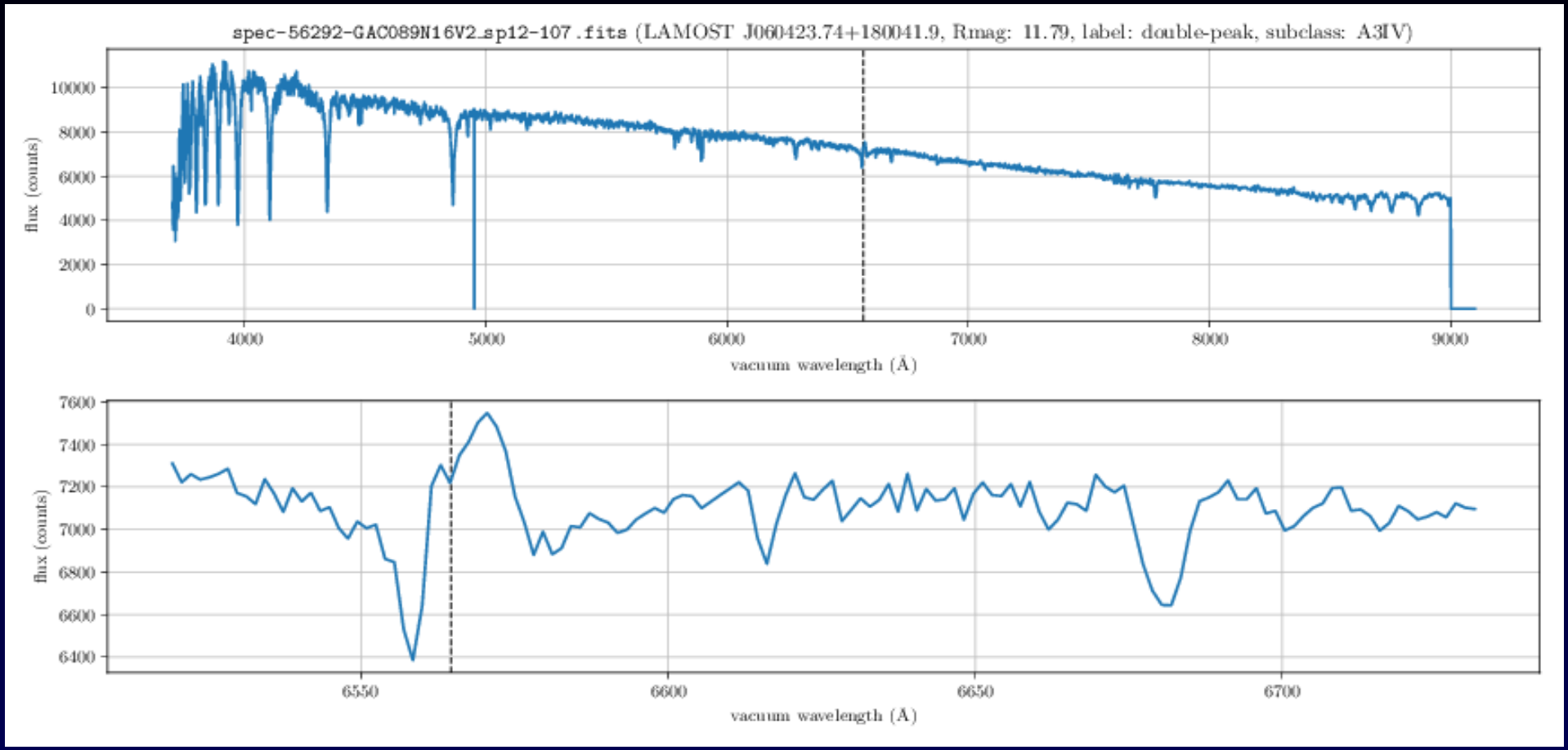
Results – Interesting



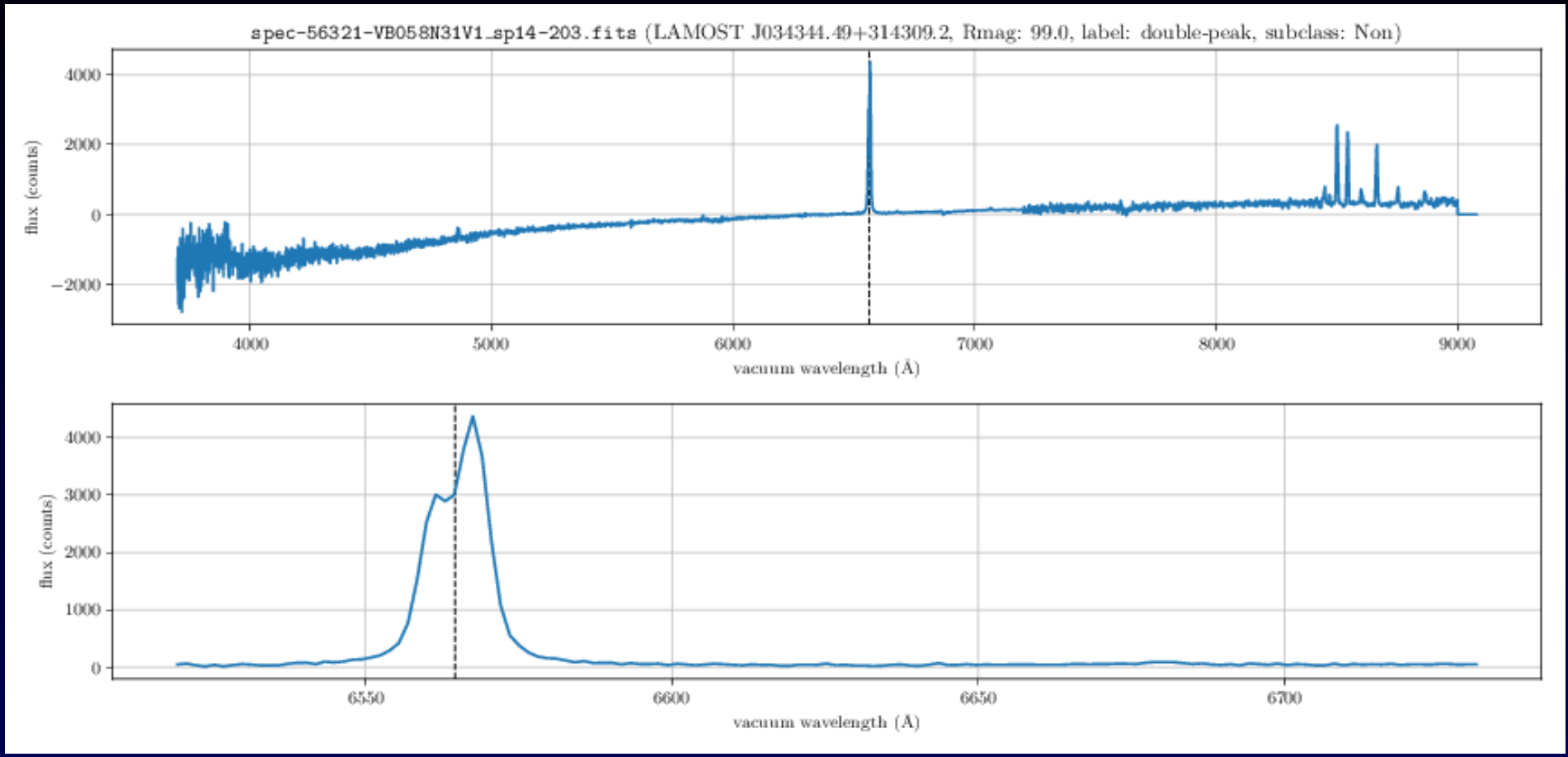
Results – Interesting



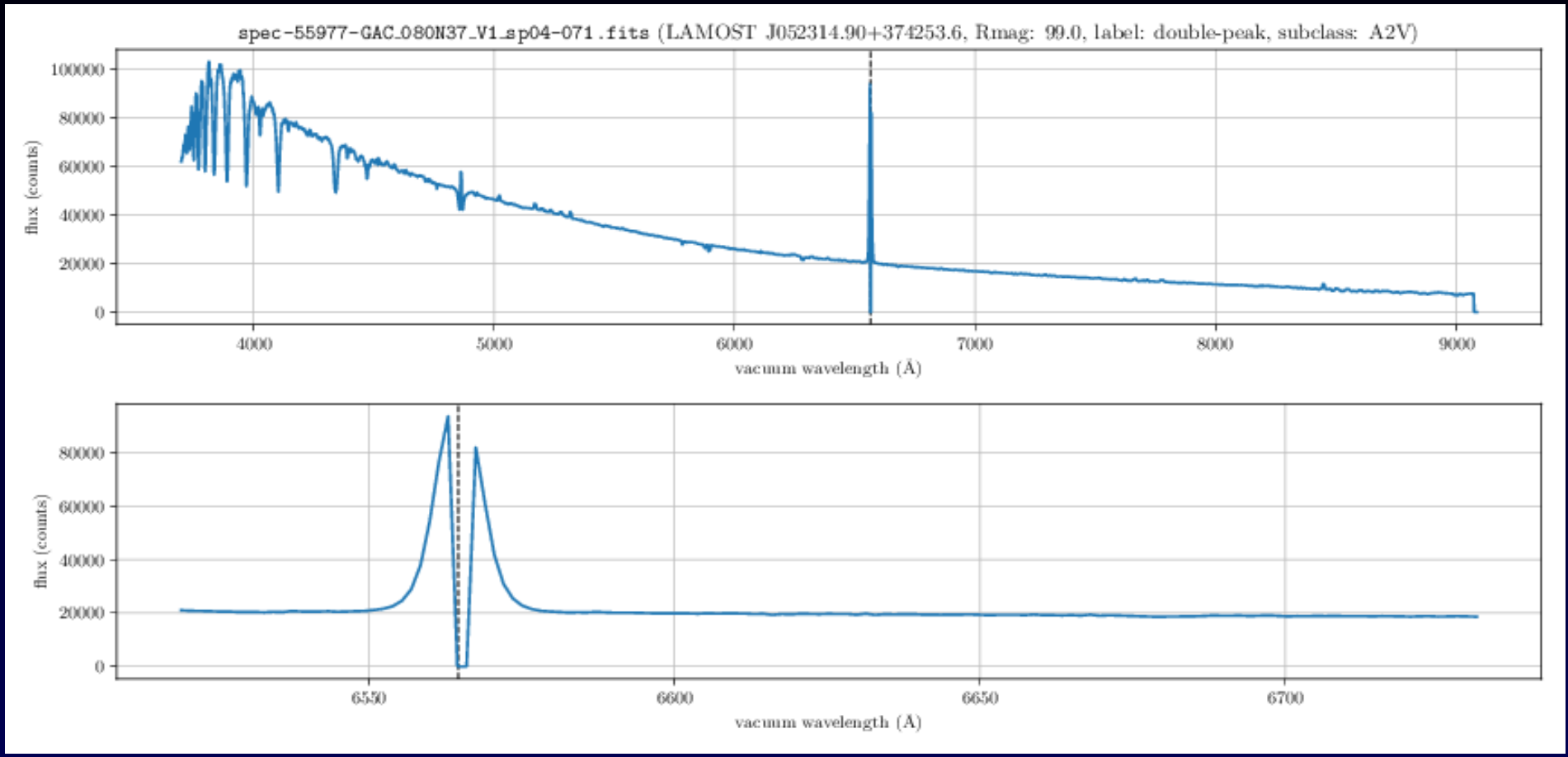
Results – Interesting



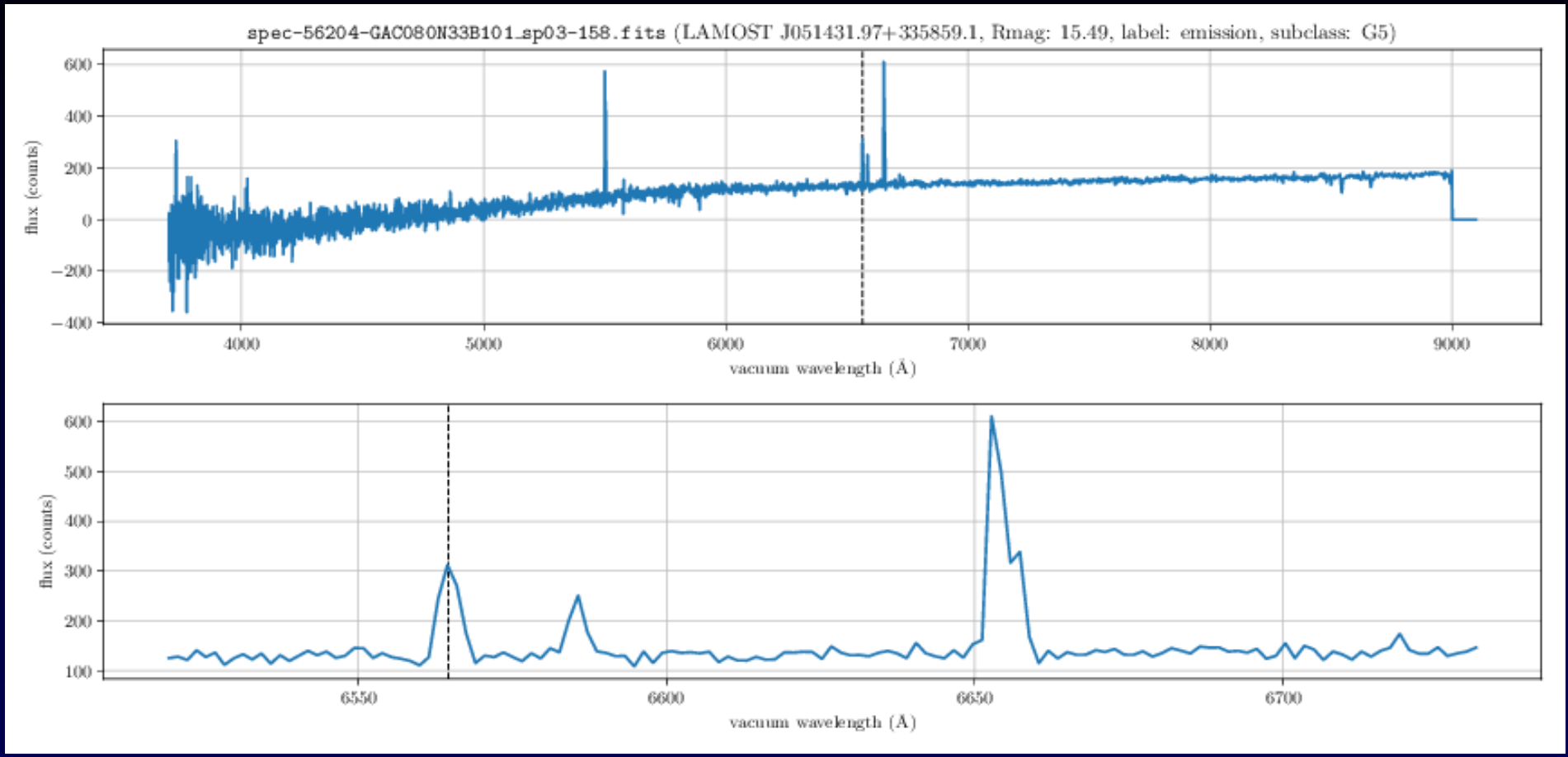
Results – Interesting



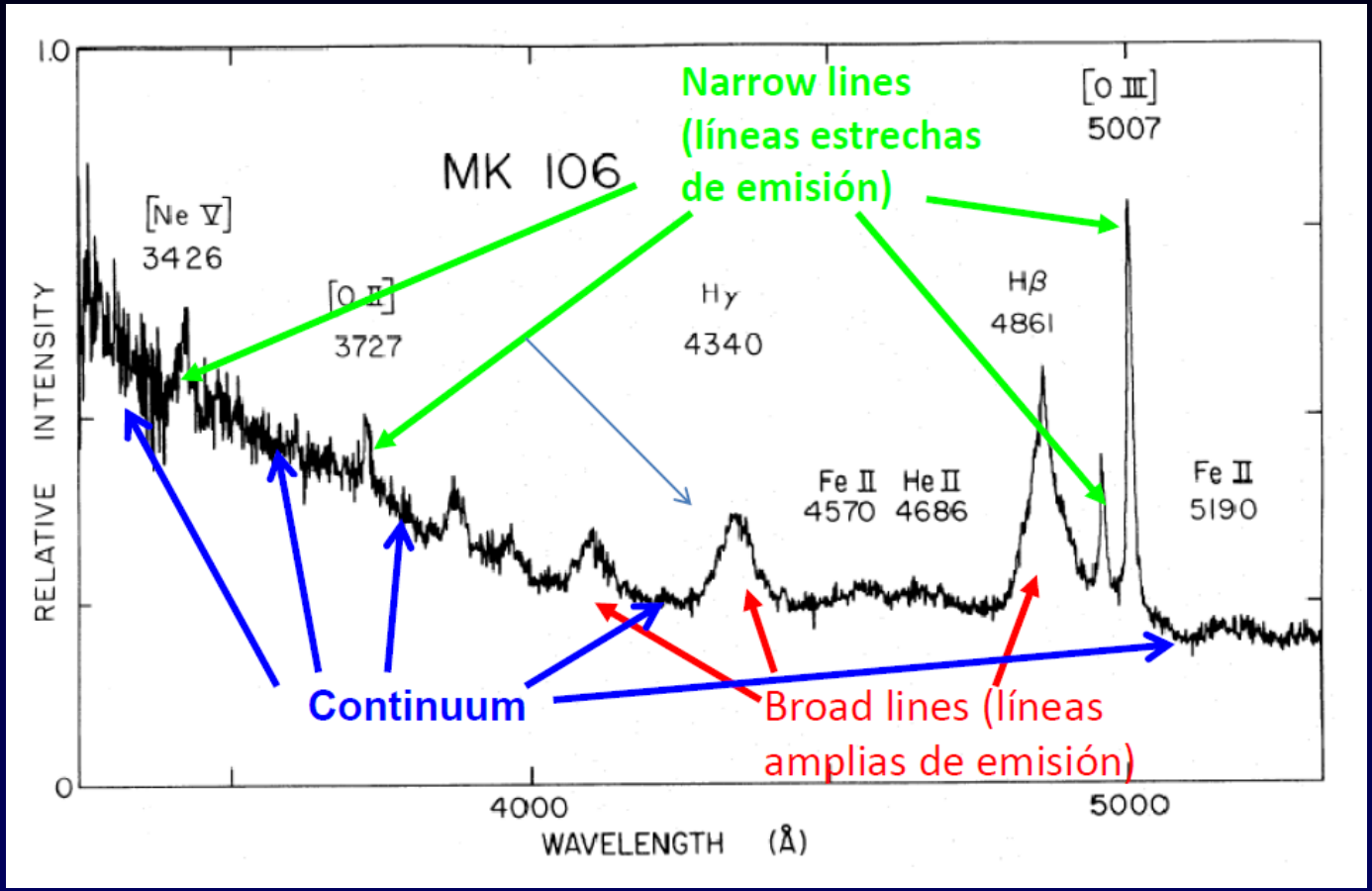
Results – Interesting



Results – Interesting

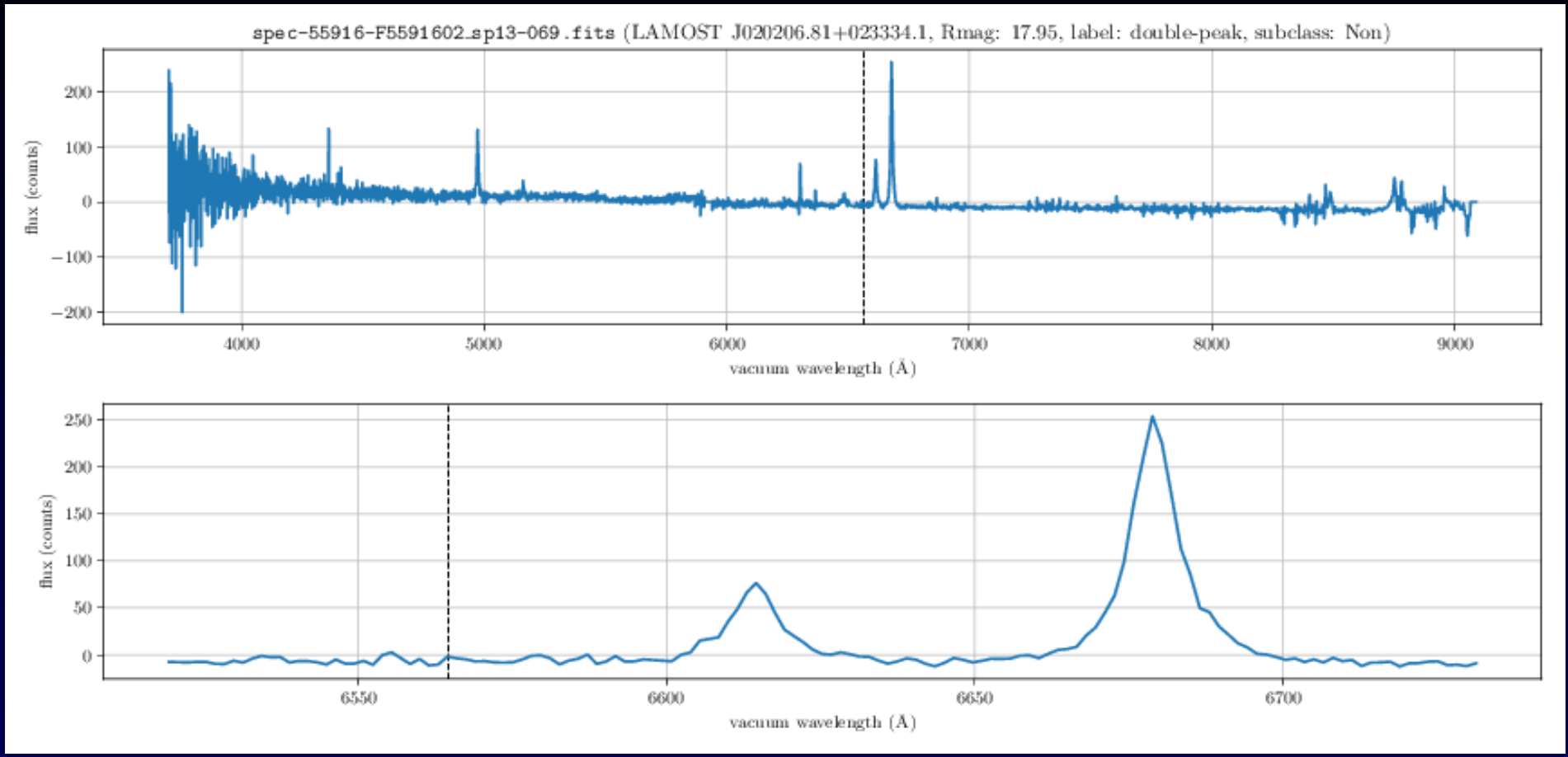


AGN Spectrum

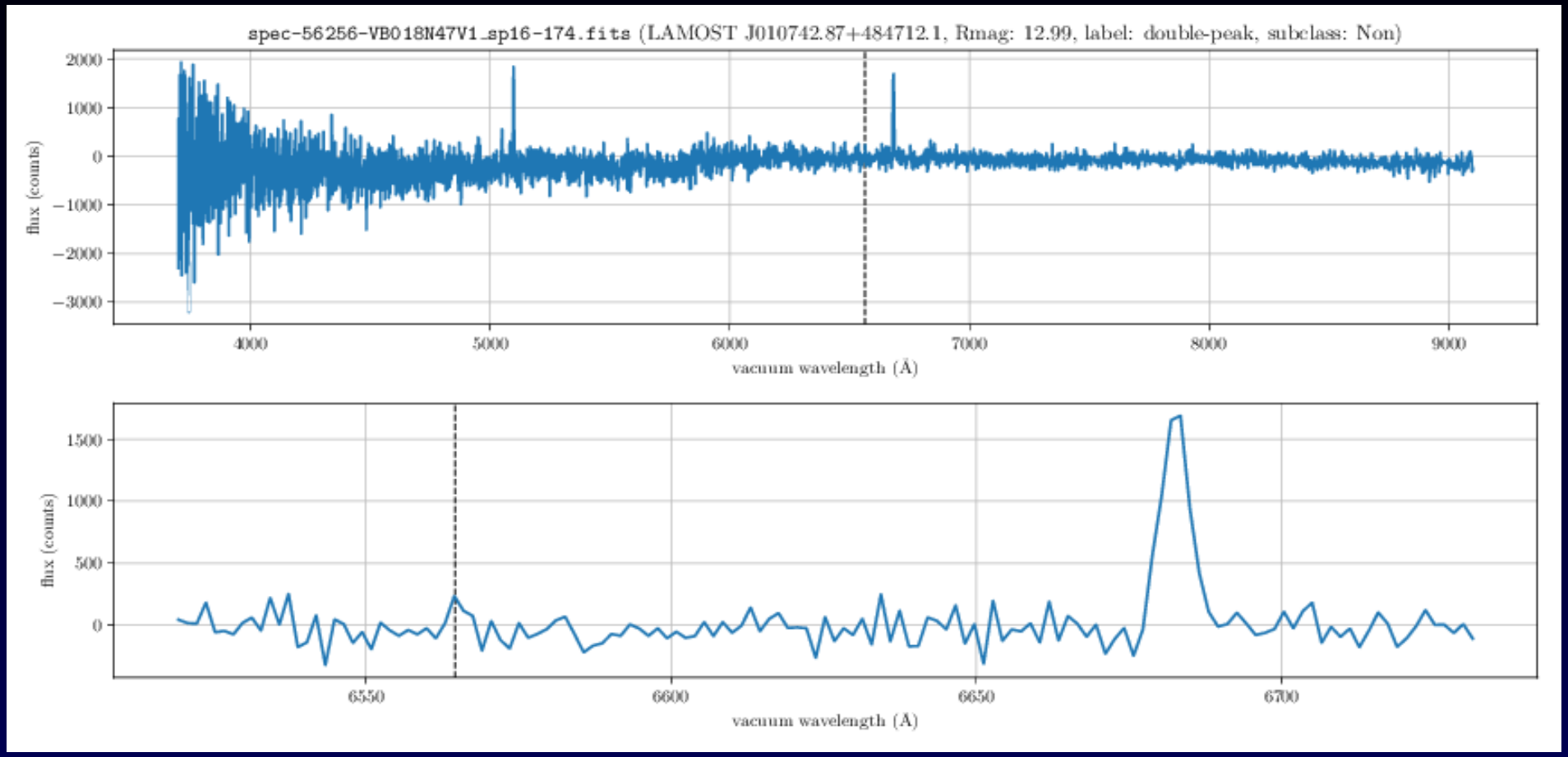


Gaskell 2009

Results – Interesting



Results – Interesting



Conclusions

- Active learning overcomes the lack of labeled data
- ML on big spectra archives may identify new interesting objects yet unknown/unexpected
- Domain adaptation gives LABELS on unlabeled set
- Crucial is interactive visualization of candidates …..
- The VO protocols + tools VERY helpful - XMATCH
- Deep learning shows its strength
- Wavelength invariance of CNN ?? (QSO search)

THANK YOU