Experimental Astronomy manuscript No.
(will be inserted by the editor)

Towards the prediction of molecular parameters from astronomical emission lines using Neural Networks

Alejandro Barrientos[1,2] · Jonathan Holdship[3,4] · Mauricio Solar[1] · Sergio Martín[5,1] · Víctor M. Rivilla[6,7] · Serena Viti[3,4] · Jeff Mangum[8] · Nanase Harada[9,10,11] · Kazushi Sakamoto[10] · Sébastien Muller[12] · Kunihiko Tanaka[13] · Yuki Yoshimura[14] · Kouichiro Nakanishi[9,11] · Rubén Herrero-Illana[5,15] · Stefanie Mühle[16] · Rebeca Aladro[17] · Susanne Aalto[12] · Christian Henkel[17,18] · Pedro Humire[17]

Received: date / Accepted: date

[1] Federico Santa María Technical University, Vicuña Mackenna 3939, San Joaquín, Santiago, Chile

[2] Joint ALMA Observatory, Alonso de Cordova 3107, Vitacura, Santiago, Chile

[3] Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands

[4] Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT, London, UK

[5] European Southern Observatory, Alonso de Cordova 3107, Vitacura, Santiago, Chile

[6] Centro de Astrobiología (CSIC-INTA), Ctra Ajalvir km 4, 28850, Torrejón de Ardoz, Madrid, Spain

[7] INAF-Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125, Florence, Italy

[8] National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903-2475, USA

[9] National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan

[10] Institute of Astronomy and Astrophysics, Academia Sinica, 11F of AS/NTU Astronomy-Mathematics Building, No.1, Sec. 4, Roosevelt Rd, Taipei 10617, Taiwan

[11] Department of Astronomy, School of Science, The Graduate University for Advanced Studies (SOKENDAI), 2-21-1 Osawa, Mitaka, Tokyo, 181-1855 Japan

[12] Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-43992 Onsala, Sweden

[13] Department of Physics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Yokohama, Kanagawa 223–8522 Japan

[14] Institute of Astronomy, Graduate School of Science, The University of Tokyo, 2-21-1 Osawa, Mitaka, Tokyo 181-0015, Japan

[15] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Magrans, E-08193 Barcelona, Spain

[16] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, D-53121 Bonn, Germany

[17] Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

[18] Astronomy Department, Faculty of Science, King Abdulaziz University, P. O. Box 80203, Jeddah 21589, Saudi Arabia

Abstract Molecular astronomy is a field that is blooming in the era of large observatories such as the Atacama Large Millimeter/submillimeter Array (ALMA). With modern, sensitive, and high spectral resolution radio telescopes like ALMA and the Square Kilometer Array, the size of the data cubes is rapidly escalating, generating a need for powerful automatic analysis tools. This work introduces *MolPred*, a pilot study to perform predictions of molecular parameters such as excitation temperature (T_{ex}) and column density ($log(N)$) from input spectra by the use of neural networks. We used as test cases the spectra of CO, HCO^+, SiO and CH_3CN between 80 and 400 GHz. Training spectra were generated with MADCUBA, a state-of-the-art spectral analysis tool. Our algorithm was designed to allow the generation of predictions for multiple molecules in parallel. Using neural networks, we can predict the column density and excitation temperature of these molecules with a mean absolute error of 8.5% for CO, 4.1% for HCO^+, 1.5% for SiO and 1.6% for CH_3CN. The prediction accuracy depends on the noise level, line saturation, and number of transitions. We performed predictions upon real ALMA data. The values predicted by our neural network for this real data differ by 13% from the MADCUBA values on average. Current limitations of our tool include not considering linewidth, source size, multiple velocity components, and line blending.

Keywords Molecular Astronomy · Molecular Parameters · Machine Learning · Neural Networks · MADCUBA · ALCHEMI

1 Introduction

The study of the molecular composition of objects in space has been a matter of extensive research since the 1930s (Swings and Rosenfeld, 1937). Through the use of spectroscopy techniques over 200 molecules have been identified in interstellar or circumstellar medium (Woon, 2020), and the number keeps growing every year. The analysis process to detect these molecules is quite complex and requires effort from observers, astrochemists and laboratory spectroscopists (Cernicharo, 2012). Yhe construction of astronomical observing facilities like the Atacama Large Millimeter / Submillimeter Array (ALMA) has created the possibility to observe the universe with an unprecedented combination of sensitivity and angular resolution. Such facilities allow astronomers to study the physical and chemical properties of the molecular gas in a variety of sources in the Universe (e.g. Nakajima et al., 2015).

The introduction of ever larger and more sensitive instruments in astronomy and the latest developments in computing performance has generated a need for tools to support the analysis of data sets (Berriman and Groom, 2011). This is illustrated by the number of new astronomical facilities that are implementing automated data reduction pipelines such as ALMA (Lightfoot et al., 2008), Vera C. Rubin Observatory (formerly LSST; Jurić et al., 2015) and E-ELT (Mach et al., 2016). Telescopes like the Square Kilometer Array (SKA), that will be deployed in the very near future, will even rely only on fully automatic reduction pipelines, due to the large amounts of data they will produce (Farnes et al., 2018).

Current facilities produce crowded spectra even in sources where previous facilities were only able to detect a limited number of bright transitions of the most abundant species. Line identification and extraction of physical parameters is generally still a manual and time consuming process even making use of state of the art tools briefly described below.

Several tools have been developed over time, to assist in line analysis. These tools have varying degrees of automation. For instance CASA (McMullin et al., 2007) allows the user to connect to Splatalogue[1] catalog via casaviewer. This enables the user to idenitfy lines by overlaying transitions from the catalog on their spectra. In this case, there is no automation. Another tool is the ALMA Data-Mining Toolkit (ADMIT, Teuben, 2015) which enhances CASA with a spectral Line identification algorithm. However, it does not provide estimates of the physical parameters of the molecules identified. Other programs like RADEX (van der Tak, F. F. S. et al., 2007) and MADEX (Cernicharo, 2012) can generate models, that can be manually compared with data, although this comparison is not automated and requires the tuning of many input parameters. The XCLASS interface (Möller and Schilke, 2015) contains a program named myXCLASS which contains routines to fit models to observed spectra. The fitting model can be automated with multiple approaches as mentioned in Schilke et al. (2015). Another tool is CASSIS (Vastel et al., 2015) which computes synthetic LTE models that can be compared with observations. It contains a tool to perform physical parameter estimations. The last tool we will cover is MADCUBA (Martín et al., 2019) which is a tool for line analysis and spectroscopic work. It contains a feature called Spectral Line Identification and Modelling (SLIM) which allows the automatic fitting of molecular parameters. It allows the possibility to generate synthetic spectra programmatically, given the appropriate parameters, and also to merge several data cubes across a frequency axis. MADCUBA was selected as our primary source for training examples.

Having established that automated tools are necessary, our intention is to contribute by taking a step towards the prediction of molecular parameters using neural networks. For the sake of probing feasibility and as a pilot study, in this paper we constrain our analysis to the prediction of the parameters of excitation temperature (T_{ex}) and column density

[1] https://splatalogue.online

($log(N)$) from molecular spectra. Section 2 presents our pilot study for prediction of molecular parameters, describing the overall design, training data and methods. Section 3 describes and discusses our test results with synthetic data, with a number of parameter combinations for the neural networks. We also present our results with astronomical data, performing predictions for $log(N)$ and T_{ex} for a spectrum coming from an ALMA Large Program project. Section 4 summarizes our conclusions and comments on future work.

2 The *MolPred* prototype model

Given that astronomical spectra can be modelled from their underlying physical parameters, it should be possible to predict those parameters using a regression model. In this work, we use several neural networks which take information from the spectrum as an input. A neural network is an algorithm which learns relationships between some inputs and outputs resulting in a blackbox model (Rosenblatt, 1958).

Whilst even simple radiative transfer models require several parameters, we consider only $log(N)$ and T_{ex} in this pilot study. The column density is an important quantity as it is a measure of how much of a species is present towards an object. The intensity of a line approximately scales with the column density when lines are optically thin. The excitation temperature reflects the relative population of energy levels of a molecule (e.g., following a Boltzmann distribution at local thermal equilibrium) and helps to characterize the conditions of the gas. It primarily determines the relative line intensity among the transitions of a given species.

For this purpose, we have developed a tool called *MolPred*, which is capable of predicting said parameters for an initial sample of four molecules, provided an input spectrum. This set can be incrementally extended to more parameters and to the full set of species available in spectroscopic catalogs. The current python code of the tool is available in our GitHub repository [2]

2.1 Overall Design

An overview of the *MolPred* program is described in Figure 1. The flow starts by receiving an input spectrum for which *MolPred* will generate predictions of $log(N)$ and T_{ex}. The input spectrum is pre-processed as explained below, so the predictions for all the molecules can be done in parallel via individual neural networks. After all predictions are done, the results are collected and handled in a post-processing stage, where the resulting predictions are presented to the user.

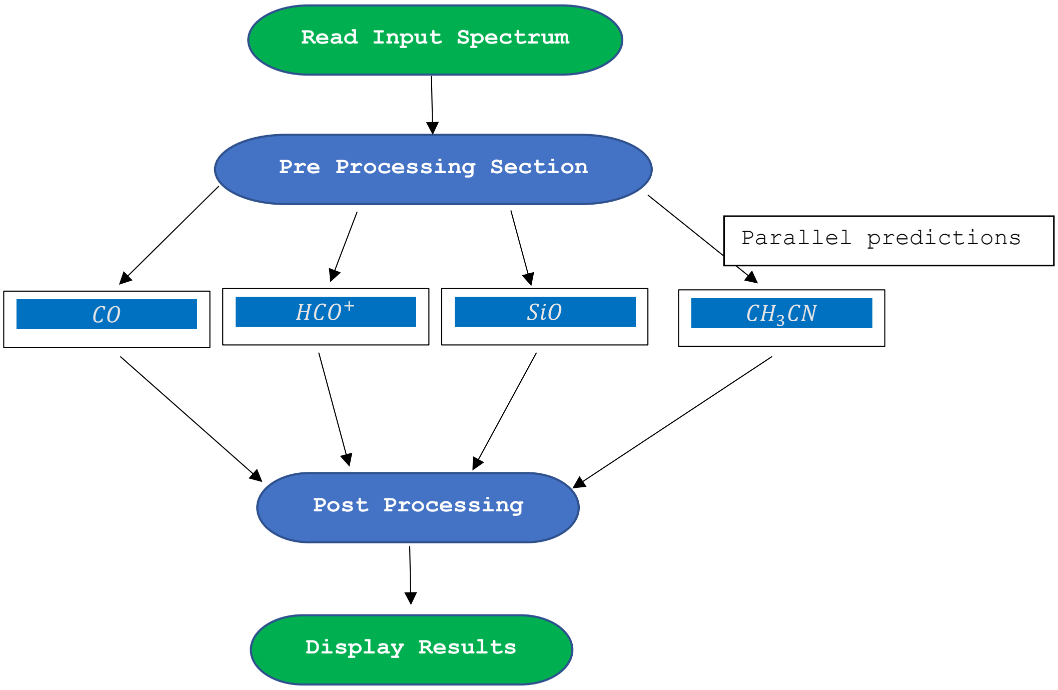

Fig. 1: A diagram showing the *MolPred* workflow. Each box indicates a process and the rectangles showing molecules indicate the neural networks.

[2] https://github.com/MadScience01/molpred

In the pre-processing stage, the input arrays for the individual neural networks are generated. The neural networks require only the peak intensity of every transition of the molecule for which it predicts. To obtain this, the intensity of the input spectrum channel closest to the rest frequency of each transition of a molecule is extracted. These intensities are then scaled to take values from 0 to 1 using the minimum and maximum values that were in the training data. The molecules and the number of transitions they have within the frequency range considered in this work are given in Table 1.

In real astronomical data, it is possible that some transitions will not be observed. If there is no channel within 10 MHz of a transition, then *MolPred* assumes the transition is missing from the data and the pre-processing stage simply inserts a zero for the intensity of that transition.

In the prediction stage, the neural networks trained to predict the $log(N)$ and T_{ex} of each molecule take the input array and produce a prediction for $log(N)$ and T_{ex}, each scaled between 0 and 1. The neural network used in this stage is the best network from the grid discussed in Section 2.3, the performance of each of these networks is discussed in Section 3. Finally, in the post-processing stage, the scaled predictions of $log(N)$ and T_{ex} are transformed to their actual values.

2.2 Training data set

The data used to train the neural networks which make up *MolPred* are a set of synthetic spectra generated by the SLIM LTE spectral model, which is part of the MADCUBA software package (Martín et al. 2019), which makes use of the spectroscopic parameters from JPL catalog (Pickett et al., 1998). We generate LTE spectra for each species, with 1 MHz resolution between 80 and 400 GHz using column densities in the range $log(N) = 12$ cm^{-2} to $log(N) = 19.9$ cm^{-2}, with steps of $log(N) = 0.1$ cm^{-2} and temperatures from 10 K, performing multiplicative increments of 30% all the way to 233 K, in this way we increase the coverage at lower temperatures.

In order to produce synthetic spectra, MADCUBA requires other parameters on top of $log(N)$ and T_{ex} considered in our pilot study. Thus these parameters were fixed as follows: output intensity units were set to Kelvin; line width and velocity were fixed to 150 and 250 km s^{-1} respectively; an emitting source size of 10$''$ was assumed. These fixed parameters were selected to match those required to fit the actual astronomical spectra from the ALMA Comprehensive High-Resolution Extragalactic Molecular Inventory (ALCHEMI, Martin et. al, in preparation), as discussed below. Since our training datasets are created in rest frequency units, it is therefore agnostic to the velocity parameter. However, our predictor cannot be directly applied to astronomical data with different parameters of line width and source size. An extension of our neural network should be required to make our tool fully useable. All training and validation spectra used in this work were then created by combining these individual molecular spectra and adding Gaussian noise with an rms between 10 and 50 mK, which we consider to be a reasonable range of values for noise in an astronomical spectrum at these frequencies.

Species	Name	Number of Transitions (80-400GHz)
CO	Carbon monoxide	3
HCO^+	Formylium	4
SiO	Silicon monoxide	8
CH_3CN	Methyl cyanide	437

Table 1: Molecules considered in this work and the number of rotational transitions from the vibrational ground state which are in the range of frequencies covered by *MolPred*.

2.3 Neural Network Training

We have used the *keras* package from the *tensorflow* library (Chollet et al., 2015) for the creation and training of the neural networks. *Keras* model files are saved after training, for later use in the prediction stages. Neural networks have a large number of hyperparameters. The hyperparameters that were varied are given in Table 2 along with the range of values trialled. For each possible combination of parameters, a model was created. Neural networks were trained for up to 1000 epochs, where an epoch is one pass of the full training dataset to the network. However, to keep training time low, we implemented an early stopping mechanism where the training would stop if the validation loss did not improve over 10 epochs.

We trained each neural network individually using spectra that contained only noise and emission from transitions of the molecule for which the network would predict. These neural networks could later be loaded together as part of the *MolPred* code to predict from full spectra. In this initial experiment we decided to use a simple sequential neural

Parameter type	**Parameter values**
Activation functions	Sigmoid, ReLU, Linear, Tanh, Swish.
Layers	Single, double, triple
Neurons	256, 1024
Molecules	CO,HCO^+, SiO, CH_3CN
Training examples	500, 1000, 2000, 4000, 8000, 16000, 32000
Noise levels	0.01 to 0.05 K
Optimizer	Adam
Loss function	MAE
Training patience	10 epochs
Maximum amount of training	1000 epochs

Table 2: A list of neural network hyperparameters which were varied and the values that were tested.

network with a maximum of 3 densely connected layers of up to 1024 neurons each. We wish to establish an initial set, as more molecules and output classes can be added later. The networks were trained to minimize the mean absolute error (MAE) between the scaled $log(N)$ and T_{ex} of an input spectrum and the predictions using the Adam optimizer (Kingma and Ba, 2017). We use the scaled predictions so that errors in $log(N)$ and T_{ex} are equally weighted as the unscaled variables differ by many orders of magnitude. To test these trained neural networks, the MAE on the scaled predictions across the entire validation set was calculated for each network. These MAE values were then compared between networks to select the best regressor to include in *MolPred*.

3 Results and Discussion

3.1 Results on Test Data

Our first task was to determine an appropriate number of training examples for the neural networks. We present the evolution of the MAE as a function of the number of training examples in Figure 2. We found that the MAE initially improves by doubling the number of training examples but beyond 16000 examples, increasing the number of training examples gives a marginal improvement. We also included the training times, for evaluation, which we found to be roughly proportional to the number of examples. To balance low MAE values with reasonable training time, we reached a maximum of 32000 examples, which we defined as our baseline for all models shown in this section. For reference, these training times were obtained using a PC with an Intel I7-8700K CPU, 32 GB of RAM, two Zotac GeForce GTX 1060 6GB video cards and a Kingston A2000 1 TB Solid State Drive - M.2 2280.

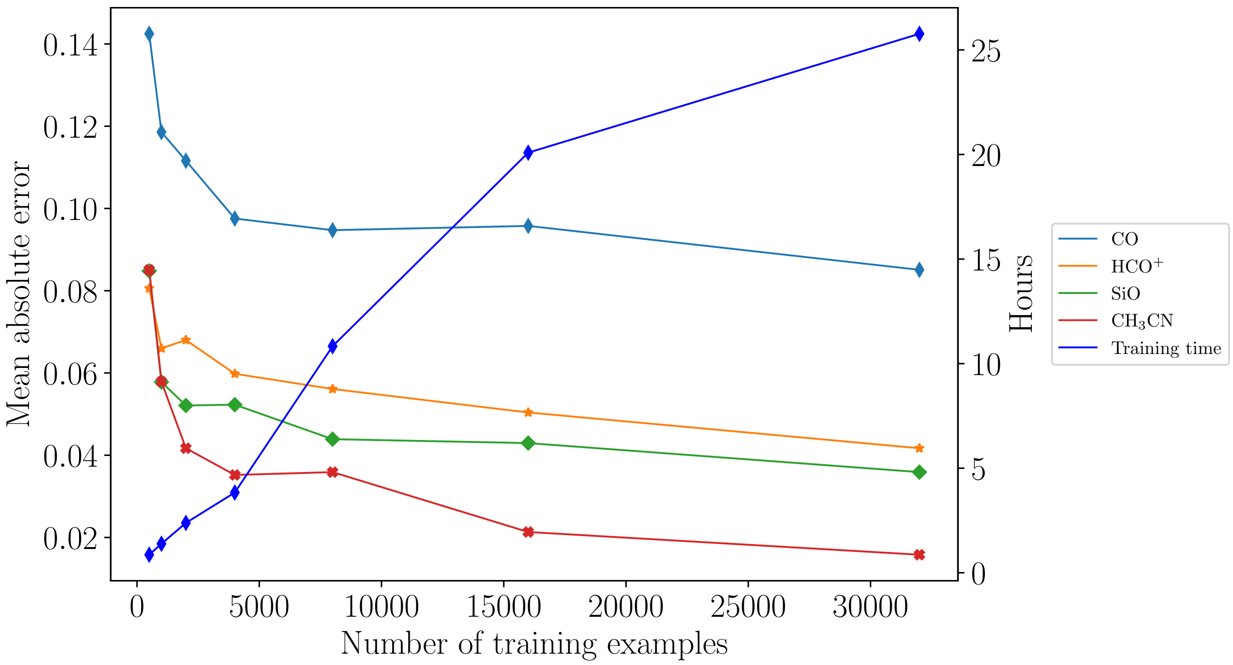

Fig. 2: Mean absolute error and training time in hours as a function of number of training examples.

Once we have established the number of training examples, we wished to observe the behaviour of the MAE in the training stage of each network. Figure 3 presents the evolution of the MAE for the neural network that performed best on each molecule. We allowed the networks to be trained for a maximum of 1000 epochs but stopped training early if the model did not improve for 10 epochs (training patience). In most models, the training stops close to 100 epochs, the lowest validation loss can be seen in the plots approximately ten epochs before the end of the plot.

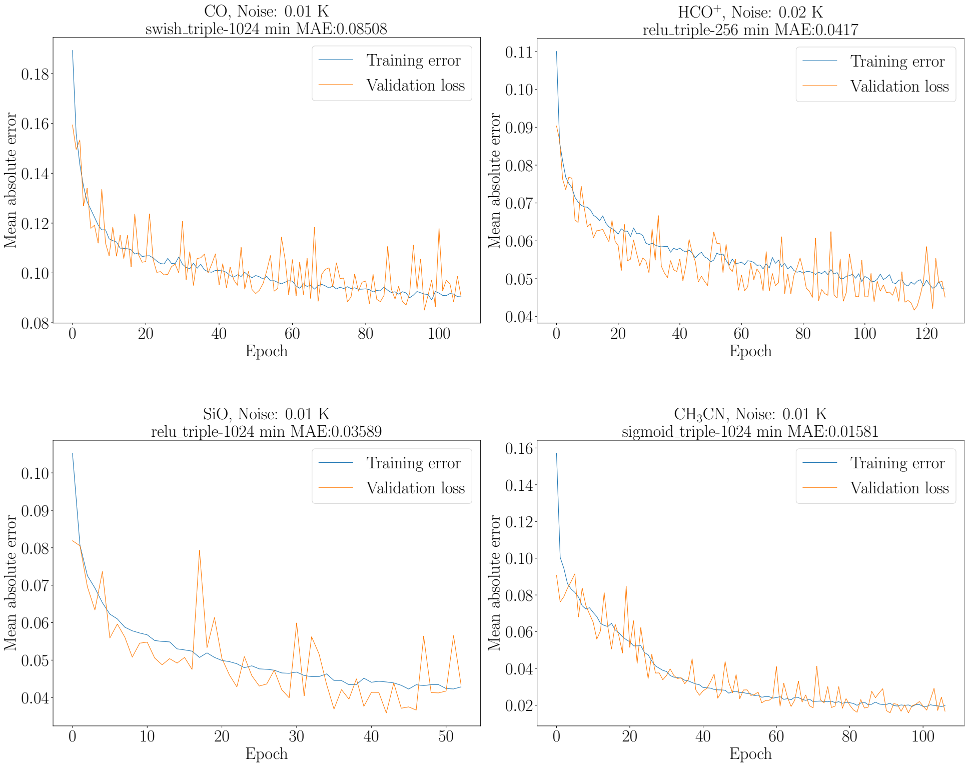

Fig. 3: Training mean absolute error over Molecule set, each figure contains the molecule name, the best neural network model, the minimum MAE and its respective noise level.

To obtain a deeper understanding of the behavior of the neural networks, we analyzed how the noise would influence the performance. In Figure 4 we can observe that the general trend is that the MAE increases with the noise. We assume the best network is the one with the lowest MAE at any noise level for that particular molecule, and use that network in *MolPred* as shown in Table 3. However, since Figure 4 shows the MAE is only weakly affected by the noise value, one could choose to train their networks with a large noise value to ensure the model is robust to the noise in real spectra.

Molecule	Noise (K)	MinMAE	Epochs	Model
CO	0.01	0.08508	96	swish_triple-1024
HCO$^+$	0.02	0.04170	116	relu_triple-256
SiO	0.01	0.03589	42	relu_triple-1024
CH$_3$CN	0.01	0.01581	96	sigmoid_triple-1024

Table 3: The MAE of the best neural network trained for each molecule and the name of the model which indicates the activation function, the number of layers (3), and number of neural per layer.

To verify the quality of the predictions on each neural network, we predicted the column density and excitation temeprature for a test set of 6400 testing examples not used during training. We then checked the error distribution of the predictions, which is shown in Figure 5. The error distribution is centrally peaked which is an important result that our choise of loss function does not guarantee. The strongly peaked distributions mean that the networks typically give small errors and are unlikely to predict an extremely incorrect value.

Having obtained the best models for each molecule, we wanted to review the prediction behaviour of the networks, which was plotted in Figure 6. We see that for all the molecular species in our pilot study, the predictions follow a

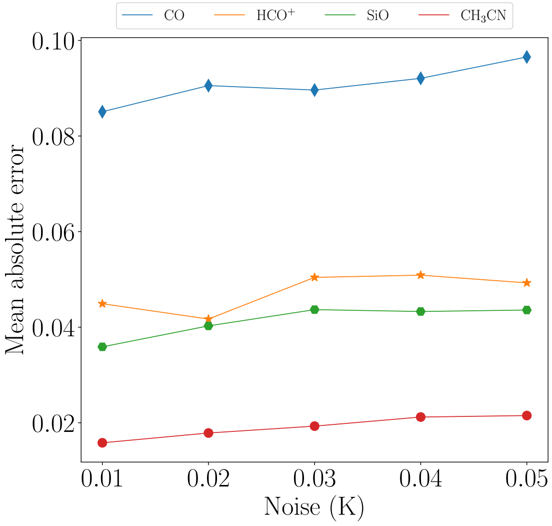

Fig. 4: Noise vs mean absolute error.

tight linear correlation with predicted values close to the true values, as also seen in the centrally peaked histograms in Figure 5. However, the prediction accuracy drops for both low and high values of $log(N)$. The $log(N)$ value at which this deviation occurs is dependent on the molecule and has a physical explanation. For low values of $log(N)$ the line intensities become close to or below the noise level, which causes the neural network to predict values that are not dependent on the true value. Moreover, this also has a low dependency on the temperature as seen in the color coding in Figure 6, as it is purely due to the lack of signal in the input spectra. Essentially, the network cannot distinguish between input spectra below a $log(N)$ threshold.

On the other hand, for high values of $log(N)$, the spectral features from the input data are saturated. That is, all transitions reach a saturation flux density once a value of $log(N)$ is surpassed (see Martín et al., 2019, for a description of line saturation). The $log(N)$ value at which saturation occurs has a strong dependency on the value of the temperature as seen in the color coding in Figure 6. The behaviour also depends on the number of transitions available for each molecule (Table 1). Thus, in the case of CH_3CN, even for high values of $log(N)$, there is still a significant number of low flux density unsaturated transitions and therefore we do not observe the effect of saturation in Figure 6.

The same physical explanation applies to the predictions of temperature parameter. Figure 7 presents the test predictions of T_{ex} against their true values. For all molecules, we can observe that for all values of T_{ex}, predictions appear dependent on the value of N. As can be seen in the figure, the predicted temperatures are closer to their real values for higher N. This has a similar explanation to the neural networks' poor performance on the column density prediction for low column densities. At low N, the intensities of the transition decrease and come closer to the noise. The noise then dominates the prediction, making the recognition of the transition impossible if the transition is below the noise level. We can compare the CO results with the results of other molecules, particularly with CH_3CN which has the largest number of transitions of the molecules in our set. These results suggest that the accuracy in the prediction of the temperature is highly dependent on the number of transitions available to construct a better characterized model of the molecule.

If we increase the noise to 0.05 (50mK) we can observe that the predictions of temperature spread over a wider range, rather than forming small groups, as in the case with lower noise. Still, the predictions remain accurate for the higher column density values (yellow/red values in Figure 8), where the spectral features are clearly identified above the noise level.

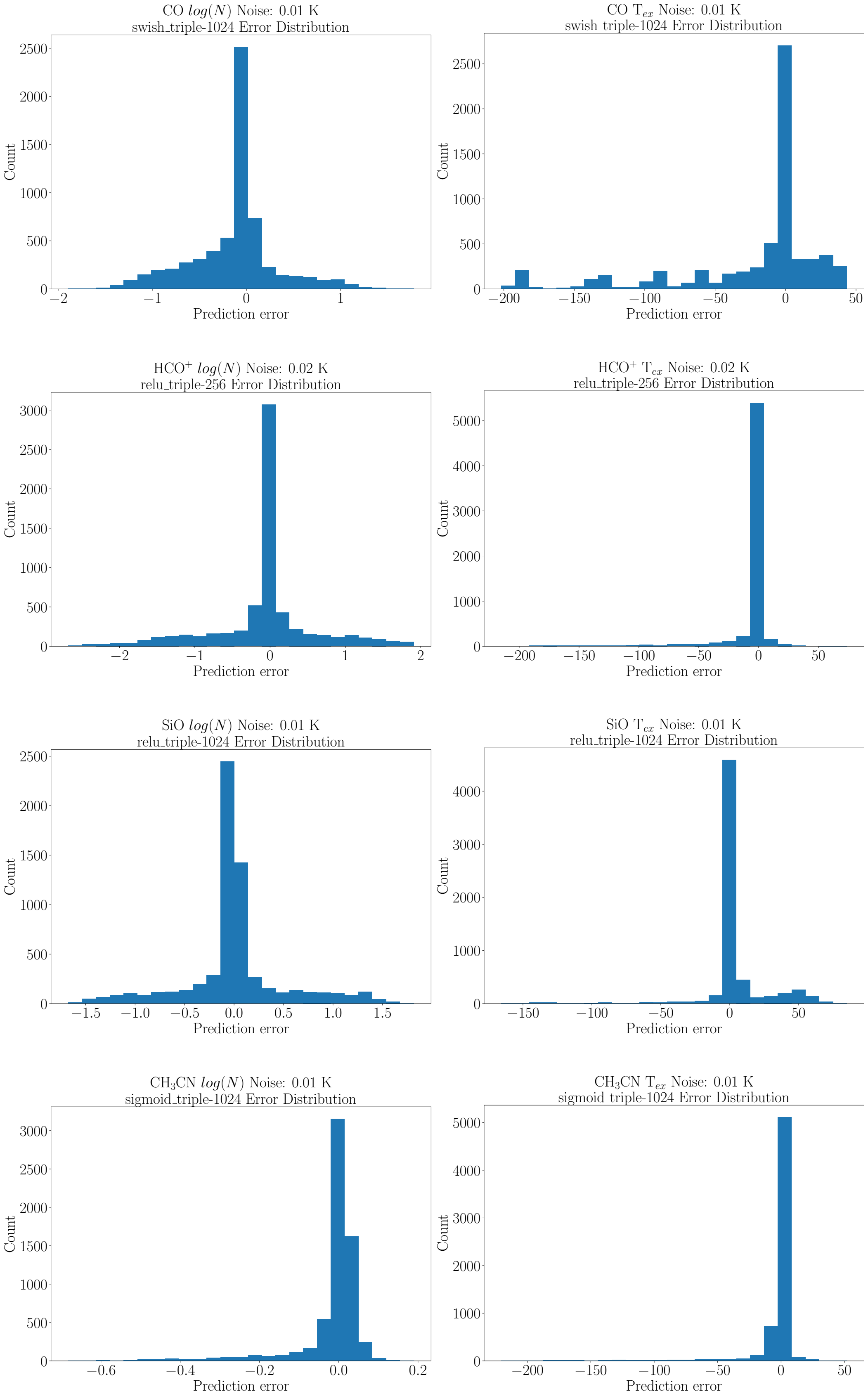

Fig. 5: Histograms showing the distribution of errors on the $log(N)$ (left column) and T_{ex} (right column) sprediction of each molecule. These all peak at zero indicating small errors are most common.

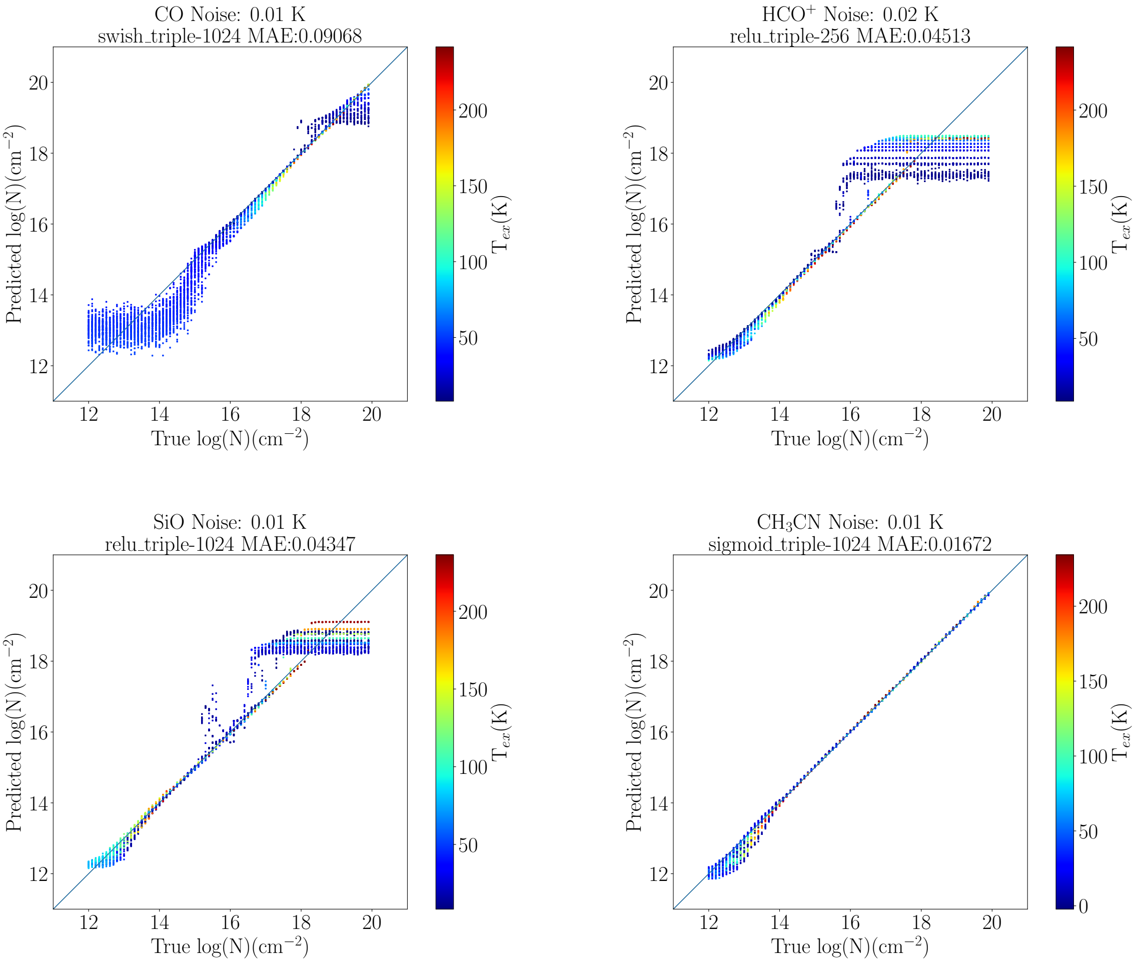

Fig. 6: Scatter plots showing the predicted $log(N)$ against its true value for all spectra in the test data. The colour scale indicates the excitation temperature of the spectrum. Each neural network was trained on spectra with a noise of 10mK for all molecules except for HCO$^+$ where the best network was trained using a noise of 20 mK.

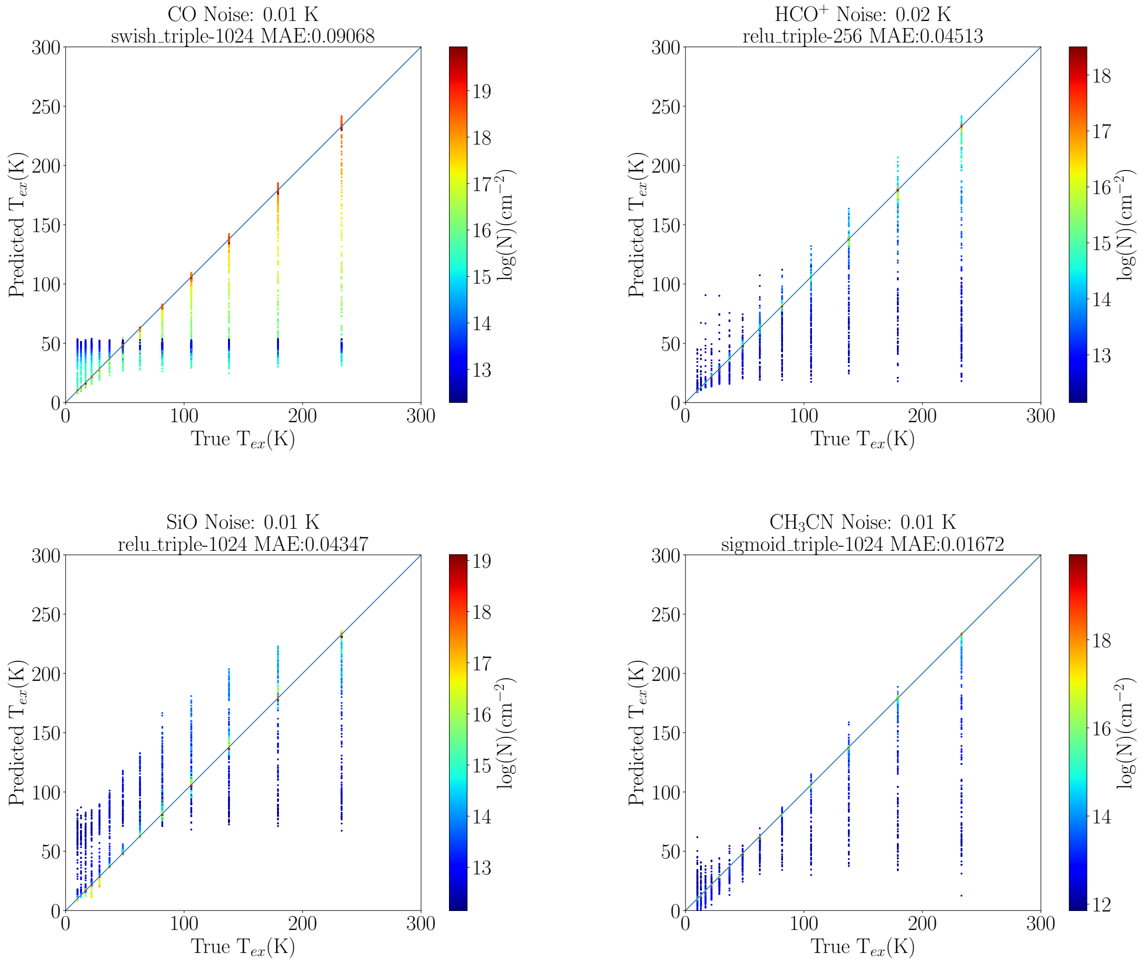

Fig. 7: Predicted T_{ex} from the best neural networks against the true value for all spectra in the test set. The column density of the test spectra is given in the colour scale. The training noise was 10 mK for all molecules except HCO^+ which was 20mK.

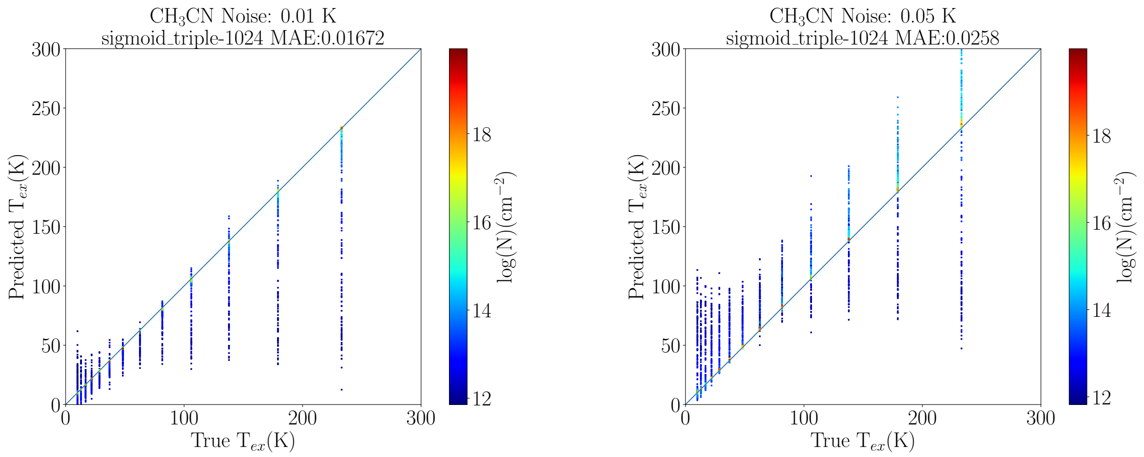

Fig. 8: Similar to Figure 7 but only showing predictions for CH_3CN. In the left plot, the predictions from a network trained with an rms noise of 10 mK is shown and on the right the noise was 50 mK.

3.2 Results on Astronomical Data

To further test *MolPred* we used data from ALCHEMI, which was one of the ALMA Large Programmes in Cycle 5. The astronomical target was the central molecular zone of the starburst galaxy NGC253. The project consists of a full spectral line survey continuously covering ALMA Bands 3 to 7. We used the low resolution spectrum from the ALMA Compact Array (Morita Array) which covers the frequency range of 125 GHz to 373.2 GHZ. It therefore contains an almost complete set of transitions from the sample molecules used in this study. The continuous coverage and uniform sensitivity makes this dataset an ideal test sample for our study. This wide band spectrum was shift to rest frequency assuming a Doppler shift of 250 km s^{-1} and used as an input to *MolPred*. *MolPred* then generated $log(N)$ and T_{ex} predictions for each molecule following the flow of Figure 1.

The *MolPred* predictions are displayed in Table 4. For comparison, we also include the fitted values using the MADCUBA AUTOFIT packaged used for the actual spectroscopic analysis of the ALCHEMI data. The *MolPred* predictions for $log(N)$ are within 1.5% of the ones from MADCUBA, and for T_{ex} are within 25% from the ones by MADCUBA. Thus the predictions agree very well except in the case of the excitation temperature of CH_3CN. However, we note that MADCUBA fitting algorithm did not converge unless taking into account the contribution from other molecular species blended to the CH_3CN. Despite the fact *MolPred* did not include information on other molecular "contaminants", the predictions were still reasonably close to the value predicted by MADCUBA. We now examine these results by plotting our predictions together with the closest examples to the MADCUBA predictions from our training files.

Molecule	*MolPred* $log(N)$	*MolPred* T_{ex}	MADCUBA $log(N)$	MADCUBA T_{ex}
CO	18.51	17.45	18.44	19.48
HCO^+	13.98	16.55	14.08	15.30
SiO	13.17	17.65	13.18	13.98
CH_3CN	13.39	58.95	13.27	38.95

Table 4: The $log(N)$ and T_{ex} predicted by Molpred for the ALCHEMI data alongside the the values obtained from the MADCUBA fit.

We start the analysis with CO, the simplest molecule of our set. The *MolPred* prediction gives a very similar value of $log(N)$ to the MADCUBA fit but has a slightly lower excitation temperature. The result of this difference can be seen in Figure 9 where we plot the intensities of MADCUBA spectra generated using the *MolPred* and MADCUBA predictions alongside the ALCHEMI data. Whilst both models underfit the data, the *MolPred* spectrum has a somewhat lower intensity than the others due to the low excitation temperature. Based on our analysis of the test data, we can expect the saturation effect seen in Figure 7 has affected the accuracy of predictions at these high column densities. Further, the value of the intensity passed to the *MolPred* predictor was taken at the rest frequency of the transition and does not match the peak value due to imprecise Doppler shifting. The peak intensity of the *MolPred* prediction is close to the value of the ALCHEMI data at the transition frequency for both CO transitions.

Molecule	J	Intensity *MolPred* (K)	Intensity MADCUBA (K)	Intensity ALCHEMI @ Transition Freq (K) @ (GHz)
CO	2-1	49.11	60.15	52.10 @ 230.5380
CO	3-2	93.11	126.60	96.14 @ 345.7959

Table 5: Peak intensities from spectra generated using the *MolPred* and MADCUBA predictions for CO.

For HCO^+, the analysis was done following the same process as used for CO. For this molecule, the ALCHEMI dataset contains data for 3 out of 4 transitions in the working frequency range. It is interesting that despite the missing transition, the *MolPred* predictions are close to those obtained with MADCUBA. We can see the prediction differences per transition in Table 6 and visually observe them in Figure 10. In Figure 6, we can see the saturation limit beyond which the column density cannot be predicted is 14 cm^{-2}. Therefore, we might expect the accuracy to be affected by this since the column density predicted by *MolPred* is close to this value. Despite this, the spectra from the *MolPred* predictions are good fit to the ALCHEMI dataset.

Molecule	J	Intensity *MolPred* (K)	Intensity MADCUBA (K)	Intensity ALCHEMI @ Transition Freq (K) @ (GHz)
HCO+	2-1	2.77	3.44	2.77 @ 178.3750
HCO+	3-2	6.64	8.24	5.77 @ 267.5576
HCO+	4-3	7.78	9.69	6.99 @ 356.7342

Table 6: Peak intensities from spectra generated using the *MolPred* and MADCUBA predictions for HCO^+.

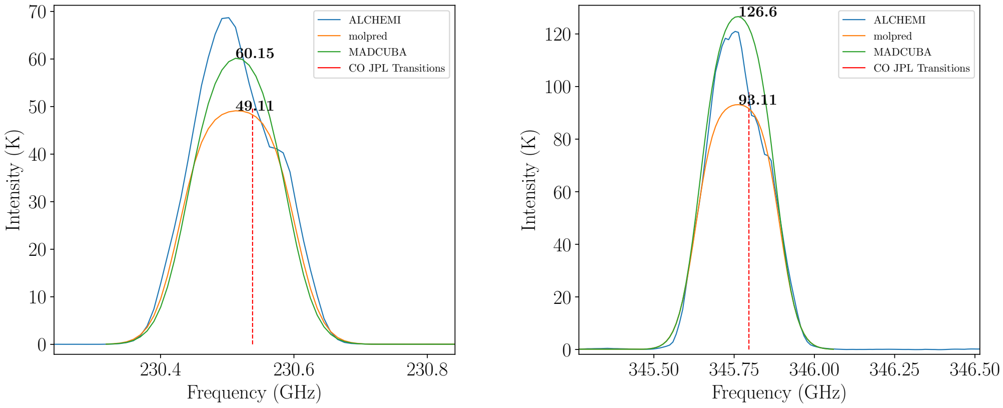

Fig. 9: CO line profiles from the ALCHEMI Data plotted in blue. Overplotted are the spectra generated from the *MolPred* predictions (orange) and MADCUBA predictions (green). For reference, we included a red vertical line to indicate the frequency of the closest transition according to the JPL Catalog. The ALCHEMI data set used in this test, only has data for 2 out of the 3 transitions in our training data, we can see the the detail of the CO(2-1) in the left figure and CO(3-2) transitions in the right figure.

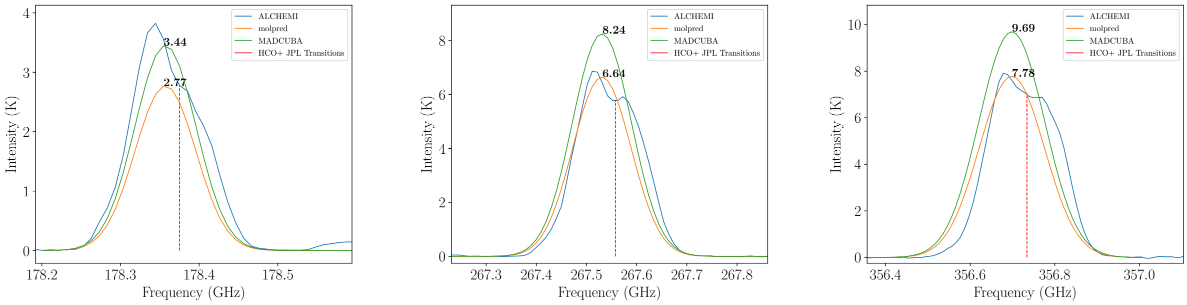

Fig. 10: Similar to Figure 9 for HCO$^+$.

The ALCHEMI dataset contains 6 out of the 8 SiO transitions in the working frequency range. Similarly to HCO$^+$, we can see the prediction accuracy is good despite the missing transitions. Intensities predicted for SiO transitions can be examined on Table 7 and seen in Figure 11. The SiO lines are quite weak and the *MolPred* fits give a small $log(N)$. As a result, the temperature prediction suffers for SiO as discussed in section 3.1. When used to generate an LTE spectrum, the high temperature predicted by *MolPred* for SiO, results in line intensities that are often too large.

Molecule	J	Intensity *MolPred* (K)	Intensity MADCUBA (K)	Intensity ALCHEMI @ Transition Freq (K) @ (GHz)
SiO	3-2	0.08	0.09	0.10 @ 130.2686
SiO	4-3	0.17	0.15	0.13 @ 173.6884
SiO	5-4	0.22	0.16	0.14 @ 217.1050
SiO	6-5	0.22	0.13	0.17 @ 260.5180
SiO	7-6	0.17	0.08	0.15 @ 303.9270
SiO	8-7	0.11	0.04	0.08 @ 347.3306

Table 7: Peak intensities from spectra generated using the *MolPred* and MADCUBA predictions for SiO

Finally we move to the molecule with the largest number of transitions in this exercise, CH_3CN. Since there are 14 groups of transitions with the same J quantum number, we will zoom in on each group to observe the behavior in Figure 12. Predicted intensities generated for each group are described in Table 8. Since the molecule contains many transitions per group, we decided to select a group representative frequency (GRF) as the frequency that is closer to the *MolPred* and MADCUBA's prediction peaks. We included in the table, the ALCHEMI intensity at that frequency.

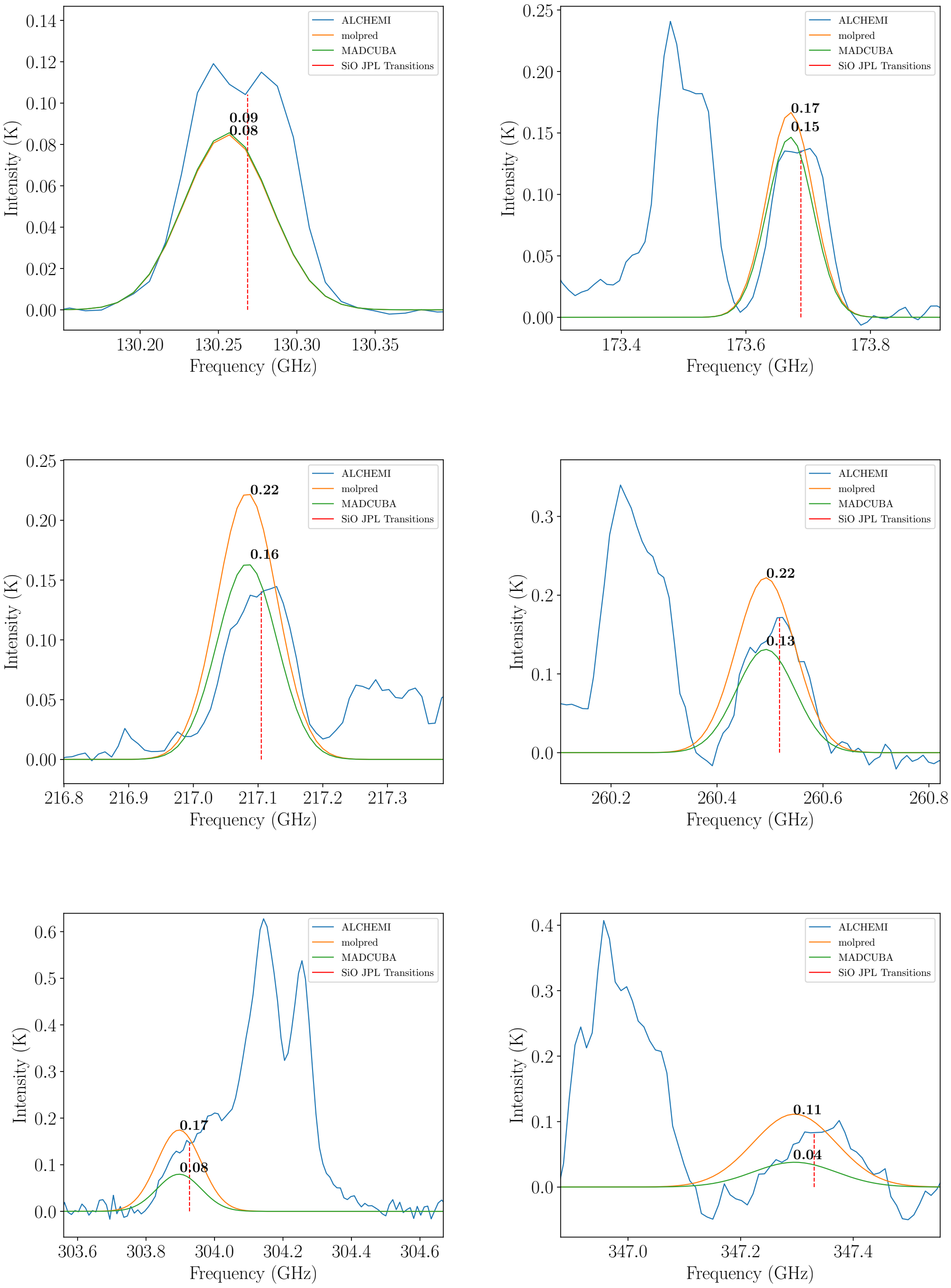

Fig. 11: Similar to Figure 9 for the SiO transitions in the ALCHEMI data.

MolPred predictions for $log(N)$ are close to 13.4 cm^{-2}, at which point the noise starts to strongly affect the predictions (see Figure 6). This may explain why the temperature predicted by *MolPred* for CH_3CN differs so strongly from the MADCUBA prediction. We can see the effect of this inaccuracy on the intensities of several transitions, where the intensities derived from the *MolPred* prediction, are higher than the ALCHEMI intensity.

Figure 13 summarizes the prediction differences seen in previous tables. In many cases, *MolPred* does very well, giving predicted intensities that are closer to the data than MADCUBA. Where the results appear significantly different, these differences can be understood based on the way the parameters are obtained. In the case of MADCUBA, the whole entire spectrum is used to fit a comb of Gaussian profiles at the frequencies of the molecular transitions. This includes a simultaneous fit to the width of the line, which can help constrain the effect of opacity mentioned in Section 3.4 of Martín et al. (2019), and the fit may be more robust to individual channel variations due to line shape or noise. On the other hand, our neural networks are trained with single intensity values for each transition. We can see how in

Molecule	J	Intensity *MolPred* (K)	Intensity MADCUBA (K)	Intensity ALCHEMI @ GRF (K) @ (GHz)
CH_3CN	$7-6$	0.07	0.08	0.11 @ 128.7577
CH_3CN	$8-7$	0.11	0.11	0.11 @ 147.1499
CH_3CN	$9-8$	0.16	0.15	0.12 @ 165.5415
CH_3CN	$10-9$	0.21	0.18	0.07 @ 183.9320
CH_3CN	$11-10$	0.27	0.20	0.12 @ 202.3204
CH_3CN	$12-11$	0.33	0.22	0.10 @ 220.7090
CH_3CN	$13-12$	0.37	0.22	0.13 @ 239.0968
CH_3CN	$14-13$	0.42	0.21	0.08 @ 257.4830
CH_3CN	$15-14$	0.44	0.20	0.07 @ 275.8678
CH_3CN	$16-15$	0.46	0.17	0.07 @ 294.2514
CH_3CN	$17-16$	0.46	0.15	0.09 @ 312.6336
CH_3CN	$18-17$	0.45	0.12	0.08 @ 331.0143
CH_3CN	$19-18$	0.43	0.10	2.6 @ 349.3450
CH_3CN	$20-19$	0.40	0.07	0.04 @ 367.0777

Table 8: Peak intensities from spectra generated using the *MolPred* and MADCUBA predictions for CH_3CN. The ALCHEMI intensity at the Group Representative Frequency is also shown.

the particular cases of CO and HCO^+, the prediction from *MolPred* follows very closely the intensity at the reference frequencies.

It is important to indicate that the training examples were calculated for a velocity of 250 km s^{-1} which is slightly different from the source velocity and thus the channel closest to each transition's rest frequency is not the peak intensity. This misalignment affects our neural network predictions since it uses the intensity of a single channel instead of the integrated emission. The results from Figures 11 and 12 suggest that a higher number of transitions may contribute to construct a better characterized model of the molecule. Figure 12 also shows the strong impact of contamination from brighter transitions from other species on the fit results. This is severely affecting $19-18$ transition of CH_3CN. Despite its limitations, the *MolPred* predictions are close to the MADCUBA fits for our astronomical data test. We believe that its performance can be further tuned moving to a model which uses more information from the spectrum such as the full line profile or an integrated intensity.

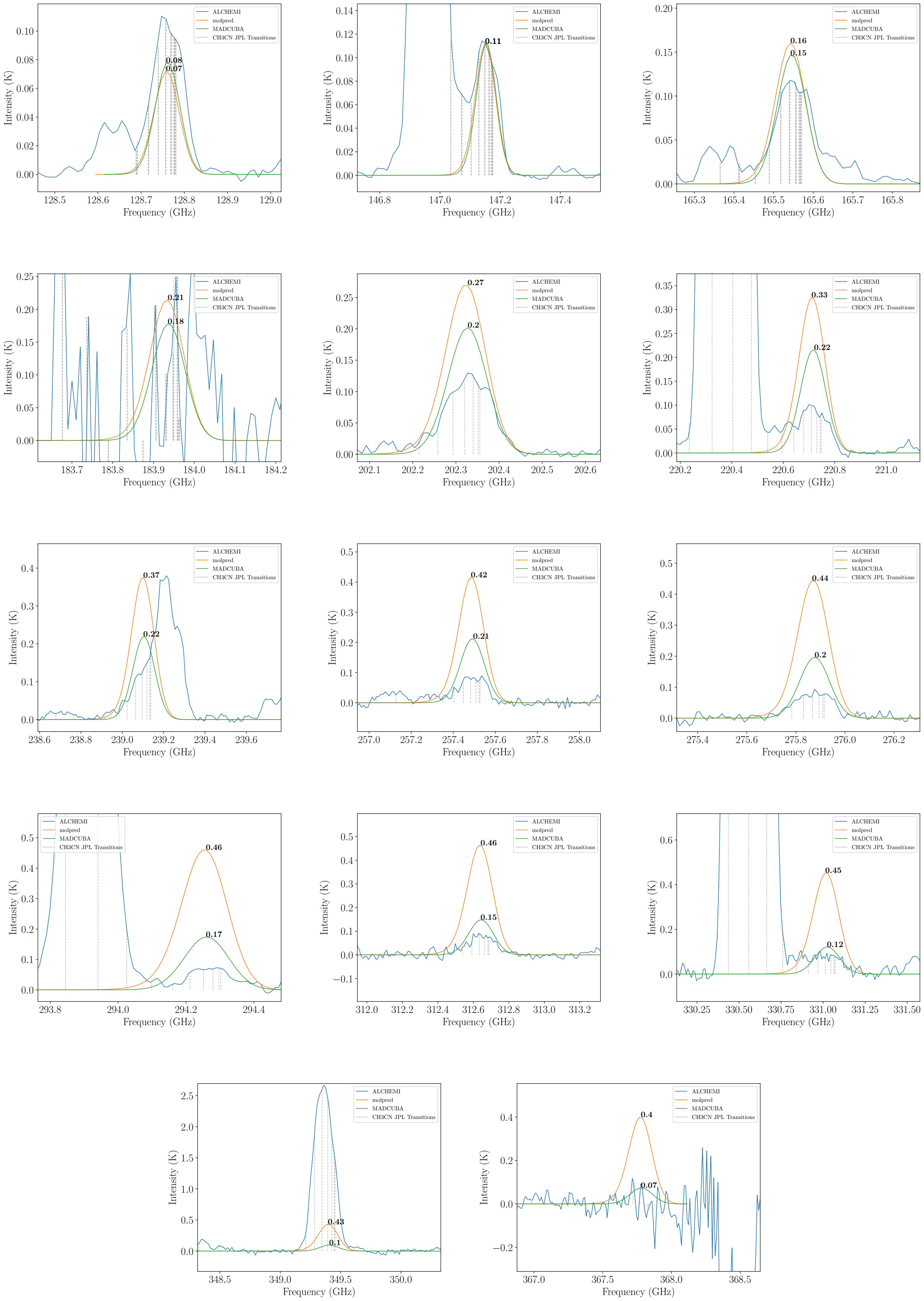

Fig. 12: Similar to Figure 9 for the CH_3CN transitions in the ALCHEMI data. The transitions are grouped by the J quantum numbers and the frequency of each transition are indicated by silver lines.

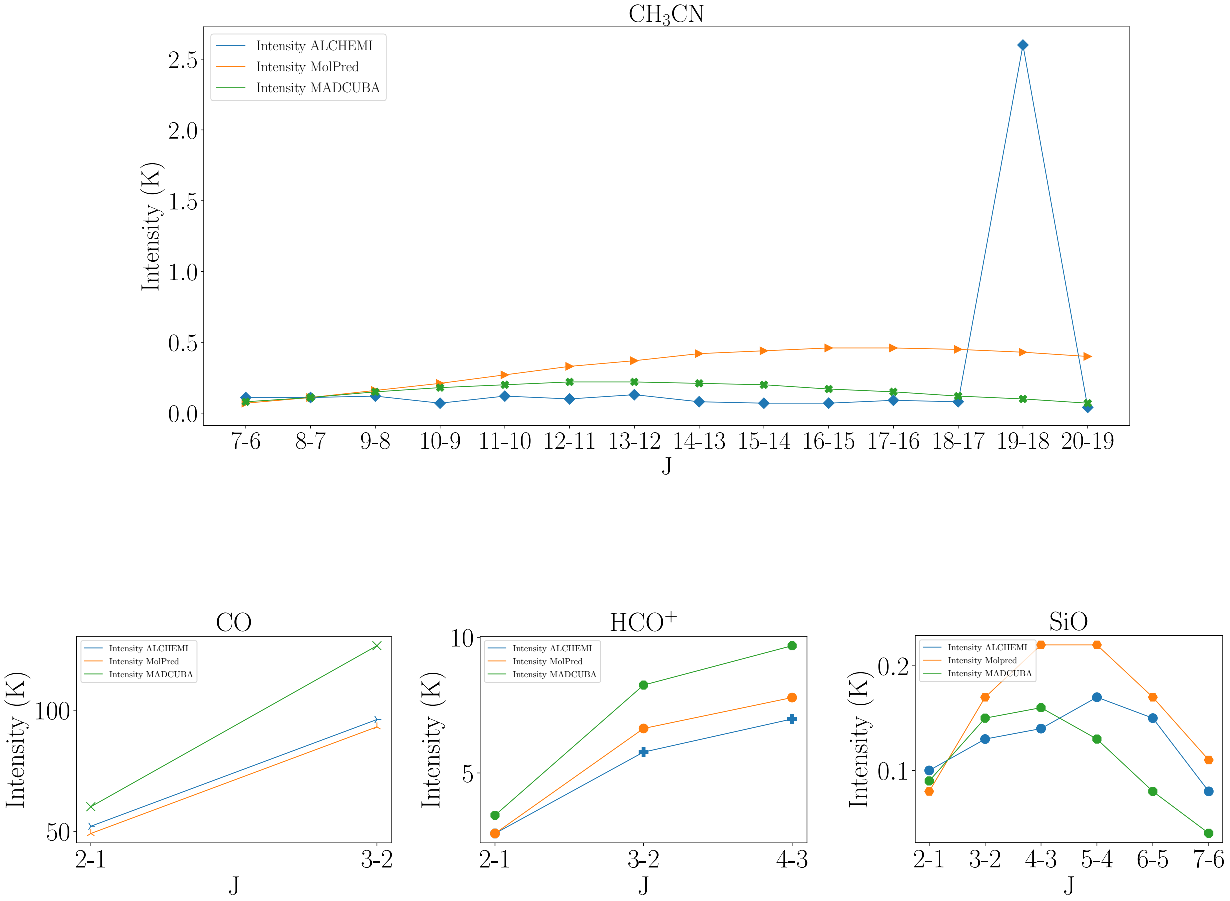

Fig. 13: Real and predicted intensities as a function of J from MolPred (orange), MADCUBA (green) and the ALCHEMI data (blue). Intensities are at the rest frequency of the transitions.

4 Conclusions and future work

We described a software package called *MolPred* which is able to extract the peak intensities of molecular transitions from spectra and use them to estimate the column density and temperature of the molecule. It is able to do this for CO, HCO$^+$, SiO and CH$_3$CN with a mean error of 1-9% on the predicted values when evaluated on synthetic data.

Molpred was also shown to perform well on real astronomical data. A spectrum from NGC 253 was processed by *MolPred* and the values obtained were similar to those found using the MADCUBA software package. This was despite the fact that some molecular transitions required for the networks were missing from the data. The predicted values of $log(N)$ and T$_{ex}$ from *MolPred* were within 13% of those found using MADCUBA on average. The differences with MADCUBA are understood, and mostly related to the fact that *MolPred* is using single values per transition training and spectrum analysis rather than the whole spectral profile.

Further work should include increasing the number of molecules and predicting a broader range of physical parameters such as the line width and source size, for which the whole line profile should be used. The number of molecules should be extended, and the capacity to predict highly blended spectrum should be explored. Another parameter to explore would be multiple velocity components. However, this will require further tuning of the networks as three of the four best neural networks in this work had three layers of 1024 nodes. This would be prohibitively large if many molecules were considered, each needing their own neural network. Therefore, either these networks must be greatly reduce or networks must be trained which can solve for more than one molecule.

Acknowledgements A.B. wishes to thank Dr. Diego Mardones for his contribution to the early stages of this work. Also, to acknowledge support from the Federico Santa María Technical University General Directorate for Research and Postgraduate Studies (DGIP). JH and SV are funded by the European Research Council (ERC) Advanced Grant MOPPEX 833460. V.M.R. acknowledges support from the Comunidad de Madrid through the Atracción de Talento Investigador Modalidad 1 (Doctores con experiencia) Grant (COOL: Cosmic Origins Of Life; 2019-T1/TIC-15379; PI: V.M. Rivilla).

References

Berriman, G.B., Groom, S.L., 2011. How will astronomy archives survive the data tsunami? Commun. ACM 54, 52–56. URL: http://doi.acm.org/10.1145/2043174.2043190, doi:10.1145/2043174.2043190.

Cernicharo, J., 2012. Laboratory astrophysics and astrochemistry in the herschel/alma era. EAS Publications Series 58, 251–261. URL: https://doi.org/10.1051/eas/1258040, doi:10.1051/eas/1258040.

Chollet, F., et al., 2015. Keras. URL: https://github.com/fchollet/keras.

Farnes, J., Mort, B., Dulwich, F., Salvini, S., Armour, W., 2018. Science Pipelines for the Square Kilometre Array. arXiv e-prints , arXiv:1811.08272arXiv:1811.08272.

Jurić, M., Kantor, J., Lim, K., Lupton, R.H., Dubois-Felsmann, G., Jenness, T., Axelrod, T.S., Aleksić, J., Allsman, R.A., AlSayyad, Y., Alt, J., Armstrong, R., Basney, J., Becker, A.C., Becla, J., Bickerton, S.J., Biswas, R., Bosch, J., Boutigny, D., Carrasco Kind, M., Ciardi, D.R., Connolly, A.J., Daniel, S.F., Daues, G.E., Economou, F., Chiang, H.F., Fausti, A., Fisher-Levine, M., Freemon, D.M., Gee, P., Gris, P., Hernandez, F., Hoblitt, J., Ivezić, Ž., Jammes, F., Jevremović, D., Jones, R.L., Bryce Kalmbach, J., Kasliwal, V.P., Krughoff, K.S., Lang, D., Lurie, J., Lust, N.B., Mullally, F., MacArthur, L.A., Melchior, P., Moeyens, J., Nidever, D.L., Owen, R., Parejko, J.K., Peterson, J.M., Petravick, D., Pietrowicz, S.R., Price, P.A., Reiss, D.J., Shaw, R.A., Sick, J., Slater, C.T., Strauss, M.A., Sullivan, I.S., Swinbank, J.D., Van Dyk, S., Vujčić, V., Withers, A., Yoachim, P., LSST Project, f.t., 2015. The LSST Data Management System. ArXiv e-prints arXiv:1512.07914.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Lightfoot, J., Kosugi, G., Wyrowski, F., Zapata, L., Muders, D., Boone, F., Tsutsumi, T., Davis, L., Wilson, C., Shepherd, D., 2008. ALMA Pipeline Heuristics, in: Argyle, R.W., Bunclark, P.S., Lewis, J.R. (Eds.), Astronomical Data Analysis Software and Systems XVII, p. 573.

Mach, M., Köhler, R., Czoske, O., Leschinski, K., Zeilinger, W.W., Kausch, W., Ratzka, T., Leitzinger, M., Greimel, R., Przybilla, N., Schaffenroth, V., Güdel, M., Brandl, B.R., 2016. Data reduction software for the Mid-Infrared E-ELT Imager and Spectrograph (METIS) for the European Extremely Large Telescope (E-ELT), in: Software and Cyberinfrastructure for Astronomy IV, p. 991327. doi:10.1117/12.2232436.

Martín, S., Martín-Pintado, J., Blanco-Sánchez, C., Rivilla, V.M., Rodríguez-Franco, A., Rico-Villas, F., 2019. Spectral line identification and modelling (slim) in the madrid data cube analysis (madcuba) package. Astronomy & Astrophysics 631, A159. URL: http://dx.doi.org/10.1051/0004-6361/201936144, doi:10.1051/0004-6361/201936144.

McMullin, J.P., Waters, B., Schiebel, D.and Young, W., Golap, K., 2007. Casa architecture and applications. Astronomical Data Analysis Software and Systems XVI (ADASS XVI) 376, 127–130.

Möller, T., Schilke, P., 2015. Manual for xclass-interface. https://www.astro.uni-koeln.de/wd-schilke/myXCLASS/XCLASS-Interface__No-NR__Linux__version_1.1.6.zip.

Nakajima, T., Takano, S., Kohno, K., Harada, N., Herbst, E., Tamura, Y., Izumi, T., Taniguchi, A., Tosaki, T., 2015. A multi-transition study of molecules toward NGC 1068 based on high-resolution imaging observations with ALMA. Publications of the Astronomical Society of Japan 67, 8. doi:10.1093/pasj/psu136, arXiv:1410.5912.

Pickett, H., Poynter, R., Cohen, E., Delitsky, M., Pearson, J., Müller, H., 1998. Submillimeter, millimeter, and microwave spectral line catalog. Journal of Quantitative Spectroscopy and Radiative Transfer 60, 883–890. URL: https://www.sciencedirect.com/science/article/pii/S0022407398000910, doi:https://doi.org/10.1016/S0022-4073(98)00091-0.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408.

Schilke, P., Möller, T., Comito, C., Sánchez-Monge, Á., Schmiedeke, A., Zernickel, A., 2015. Taming the Dragon: Automatic Line Fitting of ALMA data, in: Iono, D., Tatematsu, K., Wootten, A., Testi, L. (Eds.), Revolution in Astronomy with ALMA: The Third Year, p. 195.

Swings, P., Rosenfeld, L., 1937. Considerations regarding interstellar molecules. Astrophysical Journal 86, 483–486.

Teuben, P., 2015. Admit 0.5.2 documentation. http://tinyurl.com/osbu75s.

van der Tak, F. F. S., Black, J. H., Schöier, F. L., Jansen, D. J., van Dishoeck, E. F., 2007. A computer program for fast non-lte analysis of interstellar line spectra* - with diagnostic plots to interpret observed line intensity ratios. A&A 468, 627–635. URL: https://doi.org/10.1051/0004-6361:20066820, doi:10.1051/0004-6361:20066820.

Vastel, C., Bottinelli, S., Caux, E., Glorian, J.M., Boiziot, M., 2015. CASSIS: a tool to visualize and analyse instrumental and synthetic spectra., in: Martins, F., Boissier, S., Buat, V., Cambrésy, L., Petit, P. (Eds.), SF2A-2015: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics. Eds.: F. Martins, S. Boissier, V. Buat, L. Cambrésy, P. Petit, pp.313-316, pp. 313–316.

Woon, D., 2020. Lists of interstellar and circumstellar molecules. http://tinyurl.com/m6916qg.