Massive Stars Across Redshifts in the Era of JWST and Large Surveys
Proceedings IAU Symposium No. 402, 2026
A. Wofford, N. St Louis, M. García, S. Simón-Díaz & L. M. Oskinova, eds.
doi:10.1017/S1743921326106012

Building galaxies from basic stones and Neural Networks

Christophe Morisset[1,2] and Stéphane Charlot[3]

[1]Instituto de Astronomía (IA), Universidad Nacional Autónoma de México, Apdo. postal 106, C.P. 22800 Ensenada, Baja California, México
email: chris.morisset@gmail.com

[2]Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Av. Universidad s/n, 62210 Cuernavaca, Mor., México

[3]Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, F-75014, Paris, France

Abstract. Emission lines in the combined spectrum of star-forming galaxies offer important insights into the physical and chemical characteristics of the associated H II regions and the host galaxy itself. Classical photoionization models (using the CLOUDY code Ferland et al. 2017) for integrated galaxies are created by combining the entire stellar population of the galaxy to ionize a single cloud. Calculating a model for each H II region is highly time-intensive: modeling 250,000 galaxies, each with 1,000 H II regions, would take over 1,000 CPU-years! The objective is to develop an Artificial Neural Network (ANN) to replace Cloudy, enabling much faster modeling of H II regions, and allowing generating models of integrated galaxies by summing the emission of individual H II region in a reasonable time scale.

Keywords. HII regions, galaxy models, photoionization models, machine learning.

1. Methods and tools

1.1. *Irregular grid of CLOUDY models*

We compute a grid of 3,500,000 photoionization models for H II regions using the multi-purpose photoionization code CLOUDY in its version 17.03. The input space-parameter is of dimension 13: log(U), O/H, N/O, C/O, Ne/O, S/O, Ar/O, ionizing cluster age, shape-factor, hydrogen density, optical depth, dust content, filling factor. The ionizing SED is defined by BPASS binary-including library (Eldridge et al. 2017; Stanway & Eldridge 2018). For extremely large grids, we adopt a practical approach involving using randomly-distributed independent parameters, which offers an efficient way to populate datasets to train ML algorithms. The models are stored under reference HII_24 in 3MdB_17, which is a large database of photoionization models, with public access through the mySQL protocol since 2014 (http://3mdb.astro.unam.mx, see Morisset et al. 2015).

1.2. *CLOUDY emulator using ANN*

A subset of 800,000 models is defined to train the ANN. The ANN is implemented through the Tensorflow python library (http://tensorflow.org). The ANN of the 256-512-256 GELU neurons (Hendrycks & Gimpel 2016) takes 13 input parameters and predicts 14 emission line ratios + electron temperature + Hα absolute luminosity. The training time is 2 hours on a standard GPU. The ANN allows us to generate 1 million

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1743921326106012 Published online by Cambridge University Press

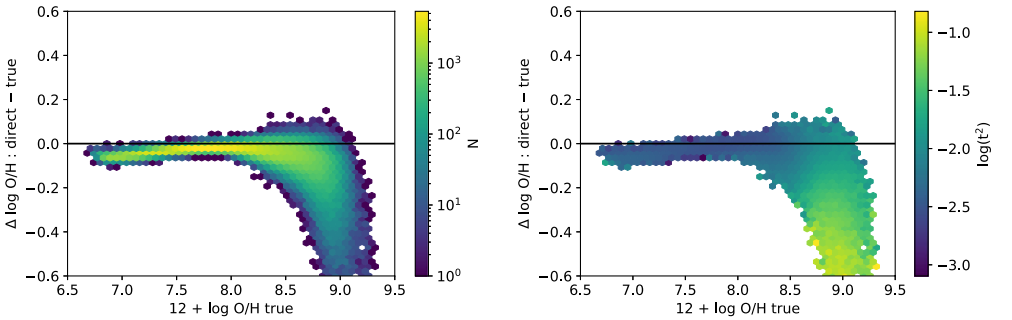


Figure 1. Distribution of the logarithmic difference between O/H derived from the direct method and true, Hα-weighted O/H as a function of 12 + log (O/H), for the subset of 207 750 composite star-forming galaxies with log([O III]λ4363/Hβ) > -3. Left: Colour-coded as a function of model number density. Right: Colour-coded as a function of root-mean-square t of H II-region temperature fluctuations. Taken from Morisset et al. (2015).

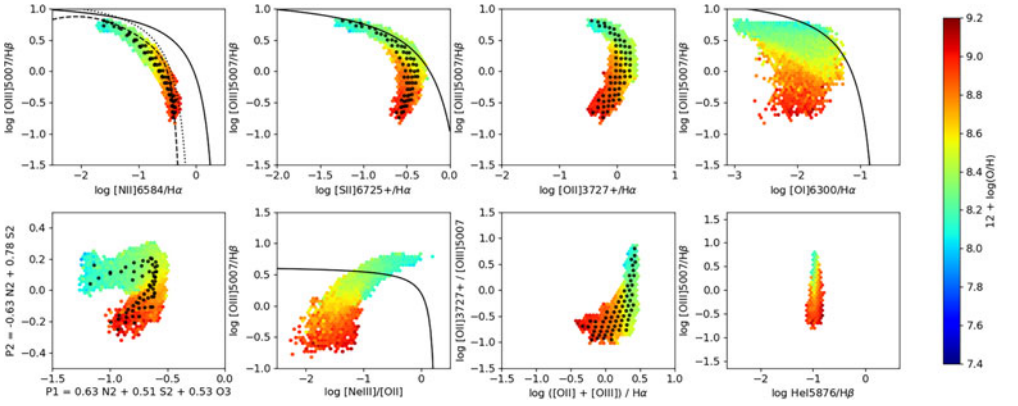


Figure 2. Plots for BPT-type diagrams for the results of best SF galaxies models. Black dots are from Curti et al. (2017). Plots are taken from Morisset et al. in prep.

models in 0.2 seconds (!). The ANN and the first results are described in paper I (Morisset et al. 2025) : `https://ui.adsabs.harvard.edu/abs/2025MNRAS.538.1884M`

2. Some results

250,000 galaxy models are produced containing between 100 and 3000 H II regions each. Global DIG emission and dust attenuation (individually defined for each H II region) are taken into account. In Fig. 1 we show that the direct-method determination of O/H in metal-rich H II regions and galaxies can strongly underestimate the true value. This is due to the combination of: 1) the contamination of [O III]λ4363 by recombination, 2) the natural electron temperature gradient in cold regions, described by Stasińska (2005), and 3) the effect of phosphorescence contamination of [O III]λ4363, see Morisset et al. in prep.

Models are compared to observations: some galaxy models are selected if they reproduce simultaneously [N II]/Hβ, [O II]/Hβ, [O III]/Hβ, and [S II]/Hβ. He I/Hβ, and [Ar III]/Hβ can also be used if present in the observational set. These models are shown in Fig. 2, in the case of fitting the stacked observations from Curti et al. (2017). The selected galaxies can then be plotted in Strong Line (SL) calibrators vs. O/H diagrams. In Fig. 3 we show the effect of changing the N/O vs. O/H relationship at the level of

https://doi.org/10.1017/S1743921326106012 Published online by Cambridge University Press

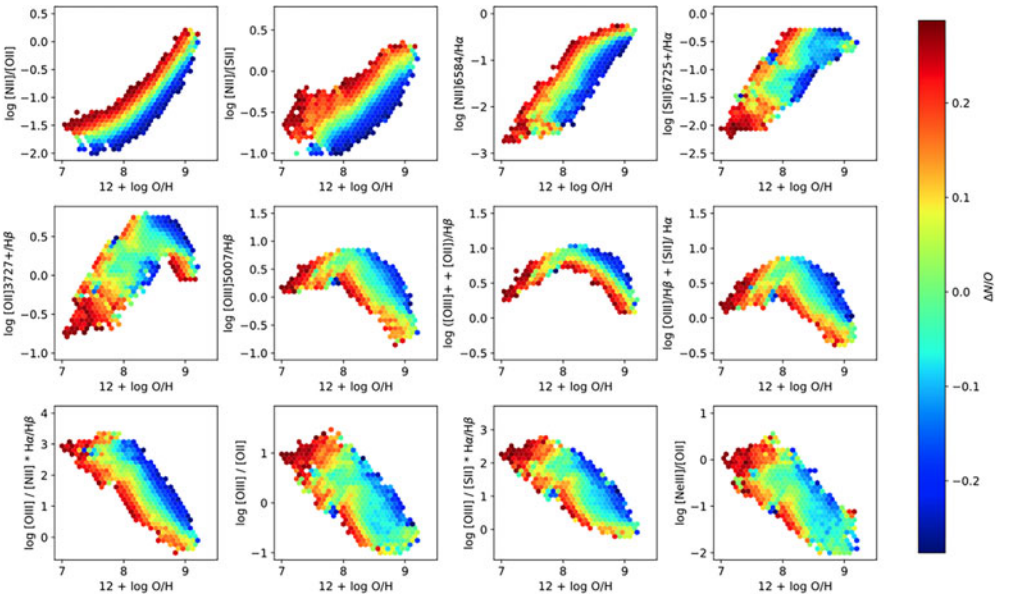


Figure 3. Strong line calibrators vs. oxygen abundance for the SF synthetic galaxies selected by fitting Curti et al. (2017) observations. The color code follows the departure from canonical N/O vs. O/H relationship. Plots are taken from Morisset et al. in prep.

the galaxy models. The scatter observed in the SL calibrators can be identified with real dispersions in the relationships between the different parameters, here N/O vs. O/H, but also log(U) vs. O/H, among others.

3. Conclusions and future work

ANN is a powerful tool to create very fast `CLOUDY` ersatz. Synthetic galaxy models obtained by summing individual H II region models allow us to explore the effects of parameter scatters. Stay tuned for paper II. The ANN can also be used for any massive production of H II region models, e.g. in an MCMC method or a Genetic algorithm, to derive physical/chemical parameters from observed line ratios. The results will soon be updated using a grid of models using new `CLOUDY` c25.00.

Acknowledgments

The initial stages of this work were conceived and developed during a CM sabbatical at the Institut d'Astrophysique de Paris (IAP) to collaborate with SC. CM gratefully acknowledges the PASPA grant that supported this sabbatical. CM acknowledges the support of UNAM/DGAPA/PAPIIT grants IN101220 and IG101223.

References

Curti M., Cresci G., Mannucci F., Marconi A., Maiolino R., Esposito S., 2017, Mon. Not. R. Astron. Soc., 465, 1384

Eldridge J. J., Stanway E. R., Xiao L., McClelland L. A. S., Taylor G., Ng M., Greis S. M. L., Bray J. C., 2017, Publ. Astron. Soc. Aust., 34, e058

Ferland G. J., et al., 2017, Rev. Mexicana Astron. Astrofis., 53, 385

Hendrycks D., Gimpel K., 2016, arXiv e-prints, p. arXiv:1606.08415

https://doi.org/10.1017/S1743921326106012 Published online by Cambridge University Press

Morisset C., Delgado-Inglada G., Flores-Fajardo N., 2015, Rev. Mexicana Astron. Astrofis., 51, 103
Morisset C., Charlot S., Sánchez S. F., Espinosa-Ponce C., Barat E., Dautremer T., 2025, Mon. Not. R. Astron. Soc., 538, 1884
Stanway E. R., Eldridge J. J., 2018, Mon. Not. R. Astron. Soc., 479, 75
Stasińska G., 2005, Astron. Astrophys., 434, 507

https://doi.org/10.1017/S1743921326106012 Published online by Cambridge University Press