Using Machine Learning to Understand Cosmic Gas Evolution

Rogelio Ochoa
Bryson Stemock
Alexander Stone-Martinez
rochoa9@nmsu.edu
bstemock@nmsu.edu
stonemaa@nmsu.edu
New Mexico State University Dept. of Astronomy
Las Cruces, New Mexico

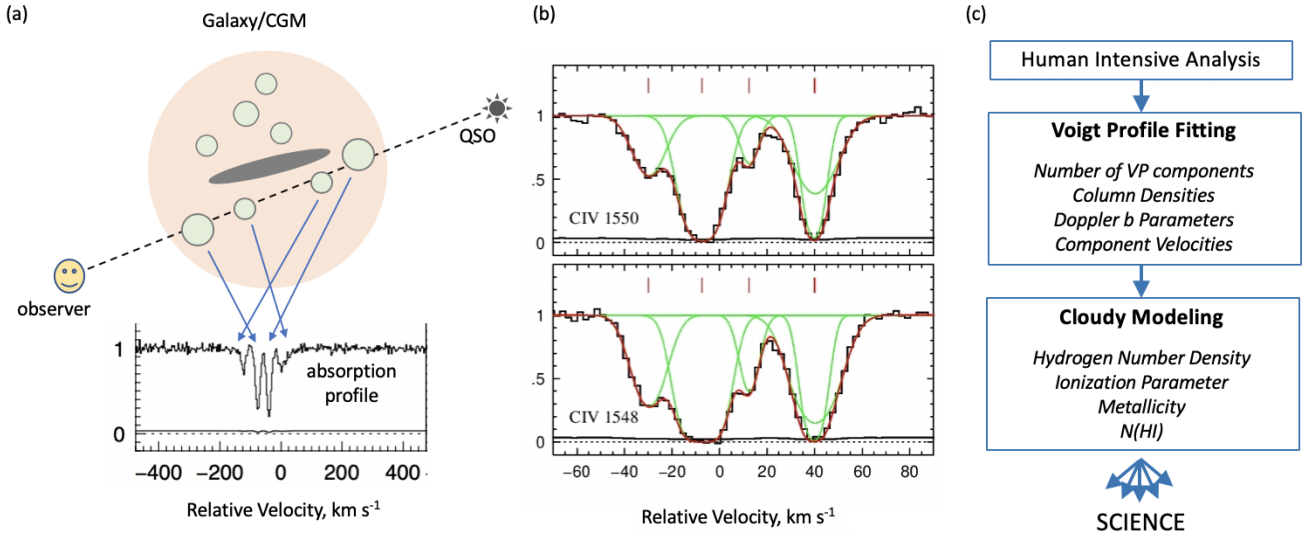

Figure 1: (a) Schematic of a quasar line of sight (dotted line) showing absorption from four distinct CGM clouds. The conventional interpretation: each absorption component corresponds to a cloud. (b) The Voigt profile (VP) model (red curve) of observed CIV $\lambda\lambda 1548, 1550$ doublet absorption profiles. Each VP component is shown as a green curves. (c) An analysis-to-science flowchart: VP fitting provides column densities that constrain ionization models, which yield physical quantities (in italics) for the science.

ABSTRACT

This project represents a small step in Bryson Stemock's Ph.D. thesis, to which Alexander Stone-Martinez and Rogelio Ochoa have contributed a significant amount already. The project focuses on an intermediate step in the thesis, the Voigt profile fitting of absorption line spectra. Synthetic spectra (i.e. the training and test sets) have been simulated already by Stemock. The goal of the project is to accurately predict physical parameters associated with absorption line spectra containing one or two clouds.

CS 519, Spring 2020, New Mexico State University

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

datasets, neural networks

ACM Reference Format:
Rogelio Ochoa, Bryson Stemock, and Alexander Stone-Martinez. 2020. Using Machine Learning to Understand Cosmic Gas Evolution. In *Proceedings of ACM Conference (CS 519).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

Since 1962, quasars (highly luminous accreting supermassive black holes in distant galaxies) have been used to probe the cosmos

[2]. When light from a quasar passes through a gas structure, that gas leaves an imprint on the spectrum of the quasar. This imprint is called an absorption line system. Detailed analysis of these absorption line systems across cosmic time provide direct insight into the physics that governs the universe and how both its large-scale structure and the galaxies that inhabit it have evolved.

In order to extract gas properties from absorption line systems, one traditionally fits the data with Voigt profiles and then uses a Markov chain Monte Carlo (MCMC) method to model the gas. This process is illustrated in Figure 1 (borrowed from our proposal to the National Science Foundation). During the process, simplifying assumptions and human subjectivity creep in, introducing some level of bias into this line of work. In addition, the typical rate at which a given researcher is able to complete this process for a single absorption line system is around 1-2 systems per week. At this rate, it is estimated that our current sample of approximately 3,500 systems would require around **70 years** of human effort.

One group member, Bryson Stemock, has taken on as his Ph.D. thesis the task of automating this process through the design, training, and implementationg of a convolutional neural network (CNN). As part of a larger group within the NMSU Astronomy Department, the remaining group members, Alexander Stone-Martinez and Rogelio Ochoa, have contributed greatly to the progress of the overall project, primarily through the design of various CNNs. The overall project goal is to create a globally-available tool that will vastly accelerate the analysis of absorption line systems and, therefore, of the evolution of the universe.

2 THE PROBLEM

Clearly, the entirety of this project is too much to accomplish in a single semester, especially with a variety of data-related subtleties that were not mentioned above which arise from the complexity of the gas we observe (and simulate). Therefore, only a minute, refined portion of the overall thesis will be tackled here. The data, which will be explored more thoroughly in Section 4 of this report, consist of one- and two-cloud absorption line systems with parameters (which were used to generate the spectra) drawn using Latin Hypercube Sampling. Each instance consists of two spectra, displayed in Figure 2, one for the MgII2796 transition and one for the MgII2803 transition. Our goal is to design a CNN that can accurately (better than R2 = 0.90) predict the log of column density, the Doppler b parameter, and the velocity position of each cloud (i.e. the output parameters from Voigt Profile Fitting).

One obstacle that will need to be addressed is the occurrence of "blending". Blending is the overlapping of absorption lines and usually refers to an extraneous absorption line impeding on a separate system. However in our case, since the CNN will be searching for two lines and may see one large line (example in Figure 3), we may run into an issue.

3 THE SOLUTION

As was mentioned in Section 2, we are designing a CNN that will take input (simulated and noiseless) absorption line systems and return the input parameters used to simulate the spectrum in the first place. B. Stemock will simulate spectra for training and testing using the code specsynth [1], which was developed and is

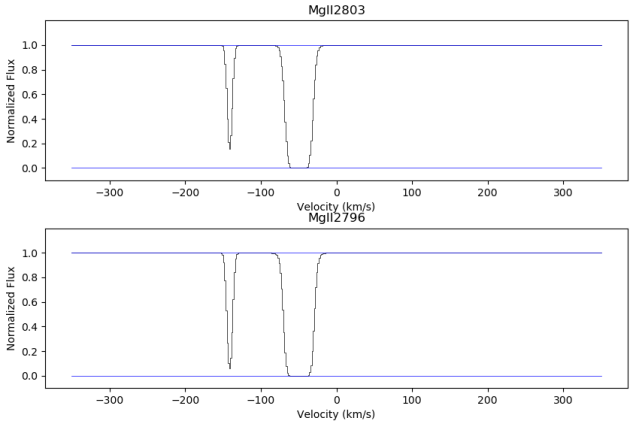

Figure 2: A sample two-cloud absorption line system. The only ions present here are from the MgII doublet at $\lambda\lambda$2796, 2803. Note that no noise is present in the data for this project.

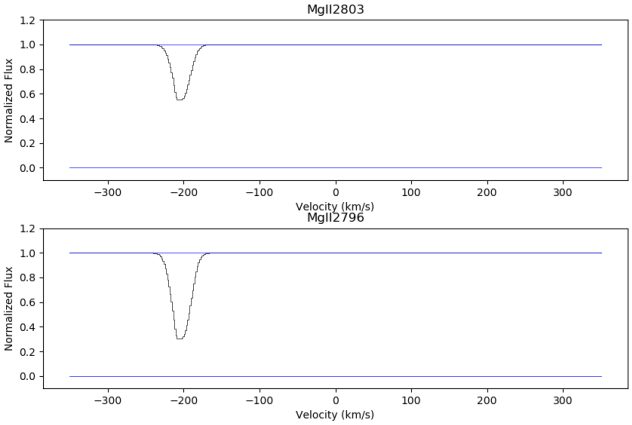

Figure 3: A sample blended two-cloud absorption line system. Note that this system contains TWO clouds on top of each other. This occurs when the clouds are very close together in velocity space.

maintained by B. Stemock's Ph.D. advisor, Christopher Churchill. Data simulation will be discussed in Section 4. One million one-cloud spectra have been simulated as well as one million two-cloud spectra to train and test a one-cloud CNN as well as a two-cloud CNN. Originally, ten thousand spectra were used, but this proved to be too few for the models to successfully return the requested parameters.

The CNN utilizes tensorflow and uses the r2_score method from sklearn.metrics to analyze the results. The architecture of our CNN uses multiple independent parallel networks that use the same input spectra and then trained to output cloud velocity, log of column density, and Doppler b parameter. The model has one independent network for each cloud with the output layers being concocted to form one Tensorflow model. An outline of our two-cloud model

architecture is shown in Figure 4. This architecture was decided via both trial/error and from theorizing that parallel networks would handle multiple clouds better

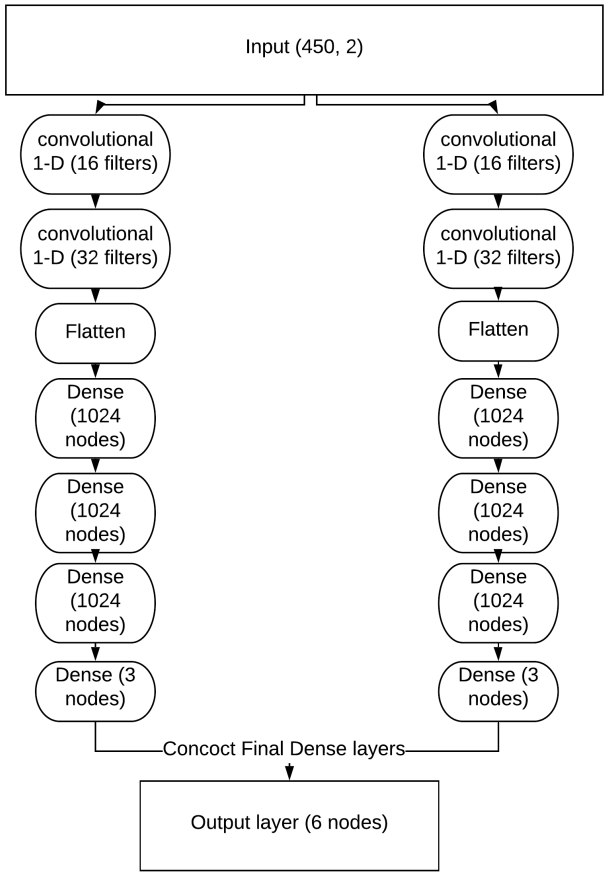

Figure 4: Architecture of our 2 cloud Tensorflow model.

4 DATA

4.1 Single Spectrum Simulation

The data consist of one million pairs of simulated spectra, which are output by specsynth. Spectra are in pairs for physical reasons. Briefly, these absorption features would result from light passing through a gas cloud with singly ionized magnesium (MgII). Due to quantum mechanical energy level splitting, we see two lines (MgII2796 and MgII2803) instead of one. Hence, this singly ionized magnesium gives us a doublet, which is represented by the pair of spectra.

To simulate a given pair of spectra, an input file is created with the redshift of the gas cloud (set to a constant 1.000 at this stage of the project for physical reasons), the ion (MgII2796 for the first step), and the desired velocity position, log of column density (logN), and Doppler b parameter. These parameters are written in as many lines as there are desired clouds. The velocity position is the line-of-sight velocity of the cloud with respect to its cosmological rest frame, the logN parameter is how many MgII ions are in our simulated line of sight through the cloud, and the Doppler b parameter deals with the internal velocity distribution of the gas cloud. These three desired parameters are then added to labels.txt. All other factors (resolution, telescope, etc.) are held constant and stored in other files. Specsynth then takes this input and uses a complex system of gas physics and absorption line optics to output spectra like those shown in Figures 2 & 3. Due to our early stage in this overall project, there is no noise added from atmospheric effects, general statistical effects, or any other physical or observational effects. Because the CNNs need a consistent input data shape, each output spectrum is exactly 450 pixels across (note that pixels are different than velocity). Finally, the output spectrum is added to the MgII2796data.txt file as a string of 450 space-separated numbers. Next, the same steps are repeated but with MgII2803 entered into the input file as the ion. The output spectrum this time is added to MgII2803data.txt.

4.2 Latin Hypercube Sampling

Now we need to simulate one million of these pairs! To do this, we need one million ordered triplets that explore the physical parameter space for velocity, logN, and b. This is achieved using Latin Hypercube Sampling (LHS), which will randomly pull N tuples from a unit hypercube of x dimensions. That is to say that, since we have a three-dimensional hypercube and one million points in that three-dimensional space, LHS is a method that will pull one million ordered triplets from a unit cube (each parameter ranges from 0 to 1). These ordered triplets will maintain a minimum distance from their nearest neighbors, yet will also be random within the small window they can inhabit. To achieve this, we use the lhs function from the pyDOE package, found at `https://pythonhosted.org/pyDOE/randomized.html`. Once one million ordered triplets have been created for each cloud (we use k hypercubes for k clouds), we use a code we wrote that loops through each ordered triplet, adds the parameters to labels.txt, creates the necessary input files, runs specsynth for each ion (MgII2796 and MgII2803), and adds the output spectra to the corresponding data file (MgII2796data.txt or MgII2803data.txt). That way, any given line in labels.txt has a corresponding spectrum in both data files for the parameters listed. The code takes the two data files as input and checks itself against the parameters in labels.txt.

5 RESULTS

Figure 5 displays the R2 score for each parameter, as well as the spread in each parameter from the y = x line, which is the line on which the test parameters are perfectly predicted by the model. The current model architecture is a living model and will continue to evolve as the overall project continues. Work is in progress to fit systems containing one, two, and three clouds, and there will eventually be work done using training and test systems with a combination of those numbers of clouds. The results from this model is are extremely promising, but there is much work left to do. The authors are currently in the process of using this model to fit observed data from the 2011 doctoral thesis of Jessica Evans. A great amount of consideration is and will continue to go into the simulation of the training data. Some such considerations include, but are not limited to, noise in the data, spectral resolution of the

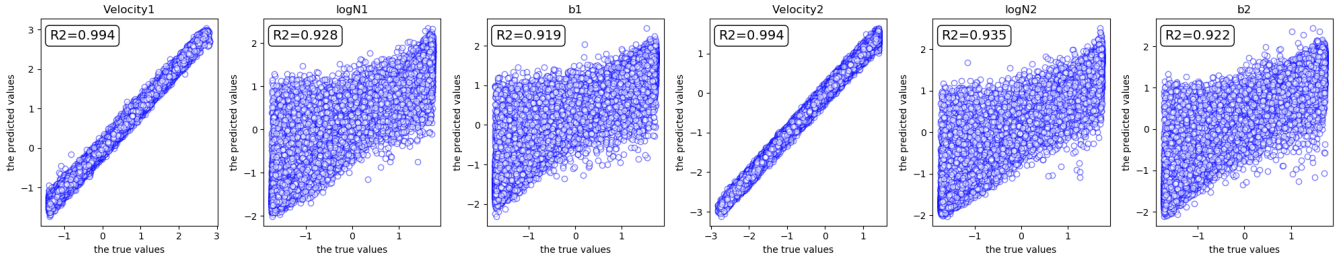

Figure 5: Results of our trained model. An RMSprop optimizer with a learning rate of 10^{-6} was utilized for this CNN. An R2 score greater than 0.919 was achieved for all labels. Scatter is evident and more extreme for the column density and the b parameters.

data, proper velocity spread and positioning, line blending, multi-phase gas physics, and generalization to observed systems outside of our own database.

6 RESULTS

The authors would like to thank the NMSU Astronomy Dept. Machine Learning Café collaboration for their efforts toward the completion of the thesis of author BS, of which this project is a subset. In addition, the authors are very appreciative to the NMSU Astronomy Dept. as a whole for the use of its resources, specifically the department gpu, and to the NMSU Discovery High Performance Computing Cluster team. Finally, and certainly not least of all, the authors are grateful to Huiping Cao and Edgar Ceh Varela for their feedback and guidance throughout the duration of this project.

REFERENCES

[1] Christopher W. Churchill et al. [n.d.]. Direct Insights Into Observational Absorption Line Analysis Methods of the Circumgalactic Medium Using Cosmological Simulations. *ApJ* 802, Article 10 ([n. d.]), 19 pages.
[2] Maarten Schmidt. [n.d.]. Spectrum of a Stellar Object Identified with the Radio Source 3c 286. *ApJ* ([n. d.]).