MACHINE LEARNING METHODS APPLIED TO INTERSTELLAR MEDIUM STUDIES

Christophe MORISSET - Instituto de Astronomía – UNAM – Mexico
chris.morisset@gmail.com

Introduction / Summary

- I'm mainly presenting here some of my recent applications of Machine Learning techniques in form of **Regressors** (thus no classifications).
- I explore the possibilities offered by Scikit-learn, Tensorflow (under Keras), and XGBoost python libraries, controlled from the AI4Neb library (Morisset et al., 2022, in prep.)

Accelerating PyNeb

- PyNeb is a python library used to analyze emission line spectra, and determine physical and chemical properties of the emitting nebulae.
  Luridiana, Morisset & Shaw 2015 A&A, 573, 42.
- [OIII]4363/5007 and [SII]6731/6716 line ratios are classically used to determine simultaneously the electron temperature and density.
- Obtaining Te, Ne for a given observation is then looking for a minimum in a 2D space.
- May be time consuming (10^{-3} secs, seems small).
- Training a small ANN (2 hidden layers of 10 neurons) to do the job takes a few seconds.
- The ANN can be saved/restored once trained.
- The determination of Te, Ne is really faster (< 10^{-6} secs).
- This is available in PyNeb since v.1.1.13 (8/2020).
- Needs AI4Neb library (on demand to C. Morisset).
- Application to MUSE observations of PNe, with Monte Carlo simulations to determine uncertainties → 10^6 line ratios.
- **Determining millions of (Te, Ne) in a few seconds.**
- See: Garcia-Rojas et al. 2021, MNRAS 510, 5444
- Using PyNeb and the ANN-accelerator in the case of PN NGC 6778:

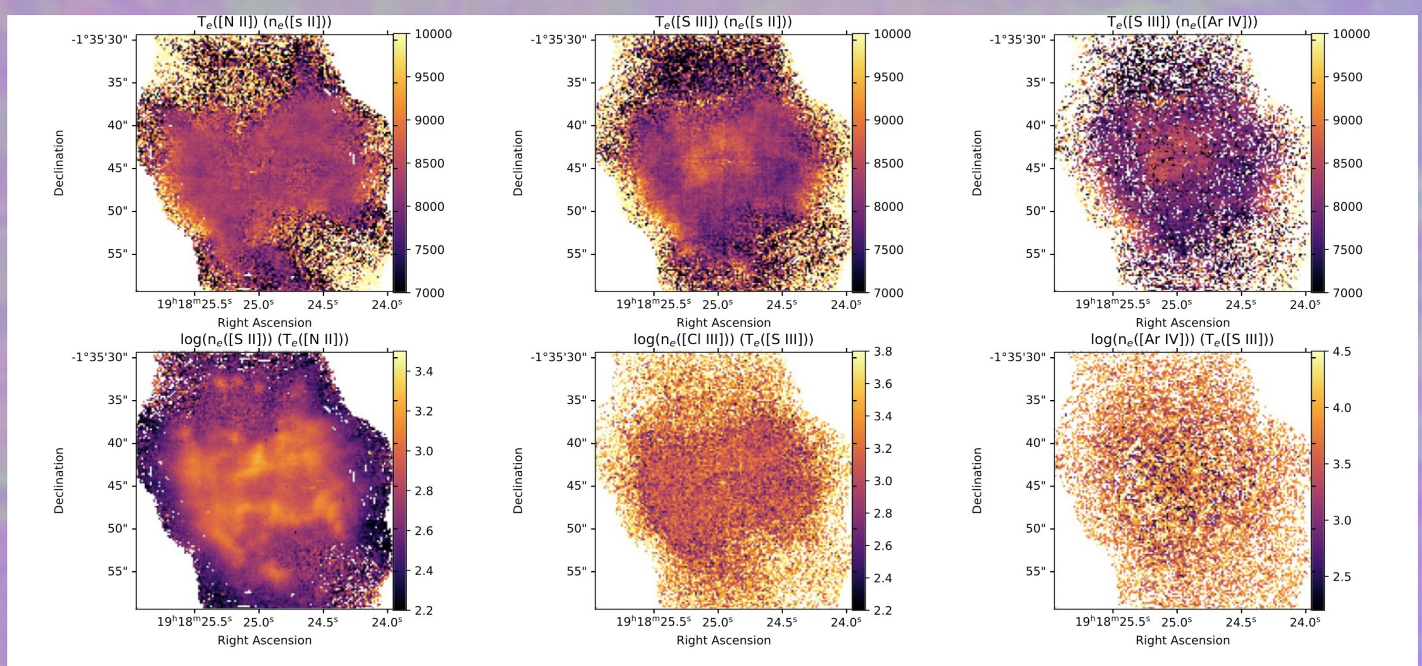
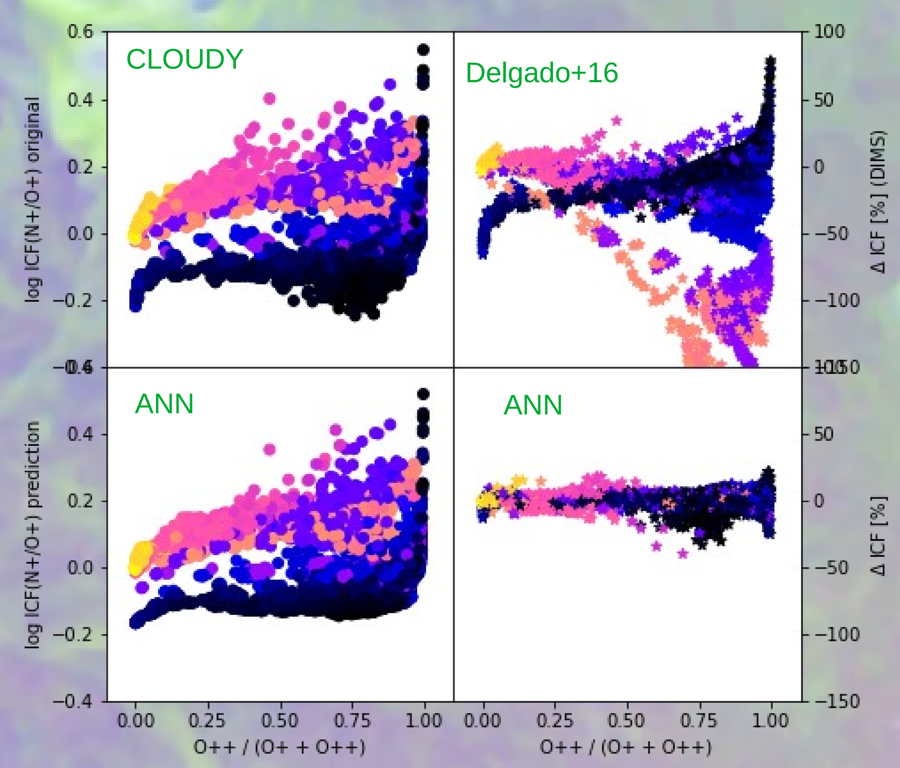


Figure 6. Electron temperature, T_e, and density, n_e, maps obtained from the combination of different temperature and density diagnostics for NGC 6778. The recombination contribution to [N II] λ5755 assuming $T_e = 4000$ K has been considered. The same maps for M 1–42 and Hf 2–2 can be found in Figs S7 and S8 of the supplementary material.

Ad-hoc determination of ICFs

- Ionization Correction Factors are used to compute the total abundance of an element taking into account the unsee ions.
- They are obtained from grid of photoionization models, and based on the previous determination of some ionic fractions.
- Example: $\frac{N}{H} = \frac{N^+}{H^+} \times ICF(N^+) = \frac{N^+}{H^+} \times \frac{O}{O^+}$
- Widely used ICFs:
  - Kingsburg & Barlow, 1994, from 10 models.
  - Delgado-Inglada, Morisset & Stasinska, 2014, from 2800 models. The models are extracted from the Mexican Million Models dataBase (3MdB, Morisset+15).
- We are now defining ad-hoc ICFs computing using Machine Learning method.
- More efficient: the RMS of the difference between the prediction and the true value is 0.1 in the case of DIMS14 formula, and 0.016 using an ANN:



ICFs from emission lines

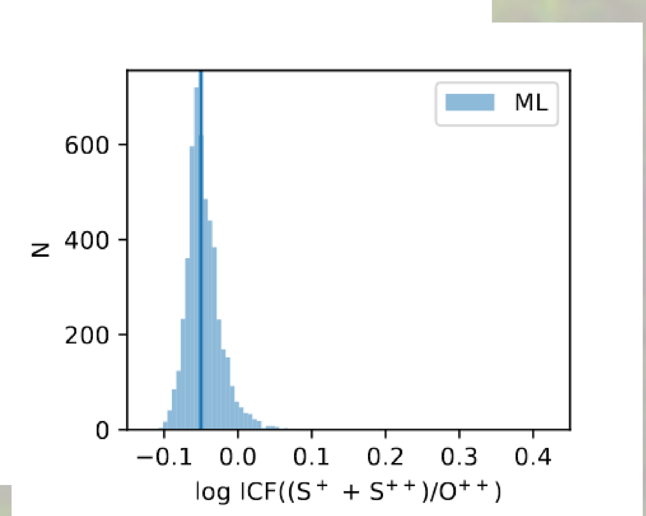
- In the case of the PN PC22, we determine 11 ICFs from 6 line ratios, using a ML method based on XGBoost.
- A Te-sensitive line ratio have been added to connect emissivities and abundances.
- Trained with 16,000 models extracted from 3MdB and "close" to PC22.
- We obtain new ICFs related to O^{++}, for this high ionization PN. They are more reliable than when based on the residual ion O^+.

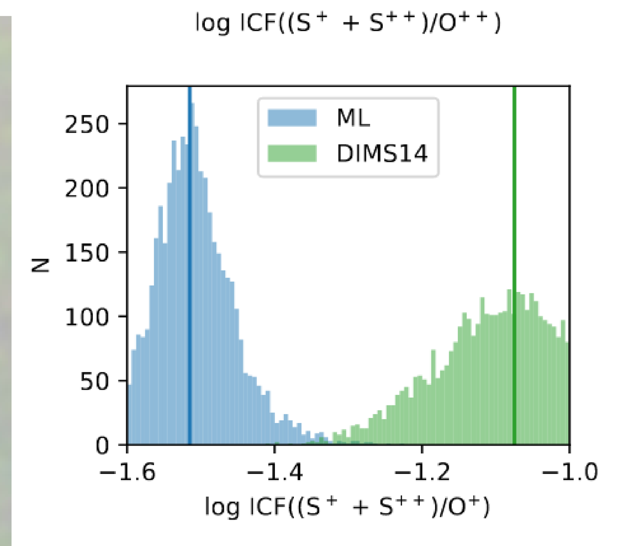
The input vector X is build from a 6D vector of the logarithmic values of the following line ratios:

- He II λ4686 / He I λ5876
- [O III] λ5007 / [O II] λ3727
- [Ne V] λλ3426, 3346 / [Ne IV] λ4726
- [Ne IV] λ4726 / [Ne III] λ3869
- [Ar V] λ6435 / [Ar IV] λλ4711, 4740
- [O III] λλ4363/5007

The output vector y is directly the set of the following ICFs (logarithmic values are used):

- $O / (O^+ + O^{++})$
- $N/O \times O^+ / N^+$
- $Ne / (Ne^{++} + Ne^{4+})$
- $Ne / (Ne^{++} + Ne^{3+} + Ne^{4+})$
- $Ne / O \times O^{++} / Ne^{++}$
- $S / (S^+ + S^{++})$
- $S / O \times O^+ / (S^+ + S^{++})$
- $S / O \times O^{++} / (S^+ + S^{++})$
- $Cl / O \times O^+ / Cl^{++}$
- $Cl / O \times O^{++} / Cl^{++}$
- $Ar / (Ar^{3+} + Ar^{4+})$

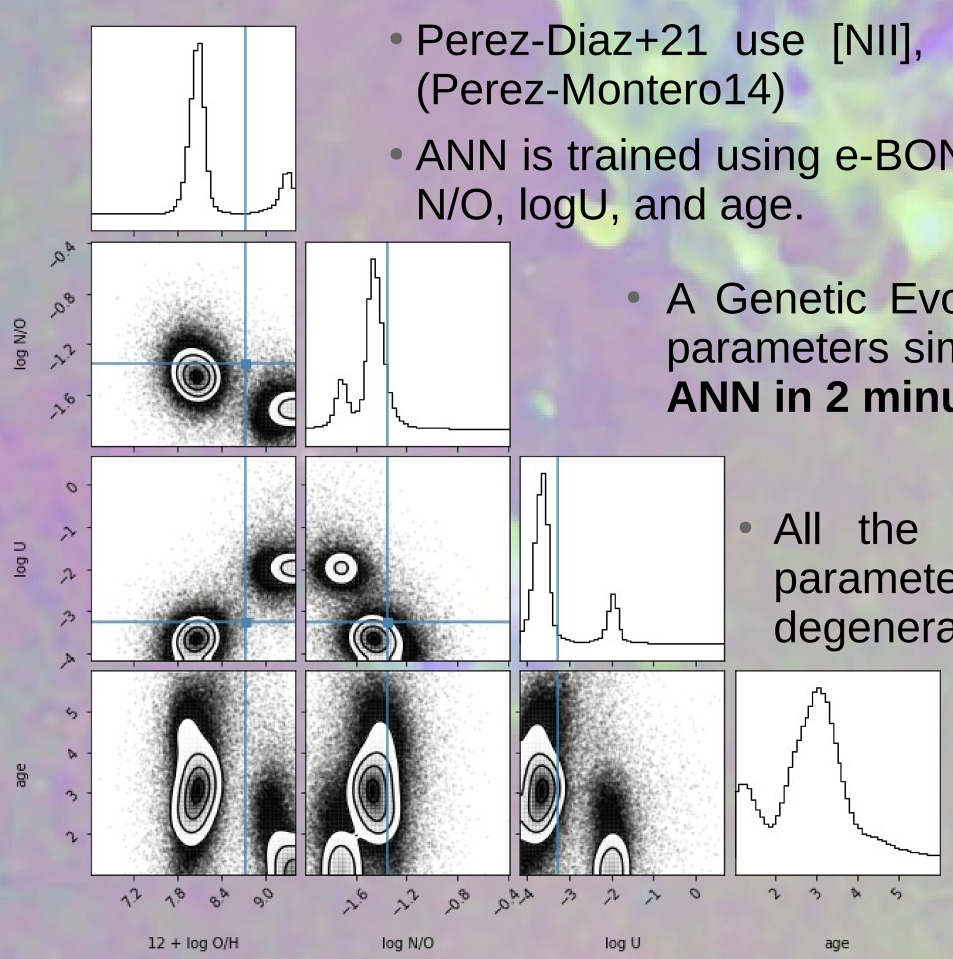




See Sabin et al 2022, MNRAS, 511, 1

HII regions: looking for best solutionS

- Looking for the sets of model parameters that leads to reproducing **simultaneously** the observed line intensities of a given object.
- We want **all the solutions**, not only the "best" one (which anyway would require a clear definition).
- **Inverse problem**: the model gives intensities from physical parameters. We want to determine the parameters from line intensities.
- Genetic method, MCMC, or even brute-force can be used. Each needs to run a lot of models: the **ANN replaces Cloudy: forward problem → very efficient**.
  - Perez-Diaz+21 use [NII], [OIII] and [SII] to determine O/H running HII_CHI_m (Perez-Montero14)
  - ANN is trained using e-BOND models from 3MdB to predict those lines, giving O/H, N/O, logU, and age.



- A Genetic Evolution model uses this ANN to look for the sets of parameters simultaneously fitting one given model. **370,000 calls to ANN in 2 minutes.**
- All the points in the contours correspond to values of parameters leading to reasonable fit to the observed data → degeneracy of O/H.
  - The "Best Model" is a meaningless concept.
  - The "weighted mean value" is rather risky.
  - Morisset et al. In prep.

Grant : UNAM/DGAPA/PAPIIT IN101220