A Generative Machine Learning Approach for Molecular Characterization of Astronomical Spectra

International Conference on Machine Learning for Astrophysics (ML4ASTRO3)

University of Malta, 2026

Momentum Program (CSIC) | AI-ITACA Project

Astrobiology Center (CAB, INTA-CSIC)

Gabriel Jaimes-Illanes

Supervisor (PI):

Victor Rivilla Rodriguez

CENTRO DE ASTROBIOLOGÍA · CAB

ASOCIADO AL NASA ASTROBIOLOGY PROGRAM

CSIC

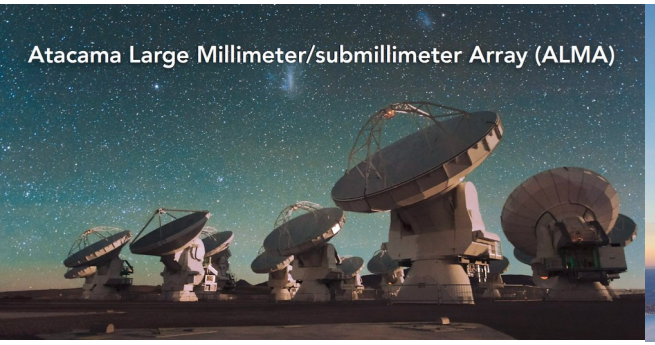
MOMENTUM CSIC

DESARROLLA TU TALENTO DIGITAL

A. Motivation using Observational Data and Species Detection

1. Observational Data and Species Detection

ULTRADEEP ASTRONOMICAL OBSERVATIONS

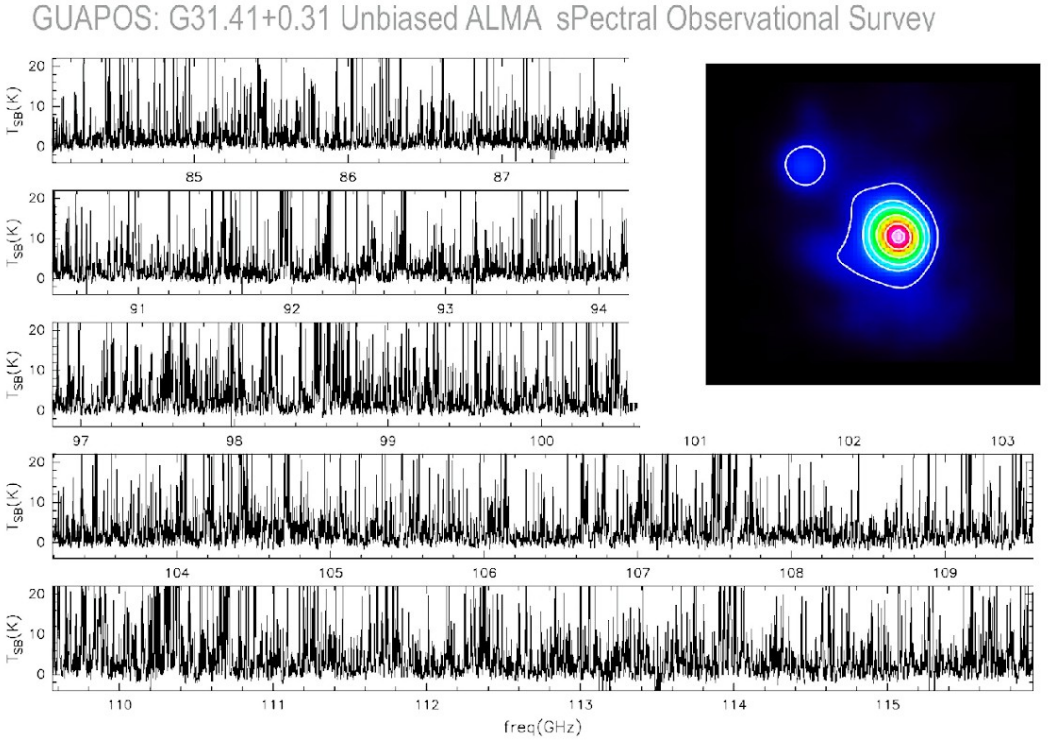


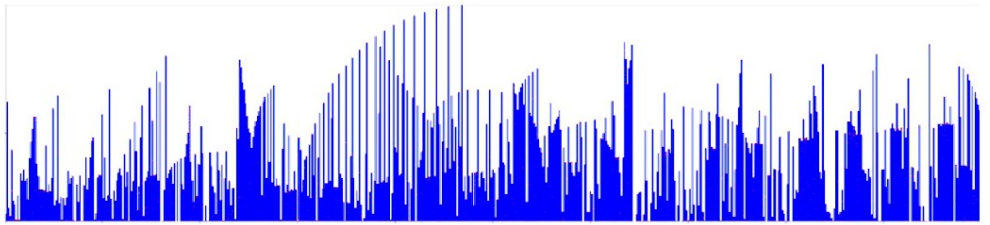
MOLECULAR SPECTROSCOPY IN THE LAB

1. Observational Data and Species Detection

ULTRADEEP ASTRONOMICAL OBSERVATIONS

MOLECULAR SPECTROSCOPY IN THE LAB





B. Using OBSERVATIONAL Spectral Cubes | Multiple VOWEL Dataset

2. MADCUBA | MACROS

Based on script developed by V.Rivilla, J.Allende and G.Jaimes.

MACROS

- Multiple SPW
- Can do Multiple Species and works by tiles of ROI
- *Quantity of Pixels depends on SNR (NOISE)*

FOR CONVERGENCE

Best Fit can be selected by:

- Chi
- DeltaLogn
- DeltaTex

INPUTS:

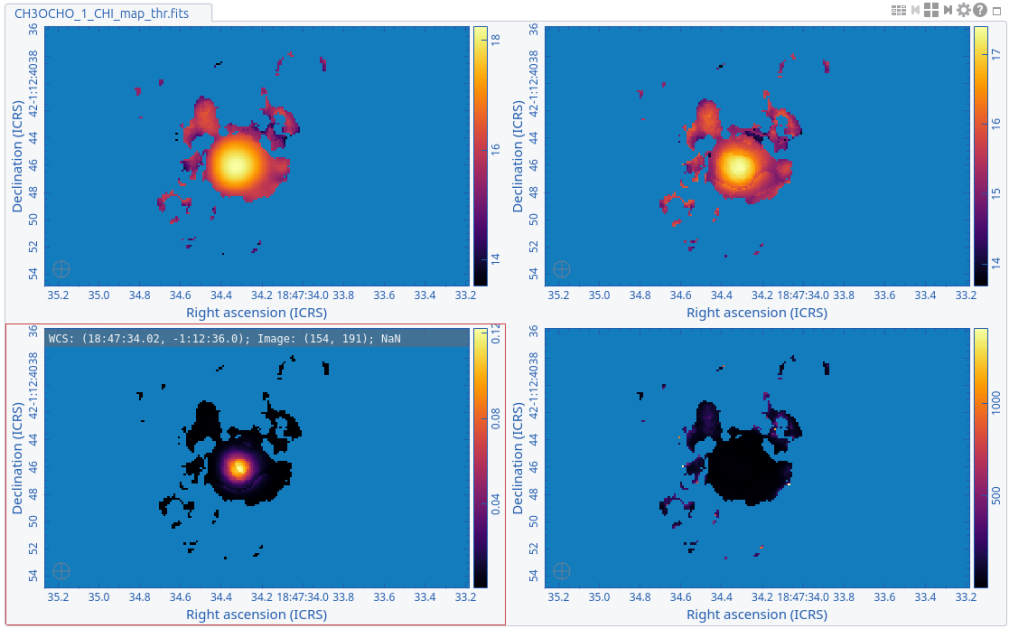
- SPW Cubes (K)
- Mom0,1,2 (K)
- TransitionList.txt
- InitialSeeds.txt

OUTPUTS

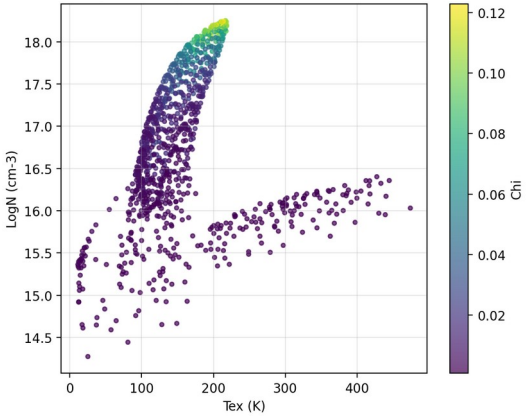
MAPS (.fits):

- LogN
- Tex
- FWHM
- Velo
- CHI

SLIM Container (.slim)



LogN (cm-2), Err LogN (cm-2), Chi and *Tex* (K) Maps (G31.41+0.31)



LogN (cm^{-2}) vs *Chi* Map

TRAINING TRIALS:

- Per SPW
- Per Frequential Region of Interest (ROI)
- Per channel

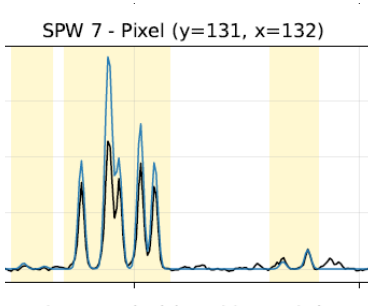
3. ML Pipeline

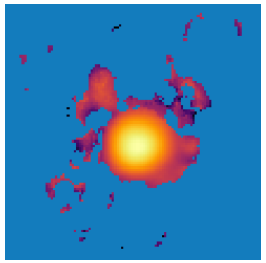
INPUTS:

MADCUBA Macros

1. Signal Components:

- **A. Synthetic**
- **B. Residuals**
  - Noise
  - Artifacts
  - Species





LogN (cm^{-2}) Map (G31.41+0.31)

2. Maps:

1) Logarithmic Column Density
2) Excitation Temperature
3) Full Width Half-Maximum
4) Velocity

ML - MODELLING

Can we generate/emulate a Synthetic+Residuals Spectral Database?

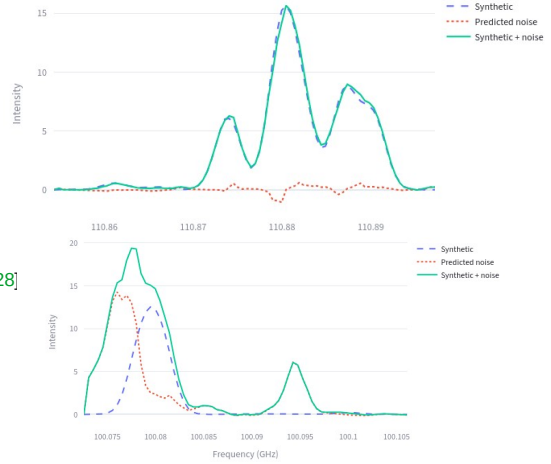
3. Methods:

Generate Synthetic + Residual Spectra → Fitting

1) Decision Tress:
   1) XGBoosting
   2) Gradient Boosting
   3) LightGBM
   4) Random Forest
2) Neural Network (NN)

```
XGB_PARAMS = {
        "n_estimators": 800,
        "learning_rate": 0.03,
        "max_depth": 6,
        "subsample": 0.85,
        "colsample_bytree":
        0.85,
        "reg_lambda": 1.0,
        "n_jobs": -1, }
```

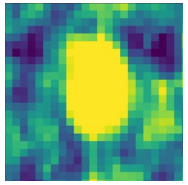
```
HIDDEN_SIZES = [512, 256, 128]
DROPOUT_RATE = 0.2
 LEARNING_RATE = 1e-3
 WEIGHT_DECAY = 1e-6
 BATCH_SIZE = 64
 NUM_EPOCHS = 20
PATIENCE = 5
LOSS_NAME = "huber"
HUBER_DELTA = 1.0
```



4. Optimization:

(60%) Training
(40%) Validation
+ Computing Cluster
+ Bootstrapping

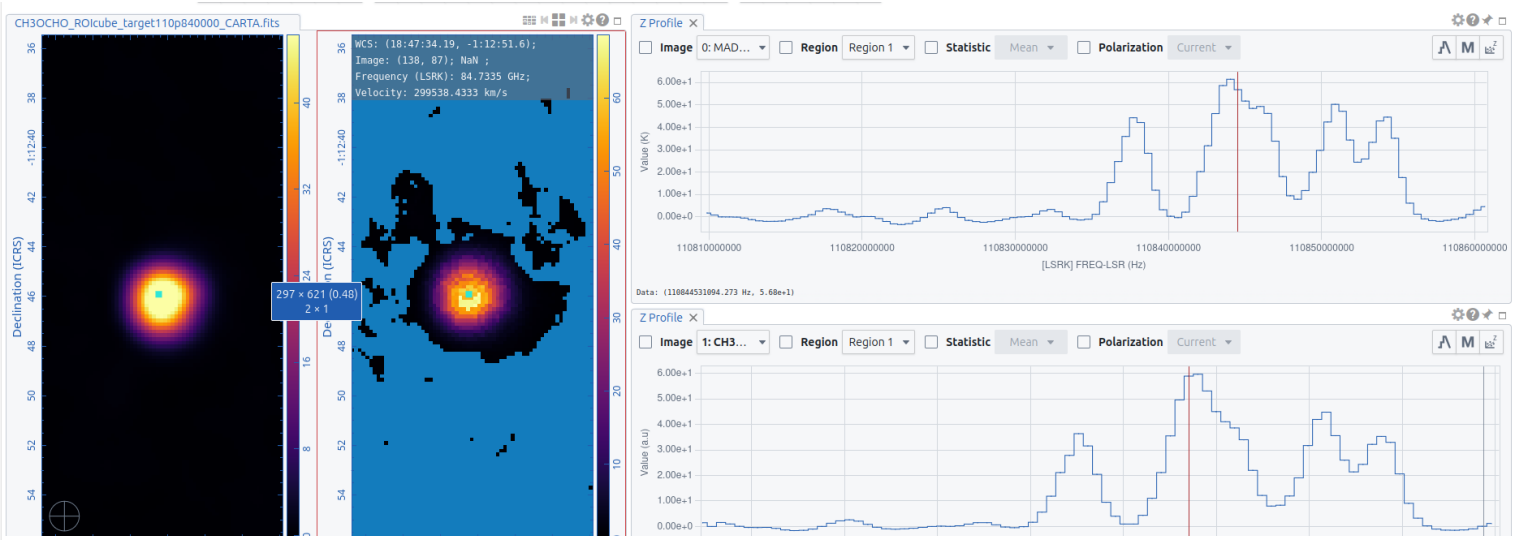
5. Test:

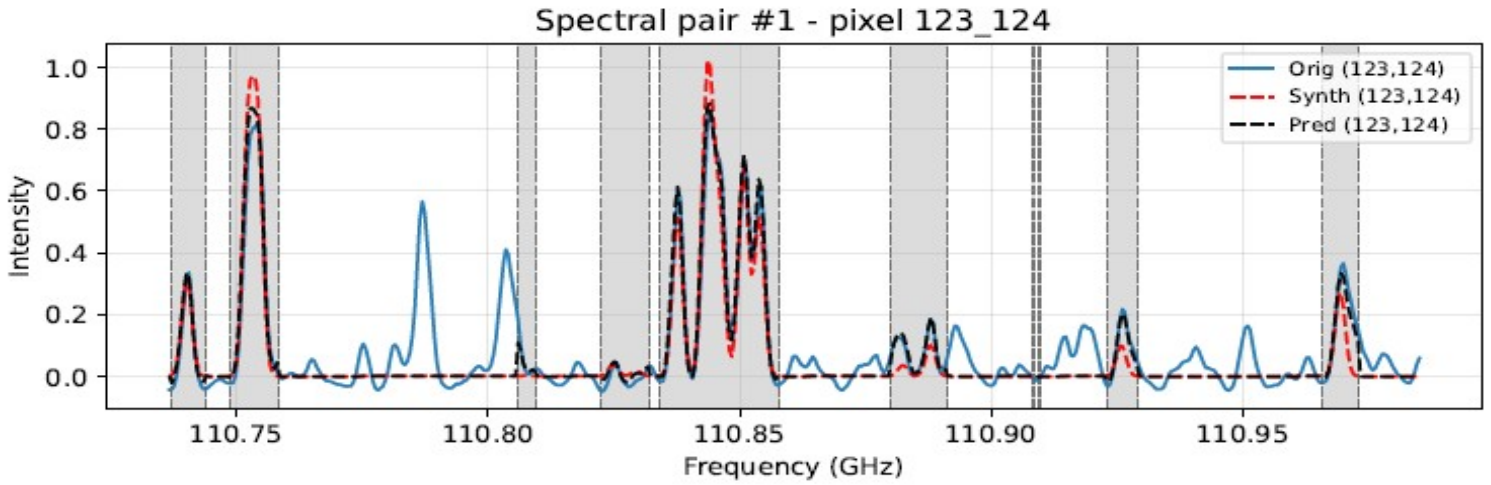


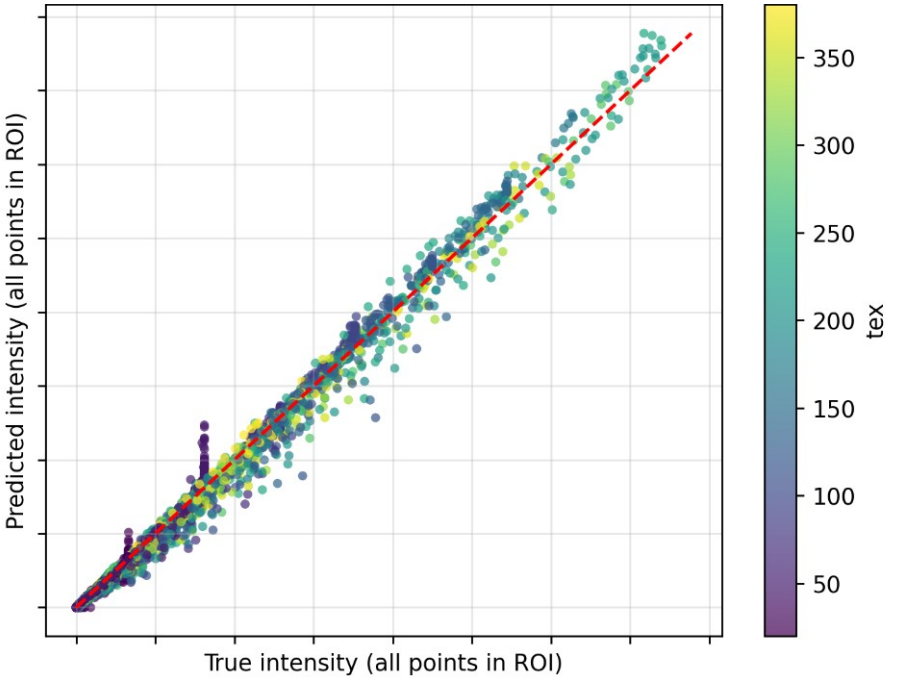
LogN (cm^{-2}) Map (W51)

4. Using OBSERVATIONAL Spectral Cubes

Generation of Synthetic + Residual Spectra vs Real Observation





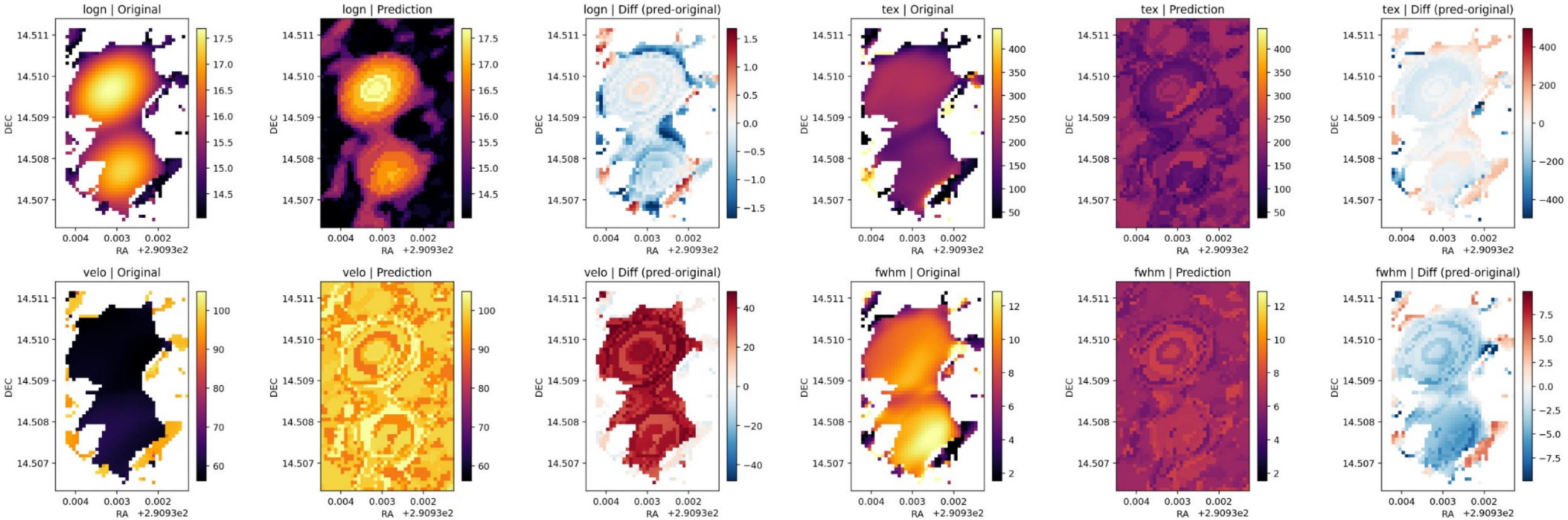


Predicted vs true intensity per ROI in Synthetic Spectra (Tex)

C. Results of Spectrum Characterization with another Source

5. Generation of Synthetic + Residual Spectra + Fitting

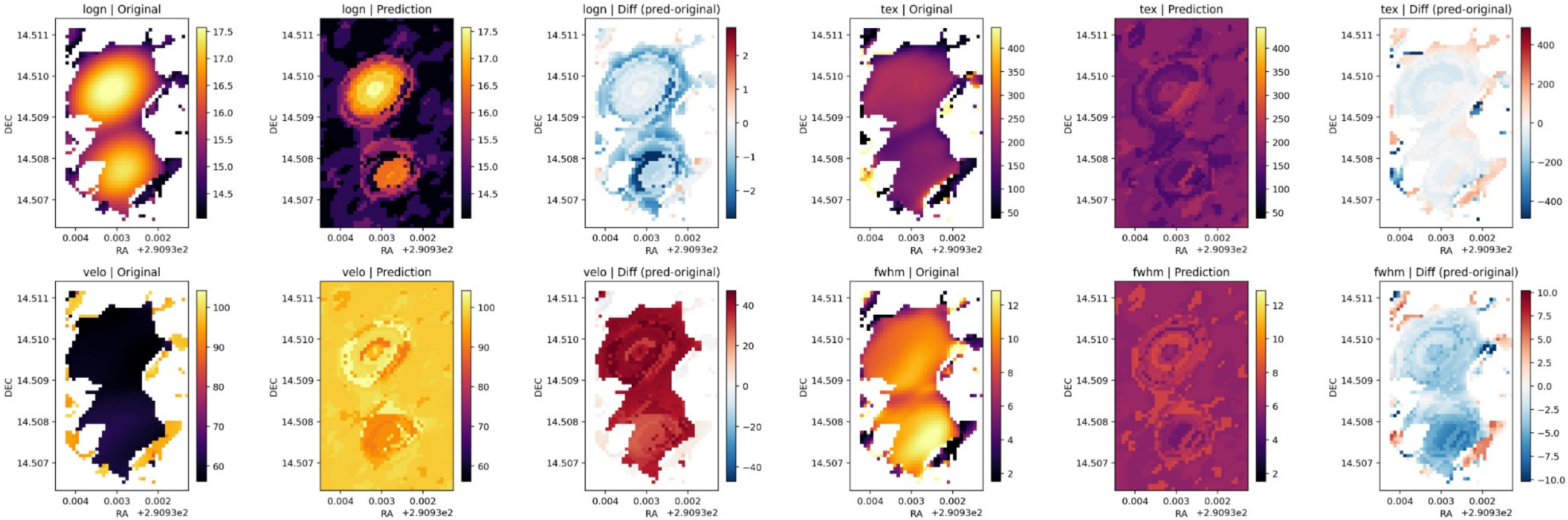
Source: W51 |Specie: C2H5OH | ROIs: 93.5, 93.57, 93.91, 106.93 GHz| Synthetic



Conventional Method | Spec Gen + Fitting | Difference | Conventional Method | Spec Gen + Fitting | Difference

5. Generation of Synthetic + Residual Spectra + Fitting

Source: W51 |Specie: C2H5OH ROIs: 93.5, 93.57, 93.91, 106.93 GHz| Synthetic+Residuals



Conventional Method | Spec Gen + Fitting | Difference | Conventional Method | Spec Gen + Fitting | Difference

“The funding for these actions/grants and contracts comes from the European Union's Recovery and Resilience Facility-Next Generation, in the framework of the General Invitation of the Spanish Government's public business entity Red.es to participate in talent attraction and retention programmes within Investment 4 of Component 19 of the Recovery, Transformation and Resilience Plan”.